Proceedings of the Tenth Annual U.S. Army Conference On Applied Statistics, 20-22 October 2004

Yasmin H. Said, Barry A. Bodt
EDITORS

Hosted by:
U.S. Army Research Laboratory

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

Cooperating Institutions:
LOS ALAMOS NATIONAL LABORATORY
GEORGE MASON UNIVERSITY
INSTITUTE FOR DEFENSE ANALYSIS
OFFICE OF NAVAL RESEARCH
INTERFACE FOUNDATION OF NORTH AMERICA, INC.

Army Research Laboratory

Aberdeen Proving Ground, MD 21005-5067

Proceedings of the Tenth Annual U.S. Army Conference On Applied Statistics, 20-22 October 2004

Yasmin H. Said, EDITOR
Department of Computational and Data Sciences, George Mason University

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

Hosted by:
U.S. Army Research Laboratory

TENTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

Agenda and Table of Contents

Monday, October 18

TUTORIAL

Quantitative Graphics and Visual Analytics for Communication and Discovery
Dan Carr, George Mason University

Tuesday, October 19

TUTORIAL CONTINUED

Wednesday, October 20

GENERAL SESSION I

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

Models for Anthrax: Antibiotics and Vaccines
Ron Brookmeyer, Johns Hopkins University (Keynote Address)

The Challenges of Streaming Data
Bill Szewczyk, National Security Administration

CONTRIBUTED SESSION I

Chair: J. Robert Burge, Walter Reed Army Institute of Research

System Engineering Approach and Metrics for Evaluating Digitization for the U.S. Army Battle Command
Jock O. Grynovicki and Jean Breitenbach, U.S. Army Research Laboratory

CONTRIBUTED SESSION IV

Chair: Douglas Tang, Uniformed Services University of the Health Sciences

CONTRIBUTED SESSION V

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

SPECIAL SESSION II: RECENT ADVANCES IN ENGINEERING STATISTICS

Chair: C.F. Jeff Wu, Georgia Tech

CONTRIBUTED SESSION VI

Chair: Lee Dewald, Virginia Military Institute

CONTRIBUTED SESSION VII

SPECIAL SESSION III: THE VOLUME-OF-TUBE FORMULA, PERTURBATION, MIXTURE MODELS AND APPLICATIONS

Friday, October 22

GENERAL SESSION IV

CONTRIBUTED SESSION VIII

Chair: Jackie Telford, Johns Hopkins Applied Physics Laboratory

GENERAL SESSION V

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

A BAYESIAN FRAMEWORK FOR STATISTICAL, MULTI-MODAL SENSOR FUSION

Michael J. Smith* Anuj Srivastava †

Abstract

We propose a framework for obtaining statistical inferences from multi-modal and multi-sensor data. In particular, we consider a military battlefield scene and address problems that arise in tactical decision-making while using a wide variety of sensors (an infrared camera, an acoustic sensor array, a human scout, and a seismic sensor array). Outputs of these sensors vary widely, from 2D images and 1D signals to categorical reports. We propose novel statistical models for representing seismic sensor data and human scout reports while using standard models for images and acoustic data. Combining the joint likelihood function with a marked Poisson prior, we formulate a Bayesian framework and use a Metropolis-Hastings algorithm to generate inferences. We demonstrate this framework using experiments involving simulated data.

1 Introduction

Tactical decision makers in the military and in homeland security are increasingly dependent upon information collected by an ever-expanding array of electronic sensors. Commanders require systems that can either formulate decisions in an automated fashion or assist in decision making by processing the available sensor data. A specific problem is to detect, track, and recognize targets of interest in a battlefield situation using imaging and other sensing devices. The widespread use of sensors such as imaging devices has made them essential tools of non-invasive surveillance of battlefields and public areas such as airports and stadiums, as well as remote locations and areas of restricted access, where additional preventive measures are needed. Usage of multiple sensors observing a scene simultaneously has become a common situation. An important question for developing automated systems is: How to fuse information from these multiple sources to learn and understand the underlying scene? In this paper, we address this problem of sensor fusion using a statistical framework, by building probability models for sensor data and scene variables, and seeking high probability solutions.

What makes the problem of fusing sensor data a difficult one? An important issue is the widely different nature of outputs generated by different sensors. For instance, an IR camera generates a 2D image, a seismic sensor measures an electromagnetic wavefront, an acoustic sensor measures an audio signal, and a human scout reports categorical data. Traditional techniques of extracting features and merging feature vectors do not apply here directly. Past research in sensor fusion has generally focused on multiple sensors of similar type, e.g. multiple cameras or multiple signal receivers, and the solutions tend to exploit this similarity. The problem of sensor fusion from completely different sensors is much more difficult. An attractive solution is to take a statistical

*Department of Mathematical Sciences, United States Military Academy, West Point, NY 10996
†Department of Statistics, Florida State University, Tallahassee, FL 32306

approach and to use joint probabilities instead of fusing data or features directly. That is, define a single inference space and use different sensor outputs to impose probabilities on this inference space. Despite differences in the nature of sensor outputs, the probabilities imposed can still be utilized individually or jointly to form scene estimates.

Some of the current ideas for fusing data from multiple sensors of similar type include the following. Viswanathan and Varshney [13] use likelihood ratio tests (LRTs) to combine the decisions of signal sensors operating in parallel; Costantini et al. [1] apply a least-squares approach to fuse synthetic aperture radar (SAR) images of different resolutions; Filippidis et al [2] study a similar problem using two SAR sensors. Rao et al [9] describe a decentralized Bayesian approach for identifying targets. Kam, Zhu, and Kalata [3] present a survey of techniques used in the problem of robot navigation including Kalman filtering, rule-based sensor fusion, fuzzy logic, and neural networks. However, rather limited attention has been focused on fusion of sensors with different modalities: Strobel et al. [12] describe the use of audio and video sensors for object localization using Kalman filtering; Ma et al. [5] use optical and radar sensor fusion for detecting lane and pavement boundaries. Some papers have focused on alternate frameworks for statistical sensor fusion: Mahler [6] develops the theory of finite-set statistics (FSST) as an extension of Bayesian methods for multiple-target tracking.

1.1 Bayesian Sensor Fusion

We take a fundamental approach to scene inference using a Bayesian formulation that is similar to the approach of Miller et al [7, 8]. Rather than extracting features, we choose to analyze the raw sensor data directly and jointly to estimate the locations and identities of target vehicles that are present. For this paper, we have avoided the difficulty of temporal registration of sensor outputs by assuming that all sensors are synchronized in time. However, our methodology obviates the need for spatial association — the fusion proceeds according to the conditional probabilities corresponding to each of the different data vectors.

We formulate the sensor fusion problem next. Consider a planar region of a battlefield containing an unknown number of targets of different types. Our goal is to use the sensor data to detect and recognize them. Let $\mathcal{D} \subset \mathbb{R}^2$ be a region of interest in a battlefield, and let X denote an array of variables describing the target positions (in $\mathcal{D}$) and types. In addition to target positions, there are a number of other variables, such as their pose, motion, load, etc, that can be of interest and, in general, one should estimate all of them. We simplify the problem by assuming these other variables to be known and fixed. In particular, we assume a fixed orientation for all target vehicles.

Table 1: Sensor Suite

Label	*Sensor*	*Nature of Operation*	*Detected Aspects*	*Output*
s_1	Infrared Camera	Low-Resolution Imager	Target Location & ID	2D Image Array (Y_1)
s_2	Acoustic Array	Audio Signal Receiver	Direction Only; No ID	1D Signal Vector (Y_2)
s_3	Scout	Human Vision	Rough Location; ID	Categorical Data (Y_3)
s_4	Seismic Array	Wave Receiver	Rough Location; Partial ID	Zone Detection (Y_4)

We cannot observe X directly; instead, we must rely on the data that the sensors generate. Sensors can typically detect only certain aspects of the scene; i.e., sensors are partial observers.

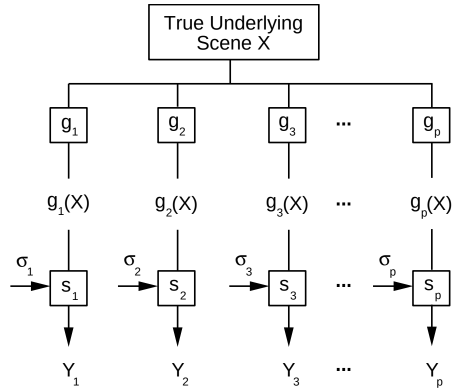


Figure 1: Sensor Data Derived from Projections of the Scene

Our goal is to use this partial and complementary information from different sensors to form a complete inference. As summarized in Table 1, an acoustic sensor array can detect the directions along which audio signals arrive from target vehicles, but it ascertains neither the targets' radial distances along those directions nor the targets' identities. A scout is trained to recognize target identities, but he has limited ability to report precise locations. Imaging sensors are also limited by their resolution, the possibility of target obscuration, and the presence of scene clutter. We assume the IR camera provides top views of the scenes using overhead shots. Despite their respective shortcomings, all of these sensors provide a means to discover the *number* of targets. In contrast, a seismic sensor is a "classifier" — it reports only target *type* (tracked vehicle, wheeled vehicle, dismounted personnel). We depend upon the complementary nature of the sensors and combine their data to conduct unified inference about the scene. Our choice of sensors is motivated by current practices and future plans of the military. In addition to the current routines of battlefield imaging using aerial (infrared) imaging and human scouting, the Army has interest in developing a variety of unmanned ground sensors (UGS) that include acoustic and seismic sensors. These UGS are advantageous over electronic/optical systems due to their low cost, low power requirement, and large detection/tracking range.

Definition 1 Bayesian sensor fusion *is a methodology for scene inference that: (i) formulates a prior distribution for the scene, (ii) constructs probability models for multiple-sensor data conditioned on the scene, and (iii) conducts unified inference about the scene using the posterior distibution of the scene given the sensor data.*

In Figure 1, we depict as projections $g_1(X), \ldots, g_p(X)$ the various aspects or attributes of the scene that our sensors $s_1, \ldots, s_p$ can detect. Each sensor is subject to observation errors σ_i in the generation of data vectors $Y_1, \ldots, Y_p$. We assume that these errors are independent so that the Y_i s are conditionally independent given X. Let $L_i(Y_i \,|\, X)$ denote the likelihood function for data vector Y_i conditioned on the scene X and let $\nu_0(X)$ denote a prior distribution on the scene X. Applying Bayes' rule and assuming conditional independence of the Y_i s given X, we obtain the posterior distribution of our interest:

$$\nu(X \,|\, Y_1, \ldots, Y_p) \;\propto\; L_1(Y_1 \,|\, X) \cdots L_p(Y_p \,|\, X)\; \nu_0(X).$$

Our methodology leads us to generate estimates $\hat{X}$ of the scene from the posterior distribution $\nu(X \,|\, Y_1, \ldots, Y_p)$. Indeed, one may distinguish different Bayesian sensor fusion schemes according to the sense in which their estimates are optimal. Several criteria such as MAP, posterior median, or MMSE, are commonly used. Techniques for producing optimal estimates according to any of these criteria are detailed in [11]. We employ Markov chain Monte Carlo methods to generate samples from the posterior distribution $\nu(X \mid Y_1, \ldots, Y_p)$. Specifically, we implement a version of the Metropolis-Hastings algorithm in a MATLAB environment. We propose a prior distribution for the scene space and probability models for the four modes of sensor data mentioned above: infrared imagery (Y_1), acoustic sensor data (Y_2), a scout's spot report (Y_3), and seismic data (Y_4). We apply our methodology to simulated battlefield scenes and obtain results that illustrate the inferential advantage to using all available sensor data.

Next, we outline major goals of this paper. (i) We propose statistical models for seismic sensor data and human scout reports, and derive their likelihood functions. (ii) Along with the established models for IR and acoustic sensors, we use these likelihood functions in formulating a fully Bayesian approach to battlefield inferences. And, (iii) we construct an MCMC solution to generating Bayesian inferences from the posterior distribution.

This paper is organized as follows. A representation of targets' positions and identities, and statistical models for two sensors leading to a joint posterior distribution are presented in Section 2. A Metropolis-Hastings algorithm to sample from this posterior is described in Section 3. Some examples of scene inferences presented in Section 4. Finally, some simulation results are illustrated in Section 5.

2 Scene Representations and Sensor Models

This section presents statistical models and representations for the scene and the sensors. Because of its modular nature, our methodology can readily accommodate different or additional models that future research may suggest.

2.1 Scene Representation and Prior Model

Let X denote the positions and target identities of vehicles present in a region of the battlefield. We represent X as a point in the space $\mathcal{X} = \bigcup_{n=0}^{\infty} (\mathcal{D} \times \mathcal{A})^n$, where $\mathcal{D} \subset \mathbb{R}^2$ is a battlefield region of interest, $\mathcal{A} = \{\alpha_1, \ldots, \alpha_M, \alpha_\emptyset\}$ is a set of M possible target types ($\alpha_\emptyset$ means that no target is present), and n is the number of targets present. Since n is not known *a priori*, we allow for all possible values of n in the construction of $\mathcal{X}$. To support follow-on Markov chain development, we discretize the battlefield region $\mathcal{D}$ along a rectangular grid: let $\mathcal{D} = \{1, \ldots, R\} \times \{1, \ldots, C\}$ with $R, C < \infty$. This allows us to use (i, j) coordinates to denote target locations. We also impose the constraint $n \leq RC$. The motivation for an upper bound on the number of targets in a fixed region of the battlefield is clear: two targets cannot occupy the same physical space. We disallow the possibility that targets stack themselves vertically; the upper bound RC generously allows for target placements at each point in the discretized region. This modifies the state space to be both discrete and finite: $\mathcal{X} = \bigcup_{n=0}^{RC} (\mathcal{D} \times \mathcal{A})^n$. We express a typical state $X \in \mathcal{X}$ as a matrix: $X = \begin{bmatrix} r_1 & r_2 & \cdots & r_n \\ c_1 & c_2 & \cdots & c_n \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{bmatrix}$, where $(r_i, c_i)^T$ are coordinates of target locations. Each column of X represents a target described by its center-mass location (row and column) and its identity (α). Let $\|X\| = n$ denote the number of columns in the state matrix X and let X_j denote the j^{th} column of X for $j = 1, \ldots, n$. For $n = 0$, let $X_\emptyset$ denote the empty state.

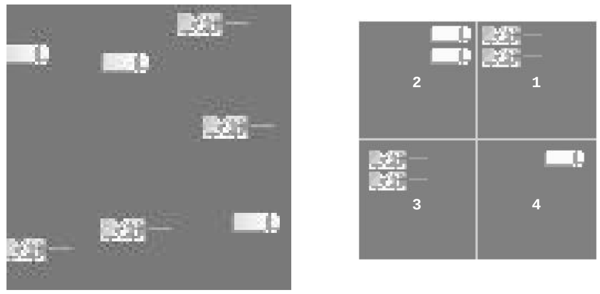


Figure 2: Left panel shows the top view of a simulated scene containing three trucks and four tanks. Right panel shows a visual rendering of scout's spot report Y_3 (right) with labeled quadrants.

We consider X to be a realization of a marked homogeneous Poisson spatial point process. In other words, we make the following collection of assumptions. Let $N \sim \text{Poisson}(\lambda|\mathcal{D}|)$ for some $\lambda > 0$ where $|\cdot|$ denotes Lebesgue measure on $\mathbb{R}^2$ and we assume that $|\mathcal{D}| > 0$. Conditioned on $\{N = n\}$, let the locations $q_1, \ldots, q_n$ of targets be distributed independently and uniformly in $\mathcal{D}$. Conditioned on the locations $q_1, \ldots, q_n$, let the target identities be assigned independently: for each location, assign identity $\alpha_j \in \mathcal{A}$ with probability $\pi_j \geq 0$ for $j = 1, \ldots, M$ where $\sum_{j=1}^{M} \pi_j = 1$. These assumptions specify a prior probability measure ν_0 defined on a σ-field of subsets of $\mathcal{X}$.

2.2 Sensor Models

Here we detail the statistical models that we have adopted for the various sensors under consideration: infrared camera s_1, acoustic sensor array s_2, human scout s_3, and seismic sensor array s_4. For s_1 and s_2, we use established models from the literature with incorporation details contained in [11]. However, this paper offers new models for s_3 and s_4 and provides detailed motivations for both.

2.2.1 Model for Scout's Spot Report

Army units conduct routine tactical operations in accordance with standing operating procedures or SOPs. Among other provisions, SOPs prescribe reporting formats that scouts use to communicate their observations to higher headquarters. Here we assume that the "spot report" format calls for a partitioning of the observed area $\mathcal{D}$ into four quadrants and that the report provides quadrant counts for each target type. See Figure 2 for an illustration. Let Y_3 denote the scout's spot report. We represent it as a vector of length $4M$ where M is the number of target vehicle identities in $\mathcal{A} \backslash \alpha_\emptyset = \{\alpha_1, \ldots, \alpha_M\}$. Each component of Y_3 belongs to $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$. We propose a hierarchical model for the conditional distribution of Y_3 given X.

To motivate the construction of the model, we may suppose that the scout sequentially answers questions that he poses to himself: *How many targets? Where are they? What are they?* He answers the first question by counting those vehicles that he can see. A reasonable model should therefore allow for a variety of cases: he sees all the vehicles that are present; he misses one or more; he "sees" one or more vehicles that are *not* present; he loses track of his count and begins repeating vehicles that he has already counted. But then, regardless of how the scout arrives at his collection of observed targets, he must decide — vehicle by vehicle — how to classify them according to quadrant and target type. Again, a reasonable model should allow for some ambiguity

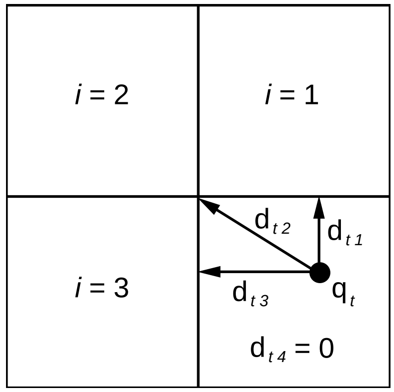


Figure 3: Distances to Nearest Quadrant Boundaries for a Fourth-Quadrant Target

in the quadrant classification of a target lying close to a quadrant boundary. The model detailed below exhibits one way to incorporate these observations about the nature of the scout's report.

Total Count: Let N_S be the total number of target vehicles that the scout observes. We model N_S as a discrete random variable taking values in $\mathbb{Z}_+$ with probability masses obtained by evaluating a Gaussian density function at these points and then normalizing. To specify the Gaussian density, we set the mean μ equal to n (the actual number of targets in the scene) and we take the variance σ^2 to be this function of the mean:

$$\sigma^2(n) = \begin{cases} \beta_0, & \text{if } n = 0; \\ n/\beta_1, & \text{if } n > 0, \end{cases}$$

where β_0 and β_1 are chosen to account for the scout's level of training, his competence, the status of his equipment, weather conditions, and other sources of error. Note that the variance increases linearly with the true target count. Let $\mathcal{G}(k) = \mathcal{G}(k \,|\, n, \beta_0, \beta_1)$ denote the probability mass that this discretized Gaussian distribution places on k. Then

$$\mathcal{G}(k) = \begin{cases} \dfrac{\exp\left(-\frac{1}{2\beta_0}(k-n)^2\right)}{\sum_{z\in\mathbb{Z}_+}\exp\left(-\frac{1}{2\beta_0}(z-n)^2\right)}, & n = 0; \\[2ex] \dfrac{\exp\left(-\frac{\beta_1}{2n}(k-n)^2\right)}{\sum_{z\in\mathbb{Z}_+}\exp\left(-\frac{\beta_1}{2n}(z-n)^2\right)}, & n > 0. \end{cases}$$

Quadrant Target Counts: Given $\{N_S = n_0\}$, we model the components of Y_3 as sums of classification counts constrained so that $\sum_{j=1}^{4M}(Y_3)_j = n_0$. The counts tally the outcomes of "generalized Bernoulli" trials. That is, each observed target corresponds to a conditionally independent trial; each trial has $4M$ possible outcomes corresponding to the scout's possible quadrant & target-type classifications. The outcome of trial t is governed by parameters $\{p_{tj}\}_{j=1}^{4M}$ that satisfy $p_{tj} \geq 0$ and $\sum_{j=1}^{4M} p_{tj} = 1$ for each $t = 1, \ldots, n_0$. These generalized Bernoulli parameters, in turn, depend upon the scene X. (Note that we have a different collection of generalized Bernoulli parameters for each vehicle. If, instead, we had one fixed collection of parameters applicable for all n_0 observed targets, then Y_3 would follow a conditional multinomial distribution given X.)

Now we describe the choice of generalized Bernoulli parameters $\{p_{tj}\}_{j=1}^{4M}$ for the t^{th} trial. Consider a target at location $q_t = (r_t, c_t)^T$ and suppose that q_t lies within quadrant i_t. Our convention is that a target's location is specified by its center-of-mass. As illustrated in Figure 3, let d_{ti} denote the distance from q_t to the i^{th} quadrant — Euclidean distance to the nearest quadrant boundary — and set $d_{ti_t} = 0$. Then, for a fixed constant $a > 0$, set

$$\tilde{p}_{ti} = \frac{\exp(-d_{ti}/a)}{\sum_{j=1}^{4} \exp(-d_{tj}/a)}, \quad i = 1, 2, 3, 4. \tag{1}$$

In words, $\tilde{p}_{ti}$ is the probability that the scout reports quadrant i as the location for target t. Now we account for the scout's reported target type. Let $I\{\alpha_t = j\}$ indicate that α_j is the identity of target t; let $I\{\alpha_t \neq j\}$ indicate that α_j is not the identity of target t. We use these indicators and a classification error parameter denoted σ_3 to split each $\tilde{p}_{ti}$: for $j = j_i = (i-1)M + 1, \ldots, iM$ with $i = 1, 2, 3, 4$, put

$$p_{tj} \;=\; (1 - \sigma_3)\, \tilde{p}_{ti}\, I\{\alpha_t = j_i\} \;+\; \frac{\sigma_3}{M-1}\, \tilde{p}_{ti}\, I\{\alpha_t \neq j_i\}.$$

In words, the scout correctly reports the target type with high probability and he is equally likely to report any of the incorrect target types.

We apply the above formulation of generalized Bernoulli parameters $\{p_{tj}\}_{j=1}^{4M}$ to each of the vehicles that the scout observes ($t = 1, \ldots, n_0$). If it happens that $n_0 = n$, where n is the correct number of vehicles, we assume that the scout observes each target exactly once and that he classifies them independently as above-described trials. In case $n_0 < n$, we assume that the scout observes and similarly classifies a proper subset of targets, where each of $\binom{n}{n_0}$ subsets is equally likely. In case $n < n_0 \leq 2n$, we assume that the scout classifies all targets that are present and that he "double counts" $n_0 - n$ targets, where each of $\binom{n}{n_0 - n}$ collections of doubly-counted targets is equally likely. Let $\lfloor \cdot \rfloor$ denote the greatest integer less than or equal to its argument. For $n_0 > 2n$, we assume that the scout repeatedly classifies each target k times, where $k = \lfloor \frac{n_0}{n} \rfloor$, and then augments this redundancy by including an equally-likely choice from among $\binom{n}{r}$ subsets where $r = n_0 \bmod k$.

Likelihood Function: As suggested earlier, the scout's target-set selection can be modeled in many ways. For the scheme described above, conditioned on $\{N_S = n_0\}$, let $T \in \mathcal{X}$ denote the array of targets that the scout observes. Let $\mathcal{P}(T)$ denote the collection of column permutations of T and let $T_o \in \mathcal{P}(T)$ denote an ordered n_0-tuple of targets (locations and identities). Then the description in this section leads to this likelihood function for Y_3:

$$L_3(Y_3 \,|\, X) = \frac{\mathcal{G}(n_0)}{n_0\,!} \sum_{T_o \in \mathcal{P}(T)} \prod_{t=(T_o)_1}^{(T_o)_{n_0}} {p_{t1}}^{(Y_3)_1}\, {p_{t2}}^{(Y_3)_2} \cdots {p_{t,4M}}^{(Y_3)_{4M}}. \tag{2}$$

The permutations arise because the scout may perform his vehicle-by-vehicle classification according to any ordering; each makes an equally-weighted contribution to the likelihood.

2.2.2 Model for Seismic Data

Open-source documentation about seismic sensors is easy to find; see, for example, "Remote Battlefield Sensor System (REMBASS) and Improved Remote Battlefield Sensor System (IREMBASS)" at the location[1]. According to such sources, a seismic sensor detects and classifies (but does not

[1] *http://www.fas.org/man/dod-101/sys/land/rembass.htm*

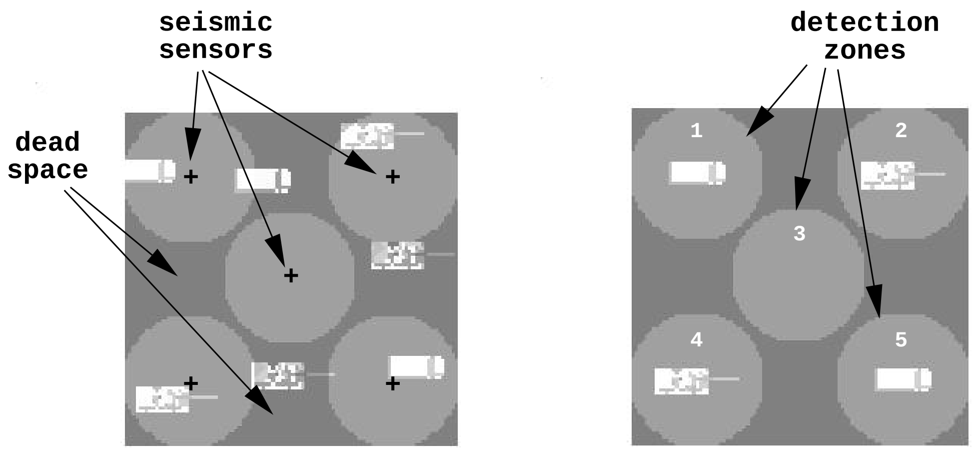


Figure 4: Same Simulated Scene with Overlay of Seismic Detection Zones (left) and Visual Rendering of Seismic Data Y_4 (right) with $k = 5$ labeled zones.

count) those targets whose ground vibrations emanate from within a circular detection zone of known radius. Depending upon the placement of the sensors, these detection zones may or may not overlap. Additionally, the battlefield region $\mathcal{D}$ may contain "dead space" where target vehicles are not detectable by any of the seismic sensors.

Any statistical model that describes data collected by these sensors should reflect certain key aspects of the sensors' behavior. First, whether or not the sensors detect the target vehicles depends upon the locations of the vehicles, the locations of the seismic sensors, and each sensor's detection zone radius. As stated in Section 2.2.1, our convention is that a target's location is specified by its center-of-mass. We assume that all seismic sensors have circular, non-intersecting detection zones with equal radii as depicted in Figure 4. Second, each sensor provides a single classification that summarizes the target-type presence in its zone. If at most one target type is present, there is no confusion. But a statistical model should contain some mechanism whereby the sensor reconciles the presence of more than one target type in its detection zone. The model that we propose offers one way to address these issues.

Assume that k seismic sensors having mutually disjoint detection zones generate a data vector Y_4 with k components (one for each sensor). Let $(Y_4)_j \in \mathcal{A}$ report the j^{th} sensor's summary of target-type presence in its detection zone. Figure 4 provides an illustration for $k = 5$. Its left panel shows the overlay of detection zones on top of the same simulated scene displayed for the scout's spot report. Note that two tanks and one truck lie in dead space — their center-mass locations are not within any of the detection zones. For this scene, the correct seismic data vector is $Y_4 = [\alpha_2, \alpha_1, \alpha_\emptyset, \alpha_1, \alpha_2]'$ where $\alpha_1 =$ tank and $\alpha_2 =$ truck. A visual rendering of Y_4 appears in the right panel where the labeling of the detection zones to match vector components is as follows: top-left is 1, top-right is 2, center is 3, bottom-left is 4, and bottom-right is 5.

Let σ_4 denote a fixed error parameter and let n_{ij} denote the number of type-α_i targets ($i = 1, \ldots, M$) that are present in detection zone $j = 1, \ldots, k$. Fix a detection zone j and let $P(\cdot)$ denote a probability measure defined as follows on all subsets of $\mathcal{A}$.

- **Case 1** If zone j is devoid of target vehicles, we allow the sensor to report correctly with high probability and we assume that the sensor is equally likely to report erroneously any of the target types:

$$P\big\{(Y_4)_j = y \,|\, n_{1j} = \cdots = n_{Mj} = 0\big\} = \begin{cases} 1-\sigma_4, & y = \alpha_\emptyset; \\ \frac{\sigma_4}{M}, & y \in \mathcal{A}; \\ 0, & \text{otherwise.} \end{cases}$$

- **Case 2** If zone j contains exactly one target type, we again allow the sensor to report correctly with high probability. However, in this case, we assume that the sensor is more likely to report an incorrect target type than to report the absence of targets; we assume that all wrong-type classifications are equally likely.

$$P\big\{(Y_4)_j = y \,|\, n_{ij} > 0 \text{ for } i = i_0 \text{ only}\big\} = \begin{cases} \frac{\sigma_4}{4}, & y = \alpha_\emptyset; \\ 1-\sigma_4, & y = \alpha_{i_0}; \\ \frac{3\sigma_4}{4(M-1)}, & y \in \mathcal{A}\backslash\{\alpha_{i_0}\}; \\ 0, & \text{otherwise.} \end{cases}$$

- **Case 3** This is most interesting — the sensor must "decide" among competing target types. Denote by $n_{\cdot j} = \sum_{i=1}^{M} n_{ij}$ the number of targets (of all types) present in detection zone j. Let $I\{\alpha_t = i\}$ indicate whether α_i is the identity of target $t = 1, \ldots, n_{\cdot j}$. Let $a > 0$ be a fixed constant and let d_t denote the distance from target t to the center of the detection zone. These distances are analogous to those depicted in Figure 3 and contribute to the classification probabilities in a manner similar to Equation 1.

$$P\big\{(Y_4)_j = y \mid 2 \le |\{i : n_{ij} > 0\}| \;\big\} = \begin{cases} \sigma_4, & y = \alpha_\emptyset; \\ (1-\sigma_4)\frac{\sum_{t=1}^{n_{\cdot j}} I\{\alpha_t = i\} e^{-d_t/a}}{\sum_{t=1}^{n_{\cdot j}} e^{-d_t/a}}, & y = \alpha_i \in \mathcal{A}; \\ 0, & \text{otherwise.} \end{cases}$$

We assume that the seismic sensors' classifications are conditionally independent given the scene. The above enumeration of cases depending on X and the assumption of conditional independence lead to the following likelihood function for Y_4:

$$L_4(Y_4 \,|\, X) = \prod_{j=1}^{k} P\big\{(Y_4)_j = y \,|\, X\big\}. \tag{3}$$

2.3 The Posterior Distribution

The likelihood functions $L_1(Y_1 \,|\, X)$ and $L_2(Y_2 \,|\, X)$ for infrared images and acoustic data (respectively) are given in [11]. Combined with the likelihood functions derived in this paper, and along with the assumption of conditional independence of the data vectors, we may now express the posterior distribution:

$$\nu(X \,|\, Y_1, Y_2, Y_3, Y_4) \;\propto\; L_1(Y_1 \,|\, X)\; L_2(Y_2 \,|\, X)\; L_3(Y_3 \,|\, X)\; L_4(Y_4 \,|\, X)\; \nu_0(X). \tag{4}$$

Although we will sometimes use the shorthand $\nu(\cdot) \equiv \nu(\cdot \mid Y_1, Y_2, Y_3, Y_4)$, we will always mean that the likelihood functions $L_i(Y_i \mid X)$ are defined (respectively) as in Equations 2 and 3 (and as in [11]) and that the prior distribution ν_0 is defined as in Section 2.1.

3 Metropolis-Hastings Algorithm

So far we have defined a posterior distribution ν on the scene space $\mathcal{X}$, and our task now is to obtain samples from the posterior distribution ν so that we may conduct scene inference. This section presents the algorithm we use to generate approximate samples from ν.

3.1 Transitions of the Markov Chain

We control the evolution of the Markov chain by restricting the one-step transitions to a class of "simple moves." Although this slows down the convergence of the resulting Markov chain, we impose the restriction because analyzing the chain is easier in this setting [4, 7, 8].

Given the current state $X^{(t)}$ at time t, we consider four fundamental types of transitions. To each type corresponds a collection of "neighboring" states (neighbors of $X^{(t)}$) — the states that can be reached from $X^{(t)}$ in one transition. We now introduce notation for these sets of neighbors.

1. The first simple move is **DEATH**. This means that we select and remove one of the current targets from the state matrix. Let $\mathcal{N}_D(X^{(t)})$ denote the neighbors of state $X^{(t)}$ under the **DEATH** transition. Define

$$\mathcal{N}_D(X^{(t)}) = \begin{cases} \{X^{(t)}_{-j} : j = 1, \ldots, n\}, & \text{if } \|X^{(t)}\| \geq 1; \\ \{X^{(t)}\}, & \text{if } \|X^{(t)}\| = 0, \end{cases}$$

where $X^{(t)}_{-j}$ denotes the matrix $X^{(t)}$ after removing column j.

2. The second simple move is **CHANGE ID**. This means that we select a current target in the state matrix and change its identity α. Let $\mathcal{N}_C(X^{(t)})$ denote the neighbors of state $X^{(t)}$ under the **CHANGE ID** transition. Define

$$\mathcal{N}_C(X^{(t)}) = \begin{cases} \{X^{(t)}_{\Delta j} : j = 1, \ldots, n\}, & \text{if } \|X^{(t)}\| \geq 1; \\ \{X^{(t)}\}, & \text{if } \|X^{(t)}\| = 0, \end{cases}$$

where $X^{(t)}_{\Delta j}$ denotes the matrix $X^{(t)}$ after changing the identity component of column j.

3. The third simple move is **ADJUST**. This means that we select a current target in the state matrix and slightly perturb its location q. Let $\mathcal{N}_A(X^{(t)})$ denote the neighbors of state $X^{(t)}$ under the **ADJUST** transition. Define

$$\mathcal{N}_A(X^{(t)}) = \begin{cases} \{X^{(t)}_{\oplus j} : j = 1, \ldots, n\}, & \text{if } \|X^{(t)}\| \geq 1; \\ \{X^{(t)}\}, & \text{if } \|X^{(t)}\| = 0, \end{cases}$$

where each $X^{(t)}_{\oplus j}$ denotes as many as eight perturbations to the location components of $X^{(t)}_j$. For example, if a current target has location $q = (i, j)$, then we permit an adjustment to $q' \in \{(i\pm1, j), (i\pm2, j), (i, j\pm1), (i, j\pm2)\} \cap \mathfrak{D}$. The symbol $\oplus$ is suggestive of this perturbation pattern of rows and columns.

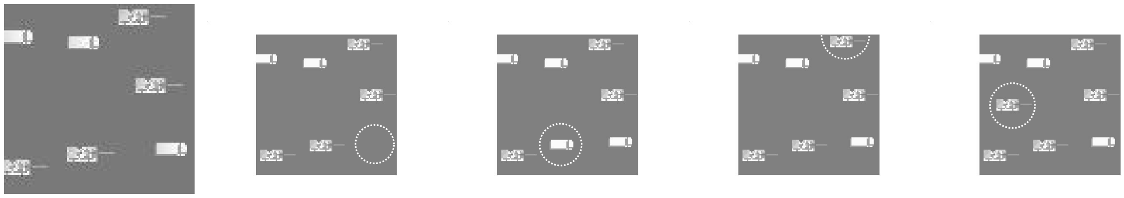


Figure 5: Simple Moves from current state (leftmost) to perform DEATH, CHANGE ID, ADJUST, and BIRTH (left-to-right).

4. The fourth simple move is **BIRTH**. This means that we augment the current state matrix by the addition of another target. Let $T_{X^{(t)}} = \{X_j^{(t)} : j = 1, \ldots, n\} \subset (\mathcal{D} \times \mathcal{A})$ denote the collection of targets represented in state matrix $X^{(t)}$ and let $\mathcal{N}_B(X^{(t)})$ denote the neighbors of state $X^{(t)}$ under the **BIRTH** transition. Define

$$\mathcal{N}_B(X^{(t)}) = \{X_\tau^{(t)} : \tau \in (\mathcal{D} \times \mathcal{A}) \setminus T_{X^{(t)}}\},$$

where $X_\tau^{(t)}$ is the augmentation of the matrix $X^{(t)}$ by one additional column τ corresponding to any "legal" target not already present: $\|X_\tau^{(t)}\| = \|X^{(t)}\| + 1$.

To help visualize the slight adjustments to a given state matrix $X^{(t)}$ contained in the sets of neighbors $\mathcal{N}_D(X^{(t)})$, $\mathcal{N}_C(X^{(t)})$, $\mathcal{N}_A(X^{(t)})$, $\mathcal{N}_B(X^{(t)})$, we present examples in Figure 5. The ADJUST example depicts a shift of the uppermost tank; the other examples are obvious. In Chapter 5, we present portions of Markov chain sample paths that exhibit incremental adjustments similar to Figure 5.

3.2 The Metropolis-Hastings Algorithm

We now state the basic algorithm that prescribes the evolution of our Markov chain; see, for example, Robert and Casella [10]. Fix a state space $\mathcal{X}$ and let ν (known as the *target* distribution) be a probability distribution on $\mathcal{X}$.

Algorithm 1 (Metropolis-Hastings)
Given the current state $X^{(t)} \in \mathcal{X}$,

1. *Generate $Y_t \sim G(y|X^{(t)})$. (G is called the proposal distribution.)*
2. *Set* $X^{(t+1)} = \begin{cases} Y_t & \text{w.p. } \gamma(X^{(t)}, Y_t); \\ X^{(t)} & \text{w.p. } 1 - \gamma(X^{(t)}, Y_t), \end{cases}$ *where* $\gamma(x, y) = \min\left\{1, \frac{\nu(y)\,G(x|y)}{\nu(x)\,G(y|x)}\right\}$.

For a large class of proposal distributions G and for $X^{(1)} \sim F$ where F is an arbitrary probability distribution on $\mathcal{X}$, this algorithm is known to generate a Markov chain with unique stationary distribution ν. For a detailed description of G based on the simple moves in Section 3.1 and for a discussion on asymptotic properties of this Markov chain, please refer to [11].

4 Conducting Scene Inference

Implementing the Metropolis-Hastings algorithm in MATLAB, we obtain an approximate sample from $\nu(\cdot \mid Y_1, Y_2, Y_3, Y_4)$. Specifically, we generate $X^{(1)} \sim \nu_0$ (prior distribution) and then observe $X^{(2)}, X^{(3)}, \ldots$ according to the Metropolis-Hastings transition kernel. After stopping the chain, we discard the first B states (sometimes called a *burn-in* period to allow time for the Markov chain to approach its stationary distribution) and we retain, for purposes of inference,

$$\{X^{(B+1)}, X^{(B+2)}, \ldots, X^{(B+R)}\}.$$

In this chapter, we describe methods for using our sample to answer a variety of questions. We denote the retained portion of the Markov chain by

$$\{X_j\}_{j=1}^{R} \quad \text{where we set} \quad X_1 = X^{(B+1)}, \ldots, X_R = X^{(B+R)}. \tag{5}$$

Letting ν denote the posterior distribution of the scene, we proceed under the assumption that $\{X_j\} \sim \nu$.

Having obtained a sample $\{X_j\}$ from the posterior distribution ν, we might wish to produce a *maximum a posteriori* estimate $\hat{X}_{\text{MAP}}$ of the scene. Such an estimate is characterized by $\hat{X}_{\text{MAP}} = \operatorname{argmax}_{X \in \mathcal{X}} \nu(X)$, that is, $\hat{X}_{\text{MAP}}$ is a *mode* of the posterior distribution. An obvious candidate to estimate $\hat{X}_{\text{MAP}}$ is the sample mode: we can simply report the state matrix that appears most frequently among $\{X_j\}$. An alternative approach abandons the previously described sample and instead uses an adjustment to the Metropolis-Hastings algorithm given earlier. The technique is known as *simulated annealing* and it provides a means to obtain MAP estimates $\hat{X}_{\text{MAP}}$; see, for example, Winkler [14].

5 Simulation Results

Now we present some experimental results demonstrating the proposed framework for Bayesian sensor fusion. In these experiments, we utilize sensor data simulated according to the models proposed.

We start with a simulated scene with corresponding sensor data in Figure 6 and construct a Markov chain to sample from the resulting posterior. Figure 7 shows periodic snapshots along a sample path of this Markov chain in $\mathcal{X}$. Before proceeding with scene inference, we make some qualitative observations about the performance of our algorithm. The top-left panel in Figure 7 depicts the initial state. Navigating through the panels in left-to-right, top-to-bottom fashion, we see the state of the chain at multiples of 100 steps. The bottom-right panel depicts the true scene. At a glance, we observe that this particular sample path evolves quite close to the true scene. Figure 8 illustrates how the *posterior energy* associated with a sample path regulates the evolution of the Metropolis-Hastings algorithm. It depicts $H(X^{(t)}) \propto -\log \nu(X^{(t)})$ plotted against $\frac{t}{25}$. The non-increasing nature of the posterior energy indicates that the Metropolis-Hastings algorithm is indeed steering the sample path toward target configurations with more and more probability mass under the posterior distribution.

6 Summary

We have presented a statistical framework for merging information from multi-modal sensors in order to generate a unified inference. To setup a Bayesian problem, we have introduced statistical

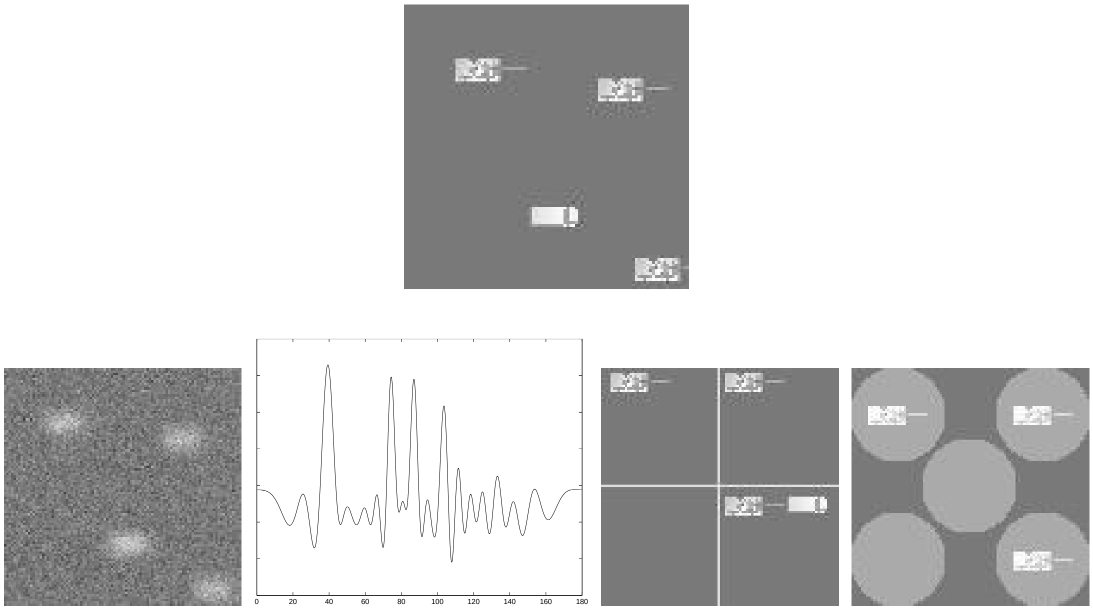


Figure 6: Simulated Scene 2 (top) and Corresponding Sensor Data (bottom) from (left to right) Infrared Camera, Acoustic Sensor Array, Scout, Seismic Sensor Array

models for two sensors - seismic sensor and human scout - and used established models for infrared camera and acoustic array. Assuming a homogeneous Poisson prior on the target placements in the scene, we formulate a posterior distribution on the configuration space, and utilize a Metropolis-Hastings algorithm to generate samples and inferences from it. Experimental results are presented for detecting and recognizing targets in a simulated battlefield scene.

References

[1] Mario Costantini, Alfonso Farina, and Francesco Zirilli. The fusion of different resolution SAR images. In *Proceedings of the IEEE, Vol 85, No 1*, pages 139–146. IEEE, January 1997.

[2] A. Filippidis, L. C. Jain, and N. Martin. Fusion of intelligent agents for detection of aircrafts in SAR images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):378–384, 2000.

[3] Moshe Kam, Xiaoxun Zhu, and Paul Kalata. Sensor fusion for mobile robot navigation. In *Proceedings of the IEEE, Vol 85, No 1*, pages 108–119. IEEE, January 1997.

[4] A. Lanterman, M. Miller, and D. Snyder. General Metropolis-Hastings jump diffusions for automatic target recognition in infrared scenes. *Optical Engineering*, 36(4):1123–1137, 1997.

[5] B. Ma, S. Lakshmanan, and A. O. Hero. Simultaneous detection of lane and pavement boundaries using model-based multisensor fusion. *IEEE Transactions on Intelligent Transport Systems*, 1(3):135–147, 2000.

[6] Ronald Mahler. *An Introduction to Multisource-Multitarget Statistics and its Applications*. Lockheed Martin, Eagan MN, 2000.

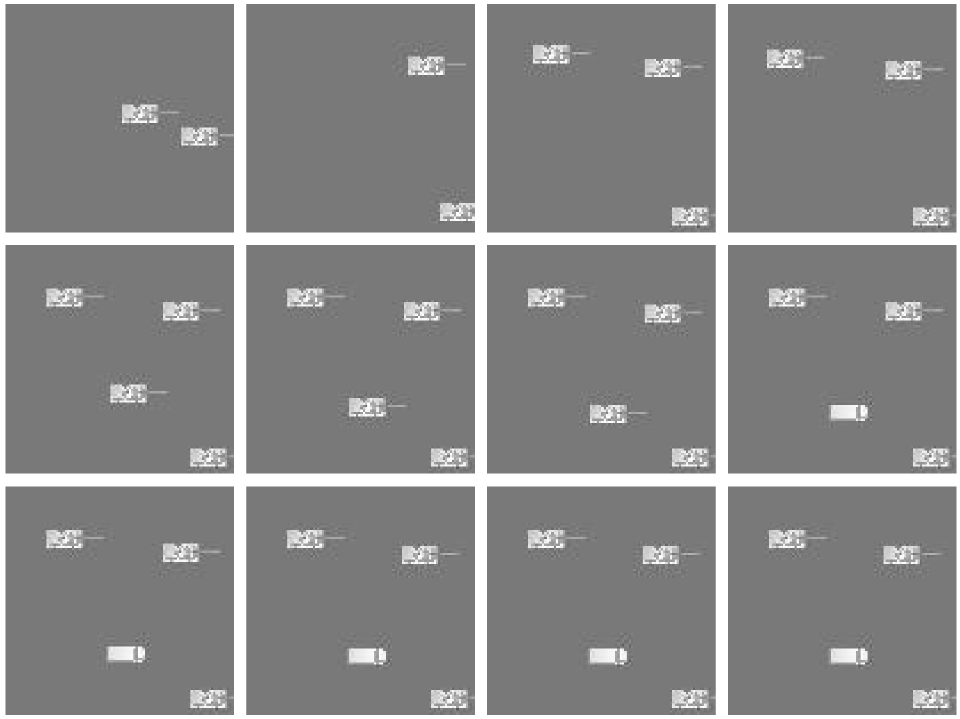


Figure 7: Evolution of Markov Chain for Simulated Scene 2: (left-to-right and top-to-bottom) $X^{(1)}, X^{(100)}, X^{(200)}, \ldots, X^{(1000)}, X_{\text{TRUE}}$

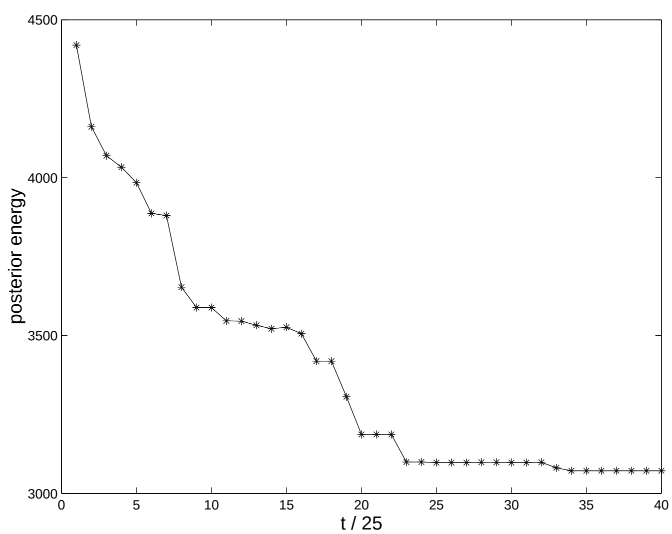


Figure 8: Posterior Energy: Evolution for the Figure 7 Sample Path

[7] M. I. Miller, U. Grenander, J. A. O'Sullivan, and D. L. Snyder. Automatic target recognition organized via jump-diffusion algorithms. *IEEE Transactions on Image Processing*, 6(1):1–17, January 1997.

[8] M. I. Miller, A. Srivastava, and U. Grenander. Conditional-expectation estimation via jump-diffusion processes in multiple target tracking/recognition. *IEEE Transactions on Signal Processing*, 43(11):2678–2690, November 1995.

[9] B. S. Rao and H. Durant-Whyte. A decentralized bayesian algorithm for indentification of tracked targets. *IEEE Transaction on Systems, Man and Cybernetics*, 23(6):1683–1698, 1993.

[10] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 1999.

[11] M. J. Smith. Bayesian sensor fusion: A framework for using multi-modal sensors to estimate target locations & identities in a battlefield scene. *PhD Dissertation, Florida State University, Tallahassee, FL*, August 2003.

[12] N. Strobel, S. Spors, and R. Rabenstein. Joint audio-video object localization and tracking. *IEEE Signal Processing Magazine*, 18(1), 2001.

[13] Ramanarayanan Viswanathan and Pramod K. Varshney. Distributed detection with multiple sensors: Fundamentals. In *Proceedings of the IEEE, Vol 85, No 1*, pages 54–63. IEEE, January 1997.

[14] Gerhard Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer-Verlag, New York, 1999.

A Bayesian Framework for Statistical, Multi-Modal Sensor Fusion

LTC Mick Smith

Department of Mathematical Sciences, United States Military Academy

Anuj Srivastava

Co-Author & Ph.D. Major Professor, Florida State University

20 October 2004

Army Conference on Applied Statistics

Acknowledgements

J. Sethuraman, FSU; Gary Krahn, USMA

ARO DAAD 19-99-1-0267; NMA 201-01-2010; USMA AN40

Outline of Presentation

- Bayesian Sensor Fusion: Definition & Motivation
- Statistical Models for Scene & Sensors
- Implementing a Metropolis-Hastings Algorithm
- Example Markov Chain with Application to Tactical Questions
- Directions for Further Work

Definition

Bayesian sensor fusion is a methodology for scene inference that:

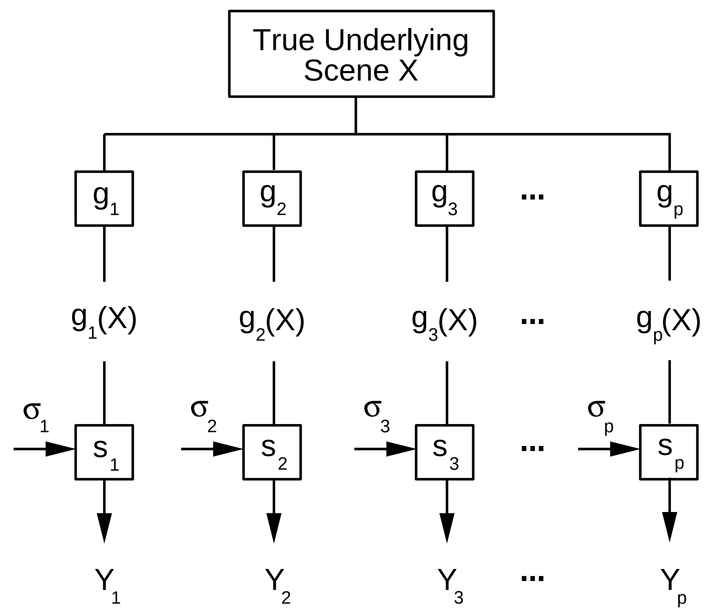
- Formulates a prior distribution for the scene.
- Constructs probability models or likelihood functions for sensor data conditioned on the scene.
- Conducts unified inference about the scene using the posterior distibution of the scene given the sensor data.

Motivation

Our Bayesian methodology for sensor fusion:

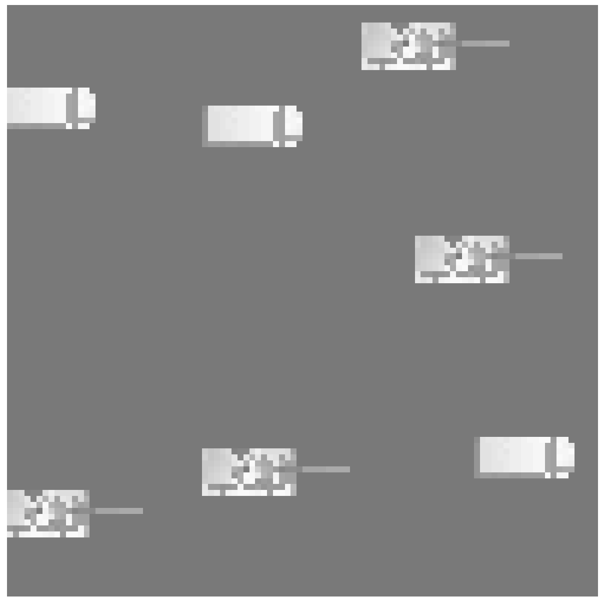
- Recognizes that sensors are partial observers of the scene.
- Exploits the complementary nature of the multi-sensor suite by merging joint probabilities.
- Affords inclusion of the commander's estimate by way of a prior distribution.

Sensors Detect Different Aspects of the Scene



- Sensors $s_1, \ldots, s_p$ observe the projections g_i of the scene X and generate data vectors $Y_1, \ldots, Y_p$.
- Data vectors are corrupted by sensor noise σ_i.

Simulated Battlefield Scene



Types of combat vehicles limited to tanks and trucks with fixed orientations.

Mathematical Description of the Scene

- We take the scene X to be a point in the space $\mathcal{X} = \bigcup_{n=0}^{\infty} (\mathcal{D} \times \mathcal{A})^n$ where:
  - $\mathcal{D} \subset \mathbb{R}^2$ is a battlefield region of interest;
  - $\mathcal{A} = \{\alpha_1, \ldots, \alpha_M\}$ is a set of M possible target types;
  - n is the number of targets present.
- After discretizing & truncating $\mathcal{X}$, a typical state in our Markov chain is a matrix with columns corresponding to target vehicles:

$$X = \begin{bmatrix} r_1 & r_2 & \cdots & r_n \\ c_1 & c_2 & \cdots & c_n \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{bmatrix}.$$

Prior Distribution: The Commander's Experience

- X is a realization of a marked homogeneous Poisson spatial point process:
  - $N \sim \mathrm{Poisson}(\lambda|\mathcal{D}|)$ for some $\lambda > 0$.
  - Given $\{N = n\}$, let the locations $q_1, \ldots, q_n$ of targets be distributed independently and uniformly in $\mathcal{D}$.
- More realistic prior distributions ν_0 on $\mathcal{X}$ are desirable.

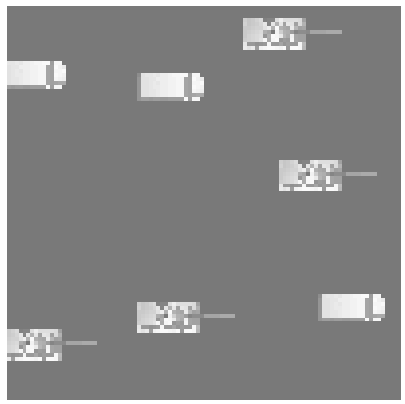
Multi-Modal, Multi-Sensor Environment

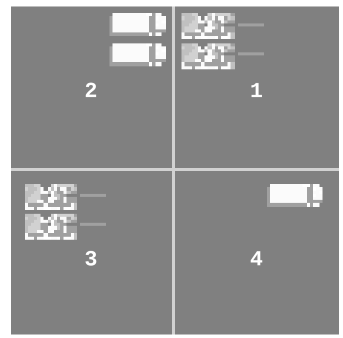
Table 1: Sensors Considered in the Paper

Label	*Sensor*	*Nature of Operation*	*Detected Aspects*	*Data Output* (Y_i)
s_1	Infrared Camera	Low-Resolution Imager	Target Location & ID	2D Image Array
s_2	Acoustic Array	Audio Signal Receiver	Direction Only; No ID	1D Signal Vector
s_3	Scout	Human Vision	Rough Location; ID	Categorical Data
s_4	Seismic Array	Wave Receiver	Rough Location; Partial ID	Local Detection

- Likelihood functions for s_1 and s_2 are adopted from published research.
- Probability models for the scout's spot report and for the seismic sensor array are newly proposed in this work & are described on the following slides.

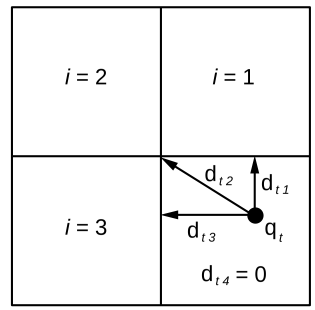
Scout's Spot Report





- Suppose that a scout reports target counts by quadrant & by type.
- To construct a likelihood function conditioned on the scene, we imagine the scout asking & answering three questions:
- How many targets? Where are they? What are they?

Scout Likelihood: How Many Targets? Where?



- Number of targets observed: $N_S \sim$ discretized Gaussian with mean n.
- Given $\{N_S = n_0\}$, require that the spot report Y_3 satisfies $\sum_{j=1}^{4M} (Y_3)_j = n_0$.
- Let d_{ti} denote distance from location q_t to quadrant i.
- Define the probability that the scout reports quadrant i as location for target t:

$$\tilde{p}_{ti} = \frac{\exp(-d_{ti}/a)}{\sum_{j=1}^{4} \exp(-d_{tj}/a)}, \quad i = 1, 2, 3, 4; \quad a > 0.$$

Scout Likelihood: What are They?

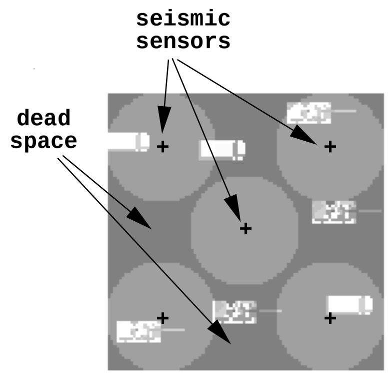
- Let $I\{\alpha_t = j\}$ indicate that α_j is the identity of target t.
- Define generalized Bernoulli parameters $\{p_{tj}\}_{j=1}^{4M}$ for the t^{th} target:

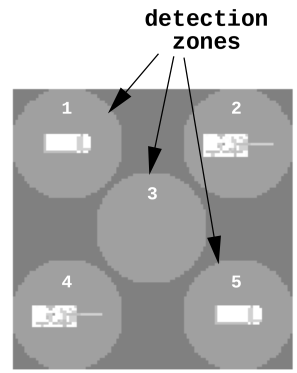
$$p_{tj} \;=\; (1 - \sigma_3)\,\tilde{p}_{ti}\,I\{\alpha_t = j_i\} \;+\; \frac{\sigma_3}{M-1}\,\tilde{p}_{ti}\,I\{\alpha_t \neq j_i\}.$$

- Parameter σ_3 is the classification error.
- In words, the scout correctly reports the target type w.p. $(1 - \sigma_3)$ and he is equally likely to report any of the incorrect target types.
- Likelihood function:

$$L_3(Y_3 \,|\, X) = \frac{\mathcal{G}(n_0)}{n_0\,!} \sum_{T_o \in \mathcal{P}(T)} \prod_{t=(T_o)_1}^{(T_o)_{n_0}} {p_{t1}}^{(Y_3)_1}\, {p_{t2}}^{(Y_3)_2} \;\cdots\; {p_{t,4M}}^{(Y_3)_{4M}}\,.$$

Seismic Sensor Array





- Seismic sensor detects & classifies targets but does not count them.
- Sensors are deployed in an array; each has a known detection-zone radius.
- Array may admit gaps of “dead space.”

Seismic Sensor Behavior

- Case 1: Zone j is devoid of targets:

$$P\big\{(Y_4)_j = y \,|\, n_{1j} = \cdots = n_{Mj} = 0\big\} = \begin{cases} 1-\sigma_4, & y = \alpha_\emptyset; \\ \frac{\sigma_4}{M}, & y \in \mathcal{A}; \\ 0, & \text{otherwise.} \end{cases}$$

- Case 2: Zone j contains exactly 1 target type:

$$P\big\{(Y_4)_j = y \,|\, n_{ij} > 0 \text{ for } i = i_0 \text{ only}\big\} = \begin{cases} \frac{\sigma_4}{4}, & y = \alpha_\emptyset; \\ 1-\sigma_4, & y = \alpha_{i_0}; \\ \frac{3\sigma_4}{4(M-1)}, & y \in \mathcal{A}\backslash\{\alpha_{i_0}\}; \\ 0, & \text{otherwise.} \end{cases}$$

Seismic Sensor Behavior

- Case 3: Sensor must "decide" among competing target types in Zone j:

$$P\Big\{(Y_4)_j = y \,\Big|\, 2 \le |\{i : n_{ij} > 0\}| \,\Big\} =$$

$$(1-\sigma_4)\frac{\sum_{t=1}^{n_{\cdot j}} I\{\alpha_t = i\} e^{-d_t/a}}{\sum_{t=1}^{n_{\cdot j}} e^{-d_t/a}}, \quad y = \alpha_i \in \mathcal{A}.$$

- Likelihood function:

$$L_4(Y_4 \,|\, X) = \prod_{j=1}^{k} P\big\{(Y_4)_j = y \,|\, X\big\}.$$

Posterior Distribution of the Scene

- We assume that, given the scene X, the sensor data vectors Y_i are conditionally independent.
- Applying Bayes' rule, we obtain an expression for the posterior distribution:

$$\nu(X) \equiv \nu(X \mid Y_1, \ldots, Y_p) \propto L_1(Y_1 \mid X) \cdots L_p(Y_p \mid X)\, \nu_0(X).$$

- To conduct inference, we must generate samples from ν.
- We do this via Metropolis-Hastings: we construct an ergodic Markov chain on $\mathcal{X}$ having stationary distibution ν.

Metropolis-Hastings Algorithm

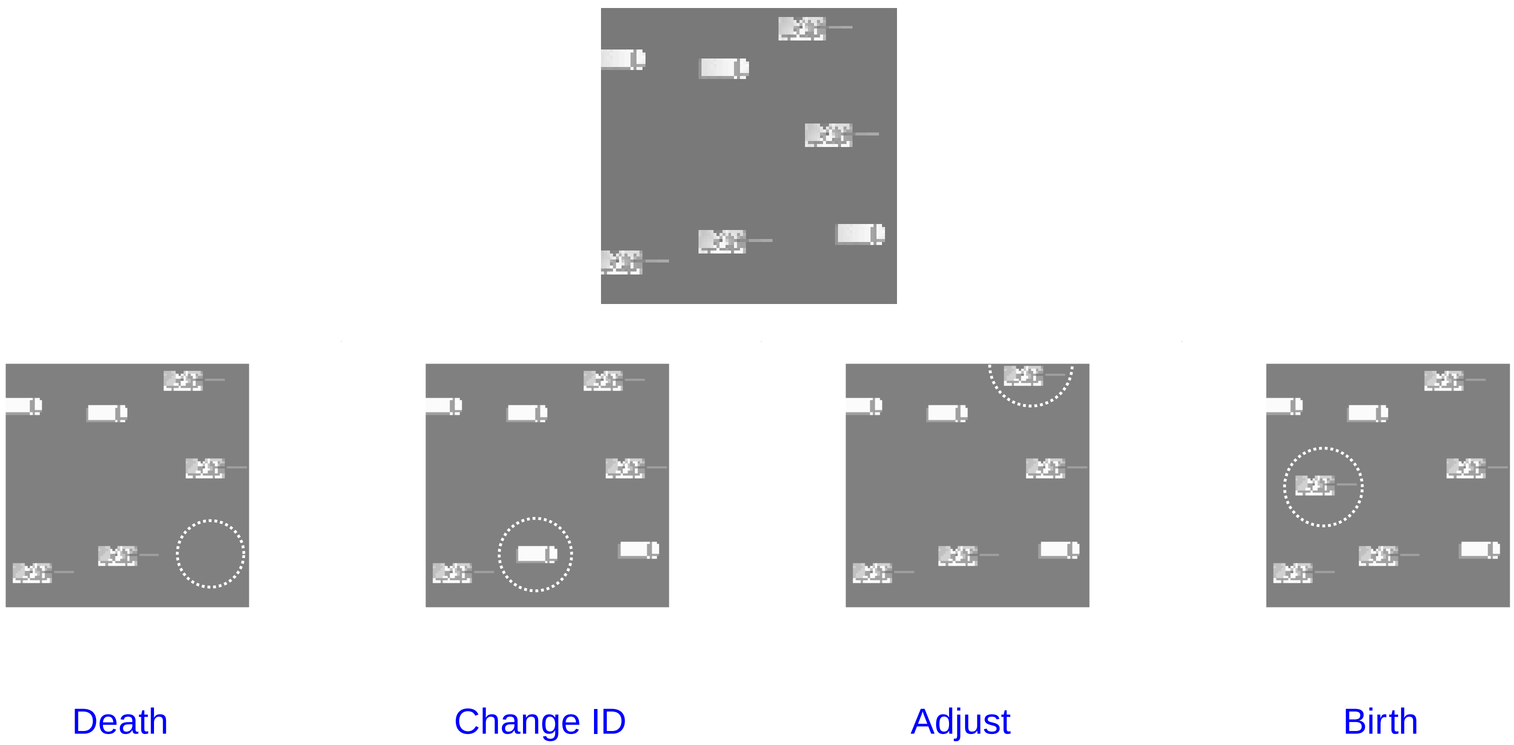
Given the current state $X^{(t)} \in \mathcal{X}$,

1. Generate $Y_t \sim G(y|X^{(t)})$. G is called the proposal distribution.
2. Set $X^{(t+1)} = \begin{cases} Y_t & \text{w.p. } \gamma(X^{(t)}, Y_t); \\ X^{(t)} & \text{w.p. } 1 - \gamma(X^{(t)}, Y_t), \end{cases}$ where

$$\gamma(x, y) = \min\left\{1, \frac{\nu(y)\,G(x|y)}{\nu(x)\,G(y|x)}\right\}.$$

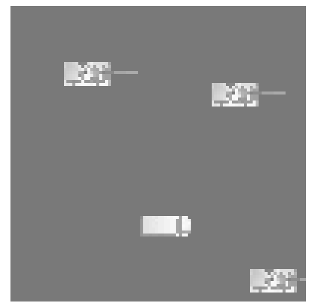
For a large class of proposal distributions G and for $X^{(1)} \sim F$ where F is an arbitrary probability distribution on $\mathcal{X}$, this algorithm is known to generate a Markov chain with unique stationary distribution ν.

Proposal Distribution: “Simple Moves”

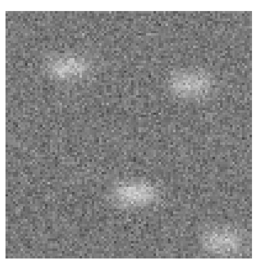


$G(y \mid X^{(t)})$ nominates a state from one of the indicated *neighborhoods*.

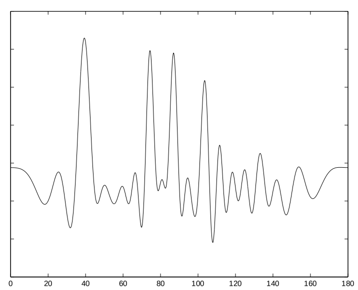
Example Scene & Sensor Data



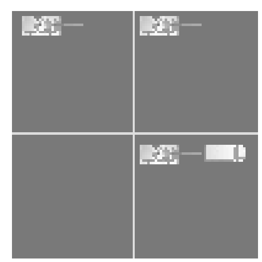
Original Scene



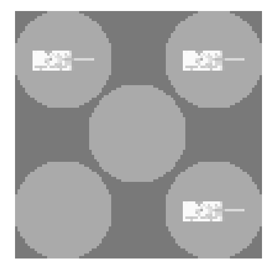
Infrared Image



Acoustic Signal

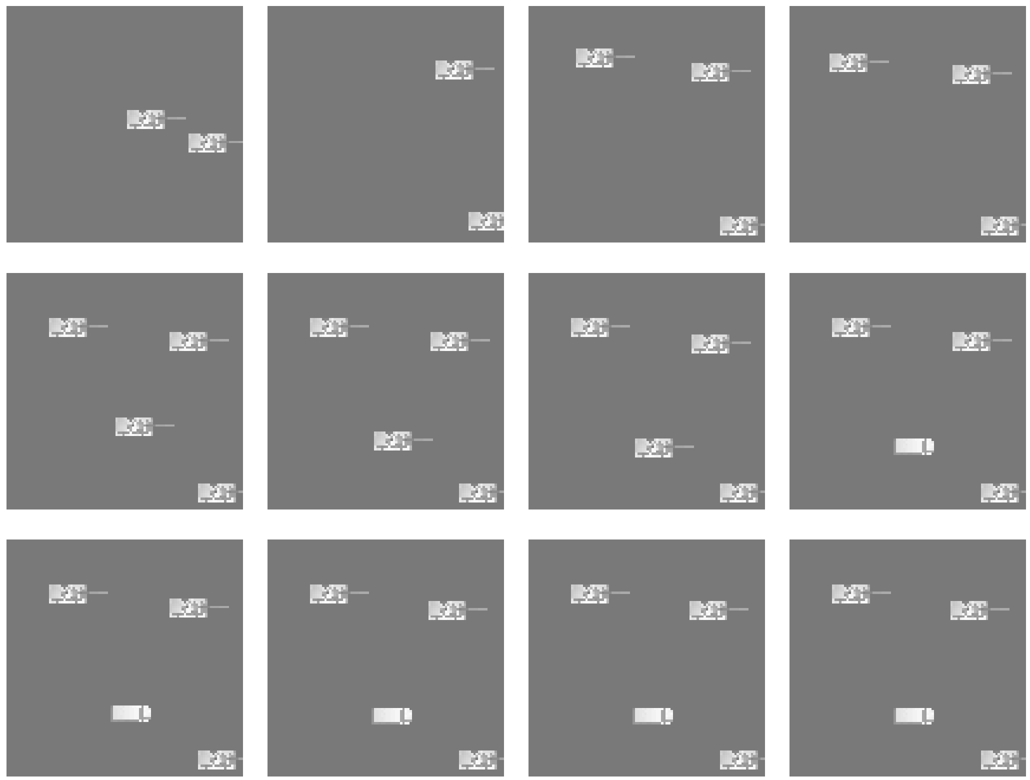


Scout's Report



Seismic Data

Evolution of the Markov Chain



Solution Original

Answering Tactical Questions

- Discard the first B states (*burn-in period*) and retain, for purposes of inference,

$$\{X^{(B+1)}, X^{(B+2)}, \ldots, X^{(B+R)}\}.$$

- Typical commander's question: *How many tanks are out there?*
- Let $A = \{X \in \mathcal{X} : \text{number of enemy tanks} \geq k\}$.
- The ergodic property of our Markov chain allows us to estimate the posterior probability of this event by $\frac{1}{R}\sum_{j=1}^{R} \mathbf{1}_A(X_j)$.
- If the commander requires this probability to be at least 0.95 (say), we may construct a simple rule based on our sample:

$$\frac{1}{R}\sum_{j=1}^{R} \mathbf{1}_A(X_j) \geq 0.95 \quad \Rightarrow \quad \text{Respond.}$$

Directions for Future Work

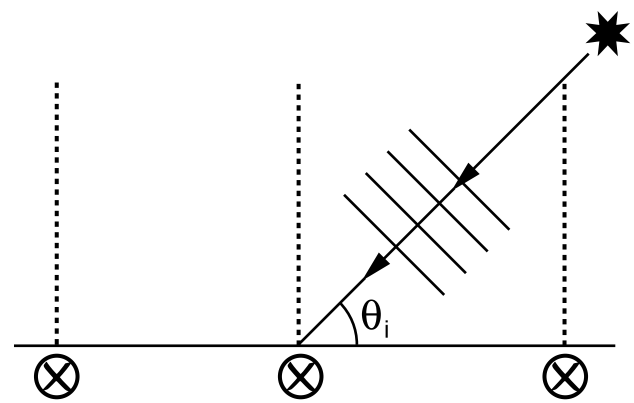
- Models for Additional Sensors — e.g., Magnetic Sensors
- Improved M-H Proposal Distribution G to Increase Acceptance Rate
- Designed Experiment to Estimate Parameters for Scout Likelihood
- Validation Using Real Data
- Recoding the Algorithm to Achieve Fast Execution

Likelihood for IR Image

$$L_1(Y_1 \mid X) = \prod_{i=1}^{rc} \frac{\big((I_0 * h)(z_i)\big)^{Y_1(z_i)}\, \mathrm{e}^{-(I_0 * h)(z_i)}}{Y_1(z_i)!}.$$

$$\hat{L}_1(Y_1 \mid X) = \frac{1}{Z} \exp\Big(-\frac{1}{2\sigma_1^2}\, \|Y_1 - I_0 * h\|_F^2\Big).$$

Likelihood for Acoustic Signal



$$L_2(Y_2 \mid X) = \frac{1}{Z} \exp\Big(-\frac{1}{\sigma_2^2} \, \|Y_2 - \sum_{i=1}^{n} d(\theta_i)\|^2 \Big).$$

$$d(\theta_i) = \big[\, 1, \exp\{-j\pi \cos(\theta_i)\}, \ldots, \exp\{-(m-1) j\pi \cos(\theta_i)\} \,\big]' \quad (j^2 = -1).$$

Neighboring States for DEATH and CHANGE ID

$$\mathcal{N}_D(X^{(t)}) = \begin{cases} \{X^{(t)}_{-j} : j = 1, \ldots, n\}, & \text{if } \|X^{(t)}\| \geq 1; \\ \{X^{(t)}\}, & \text{if } \|X^{(t)}\| = 0, \end{cases}$$

where $X^{(t)}_{-j}$ denotes the matrix $X^{(t)}$ after removing column j.

$$\mathcal{N}_C(X^{(t)}) = \begin{cases} \{X^{(t)}_{\Delta j} : j = 1, \ldots, n\}, & \text{if } \|X^{(t)}\| \geq 1; \\ \{X^{(t)}\}, & \text{if } \|X^{(t)}\| = 0, \end{cases}$$

where $X^{(t)}_{\Delta j}$ denotes the matrix $X^{(t)}$ after changing the identity component of column j.

Neighboring States for ADJUST and BIRTH

$$\mathcal{N}_A(X^{(t)}) = \begin{cases} \{X^{(t)}_{\oplus j} : j = 1, \ldots, n\}, & \text{if } \|X^{(t)}\| \geq 1; \\ \{X^{(t)}\}, & \text{if } \|X^{(t)}\| = 0, \end{cases}$$

where each $X^{(t)}_{\oplus j}$ denotes as many as eight perturbations to the location components of $X^{(t)}_j$.

$$\mathcal{N}_B(X^{(t)}) = \{X^{(t)}_\tau : \tau \in (\mathcal{D} \times \mathcal{A}) \setminus T_{X^{(t)}}\},$$

where $X^{(t)}_\tau$ is the augmentation of the matrix $X^{(t)}$ by one additional column τ corresponding to any "legal" target not already present: $\|X^{(t)}_\tau\| = \|X^{(t)}\| + 1$.

Proposal Distribution for Metropolis-Hastings

$$
\begin{aligned}
G(y \,|\, X^{(t)}) \;&=\; w_D \, \frac{1}{|\mathcal{N}_D(X^{(t)})|} \, \mathbf{1}_{\mathcal{N}_D(X^{(t)})}(y) \\
&+\; w_C \, \frac{1}{|\mathcal{N}_C(X^{(t)})|} \, \mathbf{1}_{\mathcal{N}_C(X^{(t)})}(y) \;+\; w_A \, \frac{1}{|\mathcal{N}_A(X^{(t)})|} \, \mathbf{1}_{\mathcal{N}_A(X^{(t)})}(y) \\
&+\; w_B \, \mathbb{P}_{T_{X^{(t)}}}(\tau) \, \mathbf{1}_{\mathcal{N}_B(X^{(t)})}(y),
\end{aligned}
$$

where $\mathbb{P}_{T_{X^{(t)}}}(\cdot)$ is a probability mass function on $(\mathcal{D} \times \mathcal{A}) \setminus T_{X^{(t)}}$ and where we introduce fixed positive weights satisfying $w_D + w_C + w_A + w_B = 1$.

Visual Analytics for Streaming Internet Traffic

Edward J. Wegman

George Mason University

Karen Kafadar

University of Colorado, Denver

Visual Analytics for Streaming Internet Traffic

- The following discussion is based on the following three papers
  - Wegman, E. and Marchette, D. (2003) “On some techniques for streaming data: A case study of Internet packet headers,” *Journal of Computational and Graphical Statistics*, 12(4), 893-914
  - Marchette, D. and Wegman, E. (2004) “Statistical analysis of network data for cybersecurity,” *Chance*, 17(1), 8-18
  - Kafadar, K. and Wegman, E. (2004) “Visualizing ‘typical’ and ‘exotic’ Internet traffic data,” Proceedings of COMSTAT2004.

Visual Analytics for Streaming Internet Traffic

- Introduction
- Visual Analytics
  - Analysis versus Exploration
- Block Recursion and Evolutionary Graphics
  - Waterfall plots
  - Skyline plots
  - EWMA plots

Four Stages of Data Graphics

1. Static Graphics
   - Ed Tufte’s books
   - Trellis plots, scatterplot matrices, parallel coordinate plots, most density plots
   - Most (paper) published materials
   - Perhaps some color and anaglyph stereo
2. Interactive Graphics
   - Data objects created, but underlying data untouched
   - Think of data on server, graphics on client
   - Brushing, saturation brushing, 3-D (stereoscopic) plots
   - Rocking and rotation, Dan Carr’s micromaps, cropping and cutting, linked views

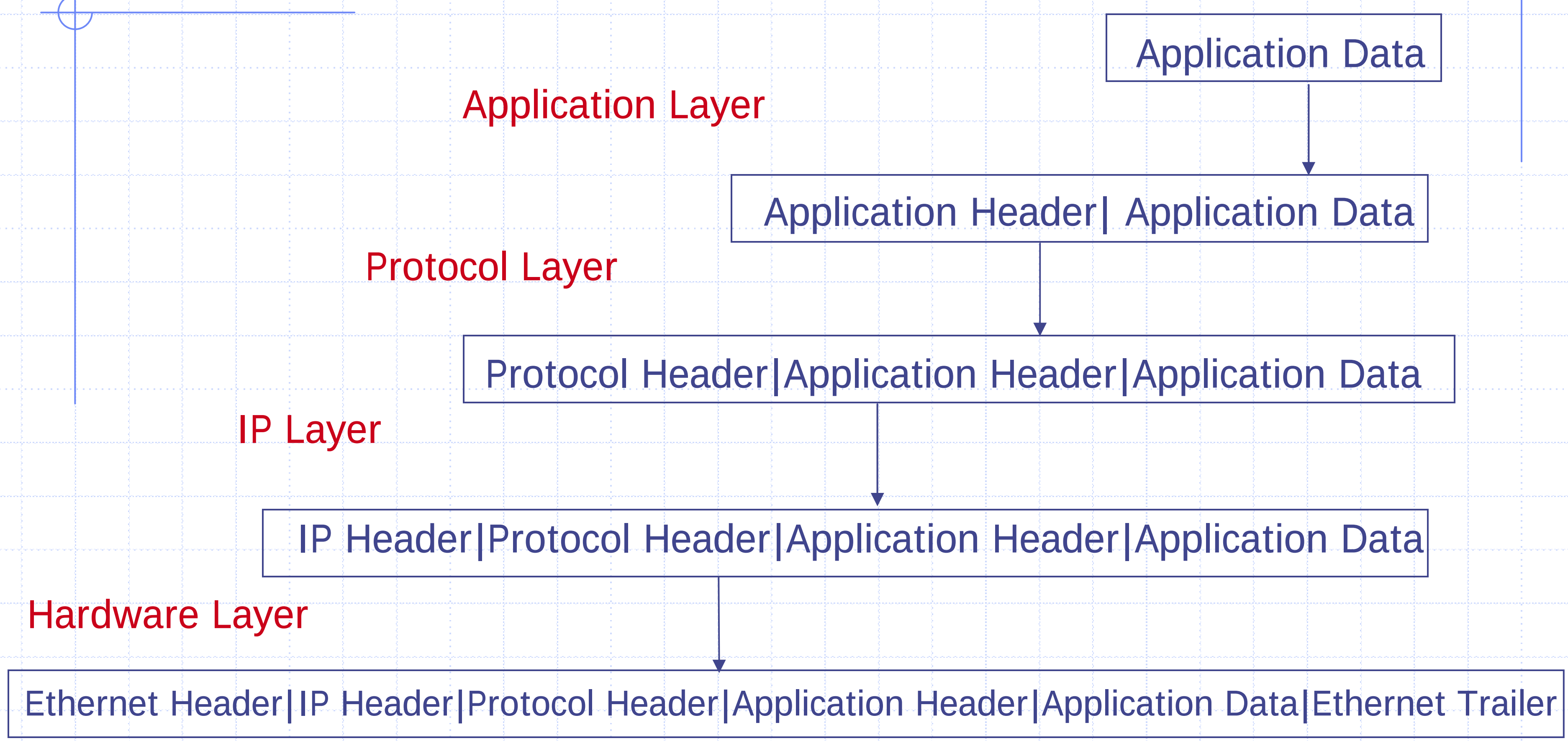
Four Stages of Data Graphics

3. Dynamic Graphics
   - Data that must be interacted with, not just client based
   - Grand tour, multidimensional grand tour, recursive or dynamically smoothed density plots, dynamic smoothing including mode trees and mode forests, Dan Carr’s conditioned chloropleth maps, pixel tours, cross corpora discovery
4. Evolutionary Graphics
   - Fixed data sets that are evolving
     - Data Set Mapping
     - Iterative Denoising
   - Streaming data
     - Recursion and Block Recursion
     - Visual Analytics
     - Waterfall, Transient Geographic Mapping, Skyline Plots

Visual Analytics for Streaming Internet Traffic - Types of Networks

- Class A – field1 identifies the network, fields2-4 identify the specific host
  - field1 is smaller than 127, e.g. 1.1.1.1
- Class B – field1.field2 identifies the network field3.field4 identifies the specific host, field3 sometimes used for subnet
  - Field1 is larger than 127, e.g. 130.103.40.210
- Class C- field1.field2.field3 identifies the network, field4 the host
  - E.g. 192.9.200.15

Visual Analytics for Streaming Internet Traffic - TCP/IP Addressing



Visual Analytics for Streaming Internet Traffic - Common Protocols

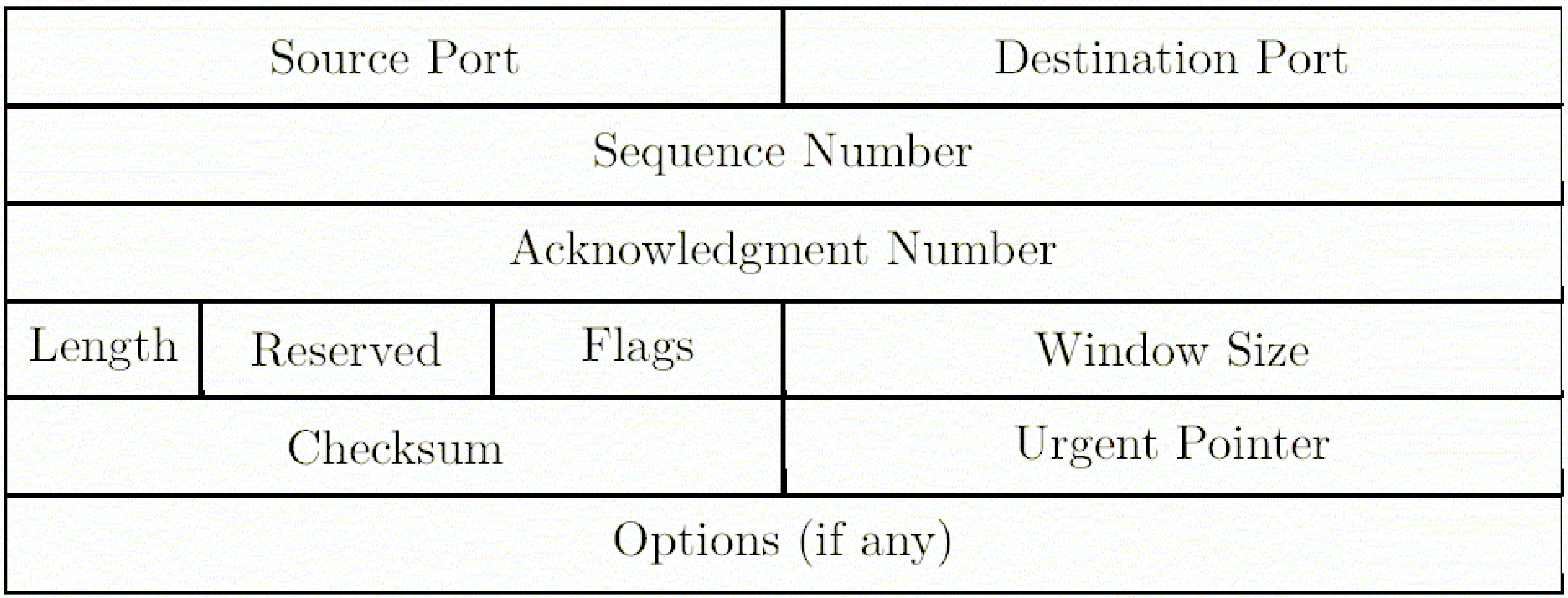
- TCP=Transmission Control Protocol
- UDP=User Datagram Protocol
- ICMP=Internet Control Message Protocol

Visual Analytics for Streaming Internet Traffic - TCP/IP Addressing

Version	Length	Type of Service	Total Packet Length	
Identification			Flags	Fragment Offset
Time to Live		Protocol	Header Checksum	
Source IP Address				
Destination IP Address				
Options (if any)				

The IP Header

Visual Analytics for Streaming Internet Traffic - TCP/IP Addressing



TCP Packet Header

Visual Analytics for Streaming Internet Traffic - TCP/IP Addressing

- Some Flag Types
  - ACK – used to acknowledge receipt of a packet
  - PSH – data should be pushed to application ASAP
  - RST – reset
  - SYN – synchronize connection so each host knows order of packets
  - FIN – finish the connection

Visual Analytics for Streaming Internet Traffic - TCP/IP Addressing

HOST 1	HOST 2
SYN	
	SYN/ACK
ACK	
PSH	
PSH	
PSH	
	ACK
	PSH
ACK	
FIN	
	FIN/ACK
	PSH
ACK	
	FIN
FIN/ACK	

Possible TCP Session

Visual Analytics for Streaming Internet Traffic - Ports

- There are some $2^{16} = 65{,}536$ ports for each host
  - Some standard services use standard ports
    - e.g. ftp – 21, ssh – 22, telnet – 23, smtp – 25, http – 80, pop3 – 110, nfs – 2049, even directv and aol have standard ports.
  - Unprotected (open) ports allow possible intrusion
    - Scanning for ports is a hacker attack strategy

Evolutionary Graphics from Streaming Internet Data

The major problem is to detect intrusions or unwanted events in streaming Internet traffic.

Evolutionary Graphics from Streaming Internet Data

1. Internet traffic is a prototypical example of streaming data and prefigures future streaming data types. I believe streaming data represents a fundamentally new data structure.
2. The papers mentioned above describe the basic protocols for Internet traffic. We look only at time stamps, destination IP, destination port, source IP, source port, number of bytes, number of packets, duration of session.
3. We ignore the data content of the packet, and seek to make inferences based only on the header data described above.

Evolutionary Graphics from Streaming Internet Data

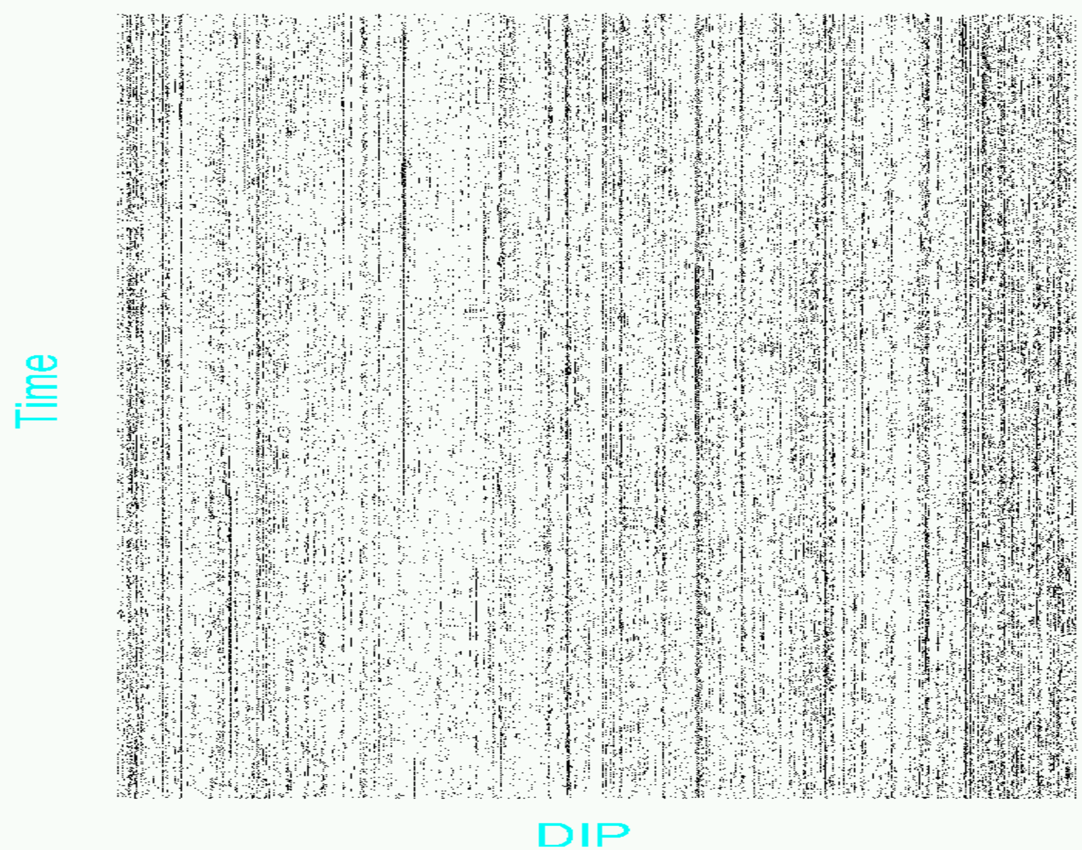
1. Streaming data arrives as such a rate that it is impossible to store the data.
   - We collect 26 terabytes of Internet header data per year.
   - Naively, we look at a data item, update a recursive algorithm, and discard the data.
2. Some suggestions we have made include:
   - Recursive formulations of counts and moments
   - Pseudo-samples based on geometric quantization
   - Recursive formulations of kernel and adaptive mixture density estimators
   - Exponentially weighted moving averages including exponentially weighted kernel smoothers.

But this talk is about **evolutionary graphics**

1. In the simplest framework, the idea is to accumulate data for a very small epoch (even instantaneously), plot the new data, and discard the old.
   - In practice for Internet traffic, the epoch may last for perhaps 10 milliseconds.
   - Our initial suggestion is a Waterfall diagram. The first epoch is plotted at the top of the graphic. As additional data are accumulated during the second epoch, the graphic for the first epoch is pushed down and the second epoch is now plotted on the top. This continues until perhaps 1000 epochs have passed. Thus the oldest epoch drops off the bottom of the page and the new replaces it at the top. The graphic evolves and is new every 10 seconds.

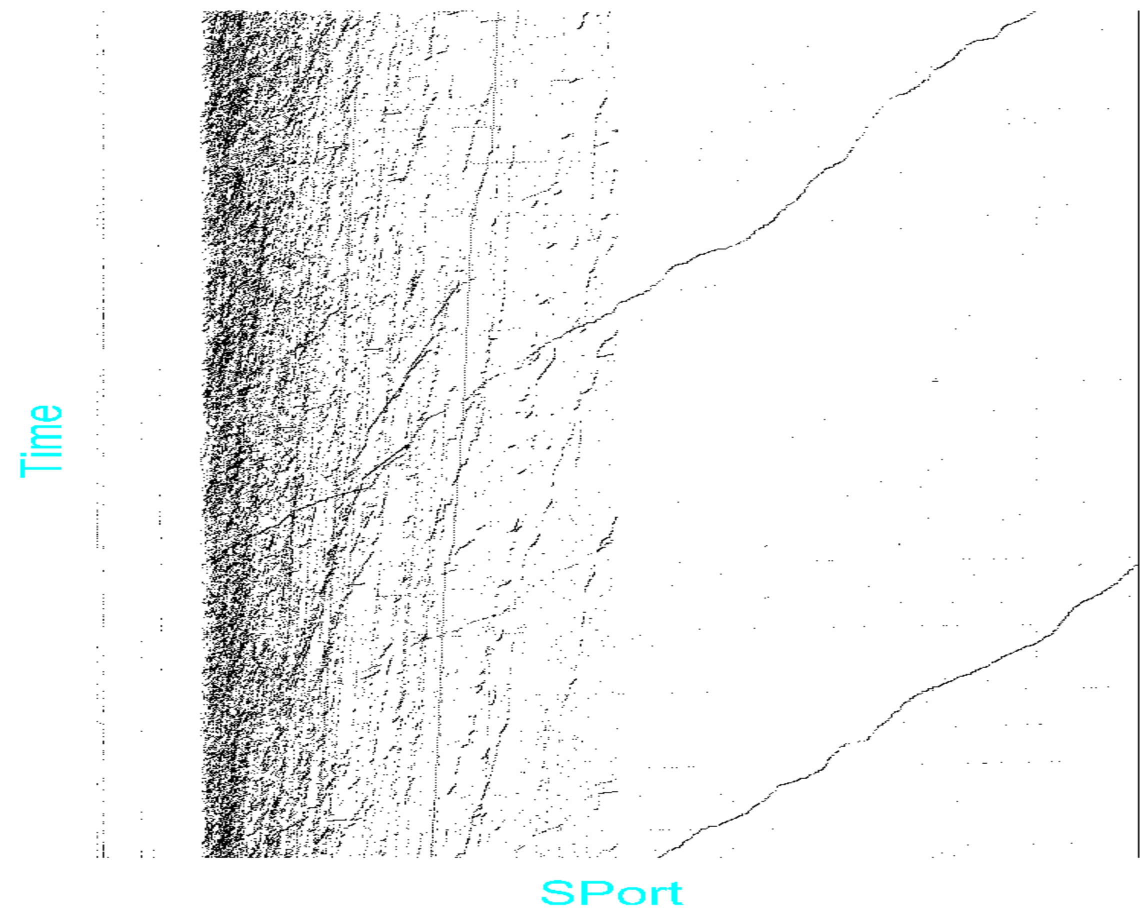
Evolutionary Graphics with Explicit Dependence on Time

Evolutionary Graphics from Streaming Internet Data



Waterfall for Destination IP versus Time for only one hour

Evolutionary Graphics from Streaming Internet Data



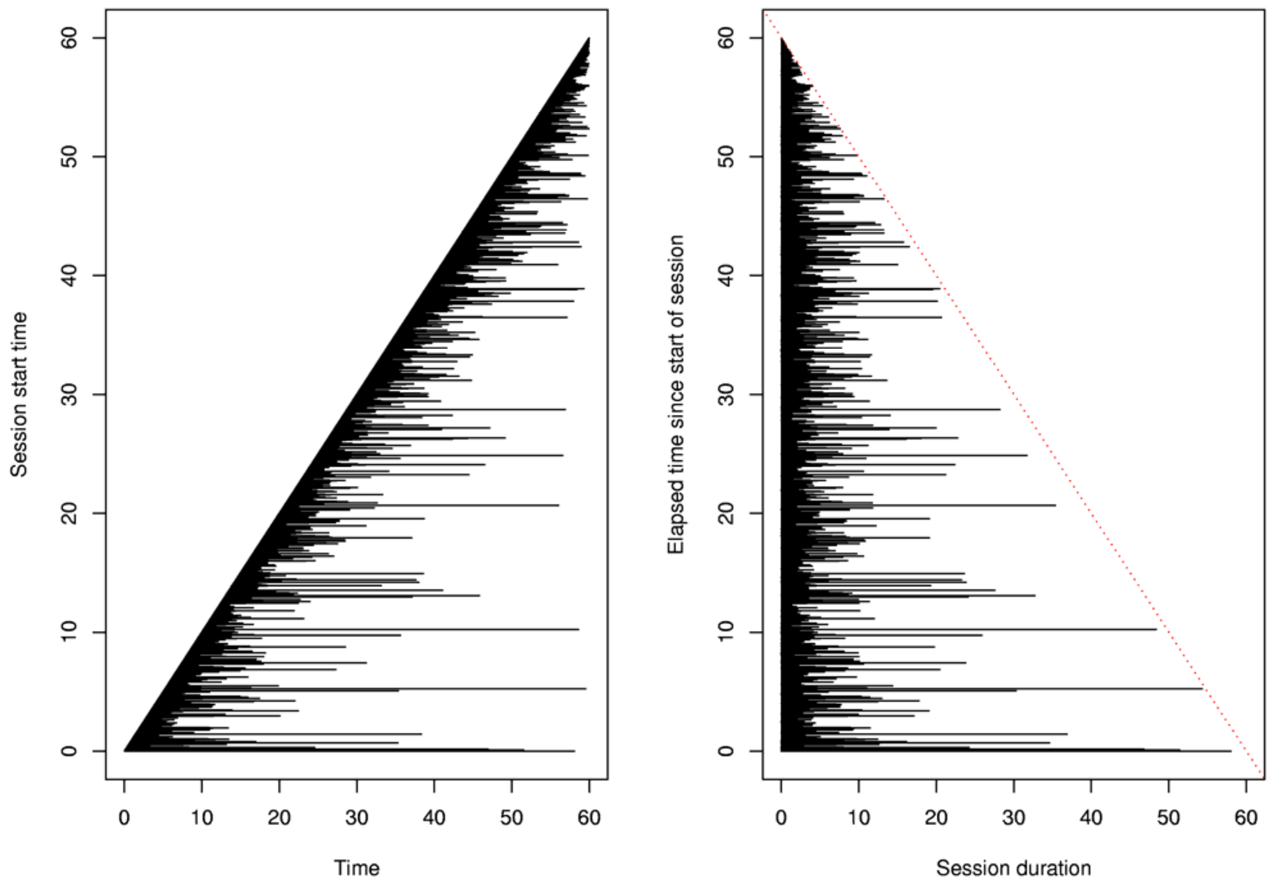
Waterfall for Source Port versus time. Diagonals are characteristics of distinct operating systems.

Evolutionary Graphics from Streaming Internet Data

The preponderance of relatively short sessions can be seen in the next figure, which displays the session durations as horizontal lines that extend from the start time to the end time. Because these data are collected in the order in which they occurred, the session start times range from time 0 (bottom line) to 59.971 (nearly the end of the hour).

The second figure shows the same information, but each line is shifted back to 0. The data are censored after one hour. With continuously monitored data, the session duration lines would continue past the censoring point. Most data are not censored because 92.3% of the sessions lasted less than 30 seconds.

Evolutionary Graphics from Streaming Internet Data



Evolutionary Graphics from Streaming Internet Data

I want to introduce two critical ideas for streaming data.

1. Block Recursion
    - Instead of adjusting the statistic or the graphic by the new observation, keep a small number of observations in a moving window.
    - Adjust the statistic or graphic by dropping off the effect of the oldest observation and inserting the effect of the newest.
    - The graphic need not explicitly how dependence on time.
2. Visual Analytics
    - Be willing to dynamically transform variables so as to exploit structure.

Examples of Block Recursion Graphics that do not Explicitly Show a Time Variable

Evolutionary Graphics from Streaming Internet Data

Most destination port numbers occur only once or twice during the hour; of the 380 distinct DPorts, 293 occurred only once, 47 occurred twice, 8 occurred 3 times, 5 occurred 4 times.

The remaining 27 ports occurred more than 5 times; the exceptional counts are DPort 80 (web, 116,134 times), 25 (mail-smtp, 6,186 times), 443 (secure web, 11,627), 554 (streaming video/audio, 200 times), and 113 (128 times).

Setting aside the “well-known” ports 0-1023, we plot the occurrence of destination ports numbered 1024 and above, which should arise more or less at random, and flag as unusual any Dport that is referenced more than 10 times.

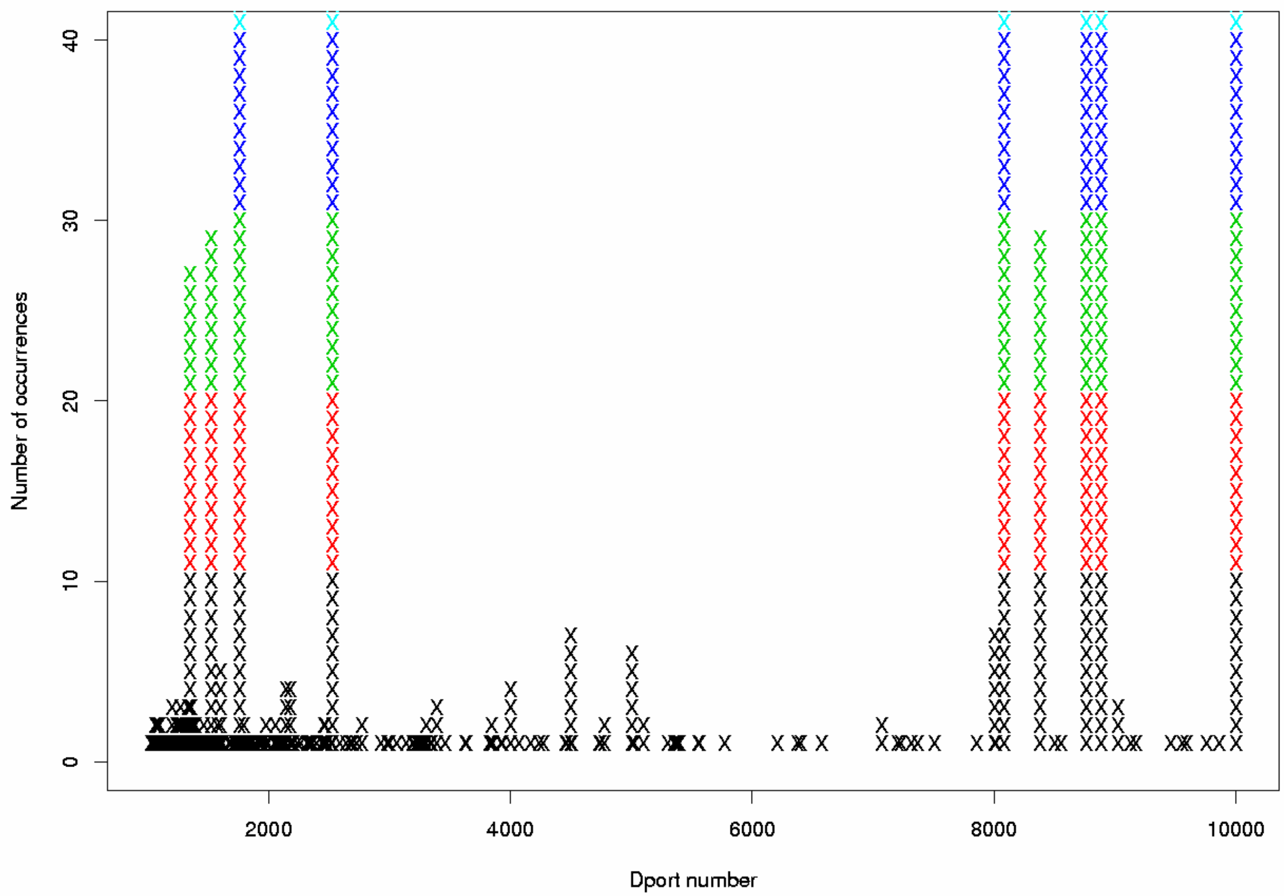
Evolutionary Graphics from Streaming Internet Data

The next figure shows such a plot; once a destination port number occurs more than 10 times, the color changes to red, indicating potentially high traffic on this destination port.

The construction of this plot resembles the tracing of a skyline, so we call it a “skyline plot.” A similar figure can be used to monitor SPort activity; however, the plot is more dense because this file contained 6,742 unique SPorts, versus only 380 distinct DPorts.

Also, while most of the 380 DPorts appeared only once in the file, a SPort typically occurred four times, with 20% of the 6742 accessed source ports occurring between 14 and 88 times.

Evolutionary Graphics from Streaming Internet Data



Animation of this plot could easily be accomplished as new data enter the block and old data depart.

Evolutionary Graphics from Streaming Internet Data

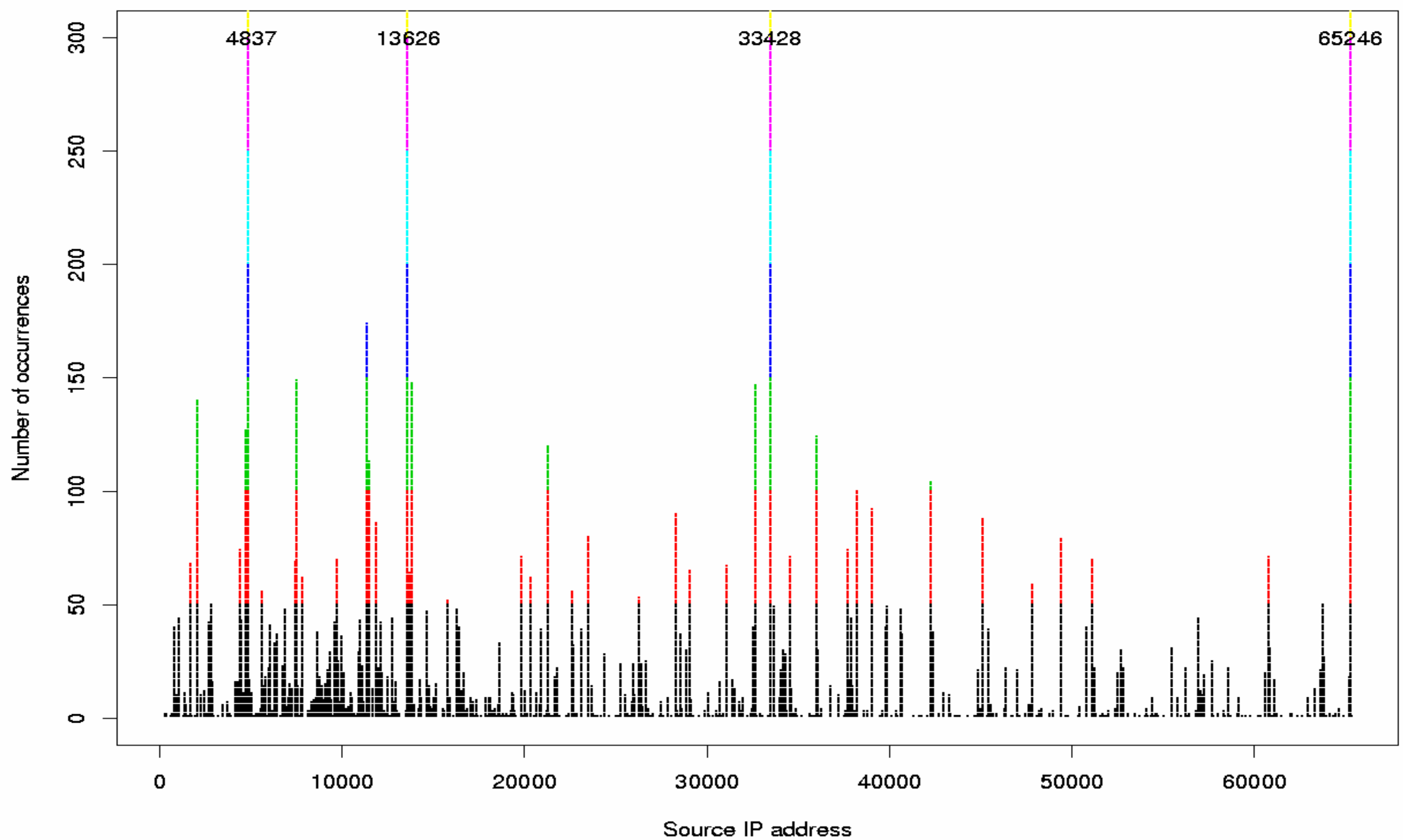
Source and destination IP addresses can be monitored also using skyline plot. In the present data file, source IP addresses are numerous (2504 unique SIPs) and frequent (the median number of occurrences is 4, and 10% occur more than 135 times).

The next figure shows this type of plot for source IP addresses in the first 10000 records, where the colors change as the number of hits exceeds multiples of 50.

Four unusually frequent source IP addresses are immediately evident in this plot: 4837, 13626, 33428,and 65246, which occur 371, 422, 479, and 926 times, respectively, in the first 10,000 sessions.

The limit for unusually frequent SIP addresses may depend upon the network and the time of day, so the limits on this “skyline” plot may need to be adjusted accordingly.

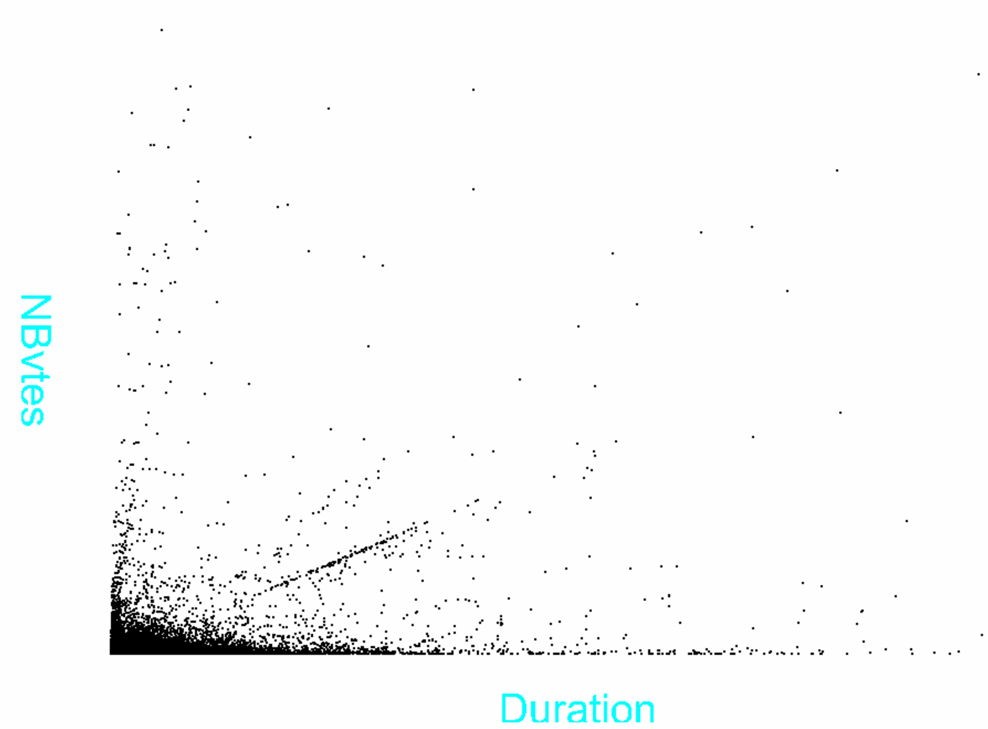
Evolutionary Graphics from Streaming Internet Data



Evolutionary Graphics from Streaming Internet Data

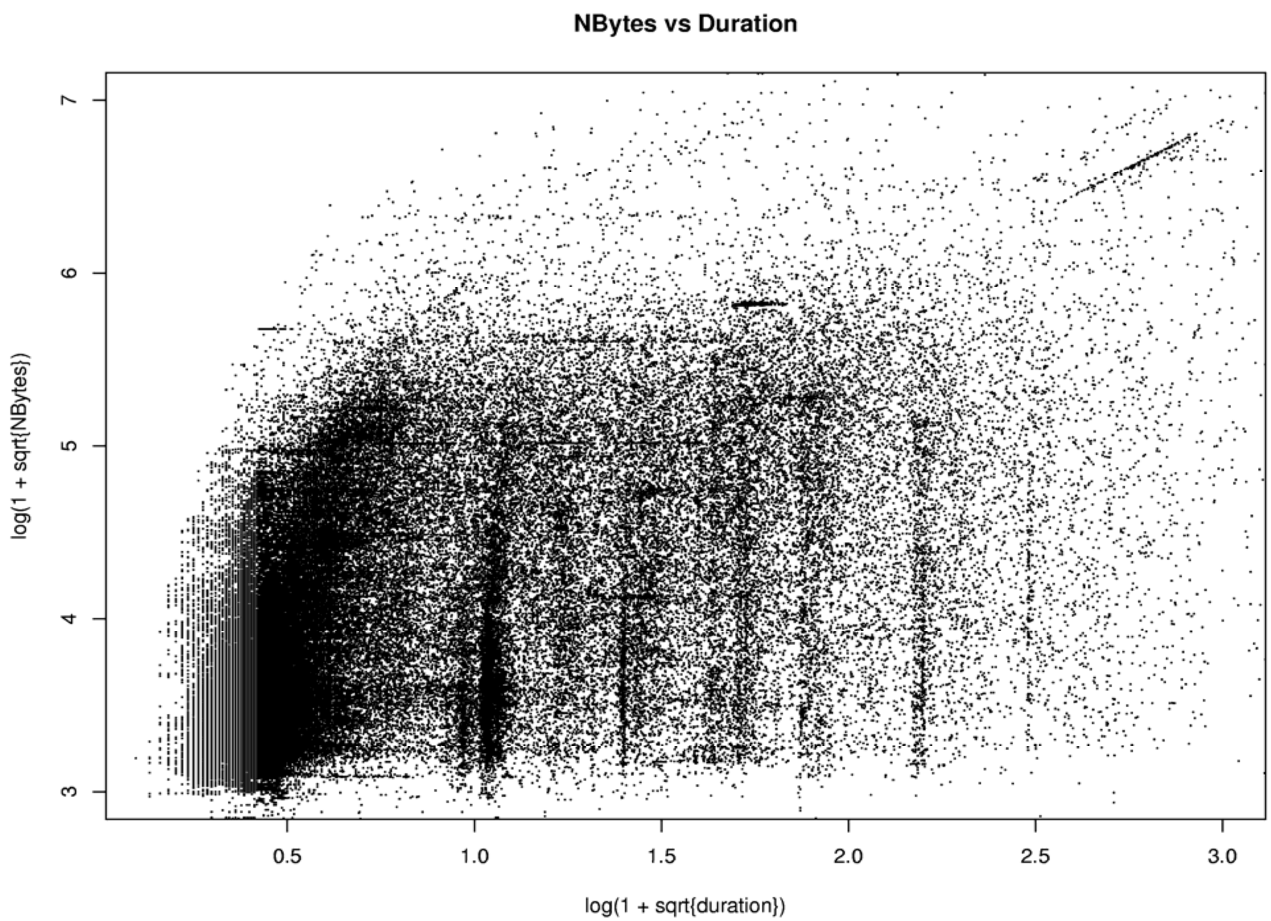
Dynamic Transformation of Variables may be Extremely Helpful in Evolutionary Graphics. In General, for Streaming Internet Data, “Size Variables” Tend to Cluster Near Zero, so LOG and/or SQRT Transformations Are Helpful for Visual Analytics

Evolutionary Graphics from Streaming Internet Data



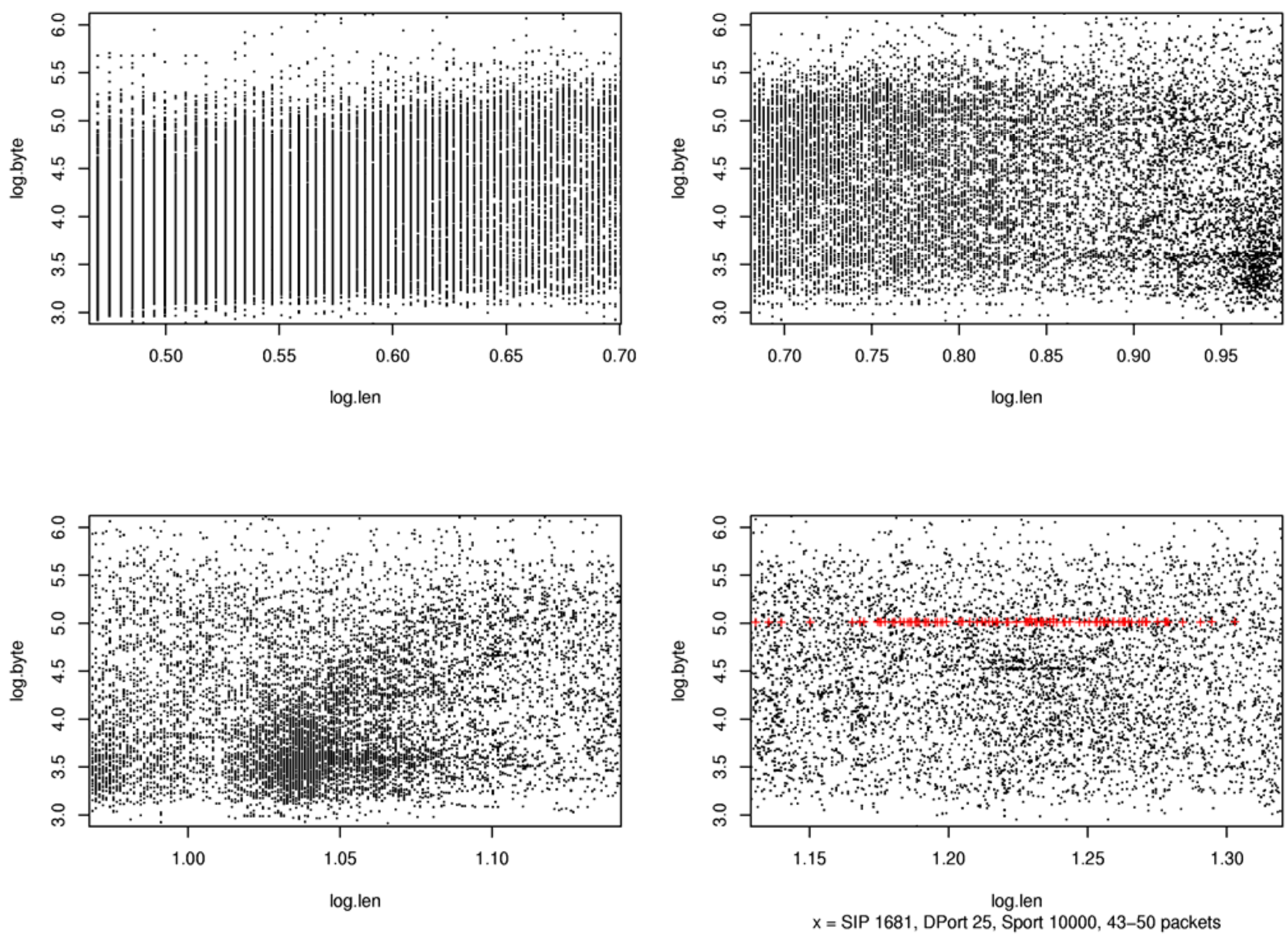
Number of Bytes versus Duration at full scale with no transformation.

Visual Analytics for Streaming Internet Traffic



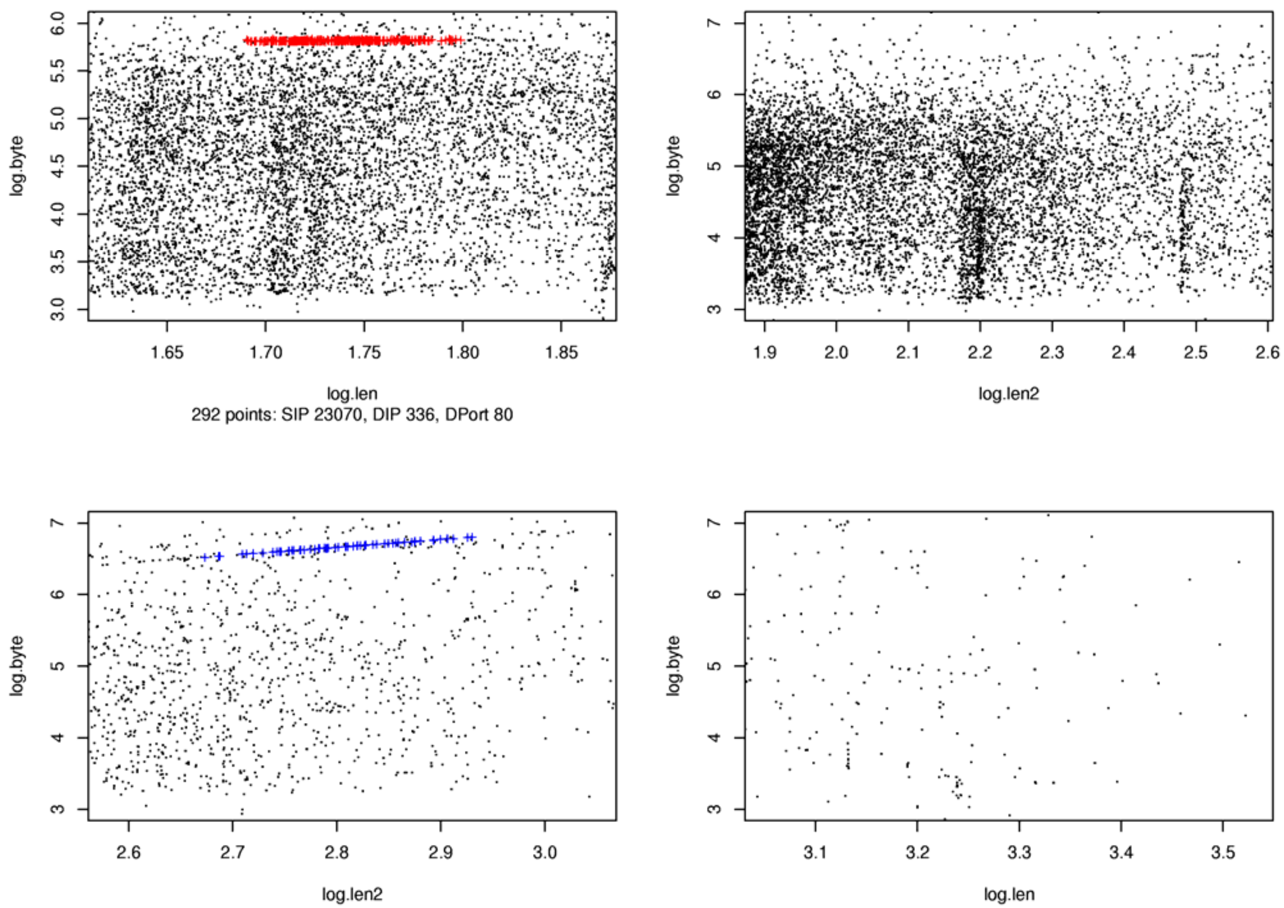
Same data as previous image with *log(1+sqrt(.))* transformation applied. Notice the additional structure visible.

Visual Analytics for Streaming Internet Traffic



Conditional plots show additional points of interest.

Visual Analytics for Streaming Internet Traffic



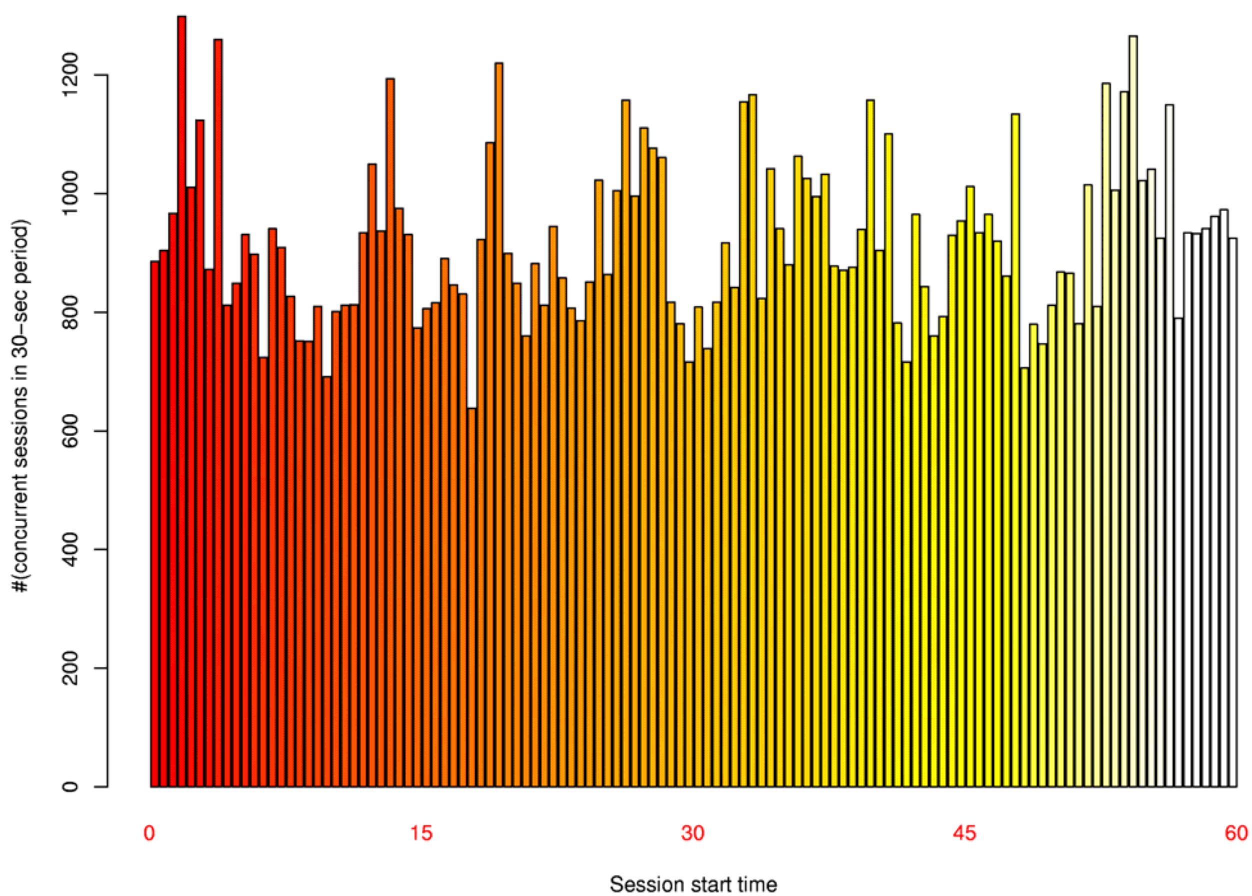
Evolutionary Graphics from Streaming Internet Data

The next figure shows a barplot of the number of active sessions during each 30-second subset of this one-hour period (a time frame of 30 seconds is selected to minimize the correlation between counts in adjacent bars).

The mean number of active sessions in any one 30-second interval during this hour is 923, and the standard deviation is about 140, suggesting an approximate upper 3-sigma limit of 1343 sessions.

A square root transformation may be appropriate. The mean and standard deviation of the square root of the counts is 30.29 and 2.23, respectively, resulting in an approximate upper 3-sigma limit of 1367, very close to the limit on the raw counts, since the Poisson distribution with a high mean resembles closely the Gaussian distribution.

Evolutionary Graphics from Streaming Internet Data



An Evolutionary Graphic

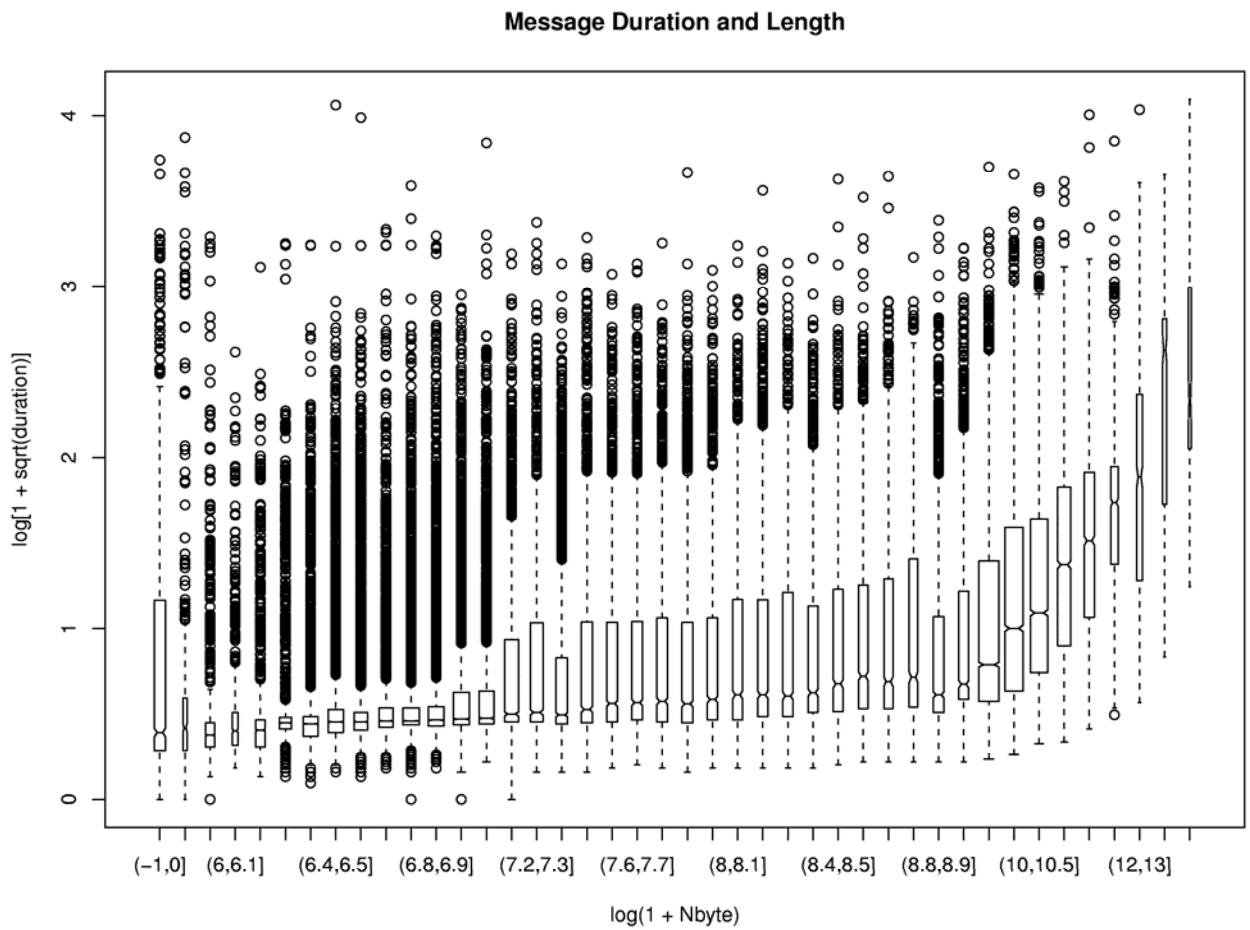
Visual Analytics for Streaming Internet Traffic

Box plots are useful for displaying the relationship between two variables. In the next plot, we plot log.len versus log.byte.

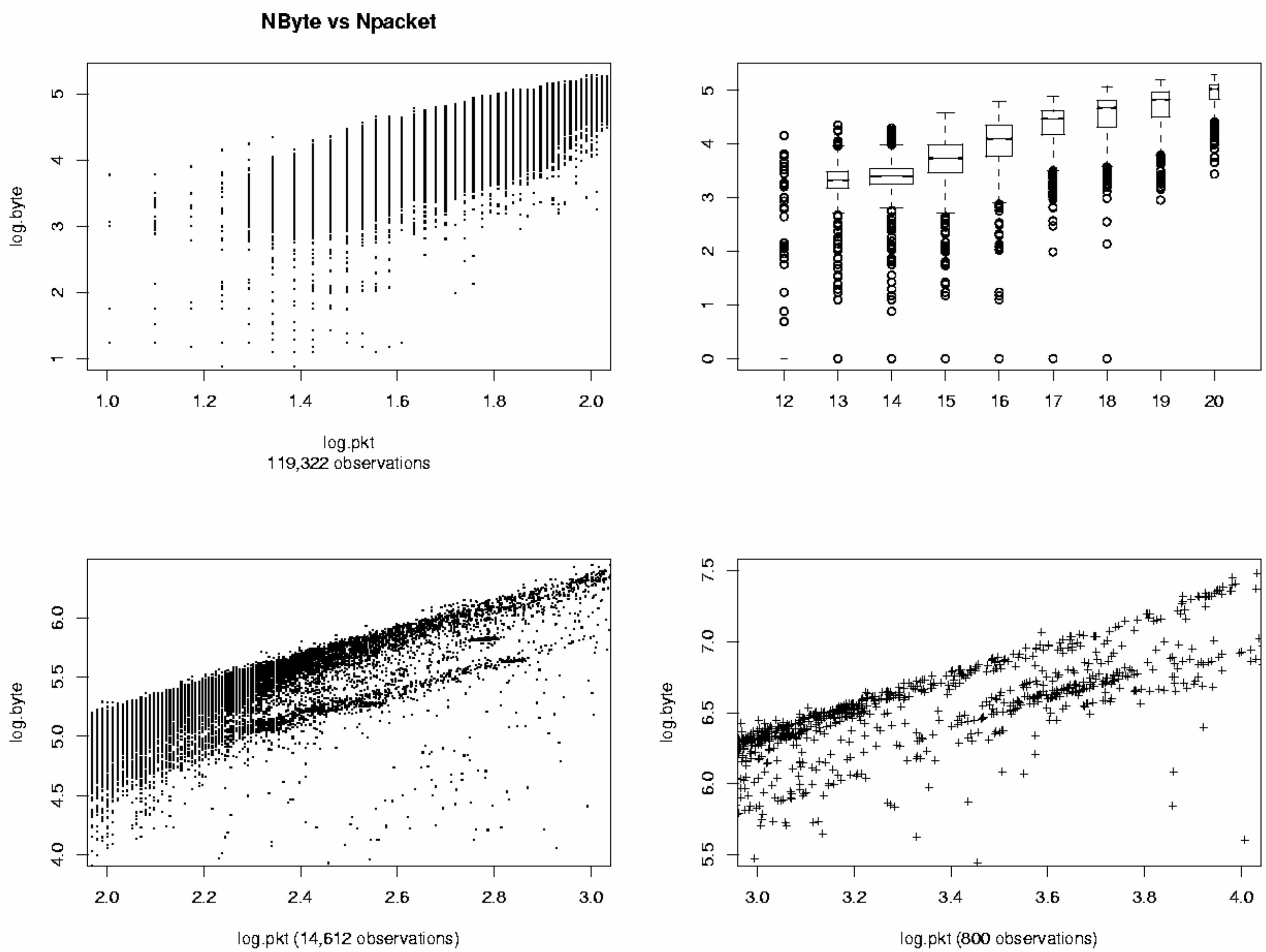
The first box contains the 2611 values for which Nbyte is zero, the next box contains 1216 values where $1 < Nbyte \leq 365$.

The distributions are clearly skewed to the left (a lot of outliers of large values. The distribution of duration as a function of Nbytes is fairly smooth and has a reasonable trend upwards.

Visual Analytics for Streaming Internet Traffic



Visual Analytics for Streaming Internet Traffic

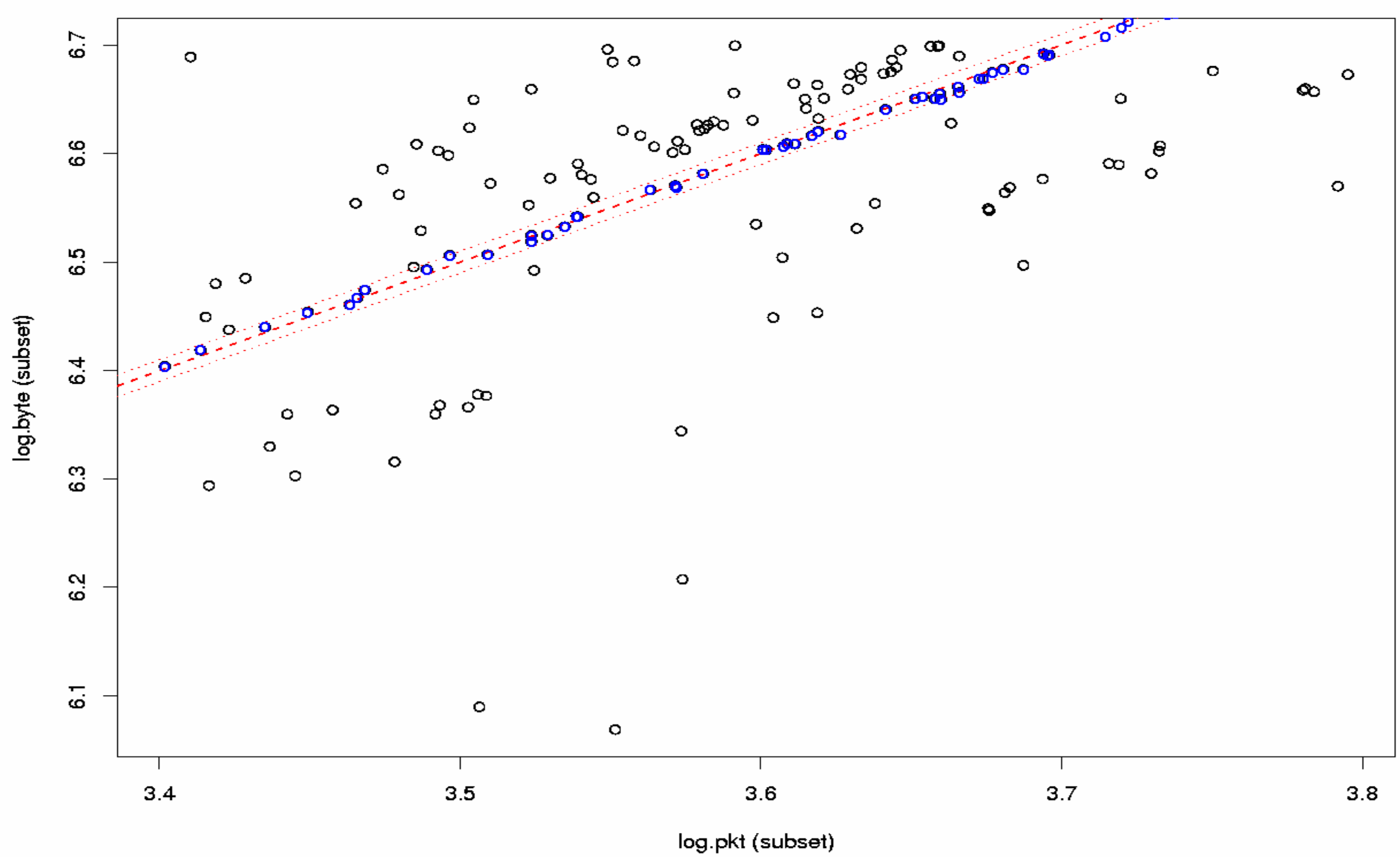


Visual Analytics for Streaming Internet Traffic

The dense set of 55 points (points in Panel (d), I.e. $3.4 < log.pkt < 3.8$) are plotted in the next figure; they lie on or very near the line $log.byte = 3 + log.pkt$, and 43 of them correspond to DPort 43.

The extent to which such a pattern could occur by chance alone should be investigated, particularly if they occur all within a few seconds of each other (these did not).

Visual Analytics for Streaming Internet Traffic



Visual Analytics for Streaming Internet Traffic

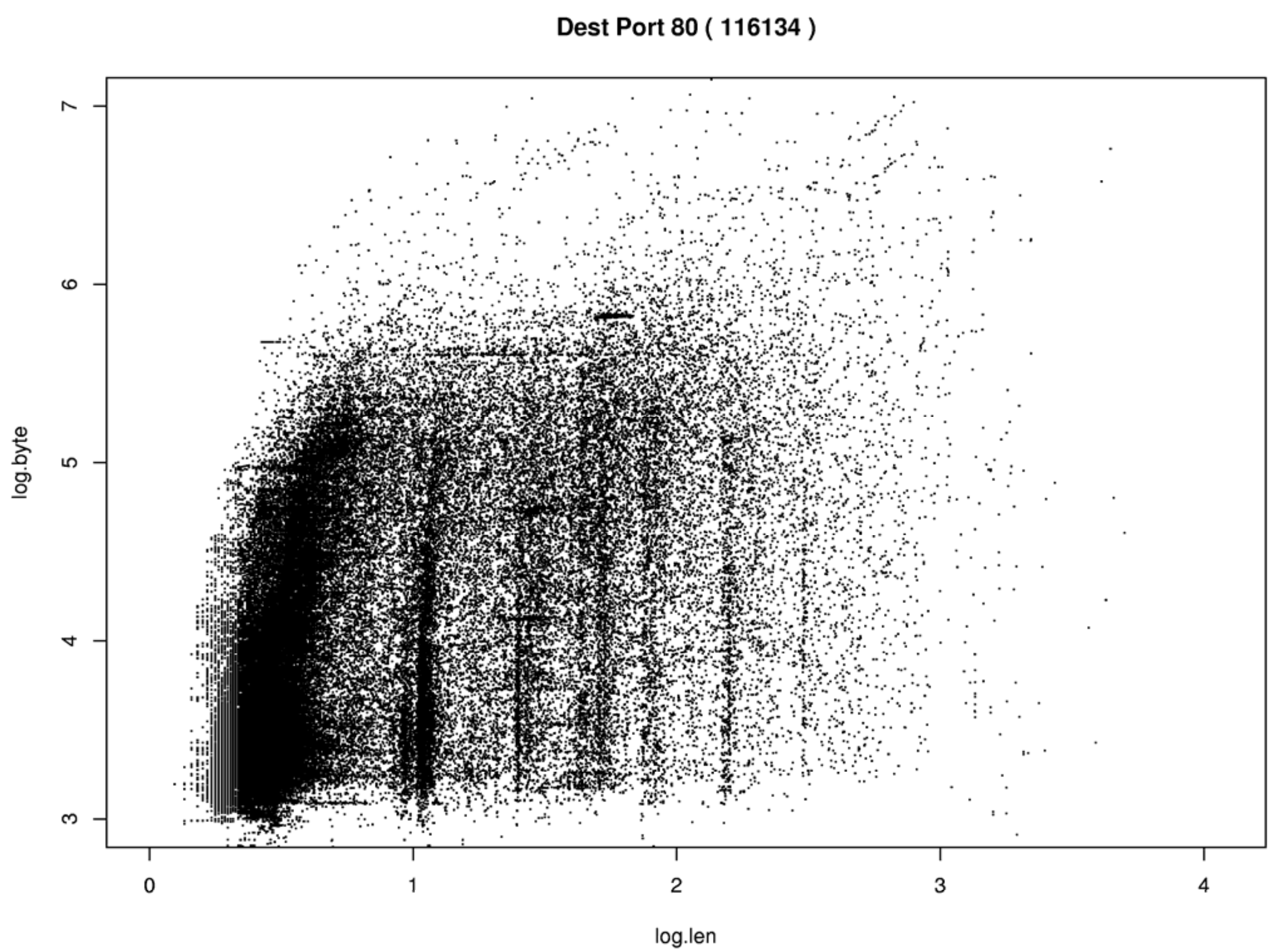
This hour of internet activity involved 380 different destination ports, DPort.

1. DPort 80 (web) is the most common, comprising 116,134 of the 135,605 records.
2. The next most common destination port is DPort 443 (secure web, https), used 11,627 times.
3. Followed by DPort 25 (mail SMTP) accessed 6,186 times.
4. Ports 554, 113, 10000, 8888 occur 200, 128, 97, 94 times, respectively.
5. Displaying all 135,605 points on one plot is not very informative, so instead we provide conditional plots according to their destination ports.

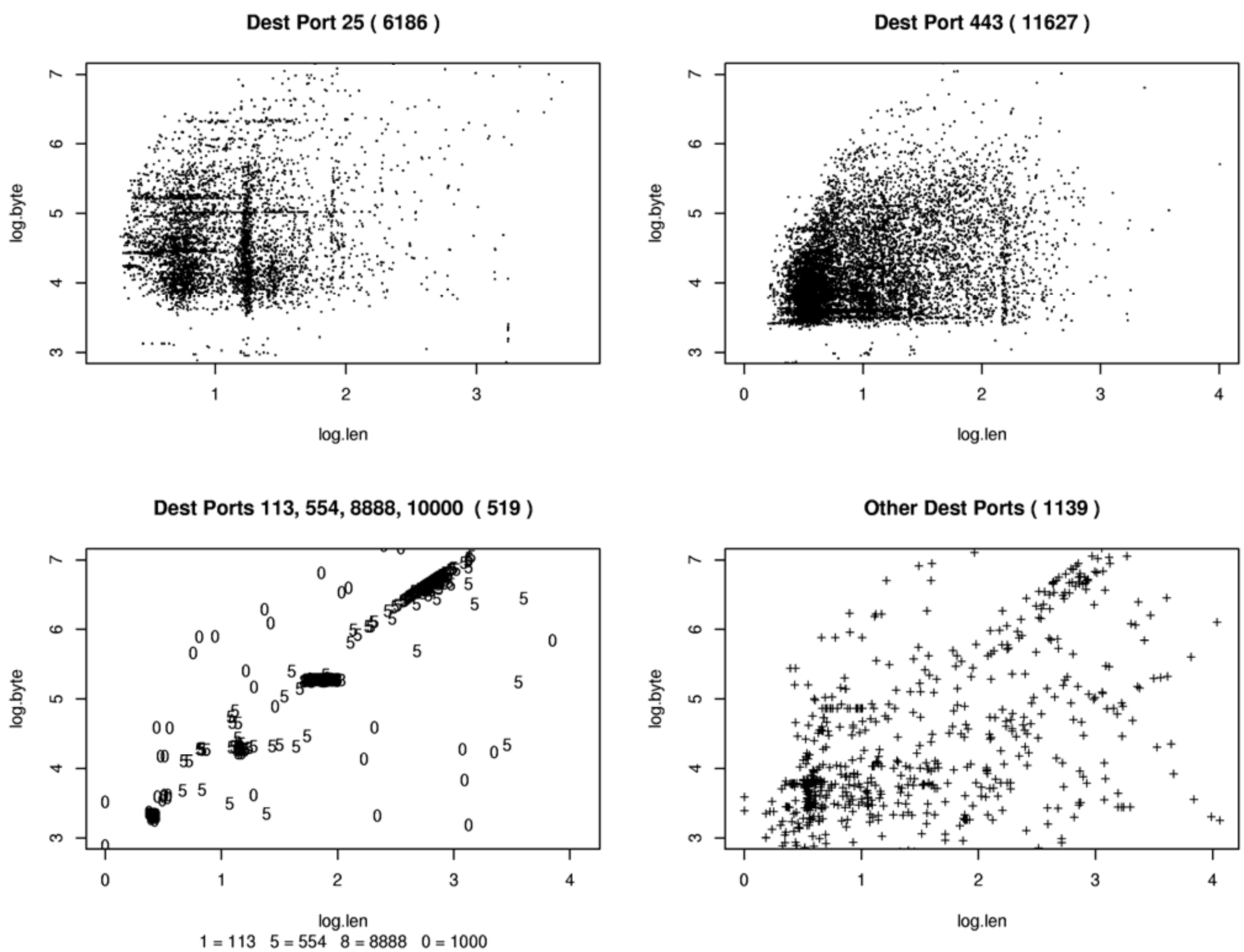
Visual Analytics for Streaming Internet Traffic

The next figure plots log.byte versus log.len for only the web sessions. A few horizontal lines of the sort noticed in previous plots appear, but otherwise no real structure is apparent.

Visual Analytics for Streaming Internet Traffic



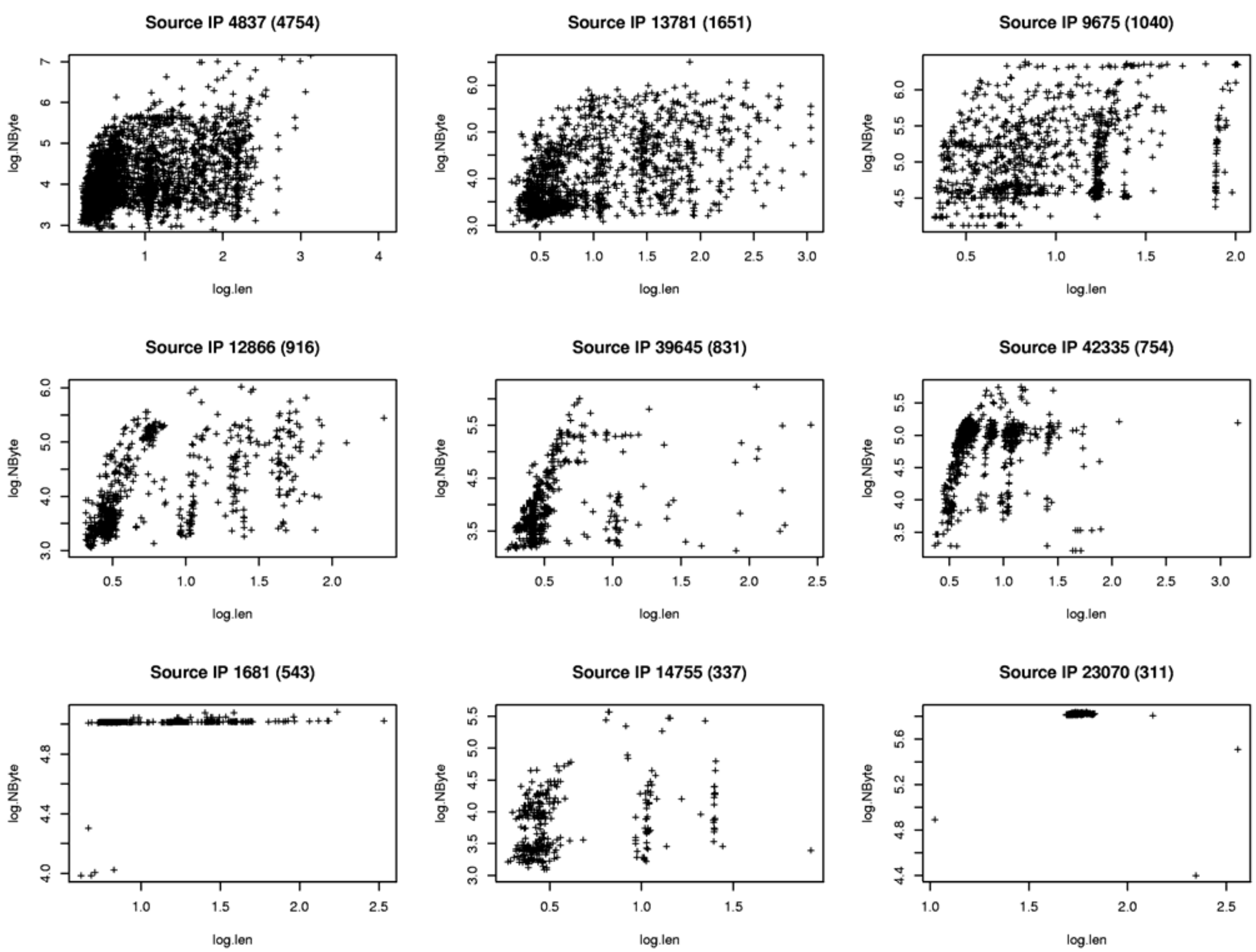
Visual Analytics for Streaming Internet Traffic



Visual Analytics for Streaming Internet Traffic

These same plots can be constructed when the data are conditioned by source IP address, SIP, as opposed to destination port number, DPort. The number of source IP addresses that may be active during a given hour of activity is likely to be very much higher than the number of destination ports; in this data set, only 380 unique destination ports were accessed, while 3548 source IPs are in the file.

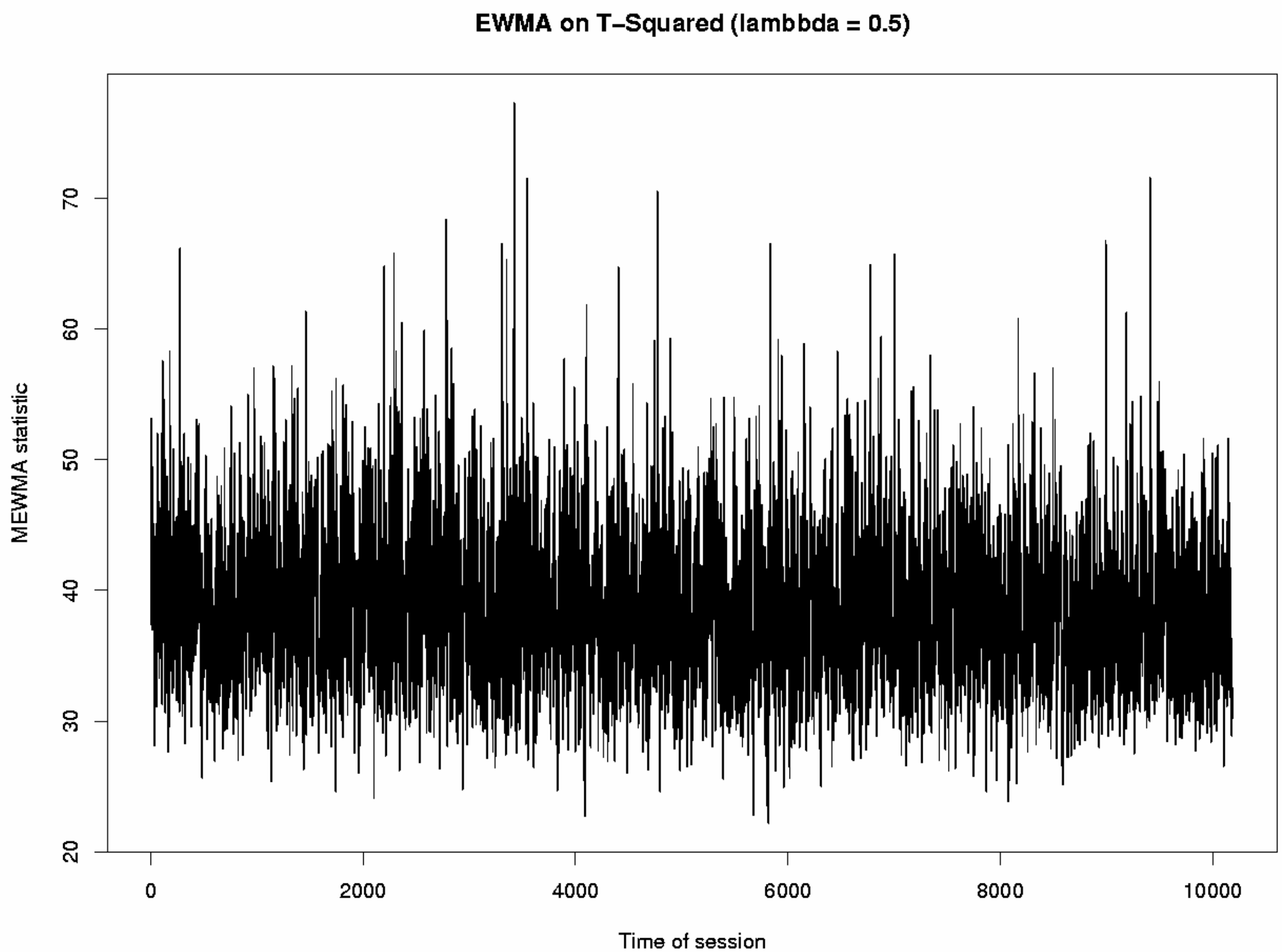
Visual Analytics for Streaming Internet Traffic



Visual Analytics for Streaming Internet Traffic

1. The three session "size" variables, log.len, log.pkt, log.byte, are somewhat correlated and are amenable to a "control chart" procedures, where the statistic being plotted is a weighted linear combination of the previously plotted variable (λ) and the current value of Hotelling's T^2 statistic ($1 - \lambda$).
2. Calculating a Hotelling's T^2 statistic on three successive observations, denoted H_t, a multivariate exponentially weighted moving average (MEWMA) chart using $\lambda = 0.5$ is shown in the next figure (last 10,202 observations only).
3. Most values (99.7%) are below 60; a successive run of observations above 60 might suggest abnormal session sizes.

Visual Analytics for Streaming Internet Traffic



This also is a evolutionary graphic, but with a distinctly visual analytic interpretation.

Visual Analytics for Streaming Internet Traffic

Acknowledgements:

Collaborators:

Karen Kafadar, David Marchette, Jeffrey Solka, Don Faxon, John Rigsby

Research Funding:

ONR, ARO, AFOSR, NSF, DARPA at one or more stages.

Visual Analytics for Streaming Internet Traffic

Contact Information:

Edward J. Wegman
Center for Computational Statistics
George Mason University, MS 4A7
4400 University Drive
Fairfax, VA 22030-4444

Email: ewegman@galaxy.gmu.edu
Phone: (703) 993-1691
FAX: (703) 993-1700

AMSAA Excellence in Analysis

A Noninformative Prior Bayesian Approach to Reliability Growth Projection

Army Conference on Applied Statistics

October 2004

Dr. Paul Ellner; paul.m.ellner@us.army.mil
J. Brian Hall; brian.hall@us.army.mil

EXCELLENCE IN ANALYSIS

Outline.

- Problem/Assumptions.
- Outline of Approach to Failure Rate Projection.
- Interpretation of Posterior Mean.
- Parsimonious Model for Expected Number of Modes by t.
- Reliability Projection using Parsimonious Model Approximation.
- Reliability Projection based on Maximum Likelihood Estimators (MLEs).
- Reliability Projection based on Method of Moments Estimators (MMEs).
- Noninformative Prior Bayesian Approach w/ A and B-Mode Classification.
- Simulation Description and Results.
- Cost versus Reliability Tradeoff Analysis.
- Extensions of Approach where not all Fixes are Assumed Delayed.
- Concluding Remarks.

Problem and Assumptions.

Problem: Assess impact of delayed corrective actions (fixes) at completion of test phase.

Model Assumptions:

1. k potential failure modes where k is large.
2. Each failure mode has constant failure rate over test phase.
3. Occurrence of failures due to modes are statistically independent.
4. Each failure mode occurrence causes system failure.
5. Corrective actions are implemented at end of test phase.
6. At least one failure mode has a repeat failure.
7. Mode failure rates are a realization of a random sample of size k from a gamma distribution with density

$$f(x) = \begin{cases} \dfrac{x^{\alpha} e^{-x/\beta}}{\Gamma(\alpha+1)\beta^{\alpha+1}} & x > 0; \\ 0 & otherwise \end{cases}$$

$$where\ \alpha > -1, \beta > 0$$

Outline of Approach.

True system failure rate after corrective actions to observed failure modes in test denoted by $\rho(T)$, given by,

$$\rho(T) = \sum_{i \in obs} (1 - d_i)\lambda_i + \sum_{i \in unobs} \lambda_i$$

1. For each observed failure mode, assess fix effectiveness factor (FEF) d_i by d_i^*. Assessment based on analysis of failure mechanism(s) that give rise to $n_i > 0$ observed failures due to mode.
2. For each $i \in obs$ and $i \in unobs$, assess unknown mode failure rate $x_i \in \{\lambda_1,\ldots,\lambda_k\}$ based on observed data o_i.
   a) Using noninformative prior, $u(\lambda_i)=1/k$ for $i=1,\ldots,k$, obtain posterior density $g(x_i|o_i)$ for X_i.
   b) Obtain expected value of posterior, denoted by $E[X_i|o_i]$. Will be in terms of k and $\lambda_1,\ldots,\lambda_k$.
   c) Express $E[X_i|o_i]$ in terms of recognizable quantities that can be represented by a parsimonious model.
   d) Statistically estimate parameters of model.
   e) Express projected failure rate in terms of estimated model parameters based on assuming k potential failure modes.
   f) Find the limit of the finite k failure rate projection as $k \rightarrow \infty$.

Posterior Mean.

Observed mode $i \in obs$ has unknown failure rate $x_i \in \{\lambda_1, \ldots, \lambda_k\}$.

Let $o_i = (t_{i,1}, \ldots, t_{i,n_i})$ denote observed sequence of cumulative failure times due to mode i, where $0 < t_{i,1} < \ldots < t_{i,n_i} \leq T$.

Let $L(o_i | x_i)$ = likelihood for o_i given x_i:

$$L(o_i \mid x_i) = \left[\prod_{l=1}^{n_i} x_i \left\{e^{-x_i (t_{i,l} - t_{i,l-1})}\right\}\right] e^{-x_i (T - t_{i,n_i})} = x_i^{n_i} e^{-x_i T} \ \ where \ \ t_{i,0} = 0.$$

Thus,

$$g(x_i \mid o_i) = \frac{\{L(o_i \mid x_i)\} u(x_i)}{\sum_{j=1}^{k} L(o_i \mid \lambda_j) u(\lambda_j)} = \frac{x_i^{n_i} e^{-x_i T}}{\sum_{j=1}^{k} \lambda_j^{n_i} e^{-\lambda_j T}} \ \ for \ x_i \in \{\lambda_1, \ldots, \lambda_k\}$$

and

$$E(X_i \mid o_i) = \sum_{l=1}^{k} \lambda_l \left\{ \frac{\lambda_l^{n_i} e^{-\lambda_l T}}{\sum_{j=1}^{k} \lambda_j^{n_i} e^{-\lambda_j T}} \right\} = \frac{\sum_{l=1}^{k} \lambda_l^{n_i + 1} e^{-\lambda_l T}}{\sum_{j=1}^{k} \lambda_j^{n_i} e^{-\lambda_j T}}$$

Posterior Mean Continued.

Consider unobserved failure mode $i \in unobs$ with unknown failure rate $x_i \in \{\lambda_1, \ldots, \lambda_k\}$.
Let o_i denote the observation that $n_i = 0$. Let $L(o_i|x_i)$= likelihood for o_i given x_i. Then,

$$L(o_i \mid x_i) = e^{-x_i T}$$

$$g(x_i \mid o_i) = \frac{\{L(o_i \mid x_i)\} u(x_i)}{\sum_{j=1}^{k} L(o_i \mid \lambda_j) u(\lambda_j)} = \frac{e^{-x_i T}}{\sum_{j=1}^{k} e^{-\lambda_j T}} \quad for\ x_i \in \{\lambda_1, \ldots, \lambda_k\}$$

$$E(X_i \mid o_i) = \sum_{l=1}^{k} \lambda_l \left\{ \frac{e^{-\lambda_l T}}{\sum_{j=1}^{k} e^{-\lambda_j T}} \right\} = \frac{\sum_{l=1}^{k} \lambda_l e^{-\lambda_l T}}{\sum_{j=1}^{k} e^{-\lambda_j T}}$$

From I. and II., since $n_i = 0$ for $i \in unobs$,

$$E(X_i \mid o_i) = \frac{\sum_{l=1}^{k} \lambda_l^{n_i+1} e^{-\lambda_l T}}{\sum_{j=1}^{k} \lambda_j^{n_i} e^{-\lambda_j T}} \text{ for } i = 1, \ldots, k,$$

Interpretation of Posterior Mean.

Let M(t) = number of distinct failure modes surfaced during test by t.

$$M(t) = \sum_{i=1}^{k} I_i(t) \text{ where } I_i(t) = \begin{cases} 1 & \text{if mode } i \text{ occurs by } T \\ 0 & \text{otherwise} \end{cases}$$

$$\mu(t) = E[M(t)] = \sum_{i=1}^{k} E[I_i(t)] = k - \sum_{j=1}^{k} e^{-\lambda_j t}. \text{ Thus } \sum_{j=1}^{k} e^{-\lambda_j t} = k - \mu(t).$$

Let $h(t) = \dfrac{d\mu(t)}{dt}$. *Thus* $h(t) = \sum_{j=1}^{k} \lambda_j e^{-\lambda_j t}$

$$h^{(1)}(t) = (-1)\sum_{j=1}^{k} \lambda_j^2 e^{-\lambda_j t}$$

$$h^{(2)}(t) = (-1)^2\sum_{j=1}^{k} \lambda_j^3 e^{-\lambda_j t}$$

$$\vdots$$

$$h^{(n_i-1)}(t) = (-1)^{n_i-1}\sum_{j=1}^{k} \lambda_j^{n_i} e^{-\lambda_j t}$$

Interpretation of Posterior Mean Continued.

For $i \in obs$, this yields,

$$E(X_i \mid o_i) = \frac{\sum_{l=1}^{k} \lambda_l^{n_i+1} e^{-\lambda_l T}}{\sum_{l=1}^{k} \lambda_l^{n_i} e^{-\lambda_l T}} = \frac{\left\{\frac{h^{(n_i)}(T)}{(-1)^{n_i}}\right\}}{\left\{\frac{h^{(n_i-1)}(T)}{(-1)^{n_i-1}}\right\}} = -\frac{h^{(n_i)}(T)}{h^{(n_i-1)}(T)}.$$

$$For\ i \in unobs,\ E(X_i \mid o_i) = \frac{\sum_{l=1}^{k} \lambda_l e^{-\lambda_l T}}{\sum_{l=1}^{k} e^{-\lambda_l T}} = \frac{h(T)}{k - \mu(T)}$$

Let $\rho^*(T)$ denote the assessed projected system failure rate based on the mode failure rate assessments $x_i^* = E(X_i|o_i)$ for $i=1,\ldots,k$, Then,

$$\rho^*(T) = \sum_{i \in obs} (1-d_i^*) x_i^* + \sum_{i \in unobs} x_i^* = \sum_{i \in obs} (1-d_i^*) \left\{ -\frac{h^{(n_i)}(T)}{h^{(n_i-1)}(T)} \right\} + (k-m)\left(\frac{h(T)}{k-\mu(T)} \right)$$

where m = number of modes surfaced by T.

Parsimonious Model for Expected Number of Modes by t.

Recall expected number of modes by t given $\underline{\lambda} = (\lambda_1,\ldots,\lambda_k)$ is

$$\mu(t;\underline{\lambda}) = k - \sum_{i=1}^{k} e^{-\lambda_i T}$$

We shall assume $\lambda_1,\ldots,\lambda_k$ is a realization of a random sample from a gamma distribution with density

$$f(x) = \begin{cases} \dfrac{x^{\alpha} e^{-x/\beta}}{\Gamma(\alpha+1)\beta^{\alpha+1}} & \text{for } x > 0; \\ 0 & \text{otherwise} \end{cases}$$

Let $\Lambda_1,\ldots,\Lambda_k$ be independent identically distributed gamma random variables with density f(x). Consider $\mu(t;\underline{\Lambda})$ where $\underline{\Lambda}=(\Lambda_1,\ldots,\Lambda_k)$. We shall approximate $\mu(t;\underline{\lambda})$ by $\mu_k(t;\alpha,\beta)=E[\mu(t;\underline{\Lambda})]$, the expected value of $\mu(t;\underline{\Lambda})$ w.r.t. $\underline{\Lambda}$. Can show (AMSAA Growth Guide)

$$\mu_k(t;\alpha,\beta) = k\left\{1-(1+\beta t)^{-(\alpha+1)}\right\}$$

Model Approximation for Posterior Mean.

We shall use $\mu_k(t;\alpha,\beta)$ and its derivatives to approximate $x_i^* = E(X_i|o_i)$ for i=1,…,k. Let $\lambda_k=E[\Lambda_1+\ldots+\Lambda_k]=k\beta(\alpha+1)$.

Define $$h_k(t;\alpha,\beta) = \frac{d\mu_k(t;\alpha,\beta)}{dt} = \frac{\lambda_k}{(1+\beta t)^{\alpha+2}}.$$

Note, $$h_k^{(1)}(t;\alpha,\beta) = \frac{(-1)(\alpha+2)\lambda_k\beta}{(1+\beta t)^{\alpha+3}}$$

$$h_k^{(2)}(t;\alpha,\beta) = \frac{(-1)^2(\alpha+2)(\alpha+3)\lambda_k\beta^2}{(1+\beta t)^{\alpha+4}}$$

$$\vdots$$

$$h_k^{(n_i-1)}(t;\alpha,\beta) = \frac{(-1)^{n_i-1}(\alpha+2)\ldots(\alpha+n_i)\lambda_k\beta^{n_i-1}}{(1+\beta t)^{\alpha+n_i+1}}$$

Let $x^*_{i,a,k}$ denote the approximation of $x_i^* = E(X_i|o_i)$ based on using $\mu_k(t;\alpha,\beta)=E[\mu(t;\underline{\Lambda})]$ to approximate $\mu(t;\underline{\lambda})$. Thus

$$x^*_{i,a,k} = \begin{cases} -\dfrac{h_k^{(n_i)}(T;\alpha,\beta)}{h_k^{(n_i-1)}(T;\alpha,\beta)} & for\, i \in obs; \\[2ex] \dfrac{h_k(T;\alpha,\beta)}{k-\mu_k(T;\alpha,\beta)} & for\, i \in unobs \end{cases}$$

Model Approximation for Posterior Mean Continued.

For i∈obs, $x^*_{i,a,k} = \frac{(\alpha + n_i + 1)\beta}{1 + \beta T}$

For i∈unobs

$$x^*_{i,a,k} = \frac{\left\{\frac{\lambda_k}{(1+\beta T)^{\alpha+2}}\right\}}{k - k\left\{1 - (1+\beta T)^{-(\alpha+1)}\right\}}$$

$$= \frac{k\beta(\alpha+1)}{(1+\beta T)^{\alpha+2}\left\{k(1+\beta T)^{-(\alpha+1)}\right\}}$$

$$= \frac{\beta(\alpha+1)}{1+\beta T}$$

Reliability Projection using the Parsimonious Model Approximation.

Let $\rho^*_{a,k}(T)$ denote the failure rate projection based on the d^*_i for $i \in obs$ and the $x^*_{i,a,k}$ for $i=1,\ldots,k$. Thus,

$$\rho^*_{a,k}(T) = \sum_{i \in obs} (1 - d^*_i) x^*_{i,a,k} + \sum_{i \in unobs} x^*_{i,a,k}$$

This yields,

$$\rho^*_{a,k}(T) = \sum_{i \in obs} (1 - d^*_i) \left\{ \frac{(\alpha + n_i + 1)\beta}{1 + \beta T} \right\} + (k - m) \left\{ \frac{\beta(\alpha + 1)}{1 + \beta T} \right\}$$

$$= \sum_{i \in obs} (1 - d^*_i) \left\{ \frac{(\alpha + n_i + 1)\beta}{1 + \beta T} \right\} + (1 - \frac{m}{k}) \left\{ \frac{\lambda_k}{1 + \beta T} \right\}$$

Reliability Projection based on the MLEs for Gamma Parameters.

Shall use MLE's for gamma parameters given data m and $\underline{n}=(n_1,\ldots,n_k)$. Let N_i denote the random variable for the number of failures due to mode i that occurs during [0,T]. Let $w(s_i;\alpha,\beta)$ denote the marginal density for the compound random variable N_i. Then [Martz & Waller]

$$w(s_i;\alpha,\beta) = \frac{T^{s_i}\Gamma(s_i+\alpha+1)}{\left\{s_i!\,\beta^{\alpha+1}\Gamma(\alpha+1)\right\}\left(T+\frac{1}{\beta}\right)^{s_i+\alpha+1}} \quad for\ s_i = 0,1,2,\ldots$$

Reliability Projection based on the MLEs for Gamma Parameters Continued.

Likelihood for (α,β) given m and $\underline{n}$ is

$$L(\alpha,\beta;m,\underline{n}) = \prod_{i=1}^{k} w(n_i;\alpha,\beta). \quad Note\ n_i = 0\ for\ i \in unobs$$

Assuming k potential failure modes, let $\hat{\alpha}_k, \hat{\beta}_k$ denote the MLEs for α, β, respectively. Also let $\hat{\lambda}_k = k\hat{\beta}_k(\hat{\alpha}_k + 1)$. Can show (Martz & Waller, Chapter 7)

$$\hat{\lambda}_k = \frac{n}{T}\ where\ n = \sum_{i=1}^{k} n_i \ \text{and}\ \hat{\beta}_k = \frac{y_k}{T}\ where \left(\frac{n}{y_k}\right)\ln(1+y_k) - \sum_{j\in obs}\sum_{i=1}^{n_j-1}\frac{1}{1+\left(\frac{iky_k}{n}\right)} = m$$

The inner sum is defined to be zero when n_j=1.
From the above equations can find $(\hat{\lambda}_\infty, \hat{\beta}_\infty, \hat{\alpha}_\infty) = \lim_{k\to\infty}(\hat{\lambda}_k, \hat{\beta}_k, \hat{\alpha}_k)$

Note $\hat{\lambda}_\infty = \frac{n}{T}$ and $\hat{\beta}_\infty = \frac{y_\infty}{T}$ where y_∞ is the unique positive solution y that satisfies $\left(\frac{n}{y}\right)\ln(1+y) = m$. It follows that $\hat{\alpha}_\infty = -1$.

Reliability Projection based on MLEs for Gamma Parameters Continued.

For i=1,…,k let $\hat{X}_{i,k}$ denote the statistical estimate of $X^{*}_{i,a,k}$ based on the MLEs $\hat{\alpha}_k, \hat{\beta}_k$ for α, β, respectively. Let $\hat{\rho}_k(T)$ denote the projected failure rate assessment obtained from $\rho^{*}_{a,k}(T)$ by replacing α, β and λ_k by $\hat{\alpha}_k, \hat{\beta}_k$ and $\hat{\lambda}_k = k\hat{\beta}_k(\hat{\alpha}_k + 1)$, respectively. Thus

$$\hat{\rho}_k(T) = \sum_{i \in obs} (1 - d_i^{*})\hat{x}_{i,k} + \sum_{i \in unobs} \hat{x}_{i,k}$$

By definition $\hat{x}_{i,k} = \dfrac{(\hat{\alpha}_k + n_i + 1)\hat{\beta}_k}{1 + \hat{\beta}_k T} \quad for\, i = 1, \ldots, k$.

Therefore,

$$\hat{\rho}_k(T) = \sum_{i \in obs} (1 - d_i^{*})\left\{\frac{(\hat{\alpha}_k + n_i + 1)\hat{\beta}_k}{1 + \hat{\beta}_k T}\right\} + \sum_{i \in unobs} \left\{\frac{(\hat{\alpha}_k + 1)\hat{\beta}_k}{1 + \hat{\beta}_k T}\right\}$$

$$= \sum_{i \in obs} (1 - d_i^{*})\left(\frac{\hat{\beta}_k T}{1 + \hat{\beta}_k T}\right)\left(\frac{(\hat{\alpha}_k + n_i + 1)}{T}\right) + (k - m)\left\{\frac{\hat{\beta}_k(\hat{\alpha}_k + 1)}{1 + \hat{\beta}_k T}\right\}$$

$$= \sum_{i \in obs} (1 - d_i^{*})\left(\frac{\hat{\beta}_k T}{1 + \hat{\beta}_k T}\right)\left(\frac{(n_i + \hat{\alpha}_k + 1)}{T}\right) + \left(1 - m/k\right)\left(\frac{n/T}{1 + \hat{\beta}_k T}\right)$$

since $k\hat{\beta}_k(\hat{\alpha}_k + 1) = n/T$

Reliability Projection based on MLEs for Gamma Parameters Continued.

Note $\hat{\alpha}_k + 1 = \frac{n}{k\hat{\beta}_k T}$. Therefore,

$$\hat{\rho}_k(T) = \sum_{i \in obs} (1 - d_i^*) \left(\frac{\hat{\beta}_k T}{1 + \hat{\beta}_k T} \right) \left(\frac{n_i}{T} + \frac{n}{k\hat{\beta}_k T^2} \right) + \left(1 - \frac{m}{k} \right) \left(\frac{n/T}{1 + \hat{\beta}_k T} \right)$$

This yields,

$$\hat{\rho}_\infty(T) = \lim_{k \to \infty} \hat{\rho}_k(T) = \sum_{i \in obs} (1 - d_i^*) \left(\frac{\hat{\beta}_\infty T}{1 + \hat{\beta}_\infty T} \right) \left(\frac{n_i}{T} \right) + \left(\frac{n/T}{1 + \hat{\beta}_\infty T} \right)$$

Reliability Projection based on MMEs for Gamma Parameters

Let $\overline{\Lambda}$ and M^2 denote the random variables that take on the values $\overline{\lambda}$ and m^2 respectively where

$$\overline{\lambda} = \frac{1}{k}\sum_{i=1}^{k}\hat{\lambda}_i \text{ and } m^2 = \frac{1}{k}\sum_{i=1}^{k}\hat{\lambda}_i^2 \text{ with } \hat{\lambda}_i = \frac{n_i}{T}$$

One can show $E[\overline{\Lambda};\alpha,\beta] = \beta(\alpha+1)$ and $E[M^2;\alpha,\beta] = \dfrac{\beta^2(\alpha+1)\left[T(2+\alpha)+ {}^{1}\!/_{\beta}\right]}{T}$

In [Martz and Waller], the MMEs for α and β, denoted by $\tilde{\alpha}_k, \tilde{\beta}_k$ respectively, are implicitly defined as follows:

$$\overline{\lambda} = \tilde{\beta}_k(\tilde{\alpha}_k+1) \text{ and } m^2 = \frac{\tilde{\beta}_k^2(\tilde{\alpha}_k+1)\left[T(2+\tilde{\alpha}_k)+ {}^{1}\!/_{\tilde{\beta}_k}\right]}{T}$$

From these equations it follows

$$\tilde{\lambda}_k = k\tilde{\beta}_k(\tilde{\alpha}_k+1) = \frac{n}{T} \text{ and } \tilde{\beta}_k = \frac{\sum_{j\in obs} n_j^2 - \frac{n^2}{k} - n}{T\cdot n}$$

Reliability Projection based on MMEs for Gamma Parameters Continued

Let $(\tilde{\lambda}_\infty, \tilde{\beta}_\infty) = \lim_{k\to\infty}(\tilde{\lambda}_k, \tilde{\beta}_k)$. One obtains

$$\tilde{\lambda}_\infty = \frac{n}{T} \text{ and } \tilde{\beta}_\infty = \frac{1}{T}\left(\frac{\sum_{j\in obs} n_j^2}{n} - 1\right)$$

Let $\tilde{\rho}_k(T)$ denote the projection for the mitigated system failure rate based on the finite k MMEs.

Then

$$\tilde{\rho}_k(T) = \sum_{i\in obs}(1-d_i^*)\left(\frac{\tilde{\beta}_k T}{1+\tilde{\beta}_k T}\right)\left(\frac{n_i}{T} + \frac{n}{k\tilde{\beta}_k T^2}\right) + \left(1 - {m}/{k}\right)\left(\frac{{n}/{T}}{1+\tilde{\beta}_k T}\right)$$

$$\tilde{\rho}_\infty(T) = \lim_{k\to\infty}\tilde{\rho}_k(T) = \sum_{i\in obs}(1-d_i^*)\left(\frac{\tilde{\beta}_\infty T}{1+\tilde{\beta}_\infty T}\right)\left(\frac{n_i}{T}\right) + \left(\frac{{n}/{T}}{1+\tilde{\beta}_\infty T}\right)$$

Noninformative Prior Bayesian Approach with A and B-Mode Classification

- *A-mode: no corrective action planned even if surfaced.*

- *B-mode: if surfaced, will be mitigated.*

Approach still applies for two failure mode categories. Apply previous procedure to set of B-modes to obtain projection of system failure rate due to the B-modes, say $\hat{\rho}_k(T)$.

The prior now pertains to $x_i \in \{\lambda_1, \ldots \lambda_k\}$, the initial failure rates of the k B-modes. Likewise, the data m and N_i pertain to the B-modes.

Then the projection for the system mitigated failure rate using MLEs is

$$\hat{\rho}_{A+B,k}(T) = \frac{N_A}{T} + \hat{\rho}_k(T)$$

and

$$\hat{\rho}_{A+B,\infty}(T) = \lim_{k \to \infty} \hat{\rho}_{A+B,k}(T) = \frac{N_A}{T} + \hat{\rho}_\infty(T)$$

where N_A denotes the number of A-mode failures in test.

The same comments apply to obtaining projections for two classifications using the MMEs for α and β.

Simulation Overview

The simulation consists of the following steps:

1. Specify inputs.
2. Generate failure rates.
3. Calculate mode failure times.
4. Calculate first occurrence times and number of failures during test for each mode.
5. Generate FEFs from beta distribution with mean = 0.80 and coefficient of variation = 0.10.
6. Calculate MTBF projections.
7. Reclassify repeat A-modes.
8. Recalculate MTBF projections.

	Simulated	Surfaced	Distribution
A-Modes	50	23	Gamma
B-Modes	100	45	Gamma

Table 1. Simulated/Surfaced Modes.

	A-Modes	B-Modes
Shape - α	3.3333	3.3333
Scale - β	0.0002	0.0002

Table 2. Gamma Parameters.

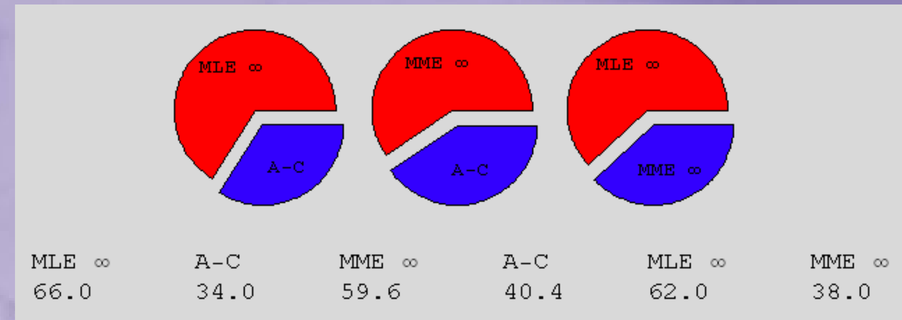
•Results obtained from simulating 1,000 tests of length 1,000 hours.

•Mode failure rates and FEFs regenerated for each test.

Simulation Results (Gamma)

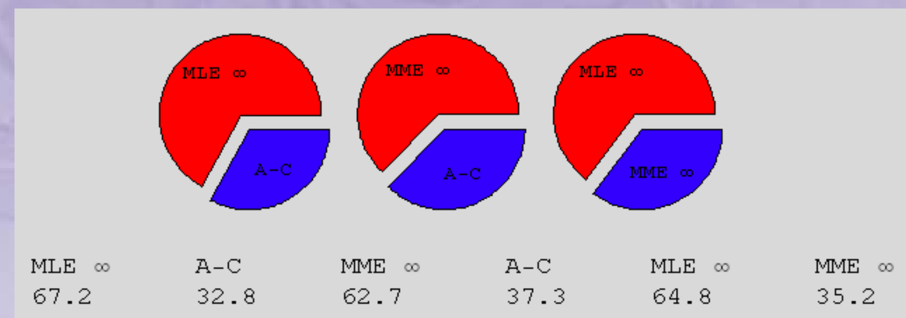
	Actual	MLE ∞	MME ∞	A-C
MTBF	14.15	13.94	13.42	13.26

Table 3. Two Categories.



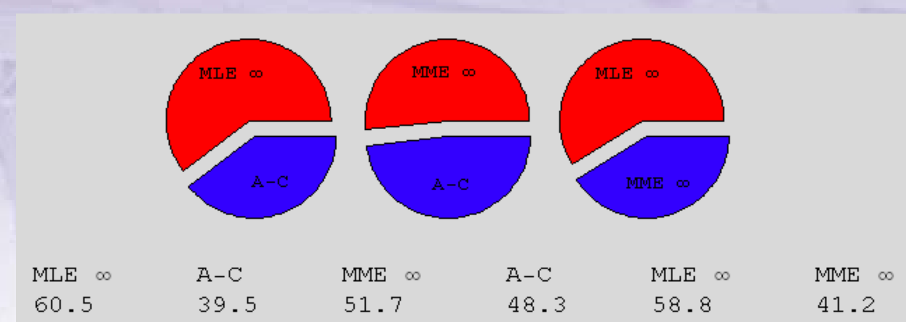
	Actual	MLE ∞	MME ∞	A-C
MTBF	14.15	13.92	13.41	13.22

Table 4. One Category.



	Actual	MLE ∞	MME ∞	A-C
MTBF	15.43	15.54	14.73	15.09

Table 5. Two Categories, after Reclassification.

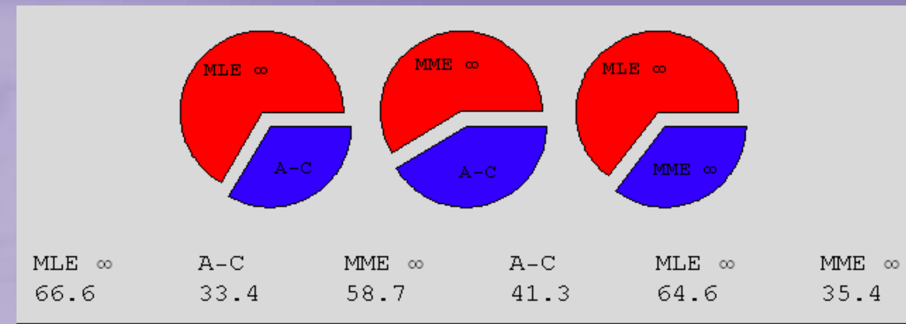


8 modes reclassified on average.

Simulation Results (Weibull)

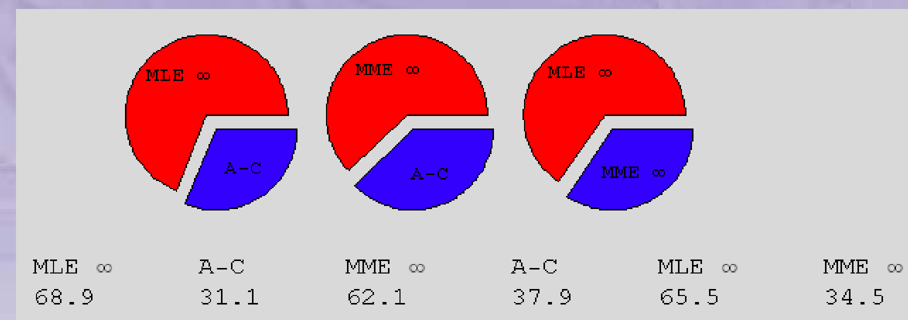
	Actual	MLE ∞	MME ∞	A-C
MTBF	14.31	13.99	13.45	13.29

Table 6. Two Categories.



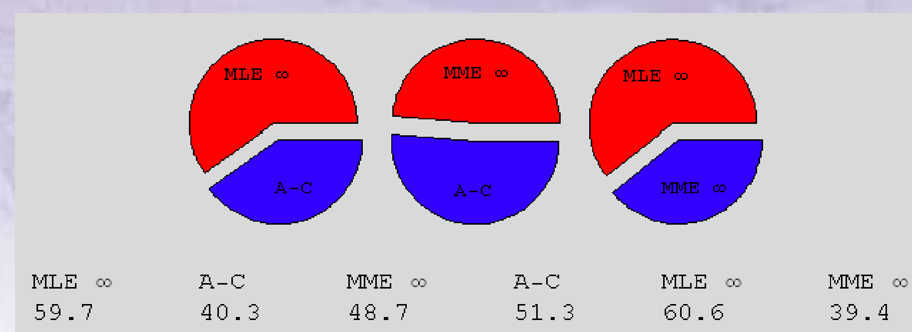
	Actual	MLE ∞	MME ∞	A-C
MTBF	14.31	13.98	13.46	13.27

Table 7. One Category.



	Actual	MLE ∞	MME ∞	A-C
MTBF	15.67	15.65	14.82	15.20

Table 8. Two Categories, after Reclassification.



8 modes reclassified on average.

Cost versus Reliability Tradeoff Analysis

Let $Z \subset obs$ be a candidate set of observed modes to receive fixes following the test phase. Based on a study of the underlying root causes of failure, fixes could be devised with associated FEF assessments d_i^* for $i \in obs$. The corresponding projection for the resulting system failure rate would be for large k (using MLEs for α and β).

$$\hat{\rho}_\infty(T;Z) = \sum_{i \in Z}(1-d_i^*)\hat{x}_{i,\infty} + \sum_{i \in obs-Z}\hat{x}_{i,\infty} + \sum_{i \in unobs}\hat{x}_{i,\infty}$$

One could also assess the cost, $c^*(Z)$, of implementing all the fixes for modes $i \in Z$.

A plot of the projected MTBF vs. associated cost for a number of selected $Z \subset obs$ would be useful in identifying a least cost solution Z to meet a reliability goal. In place of using MLEs, one could use MME based assessments.

Projection methods whose estimation procedures for a given data set treat A-mode and B-mode data differently beyond differentiation with regard to FEFs are not suitable for performing cost/reliability tradeoff analysis. Such methods include those that utilize an estimate of the expected B-mode failure rate due to the unsurfaced B-modes.

Extensions of Approach to Situations where Fixes Need Not be Delayed

I. Using the n_i for estimation.

- Unknown mode failure rate $x_i \in \{\lambda_1,\ldots,\lambda_k\}$ either generates observed data $n_i=0$ or $o_i = (t_{i,1},\ldots,t_{i,n_{i,1}}, v_i, t_{i,n_{i,1}+1},\ldots,t_{i,n_{i,2}})$ where $0 < t_{i,1} < \ldots < t_{i,n_{i,1}} \le v_i < t_{i,n_{i,1}+1} < t_{i,n_{i,2}} \le T$

  For the above data, $n_{i,1} \ge 1$ and $n_{i,1} + n_{i,2} = n_i$. The $t_{i,j}$ are the cumulative failure times for mode i and v_i denotes the time at which the fix to mode i is implemented.

- Can show

$$E(X_i \mid o_i) = -\frac{h^{(n_i)}(v_i + (1-d_i)(T-v_i))}{h^{(n_i-1)}(v_i + (1-d_i)(T-v_i))}$$

where d_i denotes realized FEF for mode i.

- The assessment of $E(X_i|o_i)$ depends on v_j and d_j^* for all $j \in$ obs.

II. Using failure mode first occurrence times.

- Unknown mode failure rate x_i either generates observed data $n_i=0$ or $o_i=t_{i,1}$ where $t_{i,1}$ is the first occurrence time for mode i.
- Can show

$$E(X_i \mid o_i) = -\frac{h^{(1)}(t_{i,1})}{h(t_{i,1})}$$

- The assessment of $E(X_i|o_i)$ will not depend on any of the v_j or d_j^*.

Concluding Remarks

Noninformative Prior Bayesian Approach useful in deriving reliability growth projection methods:

- for case where all fixes delayed,
- for situation where not all fixes need be delayed,
- potentially for deriving discrete projection methodology.

For current simulations, described procedures compare favorably to the standard adopted by the International Electrotechnical Commission (AMSAA-Crow Projection Model).

Method does not require one to distinguish for estimation purposes between A-modes and B-modes other than through FEFs.

Can also be used for case when failure modes can be split into inherent A-modes and B-modes.

Method suitable for cost versus reliability tradeoff analysis for modes that are not inherently A-modes.

Model and estimation procedures only require reference to FEFs for surfaced modes.

Comparable simulation results obtained when failure rates drawn from Weibull or lognormal distributions with same mean and variance as the gamma.

References

Broemm, William J., Paul M. Ellner, John W. Woodworth, "AMSAA Reliability Growth Guide," *AMSAA TR-652*, (September) 2000.

Martz, Harry F., and Ray A. Waller, *Bayesian Reliability Analysis*, 1982, pp 319-324.

On Optimal System Design Under Reliability and Economic Constraints[1]

Michael R. Dugas and Francisco J. Samaniego

Department of Statistics

University of California, Davis

Davis, California 95616

Abstract

Reliability Economics is a field that can be defined as the collection of all problems in which there is tension between the performance of systems of interest and their cost. Given such a problem, the aim is to resolve the tension through an optimization process that identifies the system that maximizes some appropriate criterion function (e.g. expected lifetime per unit cost). In this paper, we focus on coherent systems in n independent and identically distributed (iid) components and mixtures thereof, and characterize both a system's performance and cost as functions of the system's signature vector (Samaniego, IEEE-TR, 1985). For a given family of criterion functions, a variety of optimality results are obtained for systems of arbitrary order n. The case of an underlying exponential distribution is used to illustrate these results. Approximations are developed and justified when the underlying component distribution is unknown. In the latter circumstance, assuming that an auxiliary sample of size N is available on component failure times, the asymptotic theory of L-estimators is adapted for the purpose of proving the consistency and asymptotic normality (as $N \to \infty$) of estimators of the expected ordered failure times of the n components of the systems under study. These asymptotic results lead to the identification of optimal systems relative to a closely approximated criterion function. Proofs of the results stated herein appear in a referenced Technical Report.

I. Introduction

The emerging field of Reliability Economics (RE) is perhaps most easily defined by its goals rather than by its tools or results. The literature in Reliability Economics is quite widely scattered, and the area is yet to be unified and conceptualized as a distinct field of study. Roughly speaking, the field can be thought of as the collection of problems and frameworks in which there is tension between the performance of a group of systems of interest and their cost. In general, expensive systems perform quite well and inexpensive systems perform less well. Ideally, one would like to select the system that represents the best compromise between performance and cost. This is, in fact, often the goal of an RE analysis, though there are other goals of possible interest.

When one thinks of a particular manufactured system that one might consider purchasing, two questions naturally arise: (1) “How well does it work? and (2) “How much does it cost?” These questions are so natural that situations in which one or the other question might be deemed irrelevant would seem to be both extreme and quite rare. If money were truly “of no object”, then naturally one would purchase the system with the best performance, or if money was very tight, one might be forced to buy the least expensive system available without questioning its performance characteristics. Excluding these extreme situations, the natural strategy in procurement situations is to take both performance and cost into account. It is thus quite surprising that the mathematical and statistical underpinnings of doing so in a systematic way are, at present, in a relatively primitive state.

Exceptions exist in selected problem areas such as “warranty analysis” (see, for example, Blischke and Murthy (1996) and Singpurwalla (2004)), but general developments in Reliability Economics are at

[1] This research was supported in part by ARO Grant DAAD 19-02-1-0377

present quite sparse. In military acquisitions, for example, it is frequently the case that a particular prototype system is developed to meet certain performance and suitability goals, and that once a system meeting those goals is developed and is validated through operational testing, the system is purchased in whatever numbers the allocated budget can accommodate. Such an approach foregoes the formal investigation of optimality questions such as "What system design would give us the best performance per unit cost?" In this study, our goal is to address the question: "Is it possible to identify system designs which are optimal in some appropriate sense in the face of economic constraints?" In more common parlance, can one find the system that gives us the most bang for the buck? We discuss below our progress toward answering that question.

Let us first examine why the problem of answering the type of question posed above has heretofore resisted clean, analytical solutions. Consider the notion of "coherent systems of order (or size) n", a fundamental idea in reliability dating back to the seminal paper by Birnbaum, Esary and Saunders (1961). Coherent systems of order n are n-component systems that are monotone (i.e., the state of the system can only stay the same or improve when a component is improved), and in which every component is relevant (i.e., it actually affects system performance under some configuration of the functioning or failure of the other components).

Identifying the exact number of coherent systems of a given order is a fascinating open question. A few crude approximations exist, but all that is really known is that the number is finite but tends to be very large. The number is known to grow exponentially with n, so that, for example, there are well over a billion different coherent systems of order 30. This provides part of the explanation for the resistance seen in attempts to optimize relative to the class of all coherent systems of a given size. The problem is a discrete optimization problem in which the space to be searched is usually huge.

There is a second reason that finding analytical solutions to optimization problems would be difficult. That is that there has not been a tool available which summarizes the behavior of a system as a design parameter with respect to which one can optimize. The structure function φ (see, for example, Barlow and Proschan (1981)) which characterizes a system by the relationship between the n-dimensional vector of 1s and 0s representing the states of the n components (working or not) and the state of the system (1 or 0) is (i) awkward to compute for complex systems and (ii) too clumsy to use as an index for all coherent systems of a given size.

These two difficulties, together, have led to the reliance on "searching techniques" for seeking good (near-optimal) solutions as efficiently as possible. Genetic algorithms appear to be the favored approach in the recent literature. There is a substantial literature on the latter approach. Chapter 7 of the monograph by Kuo, Prasad, Tillman and Hwang (2001) discuss the algorithmic approach to constrained optimization problems in reliability and provide many references. For a concrete example of the use of genetic algorithms in searching for a system design that minimizes costs while achieving a fixed reliability threshold, see Deeter and Smyth (1998). An example of the use of a genetic algorithm in searching for a cost-optimal maintenance policy may be found in Usher, Kamal and Sayed (1998).

We now turn no to a discussion of some background ideas and results which provide the foundation for the approach we will take to problems of optimal system design under reliability and economic constraints.

2. Signatures and mixed systems.

In the formulation of problems in reliability economics we have studied, both of the obstacles above have been overcome, one, quite curiously, by making the space of systems of interest even larger and the other by identifying a new and useful index of that space. We'll discuss the latter issue first, as the former one follows upon it naturally. In a paper in the IEEE Transaction on Reliability, Samaniego (1985) defined the notion of "system signature". In brief, if one restricts attention to systems of order n whose components have independent and identically distributed (iid) lifetimes, then the behavior of the system's lifetime is completely determined by the underlying component lifetime distribution F and a probability vector $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ called the system's "signature". The ith component of $\mathbf{s}$ is simply the probability that the ith component failure (that is, the ith order statistic among the component failure times X_1, X_2,…, X_n) causes the system to fail. Since all n! permutations of the n failure times are equally likely under an iid assumption, the computation of **s** tends to be a manageable combinatorial problem. In Samaniego (1985), representations were given for the distribution function F_T, density function f_T, and failure rate r_T of the system lifetime T in terms of the component lifetime distribution F and the signature **s**. Although it was not applied to optimization problems at the time, it became clear over time that the signature serves as an excellent index in optimization problems involving the family of coherent systems.

The use of signatures as proxies for the corresponding system designs requires a bit of a defense. Since signatures are based on the assumption of iid component lifetimes, one might well question their relevance in problems in which components lifetimes are dependent or have different distributions. This question is taken up in Kochar, Mukerjee and Samaniego (1999), where it is argued that when comparing systems, one wants to consider a level playing field, as it is clear that a poor system with good components can outperform a good system with poor components. If, however, all components are independent and have the same distribution, then any superiority in one system's performance over another must be attributable to the system design alone. Kochar, Mukerjee and Samaniego (1999) proved a number of preservation theorem for signatures (e.g., stochastic ordering between two signatures implies stochastic ordering between the corresponding system lifetimes). These results were used to demonstrate that signatures were useful tools in the comparison of competing systems. Boland, Samaniego and Vestrup (2003) showed the notion of signature was equally applicable to comparisons among communication networks. Indeed, they established in that paper the exact relationship between Satyanarayana and Prabhakar's (1978) "dominations" and the signature of a network.

Having an index for coherent systems is, of course, not enough to render the optimization problems of interest analytically tractable. That's because one still has to contend with maximization over a large discrete space. Boland and Samaniego (2004b) proposed consideration of stochastic mixtures of coherent systems. A mixed system of order n is obtained by a randomization process over the class of coherent systems of order n. In essence, one simply picks a coherent system of order n at random according to a fixed mixing distribution. If coherent system τ_i has signature vector $\mathbf{s}^{(\mathbf{i})}$ and is chosen with probability p_i for i = 1,…, k, then it is easily seen that the mixed system $\Sigma\ p_i\tau_i$ will have signature vector $\Sigma\ p_i\mathbf{s}^{(\mathbf{i})}$. Since the k-out-of-n system (which fails upon the kth component failure) has a unit vector $\mathbf{s}_{\mathbf{k:n}}$ as its signature (with 1 as its kth element), it is clear that any probability vector **p** may be considered to be a signature. Indeed, the vector **p** is the signature of the mixed system $\Sigma\ p_k\ \mathbf{s}_{\mathbf{k:n}}$. A simple example of a mixed system or order 3 would be the system that selects the series system (whose signature is (1, 0, 0)) with probability ½ and a parallel system (whose signature is (0, 0, 1)) with probability ½. The signature of such a system is (½, 0, ½), which is different from any of the distinct signatures of the 5 possible coherent systems of order 3. Mixing clearly expands the space of available systems.

The mathematical effect of introducing mixed systems is that a complex discrete optimization problem is immediately converted into a continuous problem. We may now seek to optimize with respect to the simplex of probability vectors in an n-dimensional space. The tools of differential and integral calculus can now be brought to bear on this problem. Interestingly, we also discover, as will be explained below, that there are problems in which a certain mixed system will dominate all other systems, that is, the strategy of randomizing among a collection of coherent systems can outperform the best that can be achieved by any coherent system (a degenerate mixture) alone.

3. Optimality Criteria

Implicit in the loose description of Reliability Economics above is the existence of a criterion function that depends on both performance and cost and an optimization process for maximizing the criterion function among the class of systems under consideration. In the work we report on here, we have utilized a particular class of criterion functions with two basic properties that might be considered essential in RE problems: the criterion function should vary proportionately with measures of system performance and inversely with measures of system cost. In Samaniego (1985), it was noted that if T is the lifetime of a system in iid components and with signature **s**, then the survival function of T can written as

$$P(T > t) = \sum_{i=1}^{n} s_i \, P(X_{i:n} > t), \qquad (3.1)$$

where $X_{1:n}$, $X_{2:n}$,…,$X_{n:n}$ are the ordered failure times of the n components, from smallest to largest. It follows that

$$ET = \sum_{i=1}^{n} s_i \, EX_{i:n}. \qquad (3.2)$$

Thus, the expected system lifetime can be written as a linear combination of the elements of the signature vector. In the same vein, one could envision the expected cost of a system as a different linear combination of the components of **s**, say,

$$EC = \sum_{i=1}^{n} c_i s_i. \qquad (3.3)$$

One instance where such a linear combination arises as an appropriate representation of cost is in the "salvage model" where one assumes a fixed cost C_F for all n-component systems, and models a component cost as A and the value of a used component (salvaged from the system after the system fails) as B. Under these assumptions, the expected cost of the system is given by

$$EC = \sum_{i=1}^{n} (C_F + n(A - B) + Bi) \, s_i, \qquad (3.4)$$

which is a linear function of the elements of **s** as above. The criterion function under which the results we've obtained are derived is somewhat more general than simply the ratio of (3.2) to (3.3) or (3.4). We have sought to optimize a criterion function with would include such a ratio as a special case. Specifically, we consider, as a measure of the relative value of performance and cost, the ratio

$$m_r(\mathbf{a}, \mathbf{c}, \mathbf{s}) = \left(\sum_{i=1}^{n} a_i s_i\right) / \left(\sum_{i=1}^{n} c_i s_i\right)^r. \qquad (3.5)$$

Several remarks on (3.5) are in order. First, we note that, while setting $a_i = EX_{i:n}$ is a natural choice for the vector **a**, it is not required by our construct, and other choices might be preferred depending on the application involved. One reasonable alternative is the vector **a** with elements $a_i = P(X_{i:n} > t)$, in which case the numerator of (3.5) would simply be the system's reliability function at the point t. Further, the salvage model is but one way of motivating a sum such as $\Sigma c_k s_k$. Another would be to obtain an expert assessment of the cost of constructing a k-out-of-n system, and then set c_k equal to that cost. The justification for that choice of the vector **c** is that the mixed system represented by the signature **s** can be represented as choosing a k-out-of-n system with probability s_k and thus incurring the cost c_k with probability s_k. The expected cost of using this mixed system would be precisely $\Sigma c_k s_k$. The exponent r in (3.5) is a tuning parameter that allows one to weigh performance and cost differently. While the case when "r = 1" is of obvious interest, a large r might be required in problems in which controlling costs is essential (putting a higher value on less expensive systems) while a small r is appropriate when performance is given more weight than cost. The choice of r will vary with the application.

4. Optimality Results

Under an iid assumption on component lifetimes and under the criterion function given in (3.5), we have obtained the following results. Proofs may be found in Dugas and Samaniego (2004).

Theorem 1: When r = 1, the criterion function (3.5) is maximized by a k-out-of n system.

The result above was obtained by variational arguments. The optimal system is the k-out-of-n system with the largest ratio of a to c, that is for k such that $a_k/c_k = \max a_i/c_i$, where the maximum is taken over the values i = 1,…, n.

Theorem 2: For $r \neq 1$, the criterion function in (3.5) is always maximized by a mixture of at most two k-out-of n systems.

Theorem 2 was obtained using the tools of multivariate calculus. For each of the n(n-1)/2 possible mixed systems in contention, the best mixture of the two systems involved can be calculated in closed form. Thus, the identification of the optimal system reduces to a simple numerical comparison. It is worth noting that, when $r \neq 1$, the optimal system might be a k-out-of-n system (i.e., a degenerate mixture), but it need not be. For example, when n = 2, r =2.5, $a_i = EX_{i:n}$, F is taken to be a uniform distribution and the salvage model for costs is assumed, the mixed system with signature (2/3, 1/3) is optimal and strictly better than either of the two coherent systems of order 2.

Theorem 3. If the sequence $\{a_i/c_i,\ i = 1,\ldots, n\}$ is monotone, then the optimal system is a mixture of a series system and a parallel system, with the mixture being degenerate for r sufficiently large or small. For sufficiently large r, the series system is optimal; for sufficiently small r, the parallel system is optimal.

Theorems 1 and 2 above settle the question of finding the optimal system for the criterion function in (3.5) and for the case where the vectors **a** and **c** can be completely specified. Theorem 3 sheds light on specific circumstances under which a particular type of system design is optimal. Note that the exponential distribution satisfies the hypothesis of Theorem 3 when $a_i = EX_{i:n}$ and the salvage model is assumed.

5. Statistical Issues

The problem that remains to be addressed relates to facilitating the practical application of the results described above. The problem of identifying the mixed system that maximizes the criterion function m in (1.5) has been solved for any fixed values of the vectors **a** and **c** and the constant r. Now, the cost vector **c** and the tuning parameter r involve assessments on the part of the experimenter, and it is not unreasonable to assume that these values can be determined, with the assistance of experts, in a given application of interest. The vector **a**, on the other hand, is typically a function of the unknown underlying distribution F of the iid component lifetimes. The most natural choice for **a** is the vector of expected order statistics, with the ith element being given by $a_i = EX_{i:n}$ for i = 1, 2,…., n. We will hereafter assume, for concreteness, that this is the specification of the vector **a** that has been chosen. Our inferential results about **a** can be adapted without difficulty to alternative specifications of **a** which depend on other aspects of F.

The practical implementation of the methods above for identifying an optimal system design require that the vector **a** be estimated from data. Several questions arise: how should **a** be estimated? If the vector **a*** = $(a_1^*,\ldots,a_n^*)$ represent an estimate of the vector of expected order statistics from a "hypothetical" sample of size n from F, what are the properties of the "optimal system" corresponding to **a***? In the theorems below, we provide answers to these questions. First, we note that, in the iid framework that has been assumed, one can perform independent life-testing experiments on a set of N components and estimate the expected order statistics corresponding to a sample of size n (the order of the systems under consideration) using so-called L-statistics (i.e., linear combinations of order statistics) based on a random sample of size N. Moreover, since n is the fixed size of the systems under study while N can be freely chosen by the experimenter, one may assume that N is much larger than n. The large sample behavior of our estimators (as $N \to \infty$) of $EX_{i:n}$ for i = 1, 2,…., n will be of particular interest. It is easy to estimate the order statistic $a_i = EX_{i:n}$ consistently if one is willing to make simple but restrictive assumptions on the distribution function F. For example, under the assumption that F has bounded support, one can quite easily obtain consistent estimators $\{a_i^*\}$ and show that the signature of the approximately optimal system converges to that of the optimal system as $N \to \infty$. However, the assumptions to which we've alluded fail to apply to standard lifetime distributions F of practical interest. We thus turn to a more general framework for ensuring the desired asymptotic behavior of "optimal system" corresponding to our estimator **a***.

Stigler (1969, 1974) and Shorack (1969, 1972) developed the theory for the large sample behavior of L estimators under a variety of scenarios. Under various sets of assumptions, L-estimators are shown to be $\sqrt{N}$ - consistent estimators of their target parameter. Moreover, a suitably standardized version of the statistic will be asymptotically normal. The strongest versions of such results have been obtained by van Zwet (1980) and by Helmers (1982). Applying the tools and ideas of the theory of L-statistics, we have developed a viable theory for the approximation of optimal systems in problems of practical interest. The estimator of $\mu_{i:n} = EX_{i:n}$, for i = 1, 2, … n, that we propose for study is the L-statistic given by

$$\mu^*_{i:n} = \frac{1}{N} \sum_{j=1}^{N} w_{i,j} X_{j:N}, \tag{5.1}$$

where

$$w_{i,j} = N \int_{(j-1)/N}^{j/N} [\Gamma(n+1) / \Gamma(i)\, \Gamma(n-i+1)]\, u^{i-1} (1-u)^{n-i}\, du. \tag{5.2}$$

The following asymptotic results regarding $\mu^*_{i:n}$ have been established. Proofs may be found in Dugas and Samaniego (2004).

Theorem 4. If the underlying distribution F of the iid components of all mixed systems of order n has a finite second moment, then $\mu^*_{i:n} \xrightarrow{p} \mu_{i:n}$ as the size N of the auxiliary sample grows to ∞.

Theorem 5. If the underlying distribution F of the iid components of all mixed systems of order n has a finite third moment, and if F is nondegenerate, then

$$\sqrt{N}\,(\mu^*_{i:n} - \mu_{i:n}) \xrightarrow{D} Y \sim N(0, \sigma_i^2) \qquad \text{as } N \to \infty, \tag{5.3}$$

where

$$\sigma_i^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w_i(F(x))\, w_i(F(y))[F(\min(x,y)) - F(x)\, F(y)]\, dxdy, \tag{5.4}$$

and

$$w_i(u) = [\Gamma(n+1) / \Gamma(i)\, \Gamma(n-i+1)]\, u^{i-1} (1-u)^{n-i} \qquad \text{for } 0 \le u \le 1. \tag{5.5}$$

The results above allow one to estimate the criterion function m_r of (3.5) with arbitrary accuracy. The continuity of the function m_r in the vector **a** ensures the convergence $m_r(\mathbf{a^*}, \mathbf{c}, \mathbf{s}) \xrightarrow{p} m_r(\mathbf{a}, \mathbf{c}, \mathbf{s})$ as $N \to \infty$ for each fixed **c** and **s**, so that the value of the criterion function m_r for the approximately optimal signature **s*** converges as $N \to \infty$ to that of the optimal signature relative to the true but unknown vector **a**.

References

Barlow, R. E. and Proschan, F. (1981) Statistical Theory of Reliability and Life Testing, Silver Springs, MD: To Begin With Press.

Birnbaum, Z. W., Esary, J. D. and Saunders, S. C. (1961) "Multicomponent systems and structures and their reliability", Technometrics, 3, 55 – 77.

Blischke, W. and Murthy, D (1996) Product Warranty Handbook, New York: M. Dekker, Inc.

Boland, P. and Samaniego, F. (2004) "The signature of a coherent system and its applications in reliability", in Mathematical Reliability: An Expository Perspective, Soyer, R., Mazzuchi, T. and Singpurwalla, N. (Editors), New York: Klewer, 1 – 30.

Boland, P., Samaniego, F. and Vestrup, E. (2003) "linking Dominations and Signatures in network Reliability Theory", in Mathematical and Statistical Methods in Reliability, Lindqvist, B. and Doksum, K. (Editors), Singapore: World Scientific, 89 - 103

Deeter, A. and Smith, M. (1998). "Economic Design of Reliable Networks", IIE Transactions, 30, 1161 – 1174.

Dugas Michael R. and Samaniego, Francisco J. (2004) "New Results on Optimal System Design in Reliability Economics", Technical Report No. 488, Dept. of Statistics, Univ. of Calif., Davis.

Helmers, R. (1982) Edgeworth Expansions for Linear Combinations of Order Statistics, Mathematical Centre Tract 105, Amsterdam: Mathematisch Centrum.

Kochar, S., Mukerjee, H. and Samaniego, F. (1999) "On the signature of a coherent system and its application to comparisons among systems", Naval Research Logistics, 507 – 523.

Kohoutek, H. (1966) "Economics of Reliability" in Handbook of Reliability Engineering and Management, Ireson, W. G. and Coombs, C. (Editors), New York: McGraw-Hill.

Kuo, W., Prasad, V. R., Tillman, F. A and Hwang, C.L. (2001) Optimal Reliability Design, Cambridge University Press.

Samaniego, F. (1985) "On the closure of the IFR class under the formation of coherent systems", IEEE Transactions on Reliability, TR-34, 69 – 72.

Satyanarayana, A. and Prabhakar, A. (1978) "New topological formula and rapid algorithm for reliability analysis of complex networks", IEEE Transactions on Reliability, TR-27, 82 – 100.

Shorack, G. (1969) "Asymptotic Normality of linear combinations of functions of order statistics", Annals of Math Stat, 40, 2041 – 2050.

Shorack, G. (1972) "Functions of order statistics", Annals of Math Stat, 43, 412 – 427.

Singpurwalla, N. (2004) "Warranty: a surrogate for reliability", in Mathematical Reliability: An Expository Perspective, Soyer, R., Mazzuchi, T. and Singpurwalla, N. (Editors), New York: Klewer, 317 - 333.

Stigler, S. (1969) "Linear functions of order statistics", Annals of Math Stat, 40, 770 – 778.

Stigler, S. (1974) "Linear functions of order statistics with smooth weight functions" Annals of Stat, 2, 676 - 693.

Usher, J.., Kamal, A. and Sayed, H. (1998) Cost optimal preventive maintenance and replacement scheduling, IIE Transactions, 30, 1121 – 1128.

van Zwet, W. (1980) "A Strong law for linear functions of order statistics", Annals of Probability, 8, 986-990.

Recursive Bipartite Spectral Clustering for Document Categorization

Jeff Solka[1,2], Avory Bryant[1], and Edward J. Wegman[3]

1 – NSWCDD Code B10
2 – SCS GMU
3 – Department of Applied and Engineering Statistics GMU

Army Conference on Applied Statistics, Atlanta, 2004

Agenda

- Our treatise.
- Our datasets.
- Our features.
- Mathematical background.
- Results on the Science News data.
- Results on the ONR ILIR data.

In a Nutshell?

- What are we trying to do?
  - Develop a semi-automated system to facilitate the text data mining
    - Discovery of articles from disparate corpora that may contain subtle relationships.
    - Discovery of interesting clusters of articles.
- What is our approach predicated on?
  - The synthesis of methodologies from statistics, mathematics and visualization.
  - Use of minimal spanning trees and spectral graph theory as technological enablers.
- What are the test cases?
  - Roughly 1200 Science News abstracts that have been precategorized into 8 categories.
  - Roughly 343 Office of Naval Research In-house Laboratory Independent Research documents.

The Science News Corpus

- 1117 documents from 1994–2002.
- Obtained from the SN website on December 2002 19,2002 using wget.
- Each article ranges from 1/2 a page to roughly a page in length.
- The corpus html/xml code was subsequently parsed into straight text.
- The corpus was read through and categorized into 8 categories.

The Science News Corpus Breakdown

- Anthropology and Archeology (48).
- Astronomy and Space Sciences (124).
- Behavior (88).
- Earth and Environmental Sciences (164).
- Life Sciences (174).
- Mathematics and Computers (65) .
- Medical Sciences (310) .
- Physical Sciences and Technology (144)

The Office of Naval Research (ONR) In-House Laboratory Independent Research (ILIR) Corpus

- 343 Documents
- Obtained from ONR
- Support on-line querying and mining of their ILIR database

ILIR Corpus Breakdown

- Advanced Naval Materials (82)
- Air Platforms and Systems (23)
- Electronics
- Expeditionary/USMC
- Human Performance /Factors (49)
- Information Technology and Operations (18)
- Manufacturing Technologies (21)
- Medical S&T (19)
- Naval & Joint Experimentation
- Naval Research Enterprise Programs
- Operational Environments (27)
- RF Sensing, Surveil, & Countermeasures (27)
- Sea Platform and Systems (38)
- Strike Weapons
- Undersea Weapons
- USW-ASW (5)
- USW-MIW (17)
- Visible and IR Sensing, Surveil & Countermeasures (17)

Denoising and Stemming

- These steps are performed prior to subsequent feature extraction steps.

- Various approaches to denoising were used
  - Simplest consists of removal of all words that appear on a stopper or noise word list.
  - the, a, an, …
  - More on this later

- Stemming transforms a given word into its base
  - walking → walk
  - walked → walk

- Denoising is implemented within the current system stemming is implemented in some versions but is not in others

Document Features

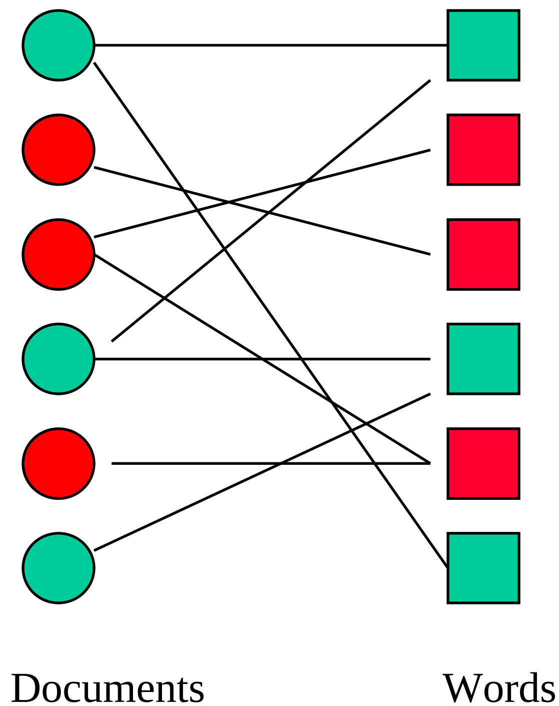
- Bigram Proximity Matrices ala Martinez 2002
  - Angel Martinez, "A Framework for the Representation of Semantics," *Ph.D Dissertation under the direction of Edward Wegman*, October 2002.

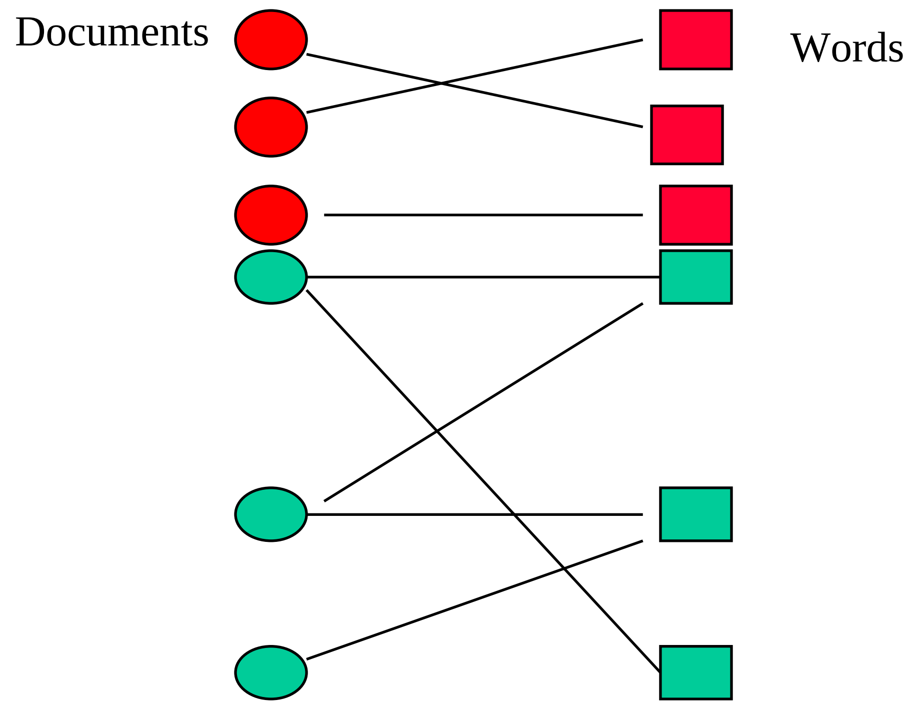
- Mutual Information Features ala Lin 2002
  - Patrick Pantel and Dekang Lin, "Discovery word senses from text," in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pgs. 613-619, 2002.

- "Normalized" term document matrices ala Dhillon 2001
  - Inderjit S. Dhillon, "Co-clustering documents and words using Bipartite Spectral Graph Partitioning," UT CS Technical Report # TR 2001-05.

Bipartite Spectral Based Clustering

- Inderjit S. Dhillon, "Co-clustering documents and words using Bipartite Spectral Graph Partitioning," KDD 2001.





Cut measures the sum of the crossing between vertex set V_1 and vertex set V_2.

$$\mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2) = \sum_{i \in \mathcal{V}_1, j \in \mathcal{V}_2} M_{ij}$$

The Graph Theoretic Formulation

Our Graph	Vertex Set	Edge Set	Edge Weights
$G = (\mathcal{V}, E)$	$\mathcal{V} = \{1, 2, \ldots, \lvert\mathcal{V}\rvert\}$	$\{i, j\}$	E_{ij}

Adjacency Matrix

$$M = \begin{cases} E_{ij}, & \text{if there is an edge } \{i, j\}, \\ 0, & \text{otherwise.} \end{cases}$$

The cut between two subsets of vertices.

$$\text{cut}(\mathcal{V}_1, \mathcal{V}_2) = \sum_{i \in \mathcal{V}_1, j \in \mathcal{V}_2} M_{ij}.$$

The cut between k subsets of vertices.

$$\text{cut}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k) = \sum_{i<j} \text{cut}(\mathcal{V}_i, \mathcal{V}_j)$$

The Document Word Bipartite Model

Our graph consisting of a vertex set consisting of documents and words along with associated edges.

$$G = (\mathcal{D}, \mathcal{W}, E)$$

The word vertices.

$$\mathcal{W} = \{w_1, w_2, \ldots, w_m\}$$

The document vertices.

$$\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$$

One strategy for setting the edge weights.

$$E_{ij} = t_{ij} \times \log\left(\frac{|\mathcal{D}|}{|\mathcal{D}_i|}\right)$$

where t_{ij} is the number of times word w_i occurs in document d_j, $|\mathcal{D}| = n$ is the total number of documents and $|\mathcal{D}_i|$ is the number of documents that contain word w_i.

$$M = \begin{bmatrix} \mathbf{0} & A \\ A^T & \mathbf{0} \end{bmatrix}$$

Adjacency Matrix – $A_{ij} = E_{ij}$, 0's reflect no word to word or document to document connections

$$\text{cut}(\mathcal{W}_1 \cup \mathcal{D}_1, \mathcal{W}_2 \cup \mathcal{D}_2, \ldots, \mathcal{W}_k \cup \mathcal{D}_k) = \min_{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k} \text{cut}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k)$$

Our Clustering Criteria

Corpus Dependent Stop Word Removal

- Stop words are removed.
- Words occurring in less than 0.2% of the documents are removed.
- Words occurring in greater than 15% of the documents are removed.
- N. B.
  - The methodology has been shown successful even if stopper words are not removed.
  - 0.2% and 15% are user “tunable” parameters.

Graph Partitioning

Given a graph $G = (\mathcal{V}, E)$, the classical graph bipartitioning or bisection problem is to find nearly equally-sized vertex subsets $\mathcal{V}_1^*, \mathcal{V}_2^*$ of $\mathcal{V}$ such that

$$\mathrm{cut}(\mathcal{V}_1^*, \mathcal{V}_2^*) = \min_{\mathcal{V}_1, \mathcal{V}_2} \mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2).$$

The graph partitioning problem is known to be NP-complete.

We will follow Dhillon and use graph spectral methods to obtain an approximate solution based on a suitably formulated objective function.

Assuring An Equitable Partition – An Objective Function

$$W_{ij} = \begin{cases} \text{weight}(i), & i = j, \\ 0, & i \neq j. \end{cases}$$

The weight for a particular vertex.

$$\text{weight}(\mathcal{V}_l) = \sum_{i \in \mathcal{V}_l} \text{weight}(i) = \sum_{i \in \mathcal{V}_l} W_{ii}$$

The weight for a set of vertices.

A figure of merit function that helps assure near equal number of points in each cluster.

$$\mathcal{Q}(\mathcal{V}_1, \mathcal{V}_2) = \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\text{weight}(\mathcal{V}_1)} + \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\text{weight}(\mathcal{V}_2)}$$

One can think of this as being analogous to the ratio of between group and within group distances in our usual statistical clustering framework.

Choice of Vertex Weights

$$\text{weight}(i) = 1$$

$$\text{Ratio-cut}(\mathcal{V}_1, \mathcal{V}_2) = \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{|\mathcal{V}_1|} + \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{|\mathcal{V}_2|}$$

$$\text{weight}(i) = \sum_k E_{ik}$$

Normalized cut.

$$\mathcal{N}(\mathcal{V}_1, \mathcal{V}_2) = \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\sum_{i \in \mathcal{V}_1} \sum_k E_{ik}} + \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\sum_{i \in \mathcal{V}_2} \sum_k E_{ik}}$$

Algorithm Bipartition

$$D_1(i,i) = \sum_j A_{ij} \quad \text{(sum of edge-weights incident on word } i\text{)},$$

$$D_2(j,j) = \sum_i A_{ij} \quad \text{(sum of edge-weights incident on document } j\text{)}.$$

$$\boldsymbol{z}_2 = \begin{bmatrix} \boldsymbol{D_1}^{-1/2}\boldsymbol{u}_2 \\ \boldsymbol{D_2}^{-1/2}\boldsymbol{v}_2 \end{bmatrix} \quad (4.13)$$

Algorithm Bipartition

1. Given $\boldsymbol{A}$, form $\boldsymbol{A_n} = \boldsymbol{D_1}^{-1/2}\boldsymbol{A}\boldsymbol{D_2}^{-1/2}$.
2. Compute the second singular vectors of $\boldsymbol{A_n}$, $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$ and form the vector $\boldsymbol{z}_2$ as in (4.13).
3. Run the k-means algorithm on the 1-dimensional data $\boldsymbol{z}_2$ to obtain the desired bipartitioning.

The singular vectors $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$ of $\boldsymbol{A_n}$ give a real approximation to the discrete optimization problem of minimizing the normalized cut.

The Left and Right Singular Vectors

$$\mathbf{A}_n \boldsymbol{v}_2 = \sigma_2 \boldsymbol{u}_2, \qquad \mathbf{A}_n{}^T \boldsymbol{u}_2 = \sigma_2 \boldsymbol{v}_2, \qquad (4.12)$$

$$\sigma_2 = 1 - \lambda_2$$

The right singular vector $\boldsymbol{v}_2$ will give us a bipartitioning of documents while the left singular vector $\boldsymbol{u}_2$ will give us a bipartitioning of the words. By examining the relations (4.12) it is clear that this solution agrees with our intuition that a partitioning of documents should induce a partitioning of words, while a partitioning of words should imply a partitioning of documents.

The curious fact is that the obtained transformation allows one to map the documents and words into the same one-dimensional space.

Algorithm Multipartition(k)

$$Z = \begin{bmatrix} D_1^{-1/2} U \\ D_2^{-1/2} V \end{bmatrix} \quad (4.14)$$

$$U = [u_2, u_3, \ldots, u_{\ell+1}], \text{ and } V = [v_2, v_3, \ldots, v_{\ell+1}], \quad \ell = \lceil \log_2 k \rceil$$

Algorithm Multipartition(k)

1. Given A, form $A_n = D_1^{-1/2} A D_2^{-1/2}$.
2. Compute $\ell = \lceil \log_2 k \rceil$ singular vectors of A_n, $u_2, u_3, \ldots u_{\ell+1}$ and $v_2, v_3, \ldots v_{\ell+1}$ and form the matrix Z as in (4.14).
3. Run the k-means algorithm on the ℓ-dimensional data Z to obtain the desired k-way multipartitioning.

How Do We Know That the Dhillon 2001 Strategy is Worthwhile - I

- Confusion Matrix Performance Measures
  - Inderjit S. Dhillon, "Co-clustering documents and words using Bipartite Spectral Graph Partitioning," KDD 2001.
  - Inderjit S. Dhillon, " Co-clustering documents and words using Bipartite Spectral Graph Partitioning," Ut CS Technical Report # TR 2001-05.
  - These were obtained using "mixtures" of MEDLINE (medical database), CISI (Institute of Scientific Information database), and CRANFIELD (document searching database) document sets along with YAHOO_K5 (Reuter News Articles from Yahoo where words are stemmed and heavily pruned) and YAHOO_K1 (Reuters News Articles from Yahoo: words are stemmed and only stop words are pruned)

How Do We Know That the Dhillon 2001 Strategy is Worthwhile - II

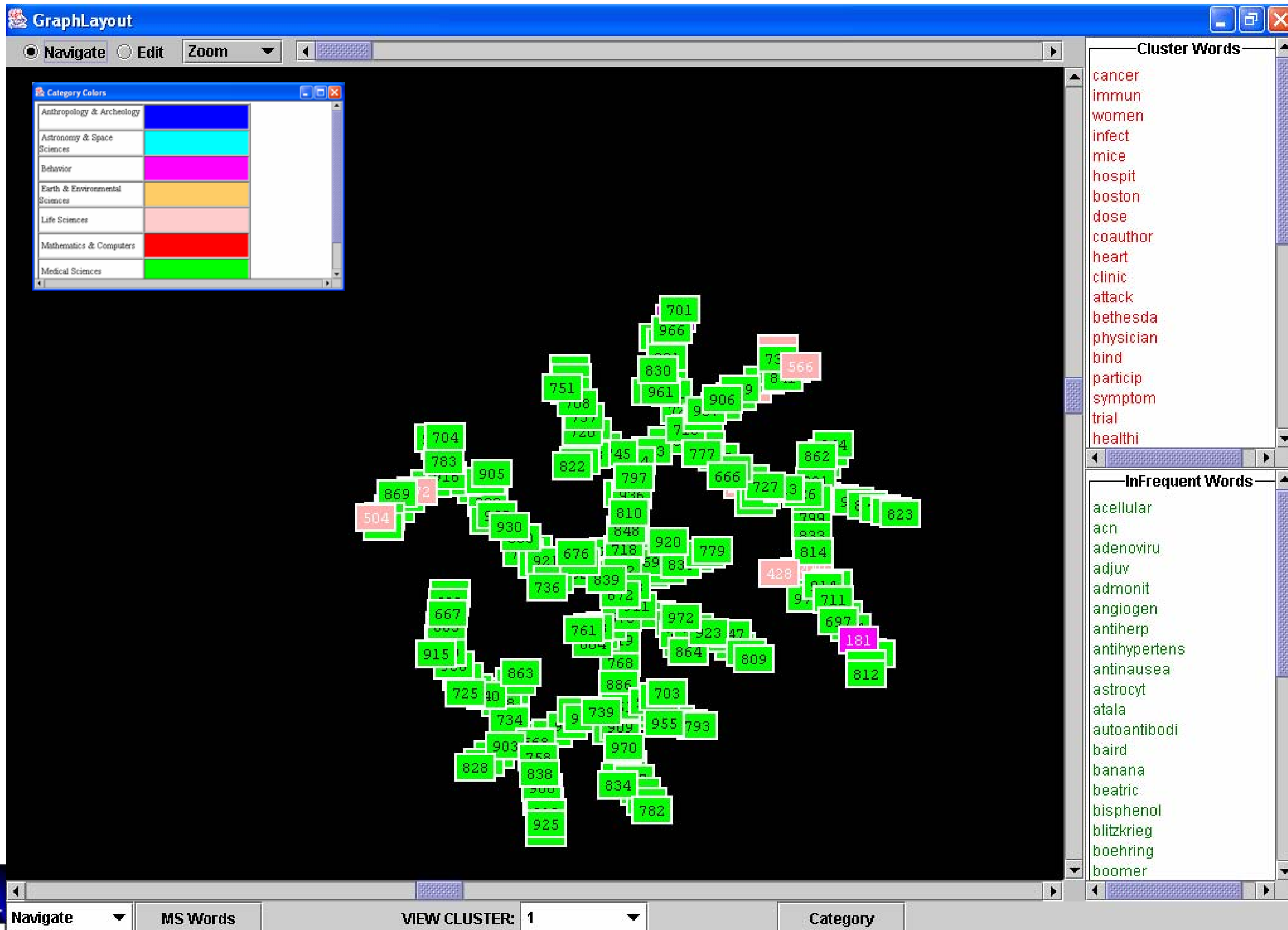
- Confusion matrix performance on the
  - Science News
  - ONR ILIR Data

- Theoretical results that insure us that the spectral based approach is a good approximation to solving the NP-compete problem.

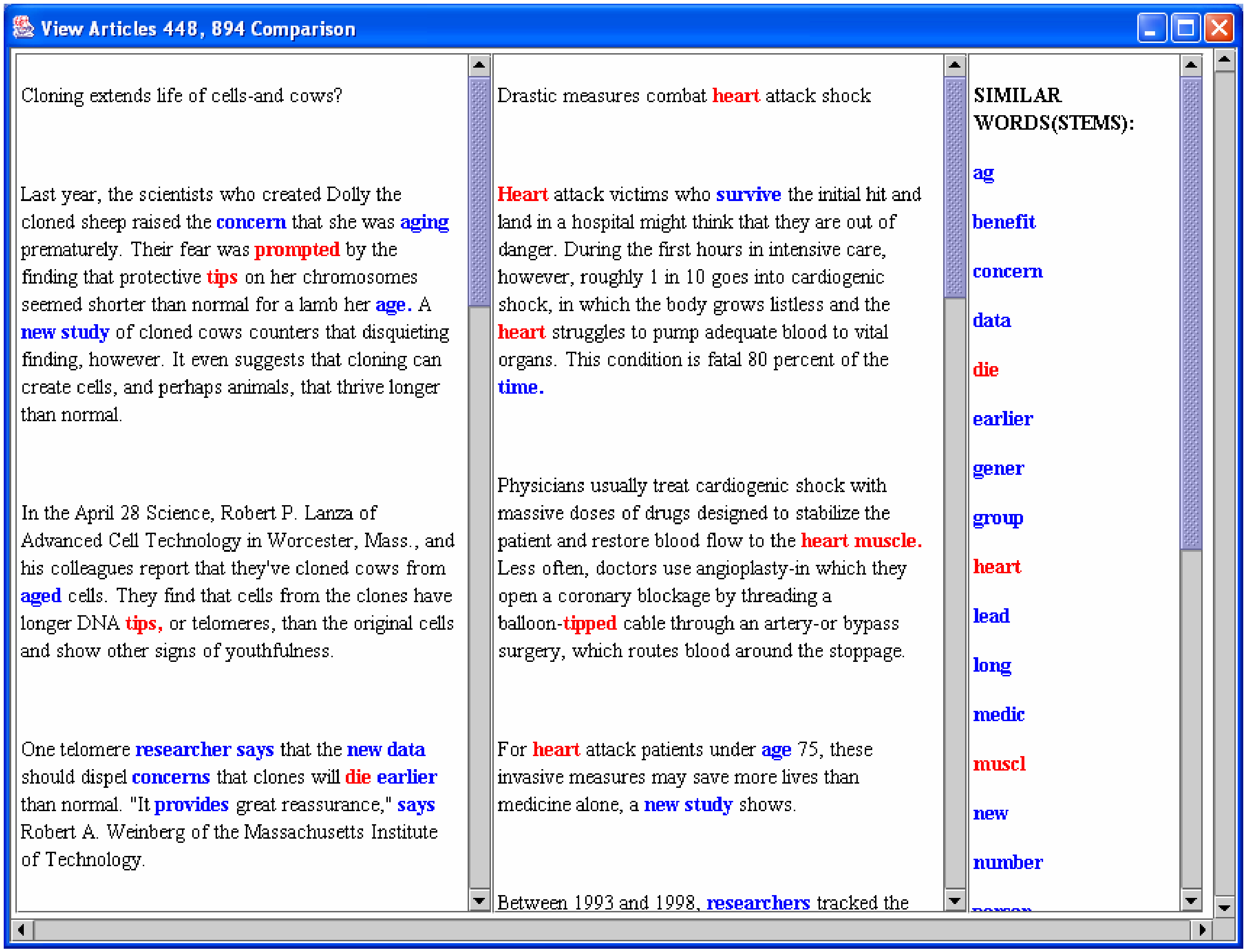
Recursive Bipartite Bipartition Methodology of Solka, Bryant and Wegman

- Alternative to the multipartition approach.
- Recursively use the bipartite bipartition methodology to obtain a multipartition of the data.
- Which cluster to split next is currently based on a simple mean distance of all observations to the centroid measure.
  - Certainly could be the subject of a more advanced statistical methodology.
- A visualization framework for exploration of the clusters (documents and words) and their associated concepts is provided.

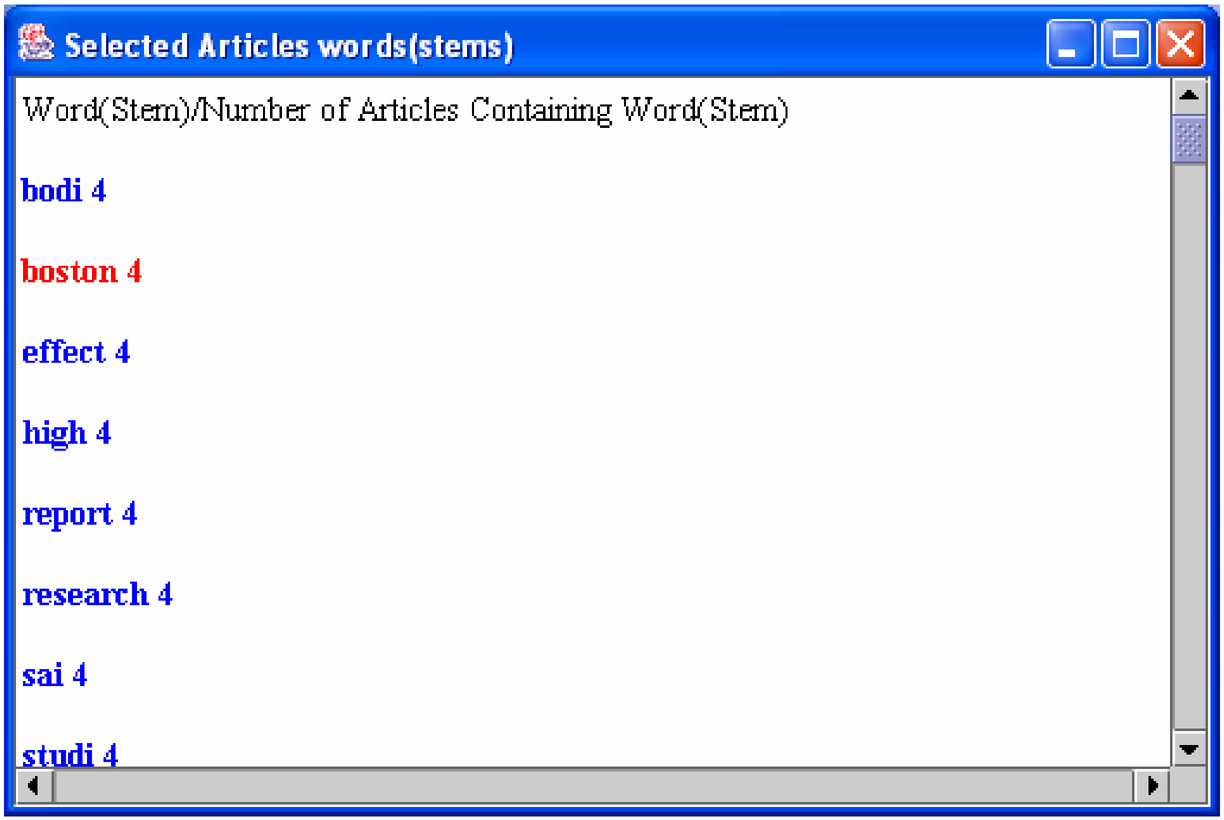
Visualization Framework - I



Visualization Framework – II (Comparison File for a Biology and Medical Sciences Article)



Visualization Framework - III (Multi-select Operation)



How Do We Measure the Quality of Our Clustering

- The clustering figure of merit is based on the ability of the methodology to match a set of user obtained categorizations.

- Deviations from these categorizations are measured via cluster:
  - Purity
  - Entropy

Purity

- A large value of purity indicates a good cluster.

$$P(D_j) = \frac{1}{n_j} \max_i (n_j^{(i)})$$

$$n_j = |D_j| \textit{ and } n_j^i \textit{ is the number of documents}$$
$$\textit{in } D_j \textit{ that belong to class } i$$

Entropy

- A small value of entropy indicates a good cluster.

$$H(D_j) = -\frac{1}{\log(c)} \sum_{i=1}^{c} \frac{n_j^{(i)}}{n_j} \log\left(\frac{n_j^{(i)}}{n_j}\right)$$

Science News Spectral Clustering Results

Average Purity Per Observation

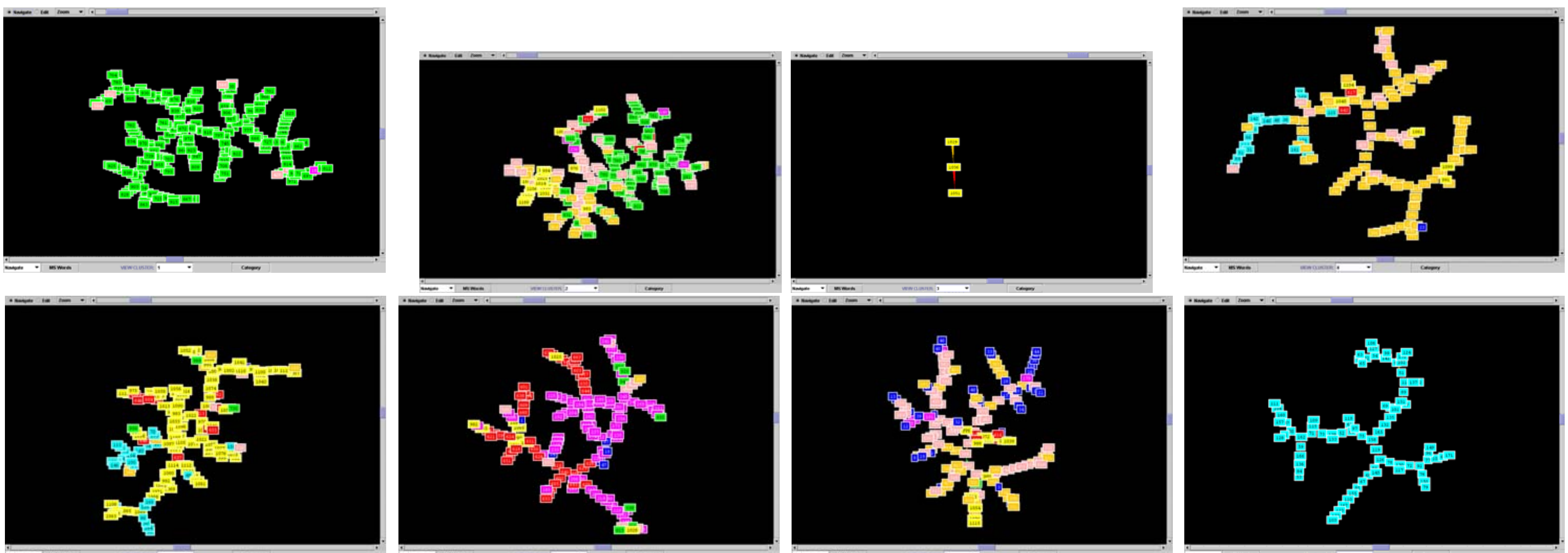
$$\hat{P} = \frac{\sum_{i=1}^{c} n_i P(D_i)}{n}$$

Average Purity Per Cluster

$$\overline{\overline{P}} = \frac{\sum_{i=1}^{c} n_i P(D_i)}{c}$$

Science News 8 Multi-partitioning

ANTHROPOLOGY & ARCHEOLOGY	ASTRONOMY & SPACE SCIENCES
BEHAVIOR	EARTH & ENVIRONMENTAL SCIENCES
LIFE SCIENCES	MATHEMATICS & COMPUTERS
MEDICAL SCIENCES	PHYSICAL SCIENCE & TECHNOLOGY



Science News 8 Multi-Partitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Cluster1	0	0	3	0	9	0	208	0
Cluster2	2	0	5	32	77	4	91	20
Cluster3	0	0	0	0	0	0	0	3
Cluster4	1	19	0	89	18	2	0	5
Cluster5	1	25	0	6	5	12	3	95
Cluster6	7	0	75	1	8	43	7	9
Cluster7	37	0	5	36	57	4	1	12
Cluster8	0	80	0	0	0	0	0	0

Class 1 is anthropology and archaeology, class 2 astronomy and space sciences, class 3 is behavior, class 4 is earth and environmental sciences, class 5 is life sciences, class 6 is mathematics and computers, class 7 is medical sciences, and class 8 is physical sciences and technology.

SciNews 8 Multi-Partitioning Purity & Entropy

PURITY

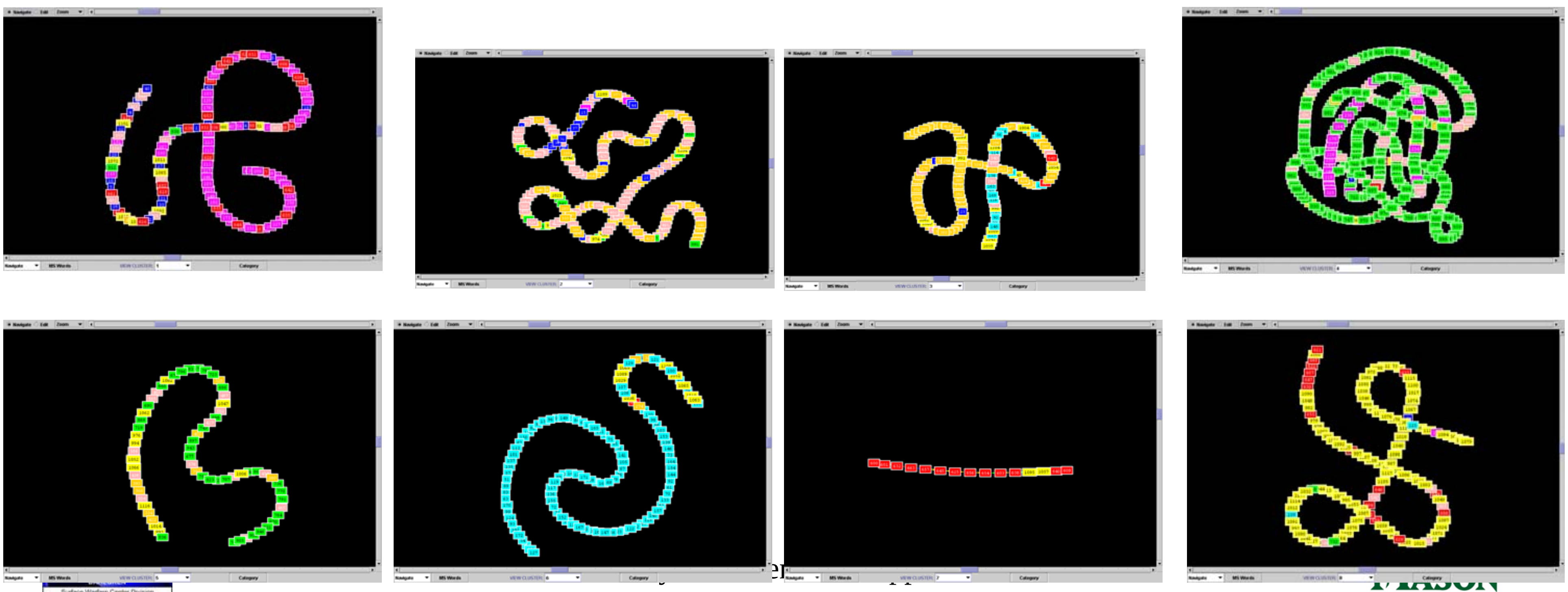
Cluster1	0.9454545454545454
Cluster2	0.3939393939393939
Cluster3	1.0
Cluster4	0.664179104477612
Cluster5	0.6462585034013606
Cluster6	0.5
Cluster7	0.375
Cluster8	1.0
Avg Purity	0.690603943409114
Avg Purity Per Observation	0.6248880931065354
Avg Aggregate Purity Per Cluster	87.25

ENTROPY

Cluster1	0.1165507016468692
Cluster2	0.6795872298749444
Cluster3	0.0
Cluster4	0.500340206242551
Cluster5	0.5515312792869759
Cluster6	0.6488882742515858
Cluster7	0.7186710223086614
Cluster8	0.0
Avg Entropy:	0.4019460892014485
Avg Entropy Per Observation	0.4810364607732902
Avg Aggregate Entropy Per Cluster	67.16471583547064

Science News 8 Recursive Bi-partitioning

ANTHROPOLOGY & ARCHEOLOGY
BEHAVIOR
LIFE SCIENCES
MEDICAL SCIENCES
ASTRONOMY & SPACE SCIENCES
EARTH & ENVIRONMENTAL SCIENCES
MATHEMATICS & COMPUTERS
PHYSICAL SCIENCE & TECHNOLOGY



Science News 8 Recursive-Bipartitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Cluster1	19	0	55	1	10	29	2	9
Cluster2	25	0	3	69	79	0	10	9
Cluster3	2	20	0	75	10	2	0	13
Cluster4	1	0	29	8	57	3	263	4
Cluster5	0	0	0	8	13	0	33	11
Cluster6	0	102	0	3	0	1	0	9
Cluster7	0	0	0	0	0	13	0	2
Cluster8	1	2	1	0	5	17	2	87

Class 1 is anthropology and archaeology, class 2 astronomy and space sciences, class 3 is behavior, class 4 is earth and environmental sciences, class 5 is life sciences, class 6 is mathematics and computers, class 7 is medical sciences, and class 8 is physical sciences and technology.

.

SciNews 8 Recursive Bi-Partitioning Purity & Entropy

PURITY

Cluster1	0.44
Cluster2	0.40512820512820513
Cluster3	0.6147540983606558
Cluster4	0.7205479452054795
Cluster5	0.5076923076923077
Cluster6	0.8869565217391304
Cluster7	0.8666666666666667
Cluster8	0.7565217391304347
Avg Purity	0.64978343549036
Avg Purity Per Observation	0.6329453894359892
Avg Aggregate Purity Per Cluster	88.375

ENTROPY

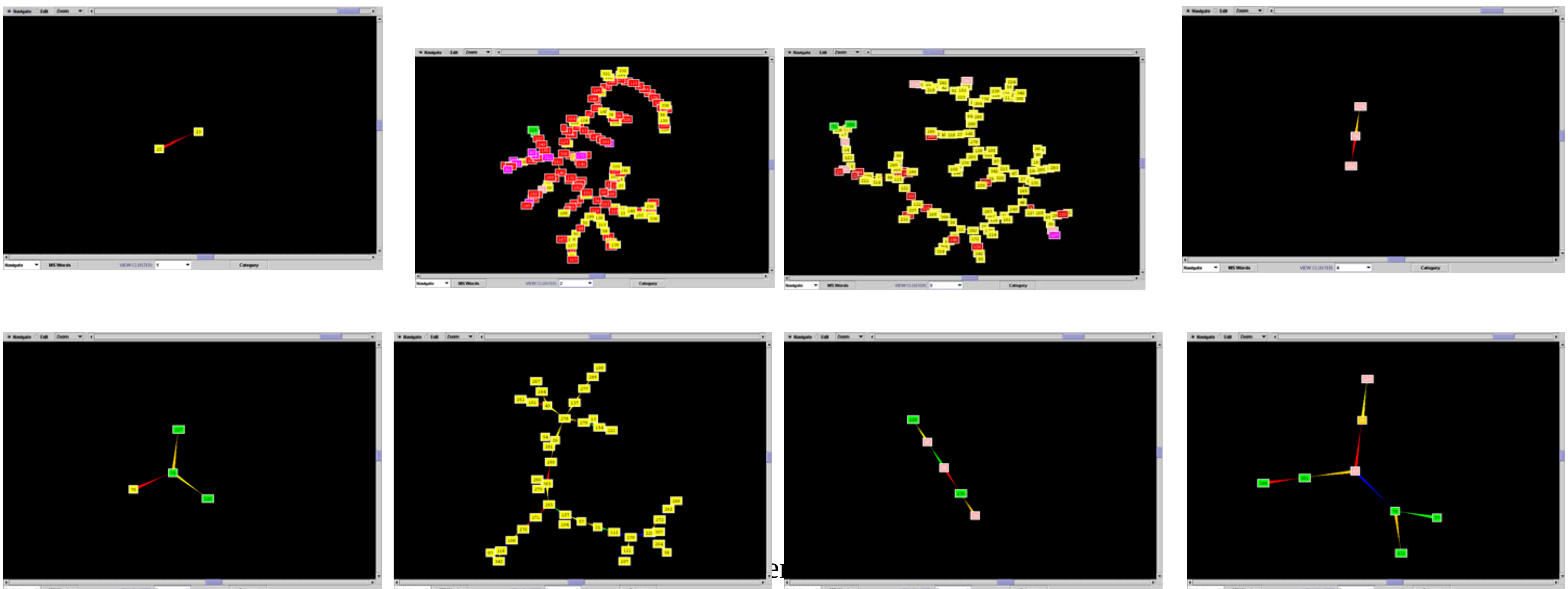
Cluster1	0.7130868633677933
Cluster2	0.6518677715369288
Cluster3	0.5645491645164644
Cluster4	0.44058717559174865
Cluster5	0.5888689116265443
Cluster6	0.21263698105033718
Cluster7	0.18883650218430179
Cluster8	0.41043119693933683
Avg Entropy	0.4713580708516819
Avg Entropy Per Observation	0.5001801770724928
Avg Aggregate Entropy Per Cluster	69.83765722374682

ONR ILIR Spectral Clustering Results

(Using Science News Categories)

ILIR 8 Multi-partitioning

ANTHROPOLOGY & ARCHEOLOGY	ASTRONOMY & SPACE SCIENCES
BEHAVIOR	EARTH & ENVIRONMENTAL SCIENCES
LIFE SCIENCES	MATHEMATICS & COMPUTERS
MEDICAL SCIENCES	PHYSICAL SCIENCE & TECHNOLOGY



ILIR 8 Multi-partitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Cluster1	0	0	0	0	0	0	0	2
Cluster2	0	0	9	2	1	83	2	46
Cluster3	0	0	1	1	5	15	2	111
Cluster4	0	0	0	0	3	0	0	0
Cluster5	0	0	0	0	0	0	3	1
Cluster6	0	0	0	0	0	0	0	43
Cluster7	0	0	0	0	3	0	2	0
Cluster8	0	0	0	1	2	0	5	0

Class 1 is anthropology and archaeology, class 2 astronomy and space sciences, class 3 is behavior, class 4 is earth and environmental sciences, class 5 is life sciences, class 6 is mathematics and computers, class 7 is medical sciences, and class 8 is physical sciences and technology.

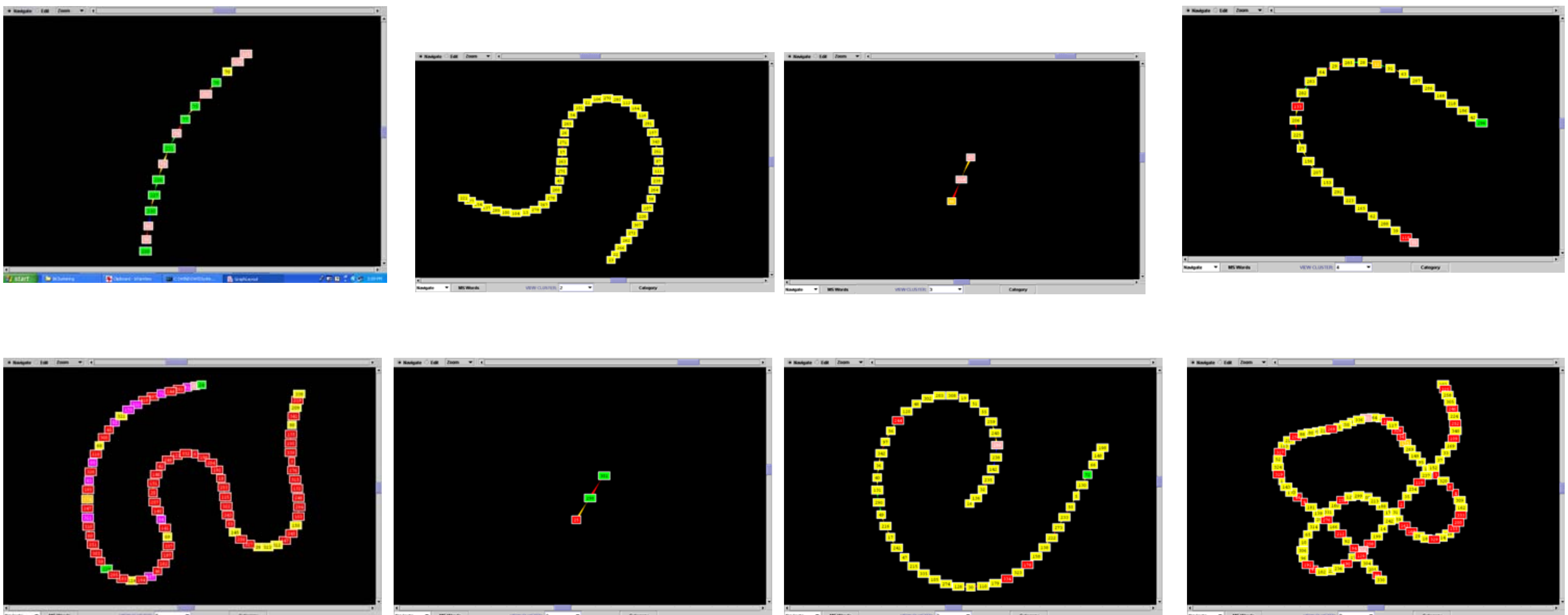
.

ILIR 8 Multi-Partitioning Purity & Entropy

PURITY		ENTROPY	
Cluster1	1.0	Cluster1	0.0
Cluster2	0.5804195804195804	Cluster2	0.48512856262450244
Cluster3	0.8222222222222222	Cluster3	0.31846159266280455
Cluster4	1.0	Cluster4	0.0
Cluster5	0.75	Cluster5	0.2704260414863776
Cluster6	1.0	Cluster6	0.0
Cluster7	0.6	Cluster7	0.32365019815155627
Cluster8	0.625	Cluster8	0.43293164689846625
Avg Purity	0.7972052253302253	Avg Entropy	0.22882475522796342
Avg Purity Per Observation	0.7376093294460642	Avg Entropy Per Observation	0.3455659119436545
Avg Aggregate Purity Per Cluster	31.625	Avg Aggregate Entropy Per Cluster	14.816138474584186

ILIR 8 Recursive-Bipartitioning

ANTHROPOLOGY & ARCHEOLOGY	ASTRONOMY & SPACE SCIENCES
BEHAVIOR	EARTH & ENVIRONMENTAL SCIENCES
LIFE SCIENCES	MATHEMATICS & COMPUTERS
MEDICAL SCIENCES	PHYSICAL SCIENCE & TECHNOLOGY



ILIR 8 Recursive-Bipartitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Cluster1	0	0	0	0	7	0	8	1
Cluster2	0	0	0	0	0	0	0	45
Cluster3	0	0	0	1	2	0	0	0
Cluster4	0	0	0	1	1	2	1	26
Cluster5	0	0	10	1	1	58	2	12
Cluster6	0	0	0	0	0	1	2	0
Cluster7	0	0	0	0	1	3	1	48
Cluster8	0	0	0	1	2	34	0	71

Class 1 is anthropology and archaeology, class 2 astronomy and space sciences, class 3 is behavior, class 4 is earth and environmental sciences, class 5 is life sciences, class 6 is mathematics and computers, class 7 is medical sciences, and class 8 is physical sciences and technology.

.

ILIR 8 Recursive Bi-Partitioning Purity & Entropy

PURITY

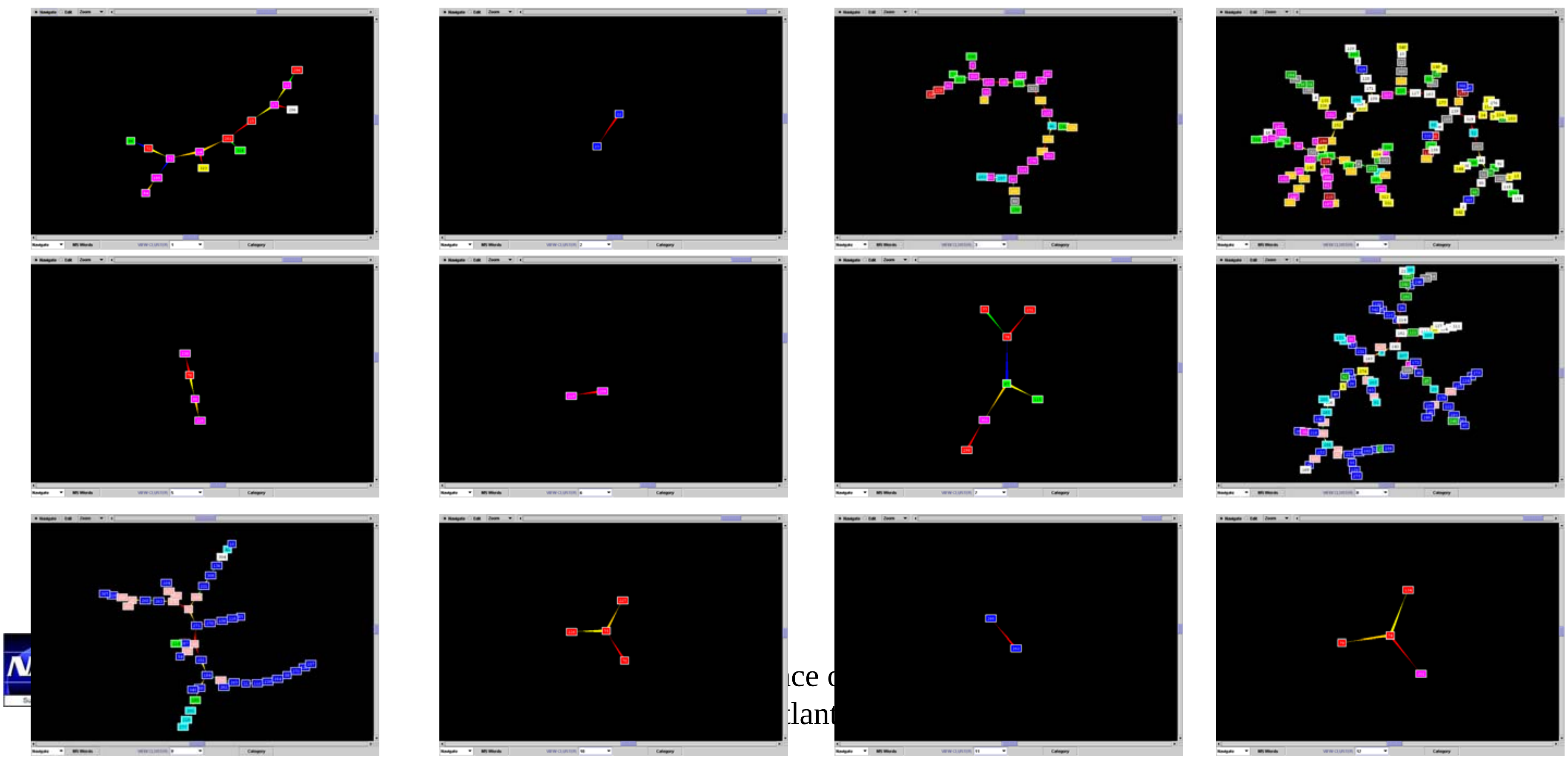
Cluster1	0.5
Cluster2	1.0
Cluster3	0.6666666666666666
Cluster4	0.8387096774193549
Cluster5	0.6904761904761905
Cluster6	0.6666666666666666
Cluster7	0.9056603773584906
Cluster8	0.6574074074074074
Avg Purity	0.7406983732493471
Avg Purity Per Observation	0.7580174927113703
Avg Aggregate Purity Per Cluster	32.5

ENTROPY

Cluster1	0.42392740719993277
Cluster2	0.0
Cluster3	0.3060986113514965
Cluster4	0.3157919672077929
Cluster5	0.4720354557082904
Cluster6	0.3060986113514965
Cluster7	0.1933754500318905
Cluster8	0.36395700045157614
Avg Entropy	0.29766056291280946
Avg Entropy Per Observation	0.3137498960545373
Avg Aggregate Entropy Per Cluster	13.452026793338288

ILIR 12 Multi-partitioning

Advanced Naval Materials	Air Platforms and Systems
Human Performance /Factors	Information Technology and Operations
Manufacturing Technologies	Medical S&T
Operational Environments	RF Sensing, Surveil, & Countermeasures
Sea Platform and Systems	USW-ASW
USW-MIW	Visible and IR Sensing, Surveil & Countermeasures



Advanced Naval Materials(1) Air Platforms and Systems(2)
Human Performance /Factors(3) Information Technology and Operations(4)
Manufacturing Technologies(5) Medical S&T(6)
Operational Environments(7) RF Sensing, Surveil, & Countermeasures(8)
Sea Platform and Systems(9) USW-ASW(10)
USW-MIW(11) Visible and IR Sensing, Surveil & Countermeasures(12)

ILIR 12 Multi-partitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Class 10	Class 11	Class 12
Cluster1	0	0	6	0	0	4	2	1	1	0	0	0
Cluster2	2	0	0	0	0	0	0	0	0	0	0	0
Cluster3	0	3	15	6	0	2	6	0	0	0	2	0
Cluster4	5	4	18	12	0	1	15	23	24	5	12	9
Cluster5	0	0	3	0	0	1	0	0	0	0	0	0
Cluster6	0	0	2	0	0	0	0	0	0	0	0	0
Cluster7	0	0	1	0	0	4	2	0	0	0	0	0
Cluster8	43	12	3	0	10	0	0	3	12	0	3	8
Cluster9	30	4	0	0	11	0	2	0	1	0	0	0
Cluster10	0	0	0	0	0	4	0	0	0	0	0	0
Cluster11	2	0	0	0	0	0	0	0	0	0	0	0
Cluster12	0	0	1	0	0	3	0	0	0	0	0	0

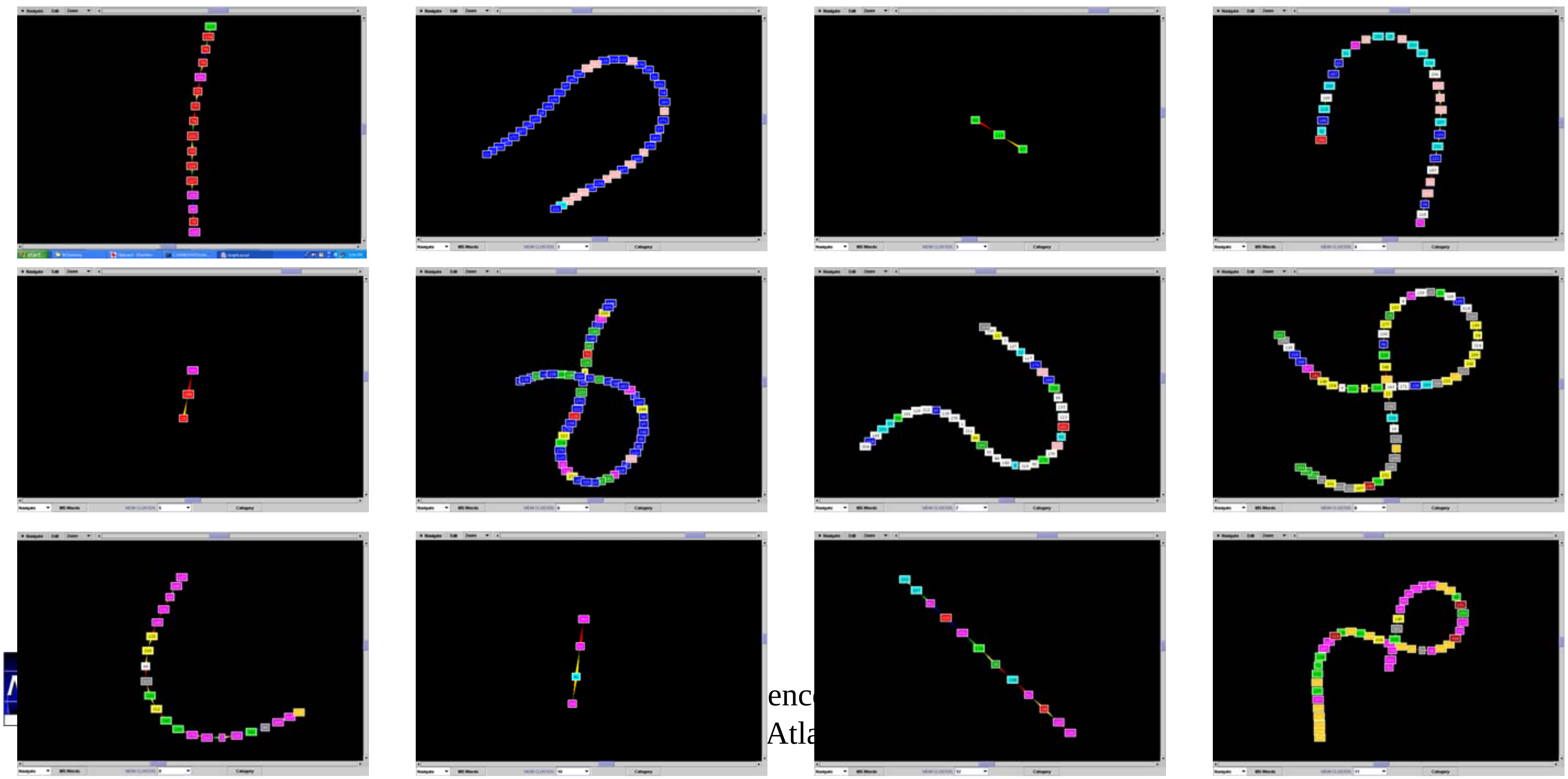
ILIR 12 Multi-Partitioning Purity & Entropy

PURITY	
Cluster1	0.42857142857142855
Cluster2	1.0
Cluster3	0.4411764705882353
Cluster4	0.1875
Cluster5	0.75
Cluster6	1.0
Cluster7	0.5714285714285714
Cluster8	0.4574468085106383
Cluster9	0.625
Cluster10	1.0
Cluster11	1.0
Cluster12	0.75
Avg Purity	0.6842602732582396
Avg Purity Per Observation	0.40233236151603496
Avg Aggregate Purity Per Cluster	11.5

ENTROPY	
Cluster1	0.5537653840548961
Cluster2	0.0
Cluster3	0.612000331799029
Cluster4	0.877077819793199
Cluster5	0.22630030977895443
Cluster6	0.0
Cluster7	0.3846019290881892
Cluster8	0.6685102839439432
Cluster9	0.4231665274665911
Cluster10	0.0
Cluster11	0.0
Cluster12	0.22630030977895443
Avg Entropy	0.330976907975311293
Avg Entropy Per Observation	0.666126132893414
Avg Aggregate Entropy Per Cluster	19.04010529853675

ILIR 12 Recursive Bi-partitioning

Advanced Naval Materials	Air Platforms and Systems
Human Performance /Factors	Information Technology and Operations
Manufacturing Technologies	Medical S&T
Operational Environments	RF Sensing, Surveil, & Countermeasures
Sea Platform and Systems	USW-ASW
USW-MIW	Visible and IR Sensing, Surveil & Countermeasures



Advanced Naval Materials(1) Air Platforms and Systems(2)
Human Performance /Factors(3) Information Technology and Operations(4)
Manufacturing Technologies(5) Medical S&T(6)
Operational Environments(7) RF Sensing, Surveil, & Countermeasures(8)
Sea Platform and Systems(9) USW-ASW(10)
USW-MIW(11) Visible and IR Sensing, Surveil & Countermeasures(12)

ILIR 12 Recursive-Bi-partitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Class 10	Class 11	Class 12
Cluster1	0	0	4	0	0	11	1	0	0	0	0	0
Cluster2	33	1	0	0	11	0	0	0	0	0	0	0
Cluster3	0	0	0	0	0	0	3	0	0	0	0	0
Cluster4	6	11	2	0	7	1	0	0	4	0	0	0
Cluster5	0	0	1	0	0	2	0	0	0	0	0	0
Cluster6	34	0	5	0	1	2	2	5	0	0	0	9
Cluster7	4	5	0	0	2	1	3	2	22	0	1	1
Cluster8	5	2	2	3	0	0	5	15	11	2	12	5
Cluster9	0	0	11	1	0	0	4	3	1	0	2	0
Cluster10	0	1	3	0	0	0	0	0	0	0	0	0
Cluster11	0	0	16	14	0	0	8	2	0	3	2	1
Cluster12	0	3	5	0	0	2	1	0	0	0	0	1

ILIR 12 Recursive Bi-Partitioning Purity & Entropy

PURITY

Cluster1	0.6875
Cluster2	0.7333333333333333
Cluster3	1.0
Cluster4	0.3548387096774194
Cluster5	0.6666666666666666
Cluster6	0.5862068965517241
Cluster7	0.5365853658536586
Cluster8	0.24193548387096775
Cluster9	0.5
Cluster10	0.75
Cluster11	0.34782608695652173
Cluster12	0.4166666666666667
Avg Purity	0.5684632674647465
Avg Purity Per Observation	0.4839650145772595
Avg Aggregate Purity Per Cluster	13.833333333333334

ENTROPY

Cluster1	0.31287377816678064
Cluster2	0.2641567106578208
Cluster3	0.0
Cluster4	0.6331562009104378
Cluster5	0.2561521449303204
Cluster6	0.5340341300193764
Cluster7	0.6340006351665233
Cluster8	0.8273794447785848
Cluster9	0.5743544330688691
Cluster10	0.22630030977895443
Cluster11	0.6308112406531853
Cluster12	0.5731120392262111
Avg Entropy	0.4555275889464219
Avg Entropy Per Observation	0.5684854885128293
Avg Aggregate Entropy Per Cluster	16.24921021332504

Future

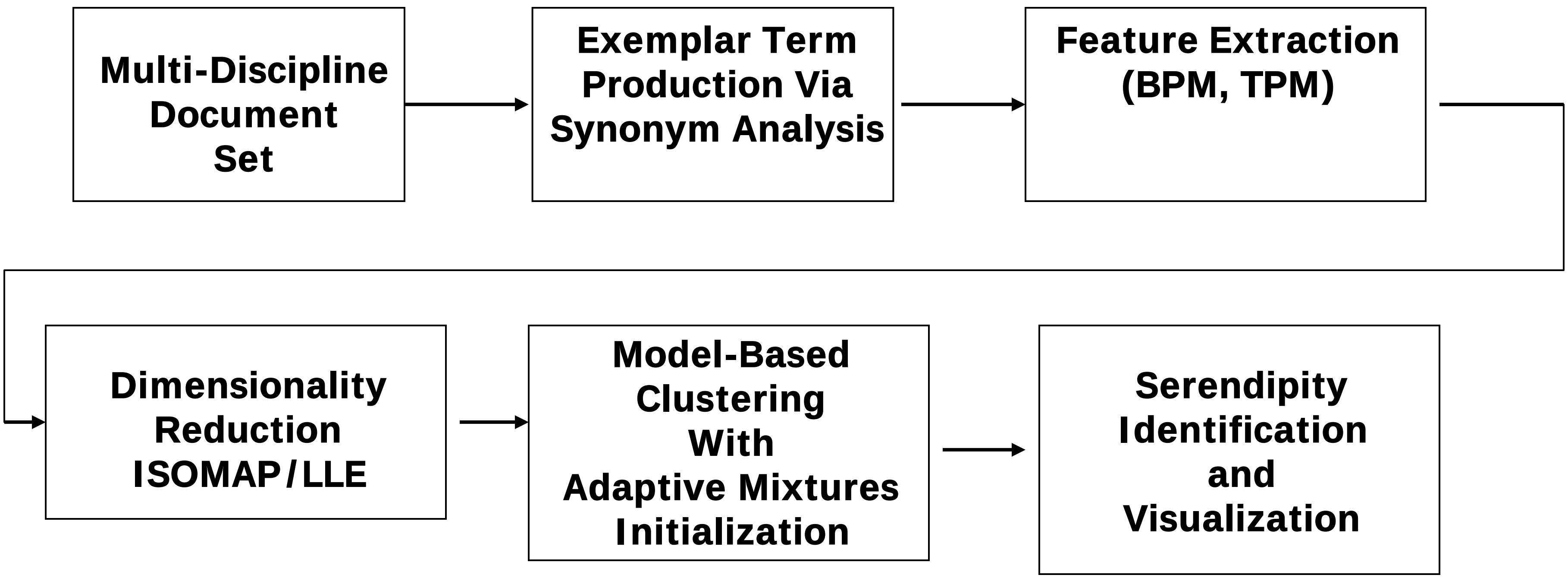
- Development of visualization frameworks that allow for simultaneous display of words and documents.
- Tree-based displays for the recursive bipartitioning tree.
- Higher dimensional visualization in the case of the multipartition algorithm.

Backup Slides

Methodology

- ILIR1:
- 12 classification categories
- Heirarchical Clustering
- Method: Average
- Tree Cut: 24

An Alternate Approach



A Paradigm

"you don't reach Serendip by plotting a course for it. You have to set out in good faith for elsewhere and lose your bearings … serendipitously."

-- John Barth, *The Last Voyage of Somebody the Sailor*

Acknowledgements

- Jim Gentle (Opportunity to speak)

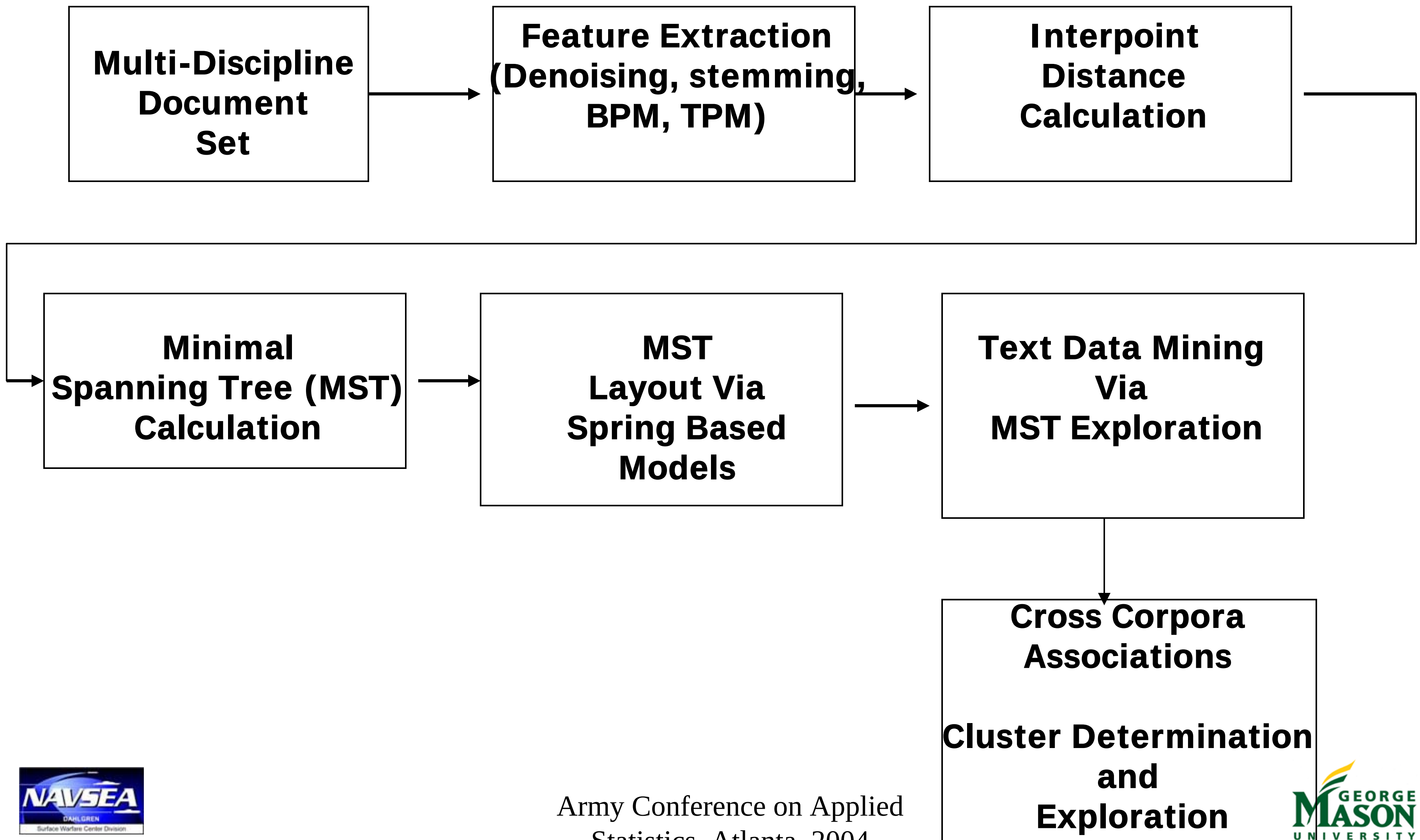
- Algotek (Funding and Program Management)
  - Anna Tsao

- Algotek Team (Helpful discussions and encouragement)
  - Carey Priebe
  - David Marchette

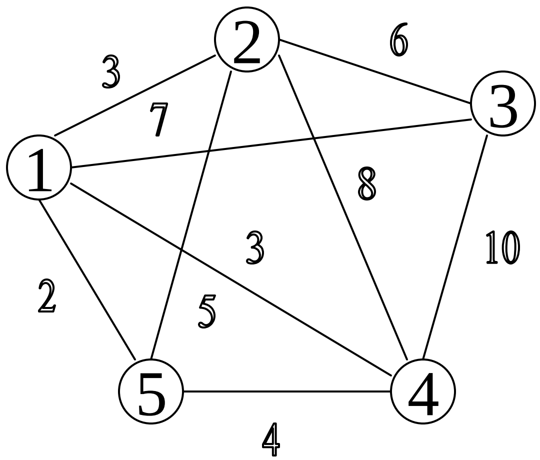
The Porter Stemming Algorithm

- "The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems. " ('official' home page for distribution of the Porter Stemming Algorithm http://www.tartarus.org/~martin/PorterStemmer)

Our Approach to be Discussed Today

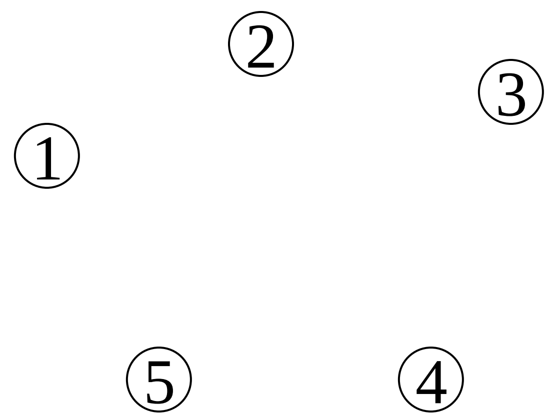


Calculation of the MST : Kruskal's Algorithm



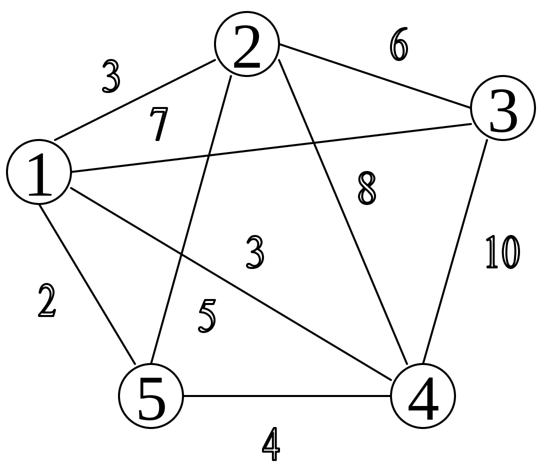
Undirected Network

2
3
3
4
5
6
7
8
10



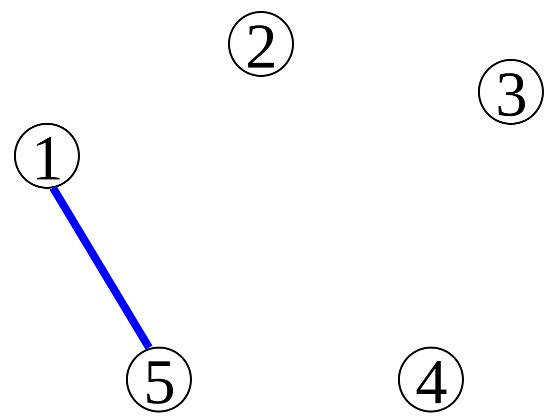
Minimum Spanning Tree

Calculation of the MST : Kruskal's Algorithm



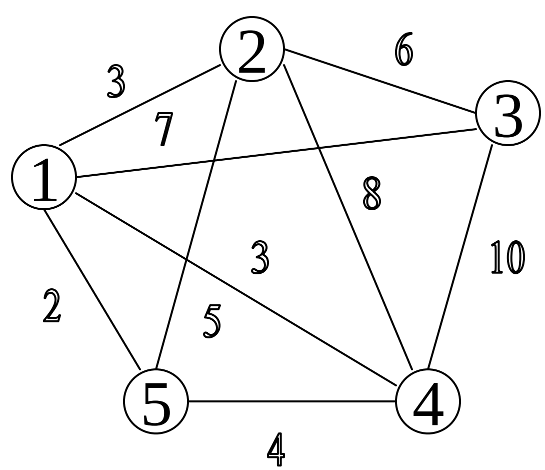
Undirected Network

3
3
4
5
6
7
8
10



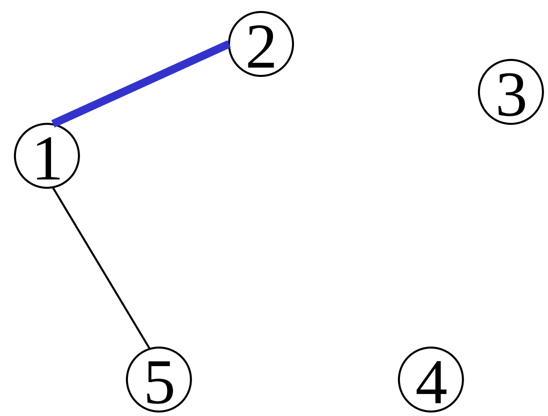
Minimum Spanning Tree

Calculation of the MST : Kruskal's Algorithm



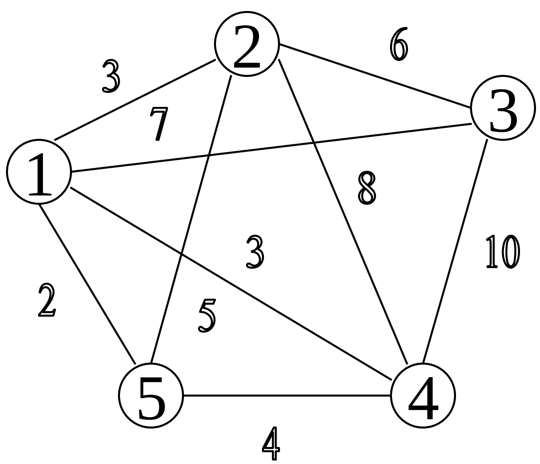
Undirected Network

3
4
5
6
7
8
10



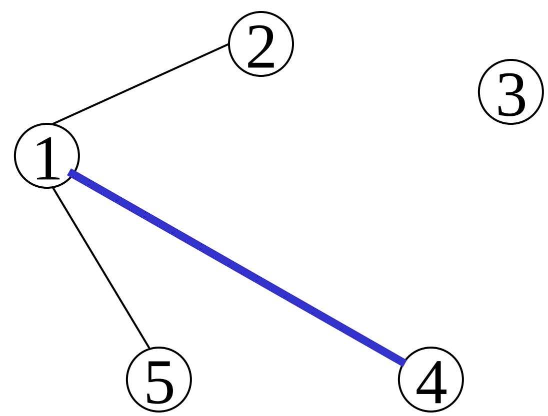
Minimum Spanning Tree

Calculation of the MST : Kruskal's Algorithm



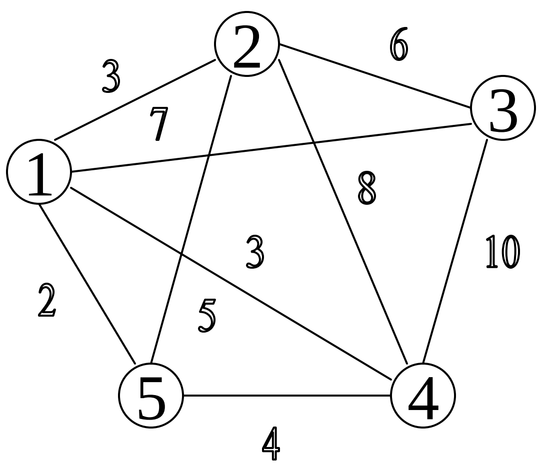
Undirected Network

4
5
6
7
8
10



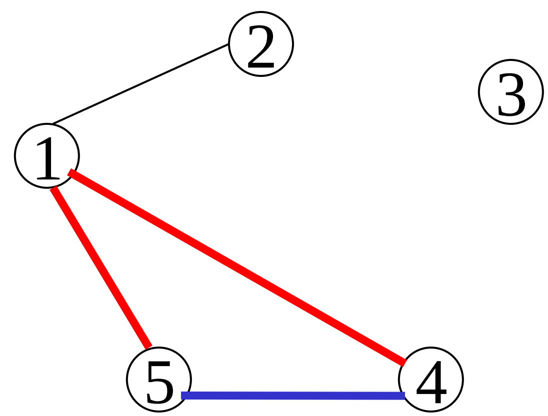
Minimum Spanning Tree

Calculation of the MST : Kruskal's Algorithm



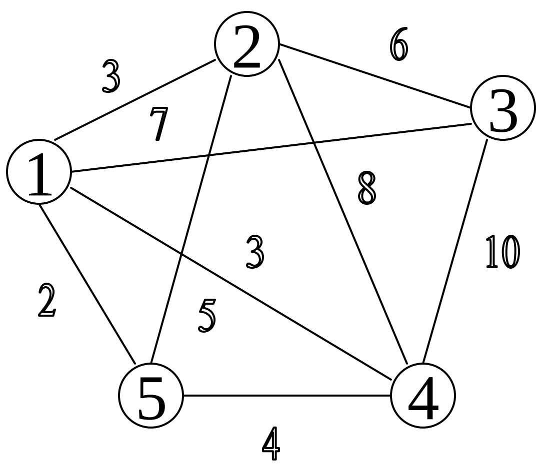
Undirected Network

5
6
7
8
10



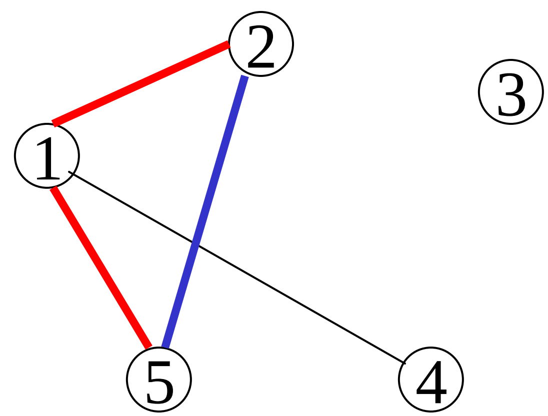
Minimum Spanning Tree

Calculation of the MST : Kruskal's Algorithm



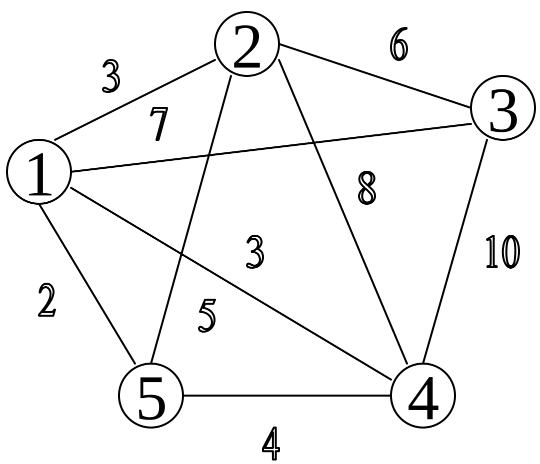
Undirected Network

6
7
8
10

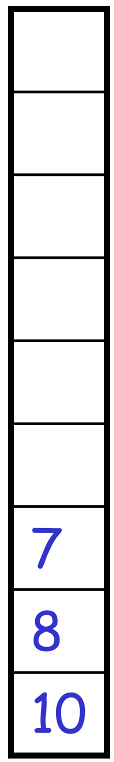


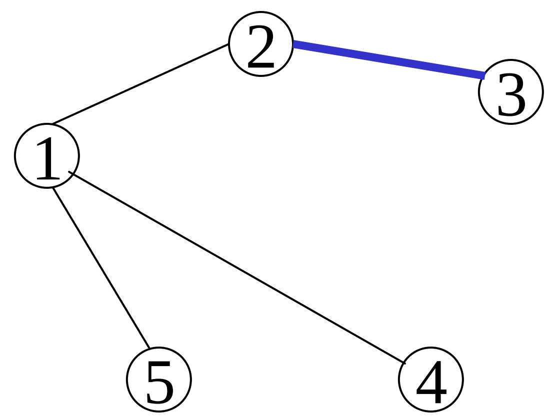
Minimum Spanning Tree

Calculation of the MST : Kruskal's Algorithm



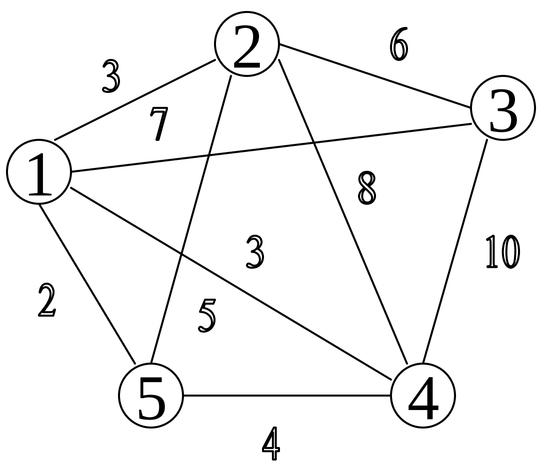
Undirected Network



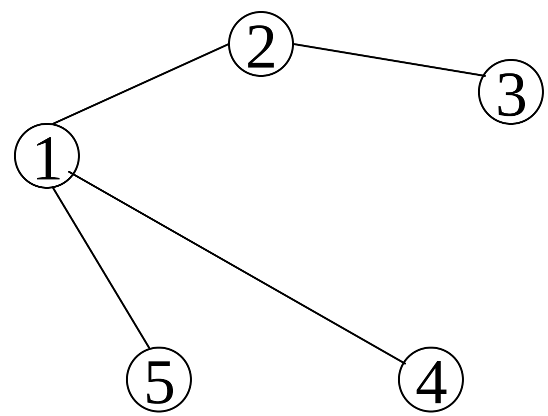


Minimum Spanning Tree

Calculation of the MST : Kruskal's Algorithm



Undirected Network



Minimum Spanning Tree

Implementation Issues (The Devil in the Details)

- BPM extraction and interpoint distance calculation:
  - Implemented in C#.
- BPM similarity and distance calculation:
  - Implemented in C#.
- MST calculation:
  - Implemented using Kruskal's algorithm in JAVA.
- Cluster calculations are performed using JAVA
- Visualization environment:
  - Implemented in JAVA.
  - Graph layout facilitated using TouchGraph.

TouchGraph

- TouchGraph is a general public license JAVA-based library for the visualization of graphs. (www.touchgraph.com)
- Graph layout in TouchGraph:
  - When a graph is first loaded, nodes start out at the center with slightly random positions, and then spread out because of node-node repulsions.
- Graph manipulation tools provided by TouchGraph.
  - Zooming.
  - Rotation.
  - Hyperbolic manipulation.
  - Graph dragging.

Equations - I

$$\boldsymbol{M} = \begin{cases} E_{ij}, & \text{if there is an edge } \{i,j\}, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathrm{cut}(\mathcal{V}_1^*, \mathcal{V}_2^*) = \min_{\mathcal{V}_1, \mathcal{V}_2} \mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2).$$

$$\mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2) = \sum_{i \in \mathcal{V}_1, j \in \mathcal{V}_2} M_{ij}.$$

$$\mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k) = \sum_{i<j} \mathrm{cut}(\mathcal{V}_i, \mathcal{V}_j)$$

$$E_{ij} = t_{ij} \times \log\left(\frac{|\mathcal{D}|}{|\mathcal{D}_i|}\right)$$

$$\boldsymbol{M} = \begin{bmatrix} \mathbf{0} & \boldsymbol{A} \\ \boldsymbol{A}^T & \mathbf{0} \end{bmatrix}$$

$$\mathrm{cut}(\mathcal{W}_1 \cup \mathcal{D}_1, \mathcal{W}_2 \cup \mathcal{D}_2, \ldots, \mathcal{W}_k \cup \mathcal{D}_k) = \min_{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k} \mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k)$$

Wrap-up

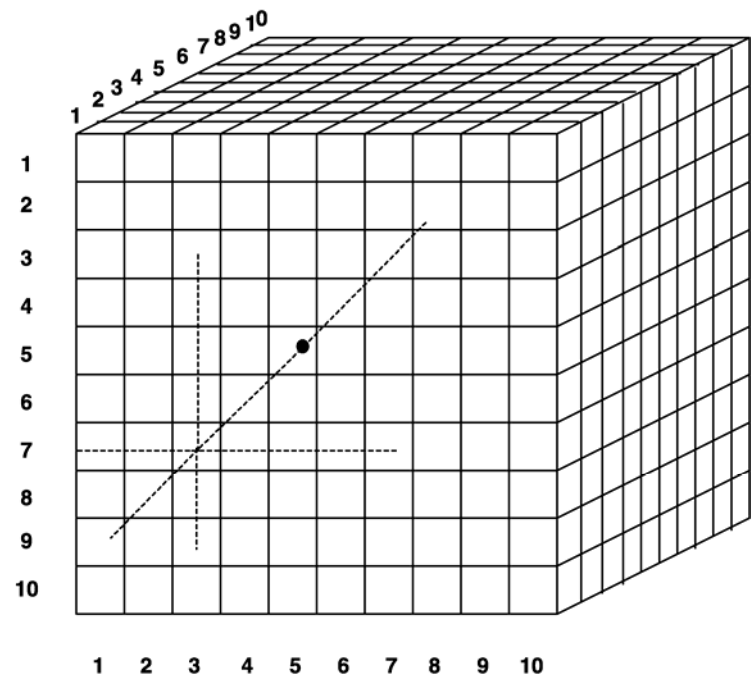
- Demonstrated a new method for cross corpora document discovery
- Method predicated on the use of BPM and the MST as a convenient foil for the exploration of the cross corpora relationships.
- This work represents the tip of the iceberg of a new area that is not only of strategic importance to the United States but also is highly relevant to all who are currently conducting research in any discipline.

Feature Extraction
(Bigram Proximity Matrix (BPM) & Trigram Proximity Matrix (TPM))

Table 2.2 Example of Bigram Proximity Matrix

	.	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				

"The wise young man <u>sought</u> <u>his father</u> in the crowd."



1 . (period)
2 crowd
3 his
4 in
5 father
6 man
7 sought
8 the
9 wise
10 young

Evidence That BPM and TPM Capture Semantic Content

- Angel Martinez, "A Framework for the Representation of Semantics," *Ph.D Dissertation under the direction of Edward Wegman*, October 2002.
  - Supervised Learning.
  - Hypothesis Tests (3 sets of tests).
  - Unsupervised Learning.
  - Supervised Learning in a Reduced Dimension Space.

Similarity Measures and Pseudometrics on the BPM

- Following Martinez (2002) we propose the use of the Ochiai measure in the case of the BPM:

$$S(X,Y) = \frac{|X \text{ and } Y|}{\sqrt{(|X||Y|)}}$$

- This is converted to a distance via:

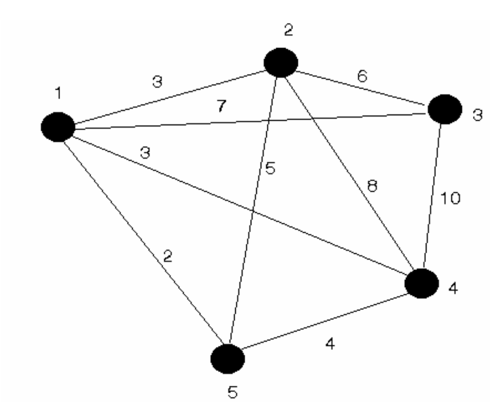
$$d(X,Y) = \sqrt{(2 - 2S(X,Y))}$$

How Do We Exploit This Interpoint Distance Matrix for Clustering?

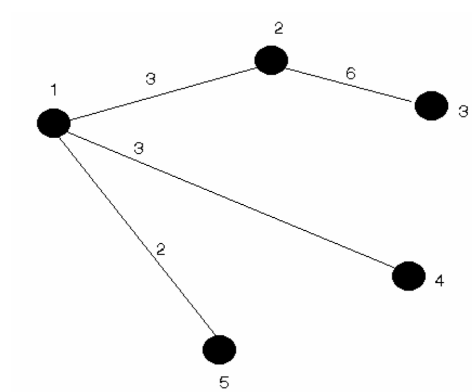
- First order exploration
  - Visualization of cluster structures
- Second order exploration
  - Exploration of cluster structures to ascertain interesting cross (within) corpora relationships

The Minimal Spanning Tree (MST): A Strategy for Effective Exploration of the Interpoint Distance Matrix and Cluster Computation

- Definition (Minimal Spanning Tree (MST)) – The collection of edges that join all of the points in a set together, with the minimum possible sum of edge values. The edge values that will be used here is the distance measures stored in our interpoint distance matrix.

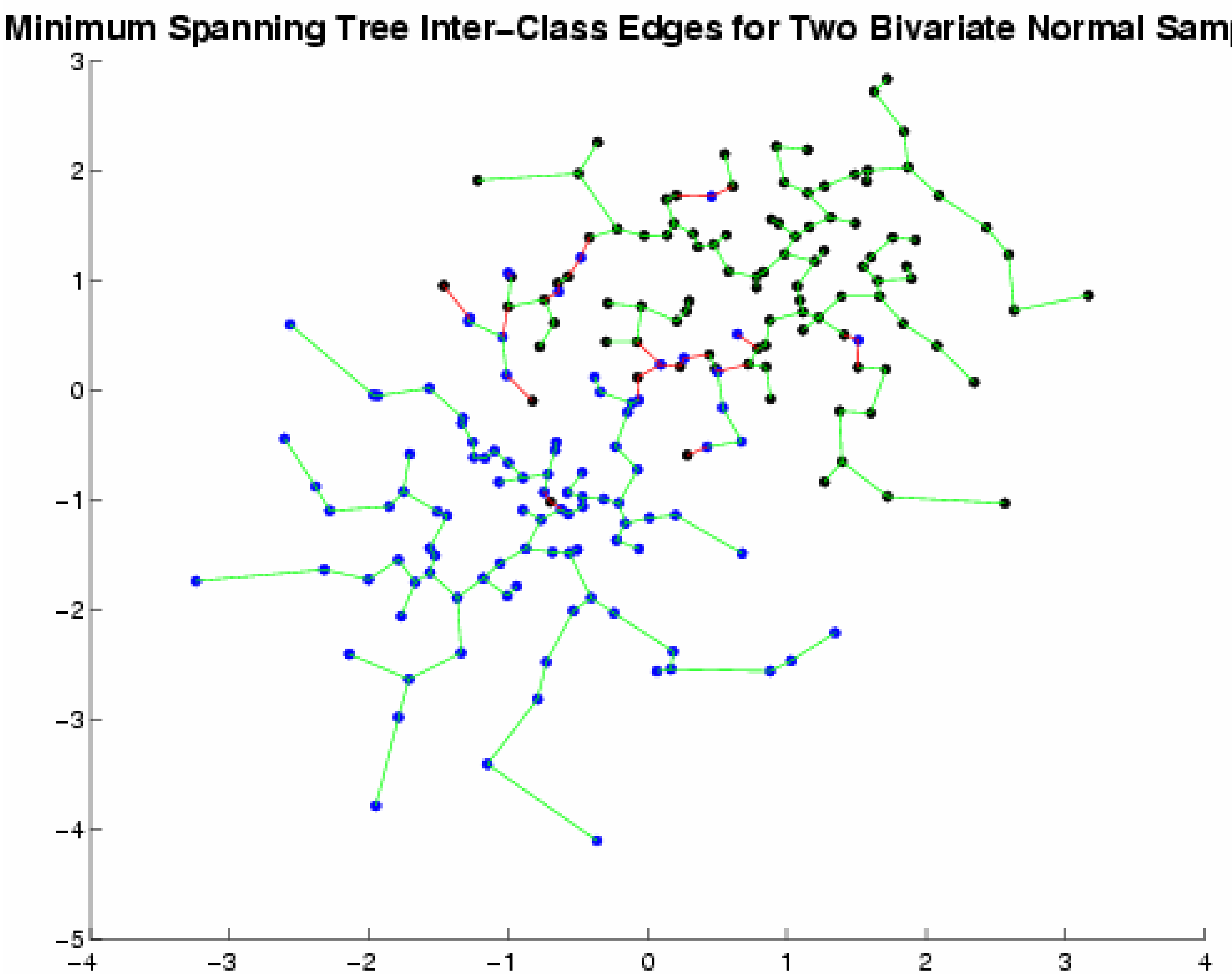


A complete graph.



Associated MST.

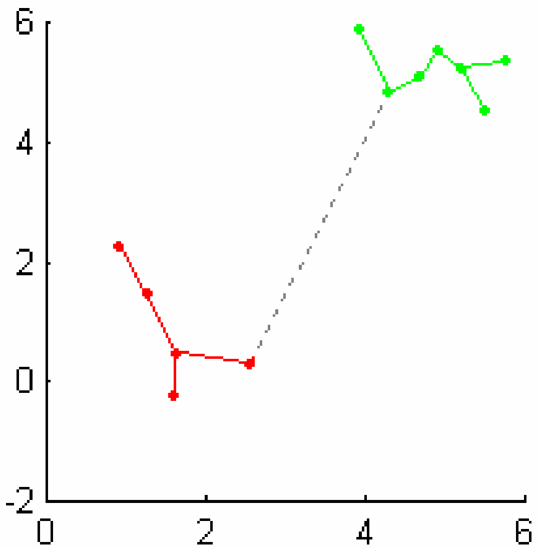
MST Classifier Complexity Characterization

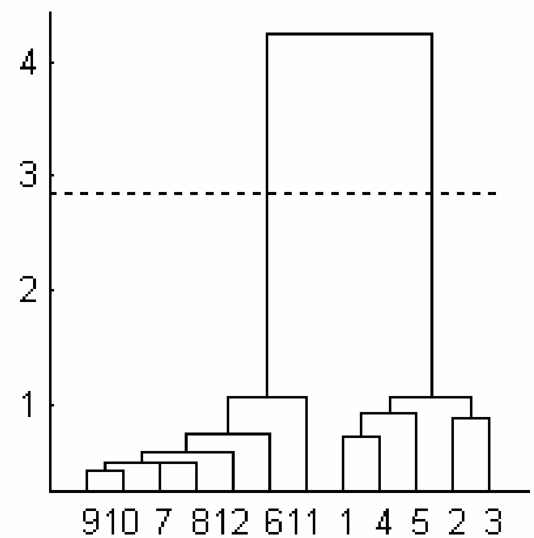


Previous work had suggested that the number of cross class edges can be used as a surrogate for classification complexity.

These cross class (corpora) edges will be used in our scheme to facilitate the cross-corpora discovery process.

MST-based Clustering





All the single-linkage clusters could be obtained by deleting the edges of the MST, starting from the largest one.

Adapted from - Course: Cluster Analysis and Other Unsupervised Learning Methods (Stat 593 E)

Speakers: Rebecca Nugent[1], Larissa Stanberry[2]

Department of [1] Statistics, [2] Radiology,

University of Washington

Applications of MST-based Clustering and Data Mining to Geospatial Data

- Diansheng Guo, Donna Peuquet, Mark Gahegan, “Opening the Black Box: Interactive Hierarchical Clustering for Multivariate Spatial Patterns,” *GIS’02*, November 8-9, 2002, McLean, Virginia, USA.

- Guo, D., D. Peuquet and M. Gahegan, “ICEAGE: Interactive Clustering and Exploration of Large and High-Dimensional Geodata” *GeoInformatica*, **7**(3): 229-253, 2003.

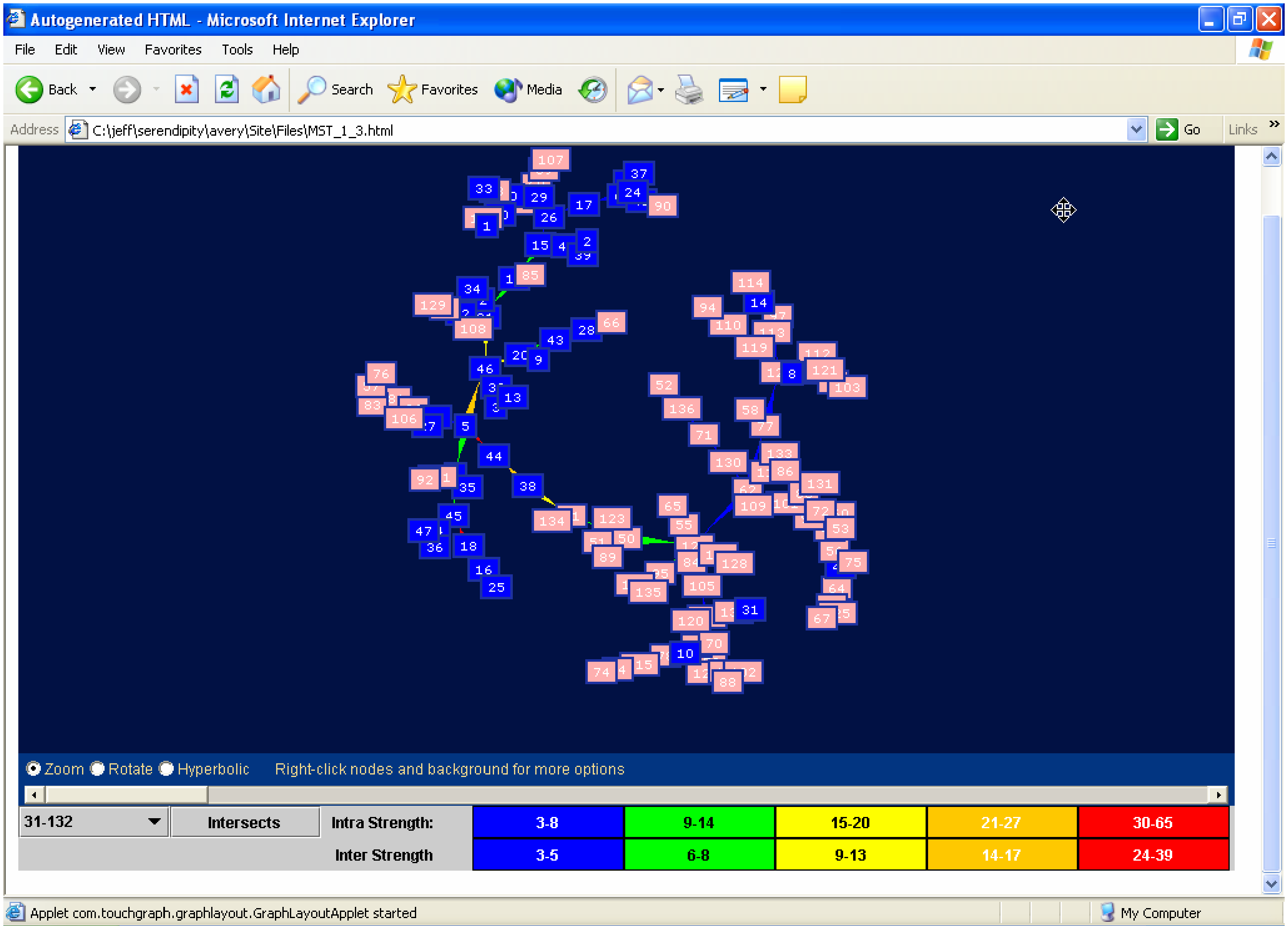
Applications of MST-based Clustering to Gene Expression Data

- Ying Xu, Victor Olman, and Dong Xu, "Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees," *Bioinformatics* Vol. 18 no. 4, pp. 536-545, 2002.

- Ying Xu, Victor Olman, and Dong Xu, "Minimum spanning trees for gene expression clustering," *Genome Informatics, 12: 24-33, 2001.*

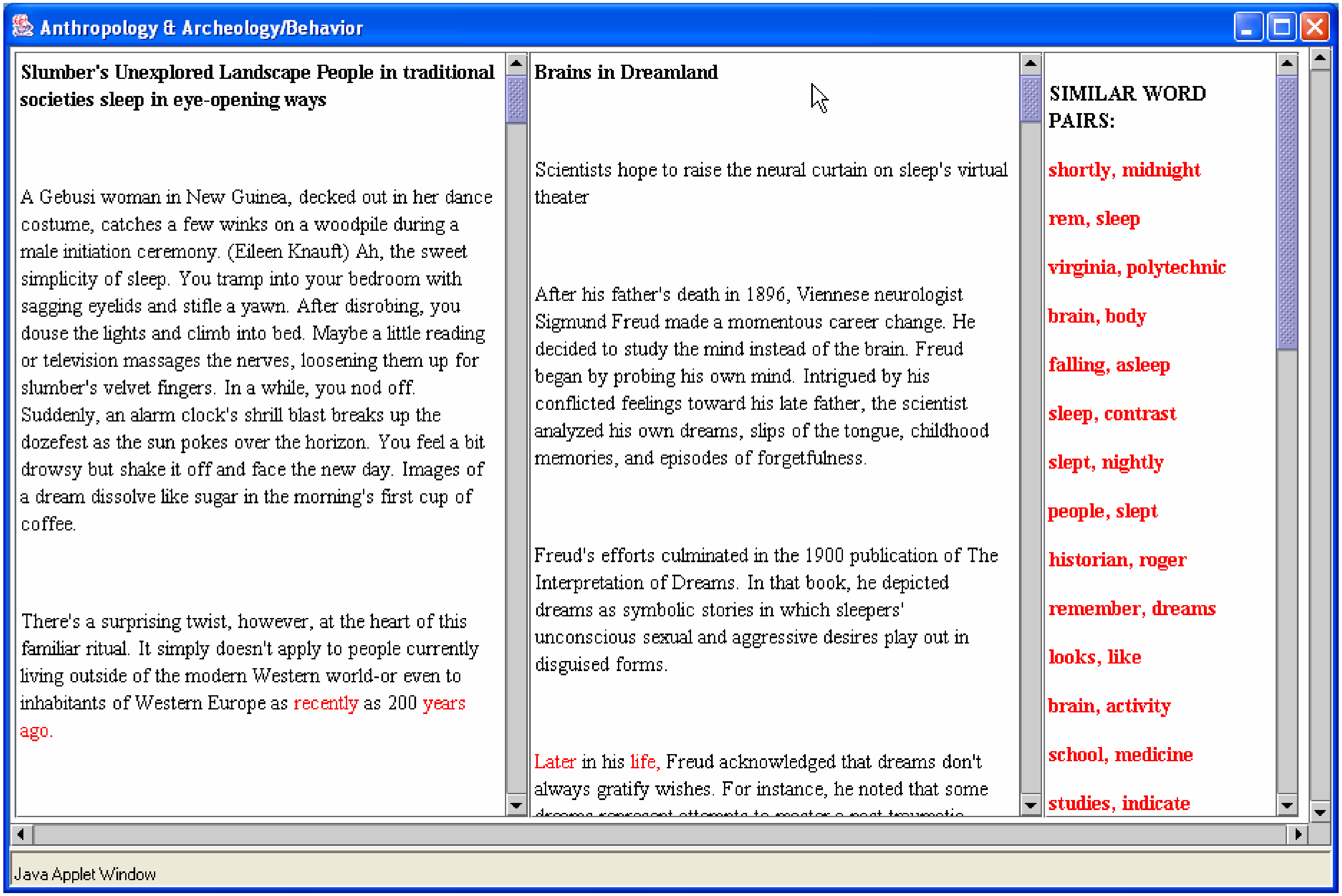
The Environment (Opening Screen)

The Environment (MST)



Blue is anthropology and archaeology. Pink is behavior.

The Environment (The Comparison File)

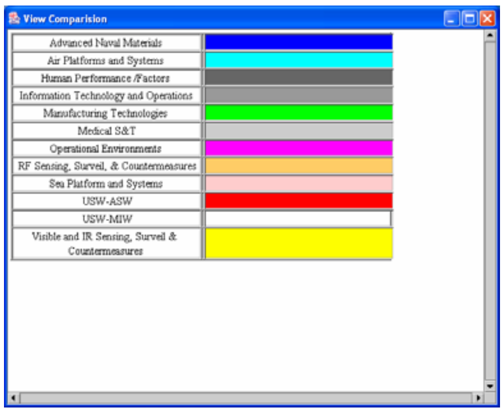


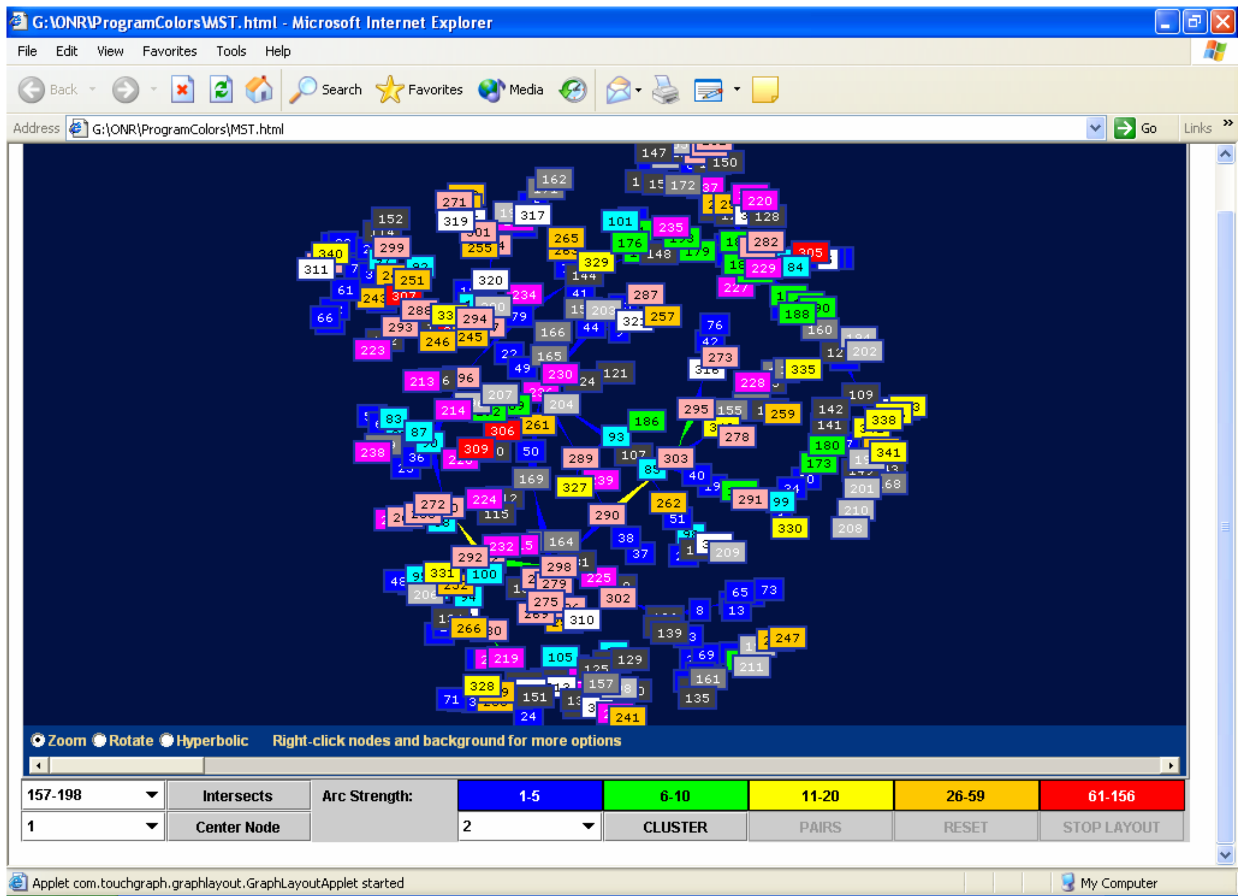
MST-Based Divisive Clustering Results on the ONR ILIR Data

Options on the Clustering Program

- Decision to cut at an edge is determine by the the edge strength/(mean of associated edges of path length k). Choose the largest value
- Nuisance parameters
  - Maximum number of clusters
  - Minimum of points per cluster
  - k

Opening Screen for ILIR Cluster Program



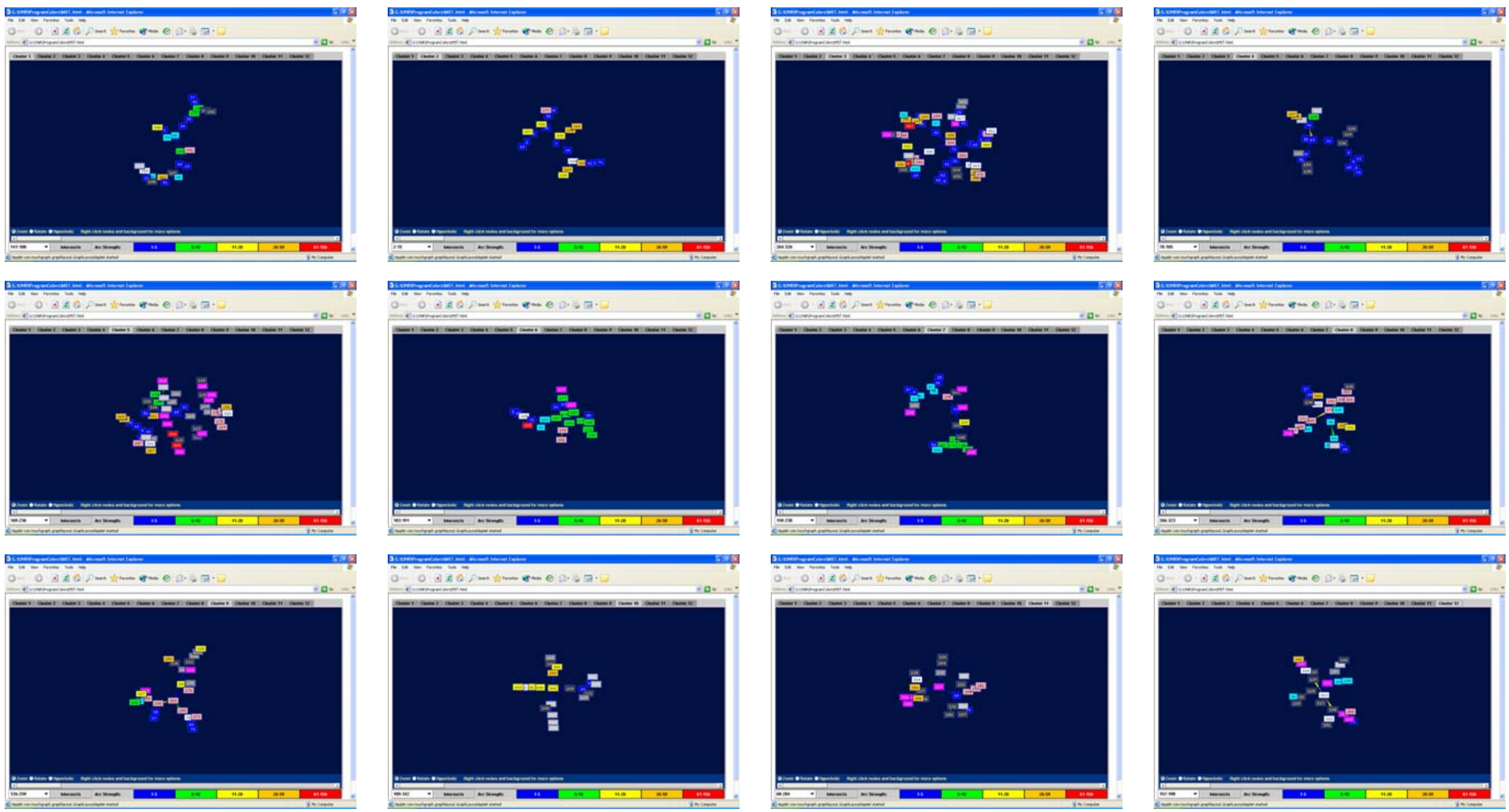


Associated Color Key

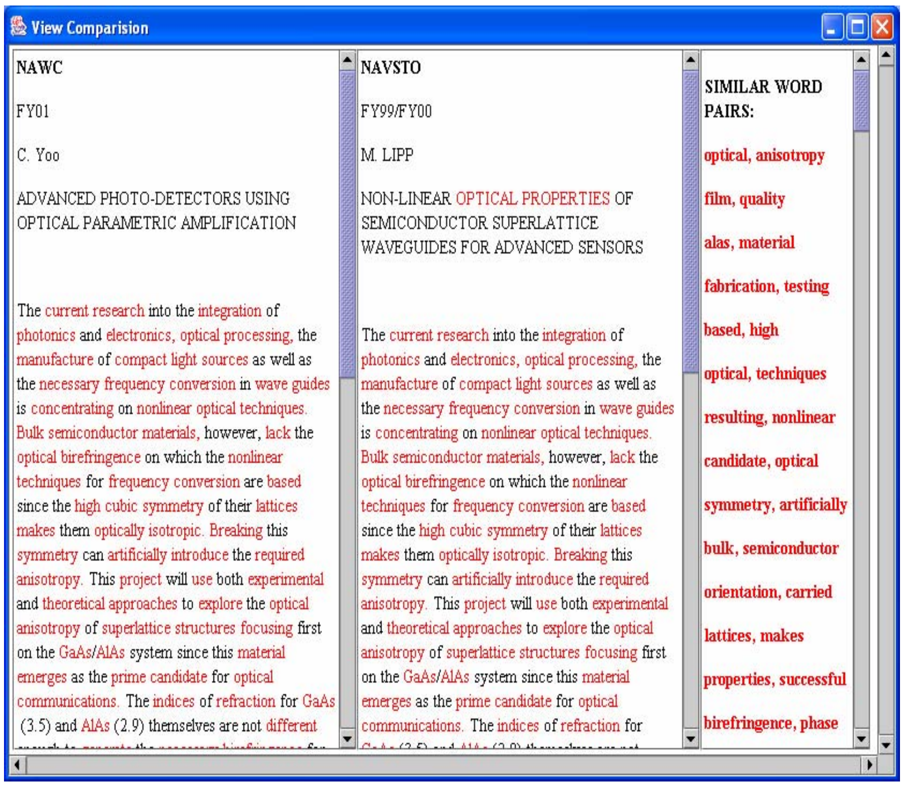
Army Conference on Applied Statistics, Atlanta, 2004

Overview of the ILIR MST Cluster Structure

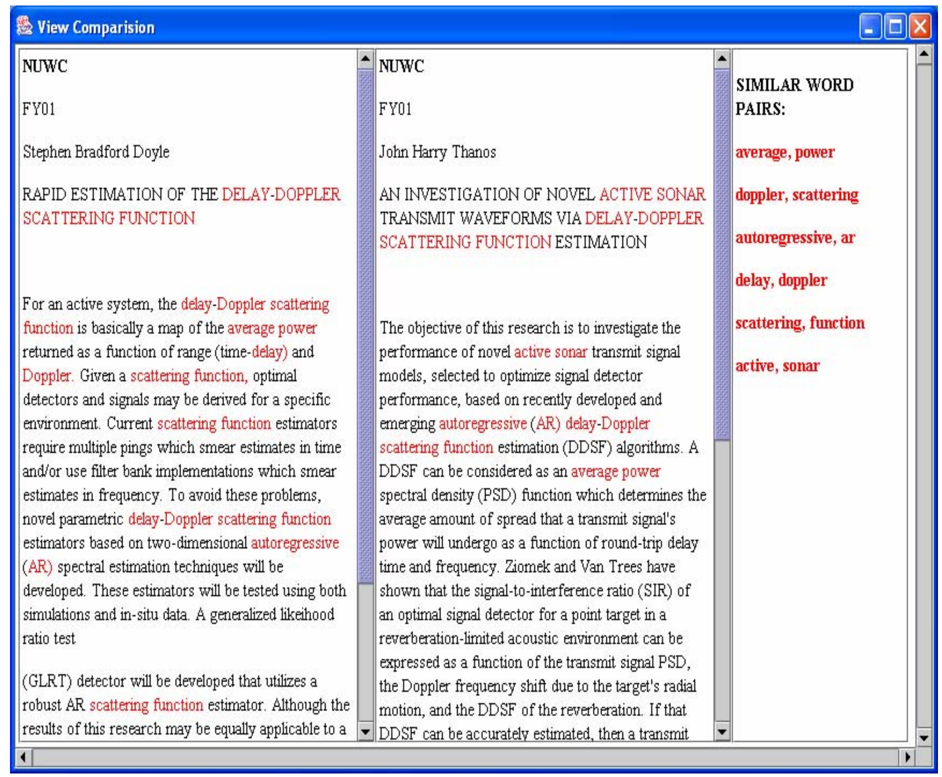
Advanced Naval Materials	
Air Platforms and Systems	
Human Performance /Factors	
Information Technology and Operations	
Manufacturing Technologies	
Medical S&T	
Operational Environments	
RF Sensing, Surveil, & Countermeasures	
Sea Platform and Systems	
USW-ASW	
USW-MIW	
Visible and IR Sensing, Surveil & Countermeasures	



Some Additional Interesting Anomalies/Discoveries in the ONR ILIR Data Made Apparent Via User Exploration of the Clusters - I

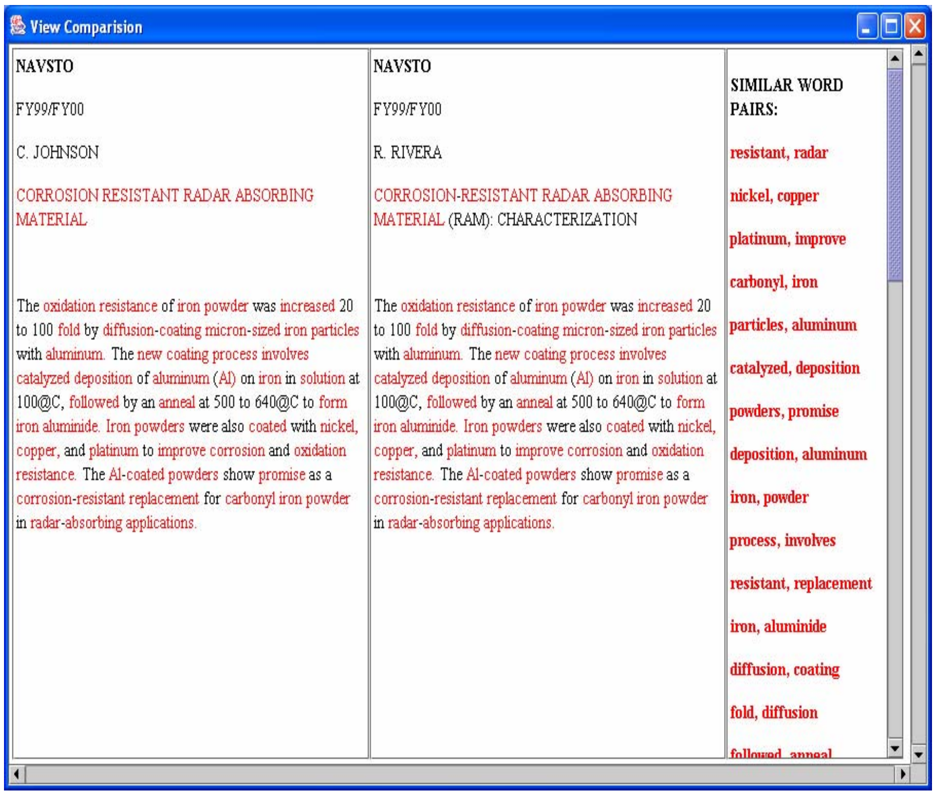


Identical Abstracts Different Author and Titles Same Year

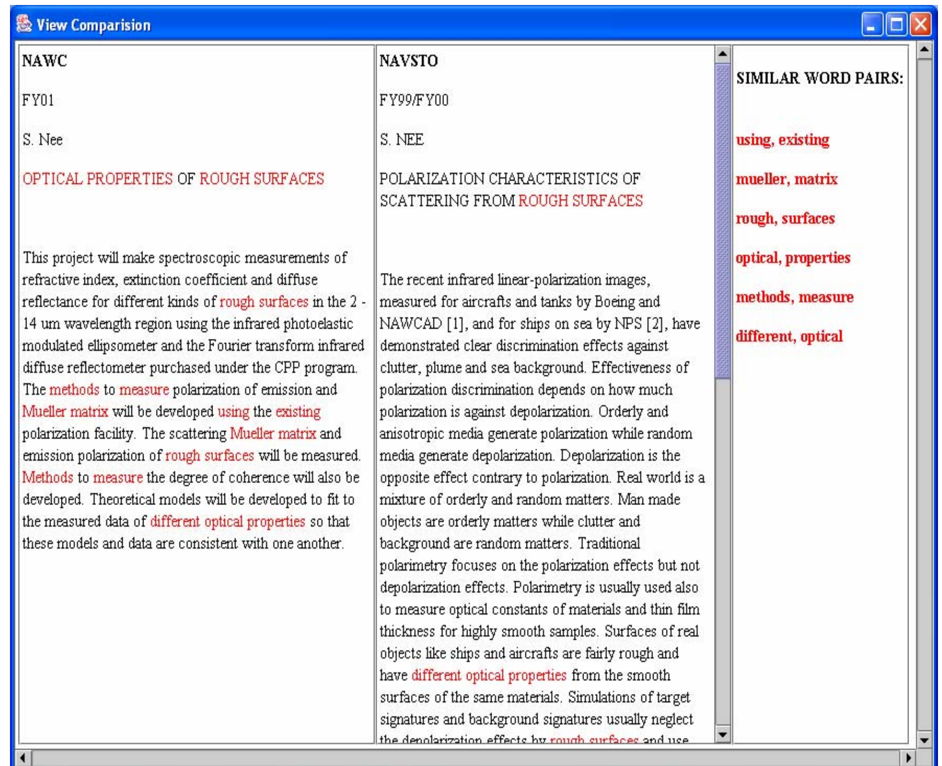


Interesting Association Between the USW-MIW and RF Sensing, Surveillance, & Countermeasures Classes

Some Additional Interesting Anomalies/Discoveries in the ONR ILIR Data Made Apparent Via User Exploration of the Clusters - II



Identical Abstracts Different Author and Titles Same Year

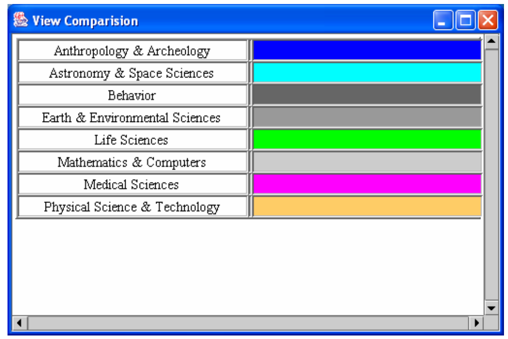


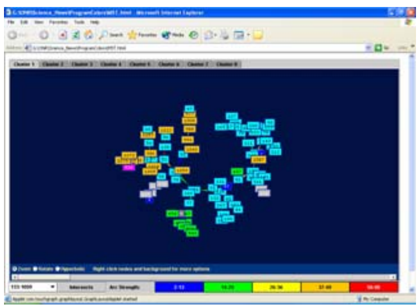
Interesting Association Between the Advanced Naval Materials and Visible and IR Sensing, Surveillance & Countermeasures Categories

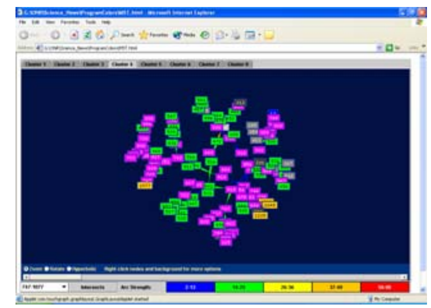
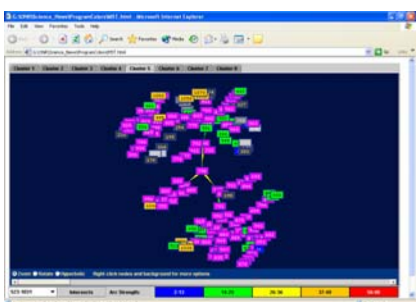
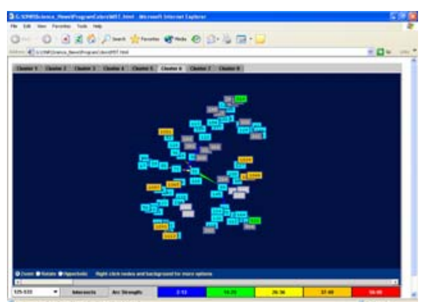
MST-Based Divisive Clustering Results on the Science News Data

Army Conference on Applied Statistics, Atlanta, 2004

Overview of the Science News MST Cluster Structure

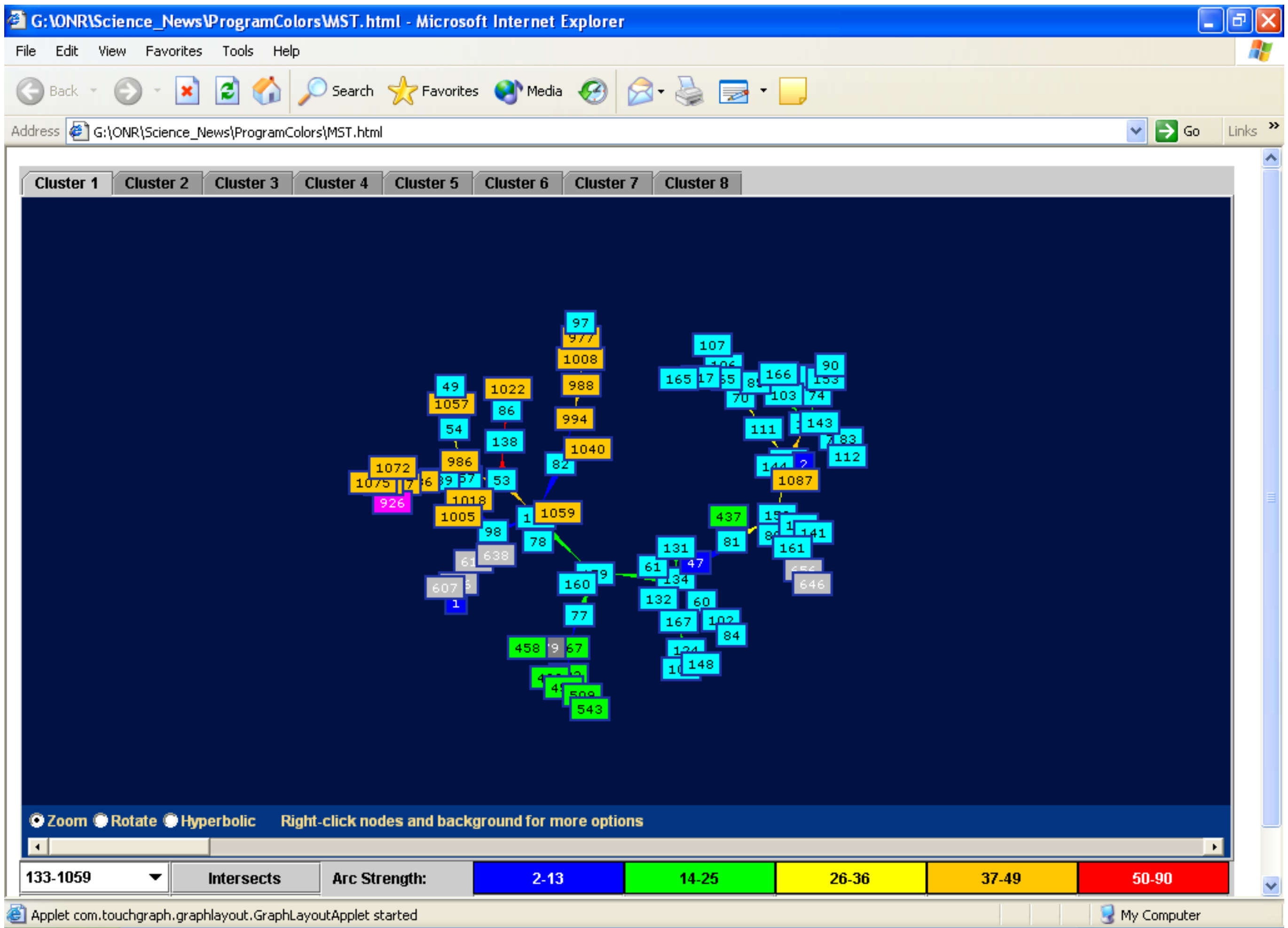




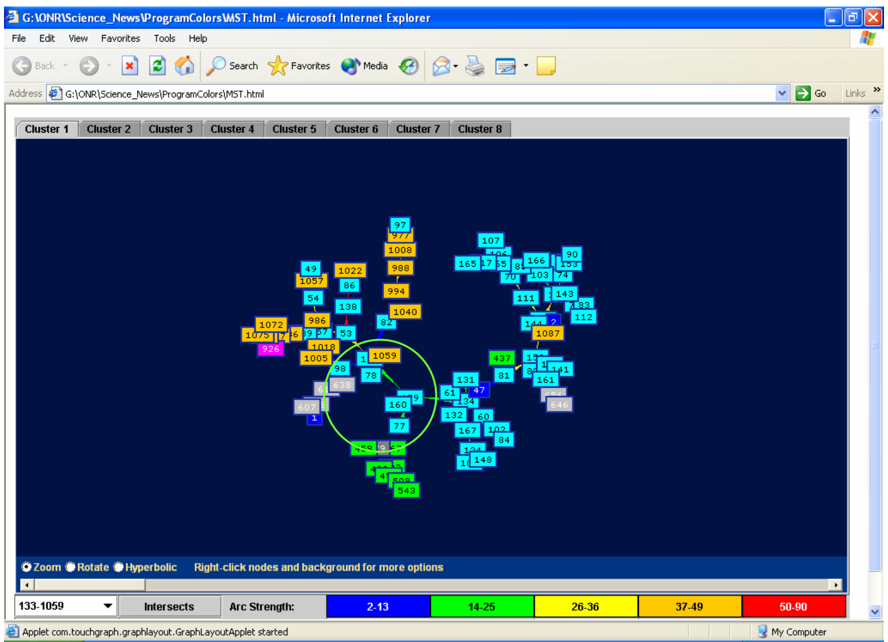
Exploration of Science News MST Cluster 1

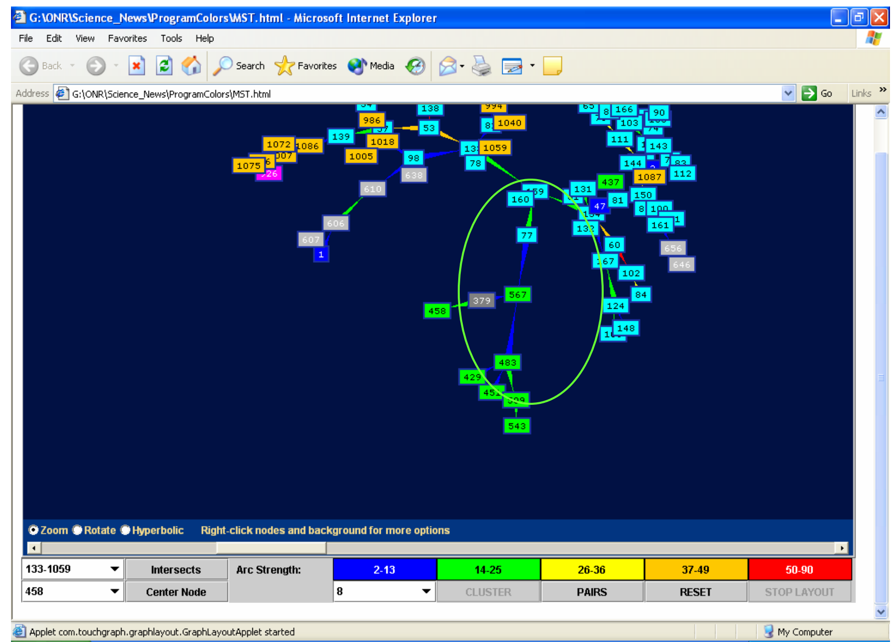
Anthropology & Archeology	
Astronomy & Space Sciences	
Behavior	
Earth & Environmental Sciences	
Life Sciences	
Mathematics & Computers	
Medical Sciences	
Physical Science & Technology	



Science News MST Cluster 1 Subcluster – Animal Behavior and Sexuality

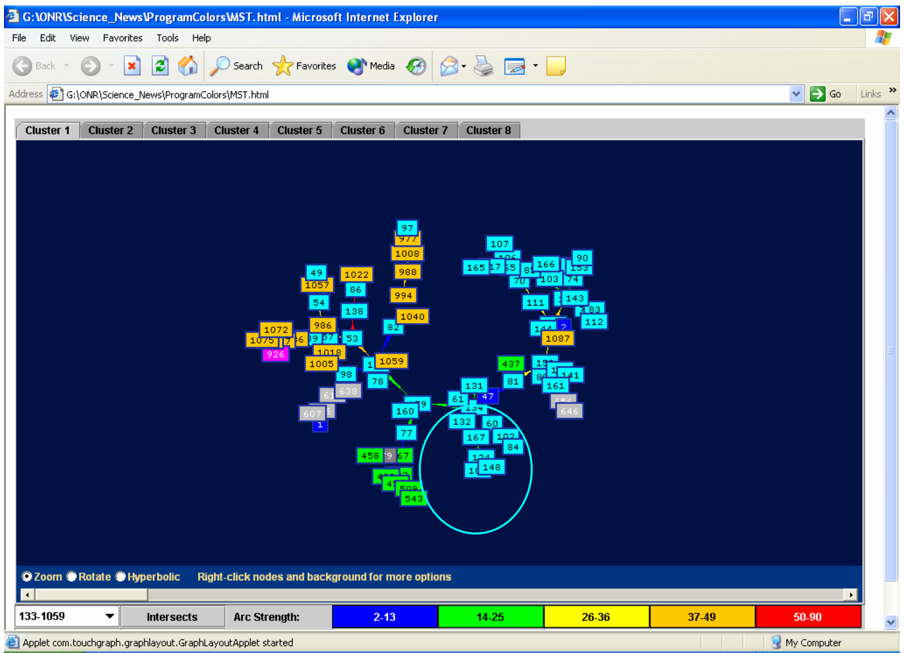
Anthropology & Archeology	
Astronomy & Space Sciences	
Behavior	
Earth & Environmental Sciences	
Life Sciences	
Mathematics & Computers	
Medical Sciences	
Physical Science & Technology	

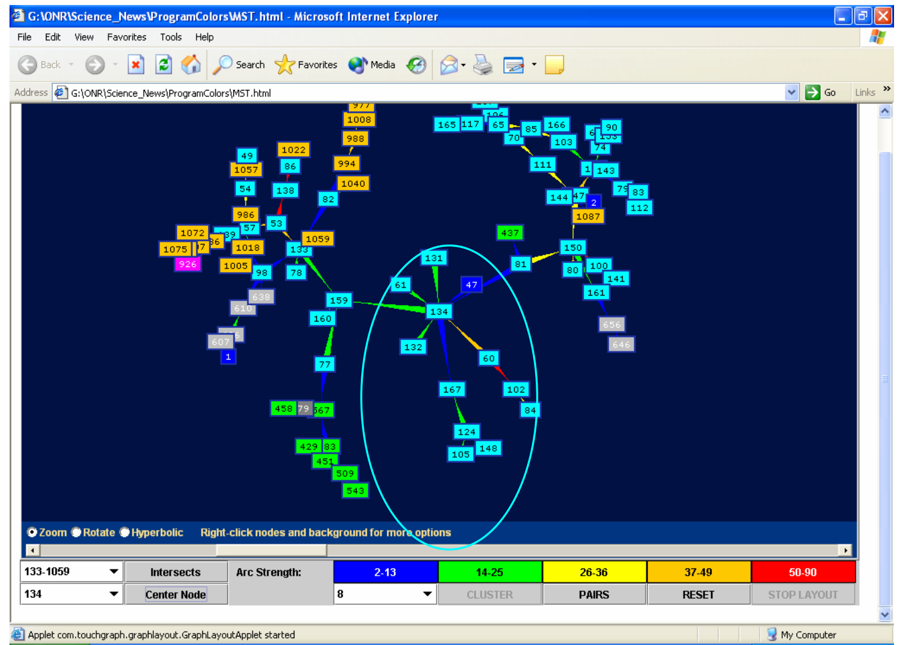




Science News MST Cluster 1 Subcluster – Infrared Camera and Its Applications to Cosmology

Anthropology & Archeology	(blue)
Astronomy & Space Sciences	(cyan)
Behavior	(dark gray)
Earth & Environmental Sciences	(gray)
Life Sciences	(green)
Mathematics & Computers	(light gray)
Medical Sciences	(magenta)
Physical Science & Technology	(orange)





Article 134 Discusses an Enabling Technology “Infrared Camera Goes the Distance”

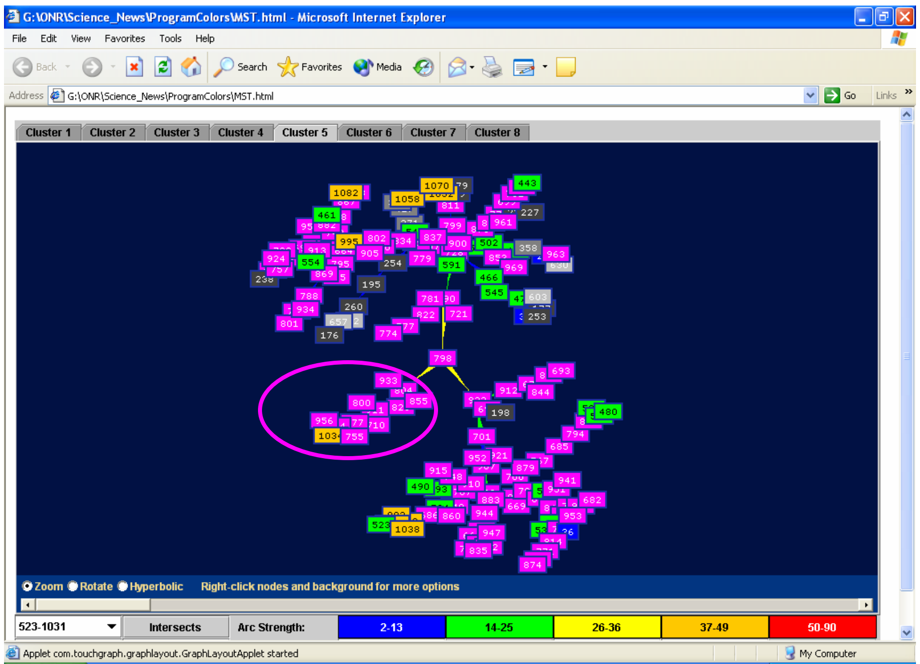
NAVSEA DAHLGREN Surface Warfare Center Division

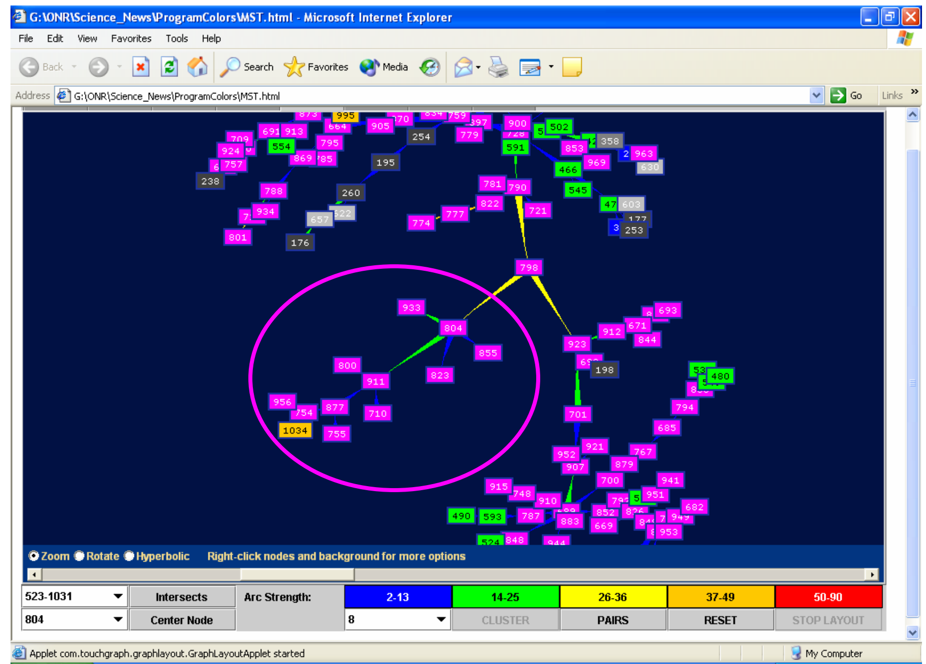
Army Conference on Applied Statistics, Atlanta, 2004

GEORGE MASON UNIVERSITY

Science News MST Cluster 5 Subcluster - Aids

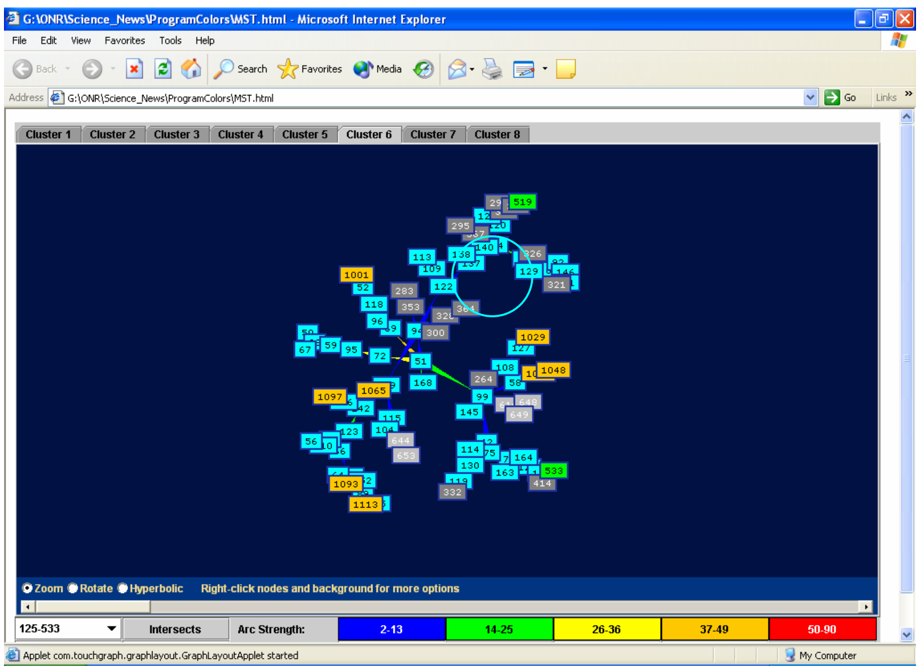
Anthropology & Archeology	
Astronomy & Space Sciences	
Behavior	
Earth & Environmental Sciences	
Life Sciences	
Mathematics & Computers	
Medical Sciences	
Physical Science & Technology	

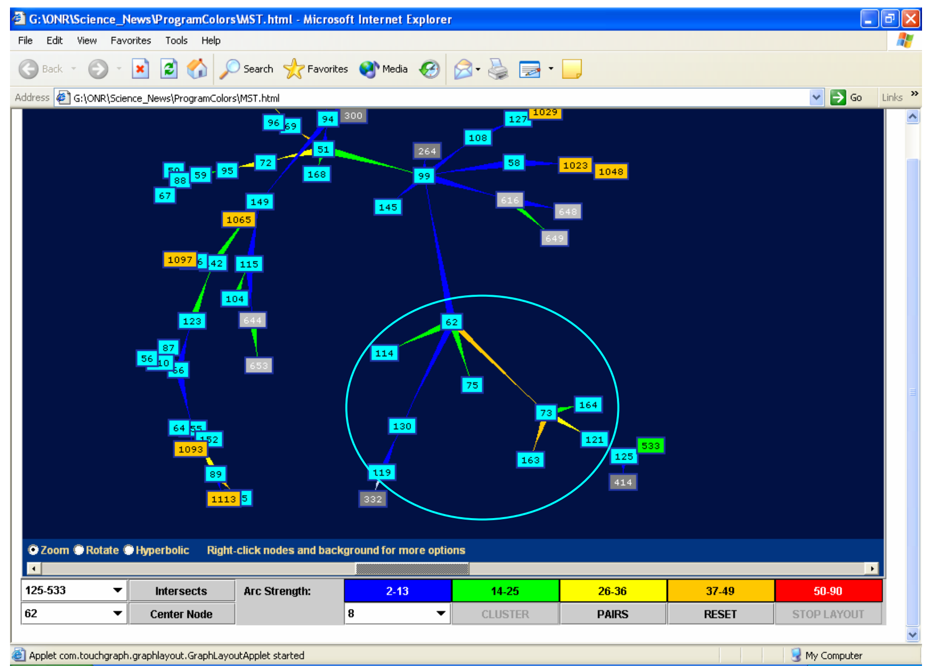




Science News MST Cluster 6 Subcluster – Solar Activity

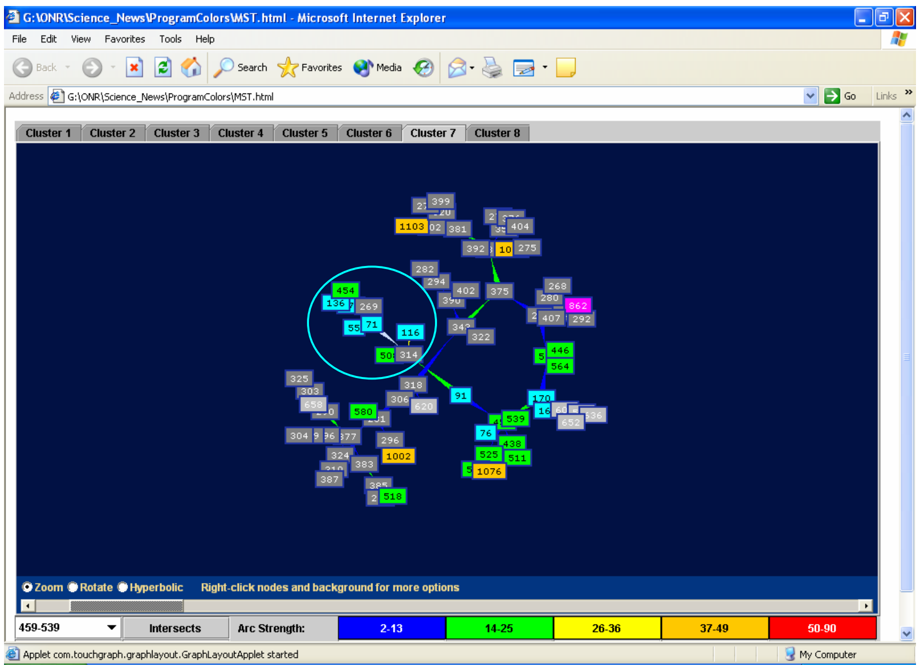
Anthropology & Archeology	
Astronomy & Space Sciences	
Behavior	
Earth & Environmental Sciences	
Life Sciences	
Mathematics & Computers	
Medical Sciences	
Physical Science & Technology	

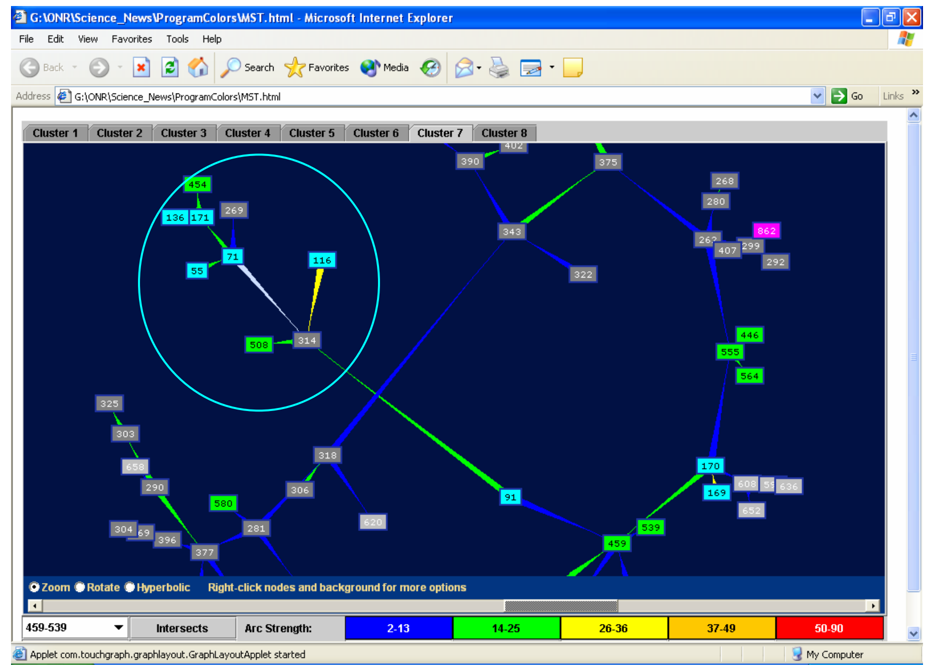




Army Conference on Applied Statistics, Atlanta, 2004

Science News MST Cluster 7 Subcluster – Evolution and the Origins of Life

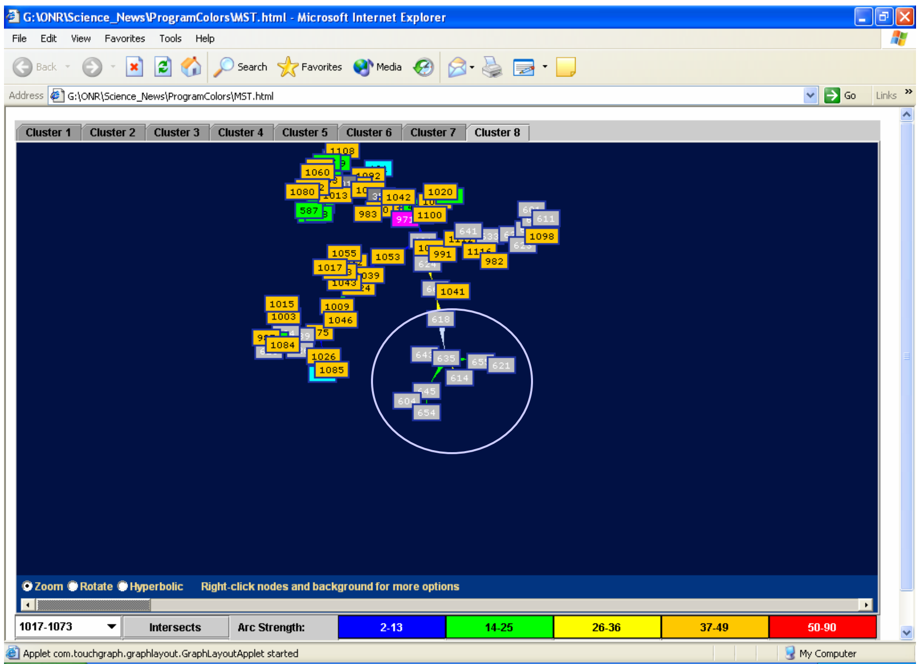


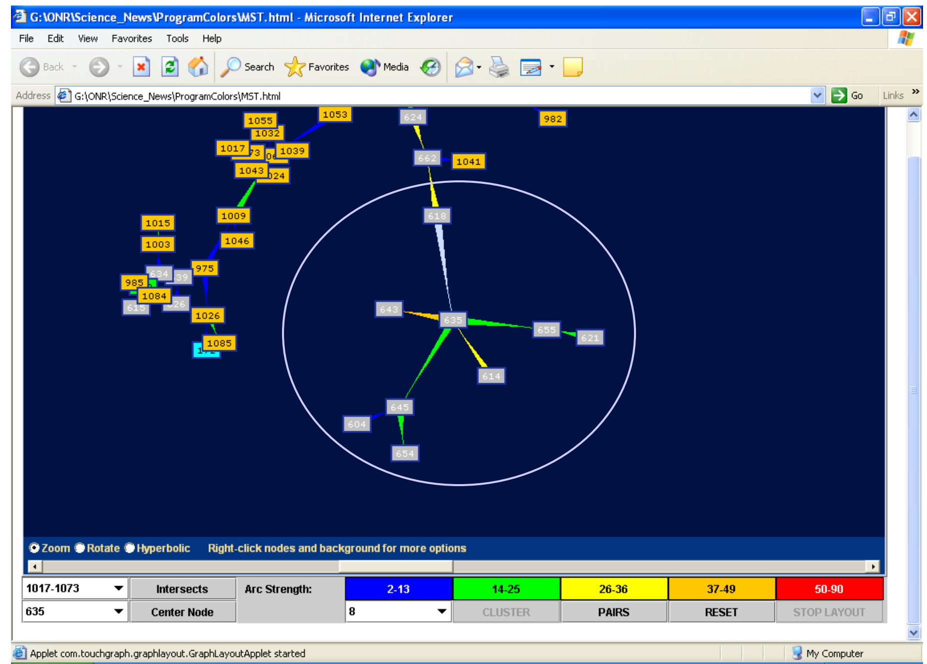


Note that cluster 7 has been rendered using a slightly different solution to the spring equations than was originally presented.

Army Conference on Applied Statistics, Atlanta, 2004

Science News MST Cluster 8 Subcluster – Artificial Intelligence





Note that cluster 8 has been rendered using a slightly different solution to the spring equations then was originally presented.

Agglomerative Clustering Results on the Science News Data

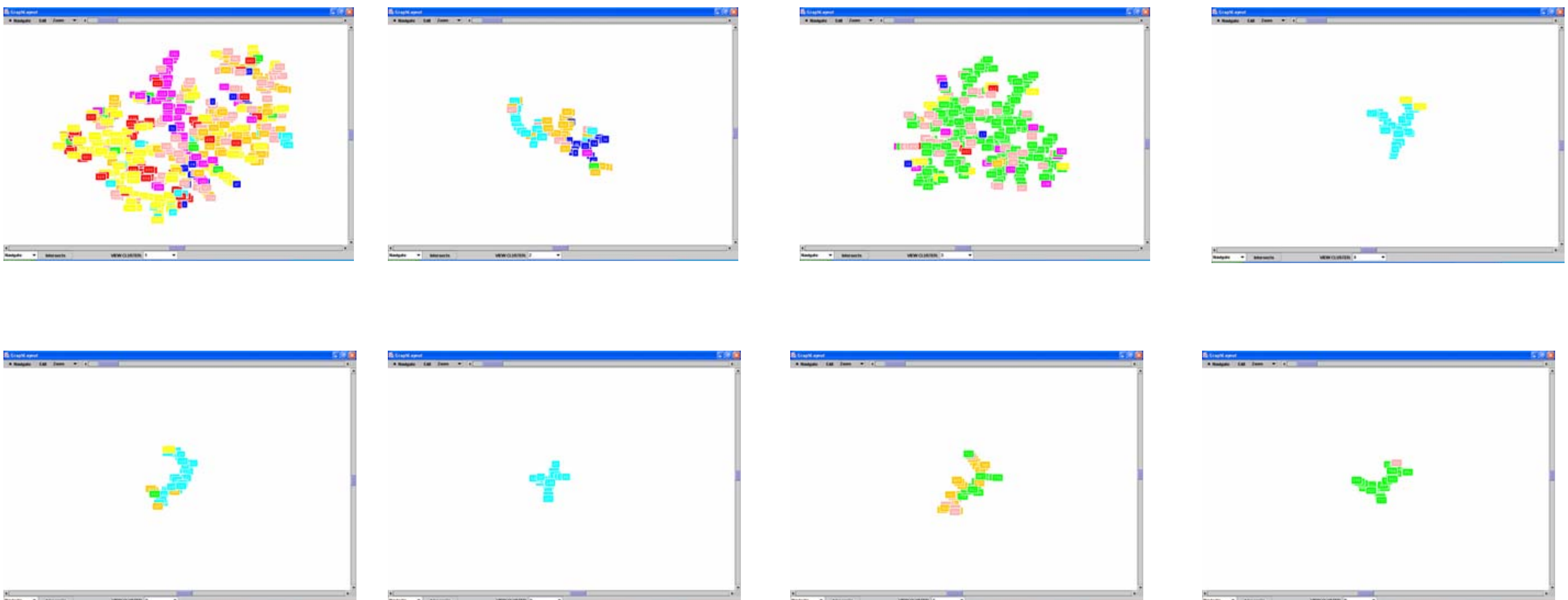
Methodology

- ScienceNews1:
  - 8 classification categories
  - Hierarchical Clustering
  - Method: Ward
    - Merge two clusters that produce the smallest variance in resultant cluster.
  - Tree Cut: 8

Science News 8 Agglomerative Clusters

Anthropology & Archeology
Behavior
Life Sciences
Medical Sciences

Astronomy & Space Sciences
Earth & Environmental Sciences
Mathematics & Computers
Physical Science & Technology



Agglomerative Clustering Results on the ONR ILIR Data

Army Conference on Applied Statistics, Atlanta, 2004

Methodology

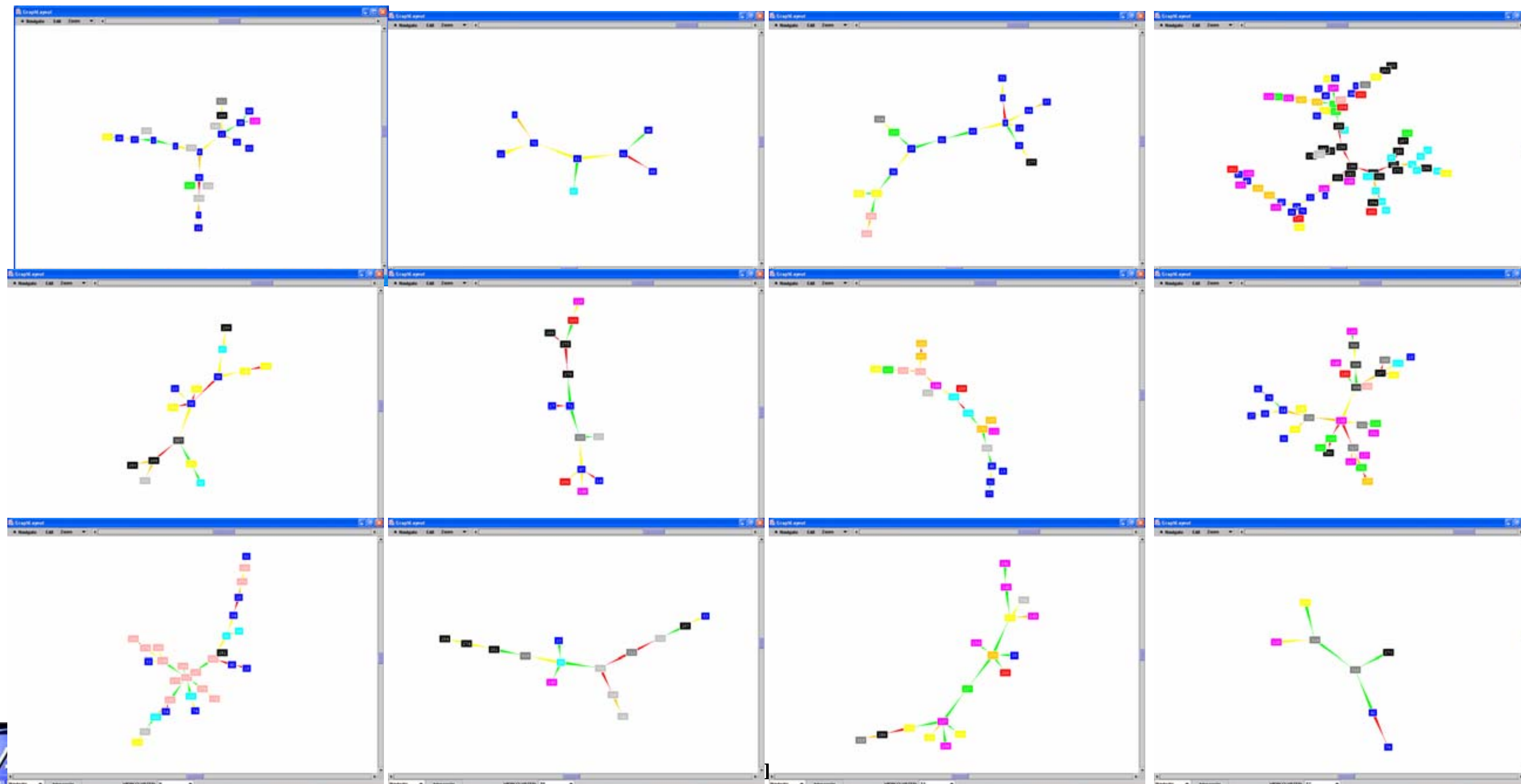
- ILIR1:
  - 12 classification categories
  - Hierarchical Clustering
  - Method: Average
    - Merge clusters with smallest average distance.
  - Tree Cut: 24

ILIR 24 Agglomerative Clusters

Advanced Naval Materials
Information Technology and Operations
Operational Environments
USW-ASW

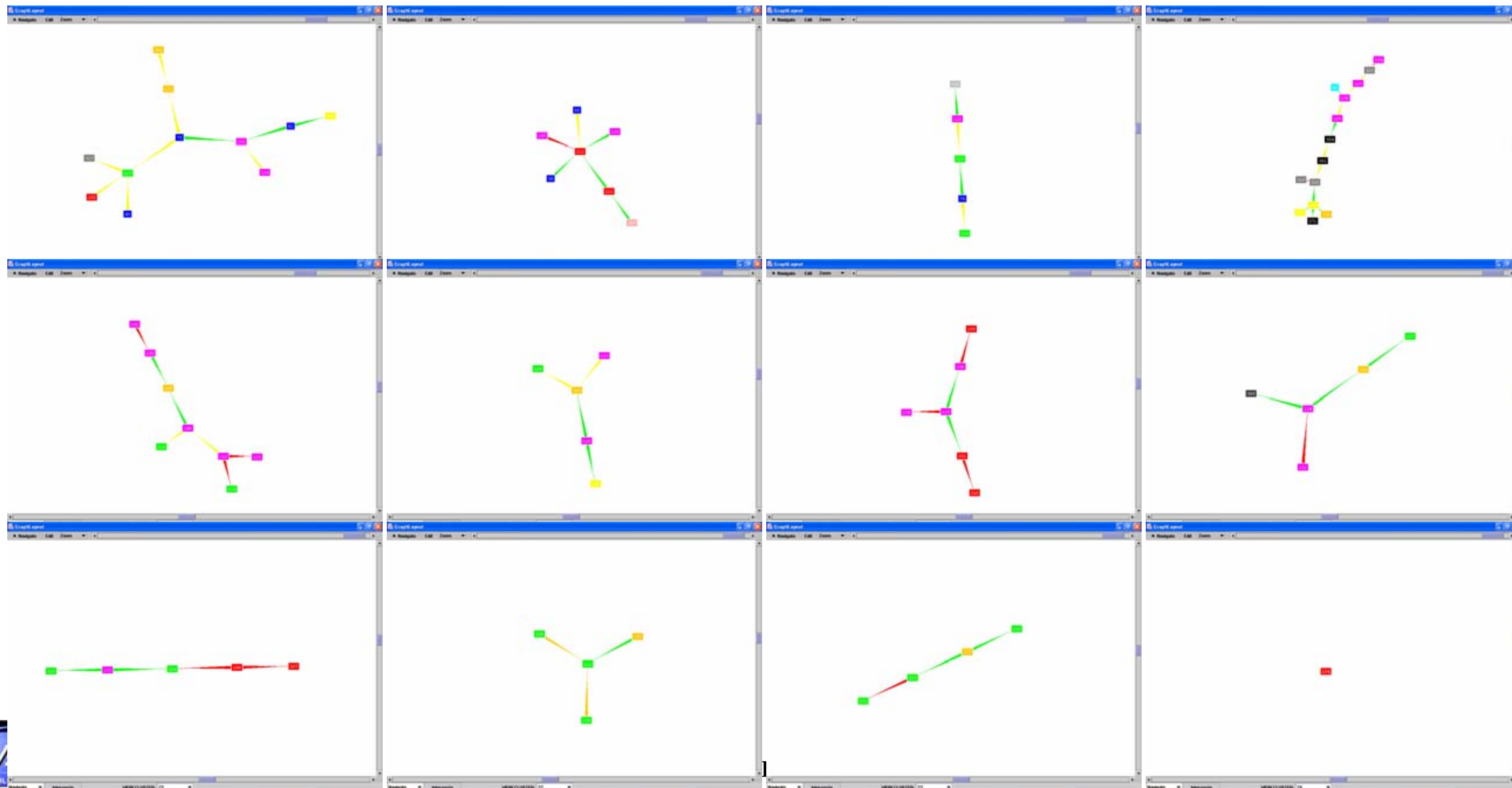
Air Platforms and Systems
Manufacturing Technologies
RF Sensing, Surveil, & Countermeasures
USW-MIW

Human Performance /Factors
Medical S&T
Sea Platform and Systems
Visible and IR Sensing, Surveil & Countermeasures



ILIR 24 Agglomerative Clusters

Advanced Naval Materials	Air Platforms and Systems	Human Performance /Factors
Information Technology and Operations	Manufacturing Technologies	Medical S&T
Operational Environments	RF Sensing, Surveil, & Countermeasures	Sea Platform and Systems
USW-ASW	USW-MIW	Visible and IR Sensing, Surveil & Countermeasures



Bipartite Spectral Based Results

Establishing the Center for Data Analysis and Statistics (CDAS) at the United States Military Academy

Rodney X. Sturdivant
Department of Mathematics Sciences
United States Military Academy, West Point, New York

1 Executive Summary

The Center for Data Analysis and Statistics (CDAS) was organized in the Department of Mathematical Sciences, United States Military Academy in January of 2004. The center is designed to provide statistical consulting and perform analysis to support researchers in the West Point community and for DoD agencies as required. In this paper we discuss the organization, mission and utility of the CDAS. The paper is designed to inform members of the DoD statistical community of the opportunities and benefits the CDAS can provide for their own agencies. We also briefly discuss completed and ongoing projects as well as lessons learned in establishing a statistical consulting service.

2 Introduction

The Department of Mathematical Sciences at the United States Military Academy (USMA) established the Center for Data Analysis and Statistics (CDAS) to address a perceived need for statistical consulting support both at USMA and throughout the Department of Defense (DoD). The CDAS was founded in January of 2004 with the primary goal of providing support to USMA researchers with statistical questions and data analysis needs. The organization is also chartered to potentially provide support to organizations outside of the West Point community as needed.

3 CDAS Organization

The CDAS is a new branch of an already existing center: the Mathematical Sciences Center of Excellence (MSCE). The MSCE provides coordination for outreach and projects for both faculty and students in the Department of Mathematical Sciences with a variety of external organizations to include an important partnership with the Army Research Laboratory (ARL). The CDAS enhances the capabilities to include a statistical component and support.

The organization of the CDAS is depicted in Figure 1. In addition to administrative leadership from a director and assistant director, the primary statistical expertise is provided by "senior faculty advisors" with Ph.D.'s in statistics or related fields. Members of the CDAS work on projects in teams (or individually) but have ready access to the senior advisors in case they need statistical support themselves.

Currently, membership and participation is completely voluntary and done in addition to normal teaching loads. The CDAS has between 15 and 20 members who have expressed an interest in working on projects. The membership is not restricted to the Department of Mathematical Sciences. The CDAS has active members from the Orthopedic Surgeon at Keller Army Community Hospital, the Department of Electrical Engineering and

Computer Science and the Department of Systems Engineering at West Point. Members include both Ph.D. and M.S. degree holders in a variety of fields to include statistics, biostatistics, epidemiology and operations research.

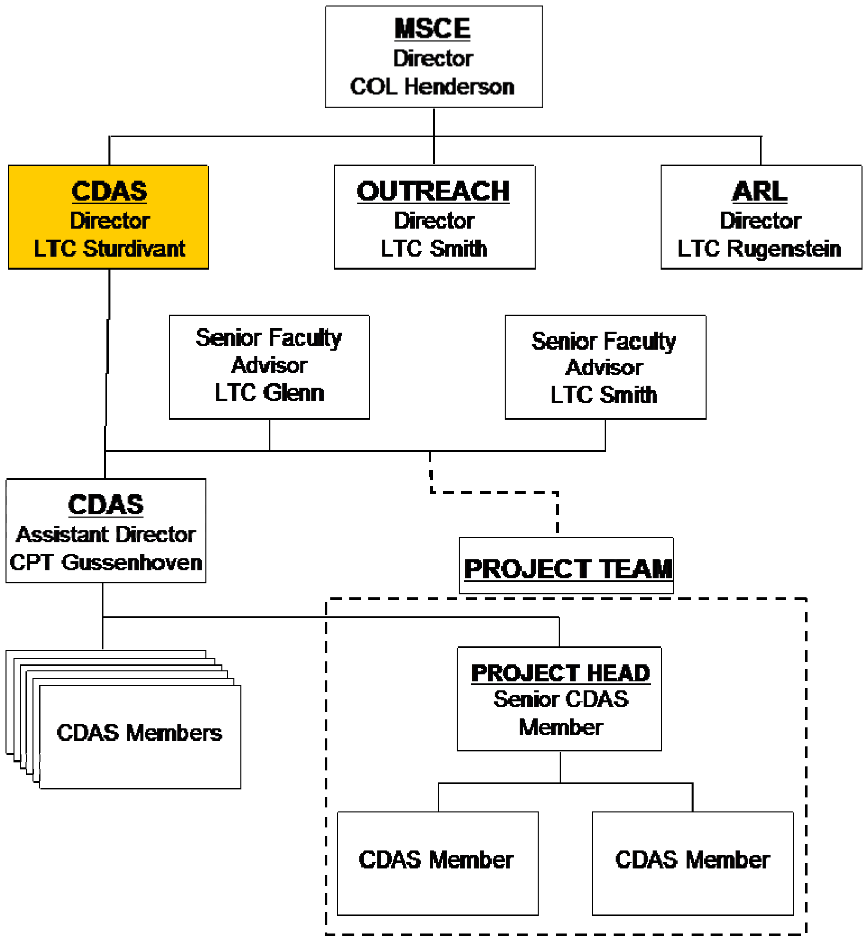


Figure 1: Organization of the CDAS

4 CDAS Mission and Projects

The CDAS is designed to provide statistical consulting and data analysis support for the West Point community first and to DoD agencies where possible. The level of support can range across a very wide spectrum. In some cases, the service could be as limited as answering a quick statistical question or acting to review/comment on a statistical approach to a problem. At the other extreme, the project might include complete data analysis done by the CDAS for a client.

In addition to provided support to other agencies, the CDAS is designed to also provide professional development opportunities for faculty members at USMA. The projects allow members to increase their statistical expertise, keep current on statistical techniques and put skills into practice. These experiences are invaluable for rotating military faculty members who will leave West Point for assignments throughout the Army – many in Functional Area (FA) 49 (Operations Research) where they will perform similar duties.

A third goal is to enhance the educational experience for cadets in our programs. This can occur in several ways. One is that the faculty projects provide insights and examples for use in the classroom. In some instances, faculty work is extended to student work in homework assignments our course projects.

A more direct impact is that cadets may become involved in the actual CDAS work. There are two mechanisms in place for such participation. The first is the MA491 course: a senior thesis conducted in the spring semester. Where the scope and timing of a client project is appropriate, the CDAS members can act as advisors for a cadet(s) thesis to work on the problem (or part of it). A second opportunity for cadets to participate is during summer Academic (AIAD). In that case, the client actually sponsors a cadet during the summer to work on a project.

While the organization is relatively new, we have already had numerous clients across the gamut of possible projects. We have provided tutoring and quick answers to statistical questions for many clients. Several much larger projects have also come to the CDAS and cross a wide spectrum of disciplines, statistical techniques and organizations. Some examples include:

- A study of juvenile recidivism in New York with the New York Military Academy; involved logistic regression and survival analysis
- Donor solicitation with the Association of Graduates (AOG) at West Point; sampling theory, ANOVA and categorical data analysis
- Football "sabermetrics" advice for a cadet project; involved ordinal logistic regression
- ACL injuries in cadets studied by the orthopedic surgeon at West Point – primarily categorical data analysis

Several projects are currently being worked and some tentative ties with organizations outside of West Point in place. We have intentionally built the organization slowly in order to ensure quality service. As a volunteer organization, our greatest challenge is encouraging participation from all members so that a few are not over-whelmed with work. We are developing a web-site and data base to help control and manage requests for statistical support. The administrative aspects of such an organization still require some work.

The other obvious challenge is to ensure we have the appropriate expertise to provide sound statistical advice and support. Most clients have brought problems somewhat foreign to the member providing the service. This leads to a need to shape client expectations as the CDAS team needs time to research the topic. On the other hand, these cases provide the very professional development opportunities we seek for our members. As a group, the CDAS has a wealth of expertise in a variety of statistical areas providing needed support to those working a project. Regular monthly meetings provide opportunities to discuss ongoing projects or have members share their own statistical knowledge to expand that of each individual member. These meetings have probably been the greatest benefit of the CDAS to date.

5 The CDAS and Other Agencies

Over time, we hope to become a ready source of statistical support throughout the DoD. In this vein, we offer several important benefits to agencies that might need statistical work done. One advantage of using the CDAS is that we can offer an "honest broker" and a "second set of eyes" on data analysis projects. In most cases, the CDAS is unaffected by the results of a statistical analysis done by agencies we might support. As a result, our confirmation of results of such analysis can provide strong support since we have no vested interest in the outcome.

A second role we might provide in time is something of a repository for statistical analysis throughout the Army. If West Point has ties to various agencies performing data analysis we might be able to help connect (as the ACAS does) those working on similar problems.

Perhaps most importantly, the CDAS can help support when either expertise or time are lacking. This support might be quick questions and advice or could be much larger in scope. The CDAS is prepared to help perform the analysis when needed. Even very large projects over longer periods of time are possible. In particular, the organizations with such needs can consider several key opportunities for support. One is the previously mentioned cadet availability. This is best during the Spring semester (January through April) with senior projects or during the summer in dedicated AIADs.

Our faculty members are also available for larger project work. This is particularly the case during the summer months when many instructors work on research projects.

Finally, membership in the CDAS is not limited to either the Department of Mathematical Sciences or United States Military Academy. We can envision a CDAS which includes statisticians from a number of organizations ready to provide support to the DoD on statistical projects. We should note here that we are actively pursuing hiring new civilian Ph.D. in statistics or related fields to infuse more expertise into the CDAS. These positions are part of an established post-doctoral fellowship (Davies Fellowship) which has been very successful for all participants over a number of years.

6 Contact Information

Information on the CDAS may be found at the web site:

http://www.dean.usma.edu/departments/math/CDAS/

A SEQUENTIAL STOPPING RULE FOR DETERMINING THE NUMBER OF REPLICATIONS NECESSARY WHEN SEVERAL MEASURES OF EFFECTIVENESS ARE OF INTEREST

October 2004

Anthony J. Quinzi
TRADOC Analysis Center
White Sands Missile Range, NM

Part I. Stopping Rule

1. Introduction

Historically, TRAC analysts have relied on a fixed-sample-size[1] procedure (the "n = 21 rule-of-thumb") to estimate the mean value μ of an output measure of battle effectiveness. For example, μ may represent the mean number of friendly losses.[2] The "n = 21 rule-of-thumb" is based on the assumptions that the replications are independent and produce a sequence of independent, identically distributed random variables $X_1, X_2, X_3, \ldots, X_n$. Confidence intervals and tests of hypothesis can then be obtained based on an application of the Central Limit Theorem, namely that for n sufficiently large, the distribution of the random variable

$$\frac{\overline{X}}{s_n/\sqrt{n}} \tag{1.1}$$

is approximately normally distributed, where $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $s_n = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$. It follows that for sufficiently large n, an *approximate* $100 \times (1 - \alpha)\%$ confidence interval for μ is given by

$$\overline{X} \pm t_{n-1,1-\alpha/2}\sqrt{\frac{s_n^2}{n}}\,, \tag{1.2}$$

where $0 < \alpha < 1$ and $t_{n-1,1-\alpha/2}$ is the upper $1 - \alpha/2$ critical point for the t distribution on $n - 1$ degrees of freedom. A value typically chosen for α is .05 yielding a confidence level of $(1 - \alpha)$ or 95%. If it is further assumed that $X_1, X_2, X_3, \ldots, X_n$ are *normally* distributed, it follows that the confidence interval (1.2) is *exact for any sample size* $n > 1$. One drawback of the fixed-sample-size approach is that the analyst has no control on the precision of the estimate $\overline{X}$.

2. Notions of precision

Law and Kelton [2000], hereafter referred to as LK [2000], define a number of ways of measuring the error in $\overline{X}$. Suppose that n replications resulted in a mean $\overline{X}$ = 99.7 when the (unknown) true value of $\mu = 100$. The *absolute error* of estimation β would be

$$\beta = \left|\overline{X} - \mu\right|$$

or 0.3. The *relative* error of estimation γ would be

$$\gamma = \frac{\left|\overline{X} - \mu\right|}{\mu}$$

or 0.003 which can be thought of as a *percentage error* of 0.3% in $\overline{X}$.

The sample mean $\overline{X}$ of a random sample of size n from a population with mean μ and standard deviation σ has a standard deviation which can be estimated by $s_n/\sqrt{n}$, where s_n is defined as above. Because s_n is a consistent estimator of the population variance, the sample mean becomes "stable" for large n. By specifying a degree of accuracy, say relative error, the

[1] That is, a fixed number n of replications

[2] Generally, analysts are not in the business of obtaining precise estimates of battle parameters for the sake of estimation alone, but rather to be able to compare these estimates across study alternatives.

analyst is able to formulate a stopping rule for a sequence of replications and be assured that with high probability (specifically, $1 - \alpha$), the sample mean has been estimated within the specified degree of accuracy.

3. The work of Cherolis

Cherolis [1992] suggested a sequential procedure based on (1.2) to determine the number of replications necessary to estimate μ with a specified degree of accuracy. The procedure is in the form of a stopping rule which can determine, after a small number of replications have been performed, how many *subsequent* replications are necessary to be able to estimate μ with a specified accuracy. One drawback of Cherolis' result is that it applies to a single measure of effectiveness. Because of recent simulation work involving the Army's Future Combat System (FCS), it is of interest to determine a stopping rule that can determine how many replications are necessary to be able to estimate a number of output parameters simultaneously.

4. The case for a single measure of effectiveness

LK [2000] suggest the following *sequential* procedure for obtaining an estimate of μ with a specified relative error of γ, $0 < \gamma < 1$, that takes only as many replications as are actually needed:

Suppose X_1, X_2, X_3, ... is a sequence of independent, identically distributed random variables. It is important to note that the random variables need not be normally distributed. These may represent, for example, the numbers of friendly losses in replications 1, 2, 3, etc.. Choose an initial number of replications $n_0 \geq 2$. The actual number chosen should depend on the amount of replication-to-replication variability. If the variability is not large, then $n_0 = 5$ replications may be sufficient. If the variability is large, then at least $n_0 = 10$ replications should be made. The choice of relative precision γ may have to be adjusted when there are not sufficient resources to perform the required number of replications.[3]

Step 1. Make n_0 replications of the simulation and set $n = n_0$.

Step 2. Compute $\bar{X}$ and the quantity $\delta(n,\alpha) = t_{n-1,1-\alpha/2}\sqrt{\frac{s^2}{n}}$, where s is defined in (1.1) and the level of confidence is $100\times(1-\alpha)$ %, $0 < \alpha < 1$.

Step 3. If $\delta(n,\alpha)/\left|\bar{X}\right| \leq \gamma'$, where $\gamma' = \gamma/(1+\gamma)$, use $\bar{X}$ as the point estimate for μ and stop. If $\alpha = .05$, for example, then the interval

$$I(.05,\gamma) = [\bar{X} - \delta(n,.05), \bar{X} + \delta(n,.05)]$$

is an approximate 95% confidence interval for μ with the desired precision. If the inequality fails, replace n by $n + 1$, make one additional replication of the simulation, go to step 2 and repeat the process.

[3] LK[2000] show that it is possible to obtain *rough estimates* (a table of values) of the number of replications required to estimate μ with desired levels γ of relative precision.

5. The case of multiple measures of effectiveness

Let $\mu_1, ..., \mu_k$ represent the means of k measures of effectiveness. For each mean μ_s, a $(1 - \alpha_s) \times 100\%$ confidence interval is determined, $s = 1, ..., k$. Suppose that $\sum_{s=1}^{k} \alpha_s = \alpha$. Then the joint probability that *all k* confidence intervals *simultaneously* contain their respective true means is at least

$$1 - \sum_{s=1}^{k} \alpha_s \ . \qquad \textbf{(5.1)}$$

This result is known as the Bonferroni inequality and (5.1) is called the Bonferroni bound. It should be noted that the α_s need not be equal. For example, given four measures of effectiveness, suppose that a 99% confidence interval were computed for μ_1, a 98% confidence interval were computed for μ_2, a 97% confidence interval were computed for μ_3 and a 96% confidence interval were computed for μ_4. In this case, it may be that the first measure ($s = 1$) is most important and so the highest level of confidence (99%) is chosen for that measure. If the confidence level is 99%, then $\alpha_1 = 1 - .99 = .01$. For confidence level 98%, $\alpha_2 = 1 - .98 = .02$, and so on. Using the Bonferroni bound (5.1), the joint probability that all 4 confidence intervals simultaneously contain their respective means would be at least [1 – (.01 + .02 + .03 + .04)] or 0.90. In order to extend the above stopping rule to include multiple measures of effectiveness, it is necessary to specify a relative precision for each measure. The 3-step procedure outlined above would have to be performed for *each* measure. At any stage of the process, it may occur that some measures require an additional replication and some not. The procedure will stop when *every* inequality in Step 3 of the above procedure holds. Because the procedure requires more data than in the case of a single measure of effectiveness, LK[2000] recommend that the number k of measures be no greater than 10.

Part II. Measures of Effectiveness

The following four measures of performance were of interest: friendly (BLUE) system (vehicle) losses, friendly individual soldier (dismounted) losses, threat (RED) system (vehicle) losses, and threat individual soldier (dismounted) losses. It was desired to apply the stopping rule to all four measures simultaneously.

Part III. Application

Because one replication of the full Caspian scenario takes approximately 60 hours of computer run time, it was recommended that the sequential procedure suggested in Part I be tested in scaled down version of the same scenario whose run time is considerably less, about 6 hours. The sequential procedure was tabulated in an Excel spreadsheet. A portion of the spreadsheet is reproduced here.

Reference: Law & Kelton Ed. 3, pp. 513-514

REP					Blue Losses Vehicles		Blue Losses Dismounts		Red Losses Vehicles		Red Losses Dismounts
1					35		32		75		110
2					43		29		83		133
3					41		32		81		122
4					32		43		82		141
5					47		24		75		115
6					38		35		81		118
7					34		36		83		122
				delta(n, a)		delta(n, a)		delta(n, a)		delta(n, a)	
	d1	2.969	80% (Each .05)	6.04	CONTINUE	6.67	CONTINUE	3.94	STOP	11.98	CONTINUE
	d2	3.707	92% (Each .02)	7.54	CONTINUE	8.33	CONTINUE	4.92	STOP	14.96	CONTINUE
	d3	4.317	96% (Each .01)	8.78	CONTINUE	9.70	CONTINUE	5.73	STOP	17.42	CONTINUE
	d4	5.959	99.2% (Each .002)	12.12	CONTINUE	13.39	CONTINUE	7.91	CONTINUE	24.05	CONTINUE
8					39		60		78		112
				delta(n, a)		delta(n, a)		delta(n, a)		delta(n, a)	
	d1	2.841	80% (Each .05)	5.01	CONTINUE	11.07	CONTINUE	3.34	STOP	10.67	STOP
	d2	3.499	92% (Each .02)	6.17	CONTINUE	13.63	CONTINUE	4.12	STOP	13.14	CONTINUE
	d3	4.029	96% (Each .01)	7.10	CONTINUE	15.70	CONTINUE	4.74	STOP	15.13	CONTINUE
	d4	5.408	99.2% (Each .002)	9.53	CONTINUE	21.07	CONTINUE	6.36	STOP	20.31	CONTINUE
9					52		31		87		118
				delta(n, a)		delta(n, a)		delta(n, a)		delta(n, a)	
	d1	2.752	80% (Each .05)	5.92	CONTINUE	9.60	CONTINUE	3.61	STOP	9.18	STOP
	d2	3.355	92% (Each .02)	7.21	CONTINUE	11.70	CONTINUE	4.41	STOP	11.20	CONTINUE
	d3	3.833	96% (Each .01)	8.24	CONTINUE	13.36	CONTINUE	5.03	STOP	12.79	CONTINUE
	d4	5.041	99.2% (Each .002)	10.84	CONTINUE	17.58	CONTINUE	6.62	STOP	16.82	CONTINUE

Explanation of Spreadsheet Calculations

1. We begin with an initial n_0 = 7 replications and test the inequality in Step 3.
2. The inequality is tested for each parameter. I tried four different experimentwise values for α: .2, ,08, .04 and 0.008 for use in the Bonferroni bound. The corresponding critical values for the *t*–distribution are listed in the next column.
3. The word "CONTINUE" appears in green if the respective inequality fails, and the word "STOP" appears in red, otherwise.
4. The quantities delta(n, a) refer to the quantities $\delta(n,\alpha)$ in the sequential procedure.

Replications were continued until n = 30, and the inequality for measure 2, friendly individual soldier losses, was never realized. It is clear that the most variable of a collection of parameters will always be the one which determines the necessary sample size. This was not a satisfactory result for the scaled down scenario, and would have been impossible to determine for the full blown scenario because of time limitations. There has to be a better way.

Questions for the Panel Members

1. Given that we are dealing with purely discrete distributions (numbers of losses) each of whose underlying distributions results from a large number of random draws in the model, should we really be considering a procedure based on the *t*-distribution which assumes population is Gaussian, especially in view of small sample sizes?

2. Should we instead be trying to estimate the underlying discrete probability mass functions [cf. Chiu, S.T. (1991) "Bandwidth selection for kernel density estimation", *Ann. Stat. Vol. 19*, pp. 1883-1905] or use some other methodology so that we might be able to improve on the Bonferroni bounds?

3. Are bootstrap methods really appropriate in a PURELY discrete context such as this?

4. Is this question crying for some sort of Bayesian approach?

REFERENCES

Chelis, George T. (1992). A Sequential Stopping Rule for Reducing Production Times during CASTFOREM Studies, Master's Thesis, New Mexico State University, Las Cruces, NM.

Chiu, S. T. (1991). Bandwidth selection for kernel density estimation, *Ann. Stat. Vol. 19*, pp. 1883-1905.

Law, Averill M. and Kelton, W. David (2000). *Simulation Modeling and Analysis*, *Third Edition*, McGraw-Hill, Boston, Massachusetts.

SEQUENTIAL STOPPING RULE FOR DETERMINING THE NUMBER OF REPLICATIONS WHEN SEVERAL MEASURES OF EFFECTIVENESS ARE OF INTEREST

Anthony J. Quinzi

October 2004

BACKGROUND

- How many replications of a scenario is enough to estimate a mean performance parameter with a specified degree of accuracy and level of confidence?
- For single measure, use the following fact:

If $X_1, \ldots, X_n$ is a random sample of size n from a normal population with mean μ, then

$$P\left(-t_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}} \leq \left| \overline{X} - \mu \right| \leq t_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}} \right) = 1 - \alpha$$

SEQUENTIAL STOPPING RULE

- Usual $100 \times (1-\alpha)\%$ confidence interval for μ

$$1 - \alpha \geq P(\left|\overline{X} - \mu\right| \leq \text{half-length}) \Rightarrow$$

$$1 - \alpha \geq P\left(\left|\frac{\overline{X} - \mu}{\mu}\right| \leq \frac{\gamma}{1 - \gamma}\right)$$

$$\left|\overline{X} - \mu\right| = \text{absolute error}$$

$$\gamma = \left|\frac{\overline{X} - \mu}{\mu}\right| = \text{relative error}$$

$$\gamma' = \frac{\gamma}{1 - \gamma} = \text{adjusted relative error to achieve}$$

a relative error of γ

SEQUENTIAL STOPPING RULE

Problem: How many replications are sufficient to achieve a given precision (γ) with confidence $100 \times (1 - \alpha)\%$?

- **Law and Kelton (1982) suggest a sequential stopping rule for estimation of the mean μ**
- ***Step 1. Make n_0 replications of the simulation and set $n = n_0$.***
- ***Step 2. Compute $\overline{X}$ and the quantity***

$$\delta(n,\alpha) = t_{n-1,1-\alpha/2}\sqrt{\frac{s^2}{n}} \quad \text{where} \quad s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

and the level of confidence is $100 \times (1-\alpha)$ %, $0 < \alpha < 1$.

- ***Step 3. If*** $\delta(n,\alpha)/\left|\overline{X}\right| \le \gamma'$, where $\gamma' = \gamma/(1+\gamma)$

  use $\overline{X}$ as the point estimate and stop. Otherwise, Replace n by (n + 1), make one additional replication and go to step 2 to repeat the process.

SEQUENTIAL STOPPING RULE

THE NEED FOR MULTIPLE MOE

- **A study of key performance parameters for the Army's Future Combat System of Systems was conducted.**
- **Following question arose: How many replications are necessary to estimate a number of mean performance parameters simultaneously?**
- **Question was important in that the Caspian scenario being used required 60 hours of real time for each replication. Post processing was a huge effort and to meet the study deadline, analysts used the results from only $n = 11$ replications.**

SEQUENTIAL STOPPING RULE

THE CASE OF MULTIPLE MOE

- Suppose $\mu_1, \ldots, \mu_k$ represent the means of *k* MOE.
- For each mean μ_s , form a $100 \times (1 - \alpha_s)\%$, where

$$\sum_{s=1}^{k} \alpha_s = \alpha$$

- Then the Bonferroni bound provides a lower bound for the joint probability that each of the *k* confidence intervals captures its respective mean.

SEQUENTIAL STOPPING RULE

STOPPING RULE FOR MULTIPLE MEASURES

- **Perform the 3-step procedure outlined above for *each* measure**
- **Stop when *every* inequality in Step 3 of the above procedure holds.**

SEQUENTIAL STOPPING RULE

APPLICATION

- A scaled-down version of the Caspian scenario (6 hours per replication) was used to test the 3-step procedure for multiple measures.
- MOE of interest were:

  1) Friendly system losses

  2) Friendly individual soldier losses

  3) Threat system losses

  4) Threat individual soldier losses
- Initial n_0 = 7 replications were run.
- Results of n = 30 replications were used.

SEQUENTIAL STOPPING RULE

		Reference: Law & Kelton Ed. 3, pp. 513-514							
REP					Blue Losses Vehicles		Blue Losses Dismounts		Red Losses Vehicles
1					35		32		75
2					43		29		83
3					41		32		81
4					32		43		82
5					47		24		75
6					38		35		81
7					34		36		83
				delta(n, a)		delta(n, a)		delta(n, a)	
	d1	2.969	80% (Each .05)	6.04	CONTINUE	6.67	CONTINUE	3.94	STOP
	d2	3.707	92% (Each .02)	7.54	CONTINUE	8.33	CONTINUE	4.92	STOP
	d3	4.317	96% (Each .01)	8.78	CONTINUE	9.70	CONTINUE	5.73	STOP
	d4	5.959	99.2% (Each .002)	12.12	CONTINUE	13.39	CONTINUE	7.91	CONTINUE

SEQUENTIAL STOPPING RULE

RESULTS

In 30 replications, stopping rule not satisfied for measure (2), the mean number of friendly individual soldier losses. Variability due primarily to large increases in losses when a squad of soldiers was mounted and the platform received a catastrophic kill.

QUESTIONS FOR THE PANEL

1. **Given that we are dealing with purely discrete distributions (numbers of losses) each of whose underlying distributions results from a large number of random draws in the model, should we really be considering a procedure based on the *t*-distribution which assumes population is Gaussian?**

QUESTIONS FOR THE PANEL

2. Should we instead be trying to estimate the underlying discrete probability mass functions [cf. Chiu, S.T. (1991) "Bandwidth selection for kernel density estimation", *Ann. Stat. Vol. 19*, pp. 1883-1905] or use some other methodology so that we might be able to improve on the Bonferroni bounds?

QUESTIONS FOR THE PANEL

3. Are bootstrap methods really appropriate in a PURELY discrete context such as this?

4. Is this question crying for some sort of Bayesian approach?

Determining A Minimal Alternatives Replication Set For Constructive Combat Simulation

Paul J. Deason, PhD

U.S. Army Conference on Applied Statistics
Georgia Tech, Atlanta
Oct 2004

Problem

- **CASTFOREM is a closed-form stochastic physics-effects based combat simulation.**
- **As more fidelity and capability to represent combat is added the single replication run time has become very large.**
- **A study consists of a Base Case and alternatives which are variations in weapons systems, etc.**
- **A usual study has 11 to 21 replications of the Base Case and each alternative**
- **A first-pass method to indicate where differences might reside using fewer replications is desired.**
- **Others have recommended the Boot Strap procedure for CASTFOREM. This is another idea.**

Assumptions

- Measures of merit are
  - Force Exchange Ratio [(Red Loss/Initial Red)/(Blue Loss/Initial Blue)]
  - Blue Loss.
- A first-pass method is wanted to determine areas of potential differences, and where further exploration may be warranted.
- Traditional hypothesis testing is not feasible due to the number of replications required for precision, and the cost in obtaining them.
- The stochastic combat simulation running one scenario is a closed system.
  - A "run number" identifies a closed set of simulation "seeds" (see next slide)
- A finite number of replication runs of the base case stochastic simulation defines the population of interest.

The Nature of CASTFOREM's Case Number Run Seeds

- **In CASTFOREM, each replication run number identifies a set of number seeds used in the stochastic simulation.**
- **If there is no change in the model, scenario or data, and the same run number (with its underlying seeds) is used in a replication, results will be identical.**
- **<u>Rather than using a random set of run numbers for producing comparison results, use a set centered on the Base Case central value</u>**
  - **The run numbers responsible for the replication run(s) producing results close to the central measures of the Base Case for the principle measure of interest can be identified.**
  - **In the example, the 9th case has the median FER, and the 7th case the median Blue Loss.**
- **Seeds identified by these case numbers can then be used as seeds to run the alternate cases.**
  - **For example, sort in ascending order the FER and the associated run numbers**
  - **Select the run number for the Median, and the three adjacent runs below and above for a total of seven run seeds**
- **The objective would be for the bulk of the limited resources available for the simulation effort to be expended in running the Base Case.**
  - **A limited set of runs would be made for the alternatives based on the run numbers identified by the Base Case's Median and adjacent runs.**

Why Select a Sample Based on the Base Case Center

- **The rationale behind selecting a sample determined by the central performance of the Base Case is to select a set that has the best possibility of exhibiting the same behavior.**
  - **Variance within the alternative sample is not of interest**
- **IF the limited sample is different enough to be identified as showing itself to be different in the most global measure of effectiveness (the most robust and resistant to change,) then there may be a strong likelihood that the alternative may in fact exhibit differences.**
- **Additional runs may then be undertaken to better understand the nature of the differences due to the alternative.**

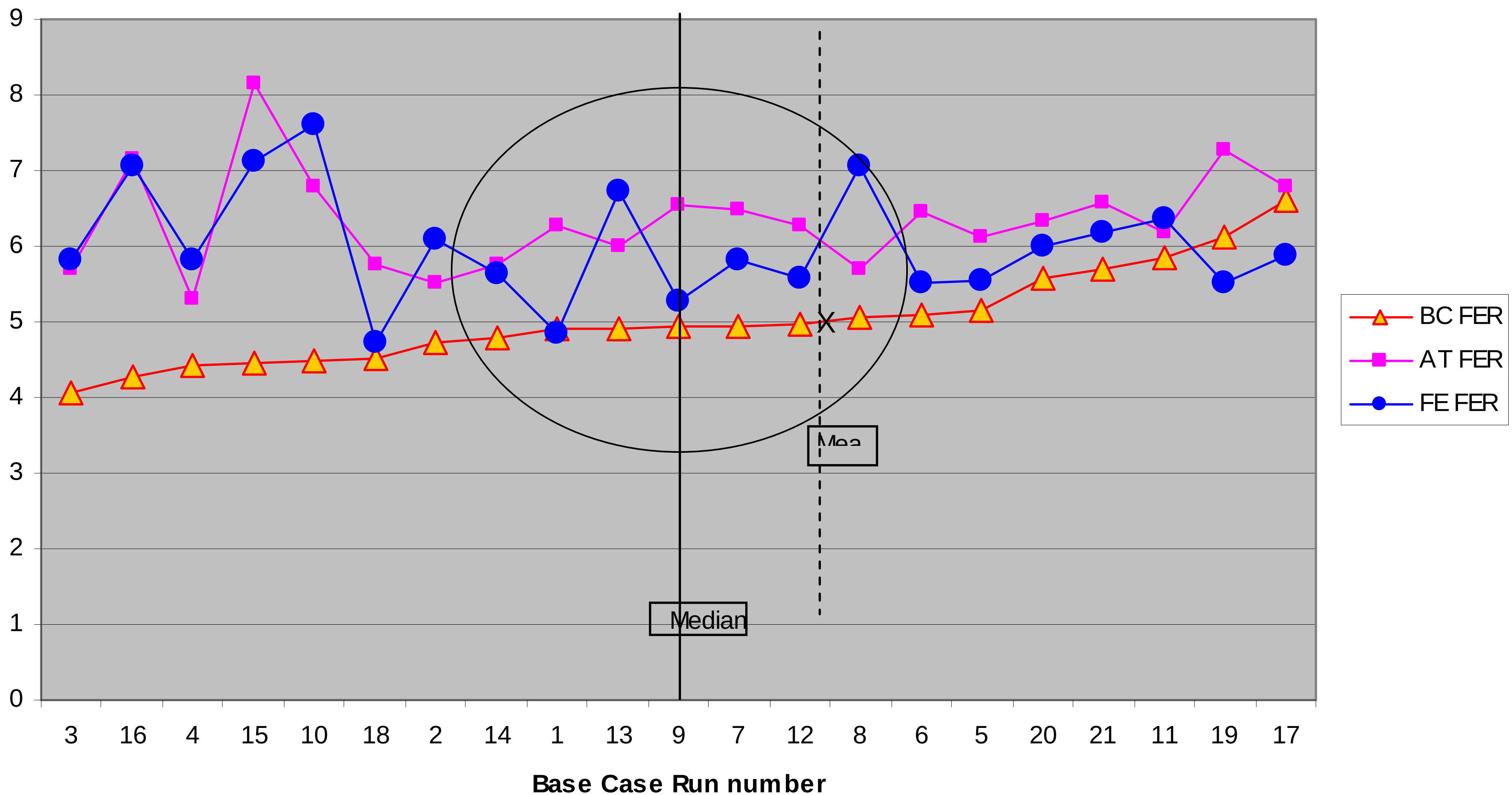
CASTFOREM Example

- **This is an unclassified example based on an Infantry Battalion in a fixed point defense against a Threat armored brigade**
  - The Base Case had the usual direct fire and artillery support
  - One alternative had a unit of advanced anti-armor (AT) weapon system
  - One alternative had a system of systems for defeating Armor deep, with multiple entities, interlinked systems of anti-tank, artillery, and other combat multipliers. (This simulation scenario was based on a field experiment)
- **Principle measure of effectiveness-- Measures of merit**
  - Force Effectiveness Ratio [(Red Loss/Initial Red)/(Blue Loss/Initial Blue)]
  - Blue Loss (interest actually in Blue survival)
  - These are non-independent measures

Method using Median

- Run Base Case numerous times (21 in this case)
- Select median 7 replications (median, three up, three down) based on:
  - Force Exchange Ratio, or
  - Blue Loss (they are not independent)
- Run the alternative using the 7 run “seeds” identified by their run numbers in the BC
- Test by:
  - Parametric (ANOVA, SNK, lsd)
  - Non-parametric (Kruskal-Wallis, Median test, Mann-Whitney)

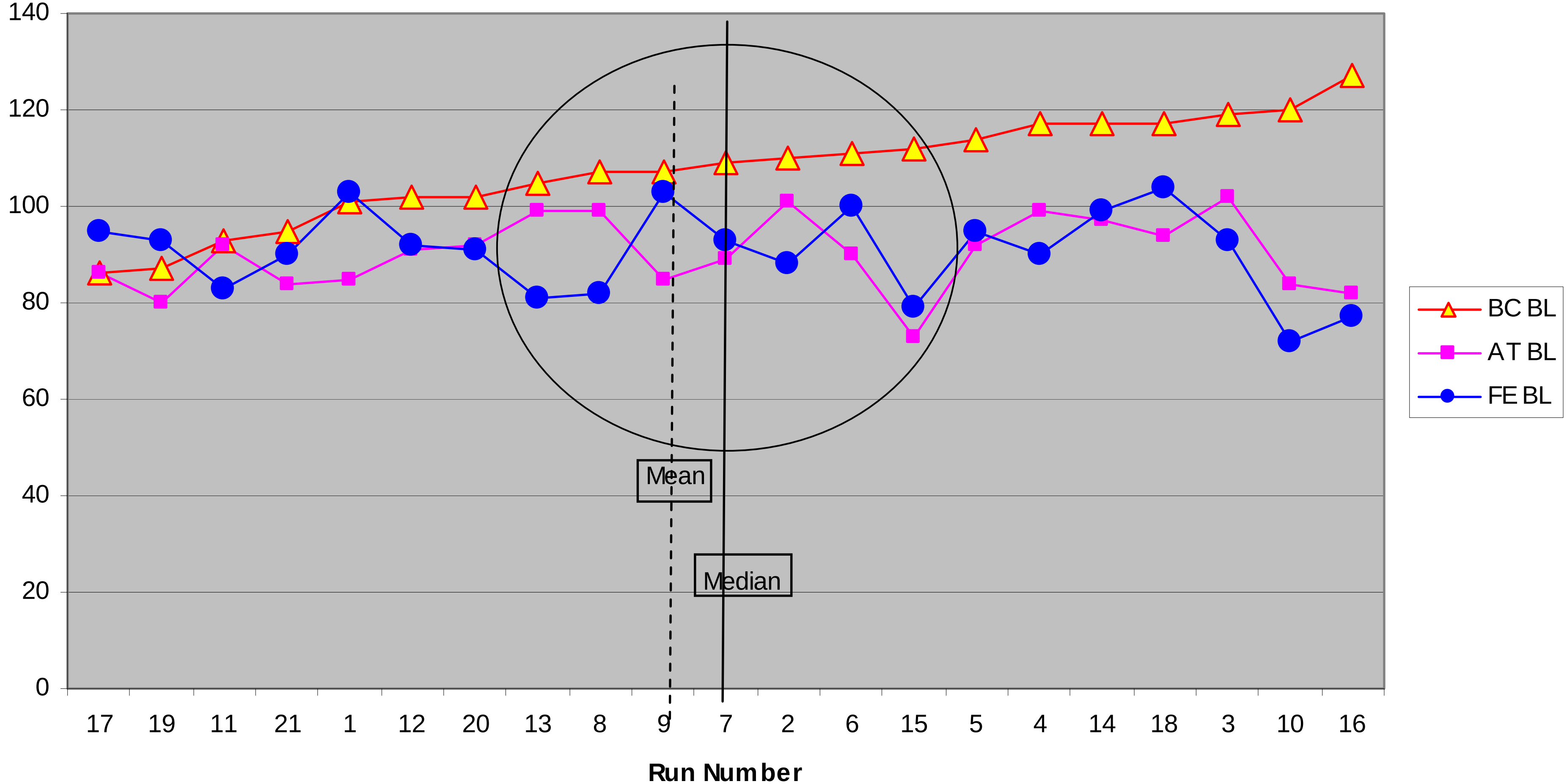
FERs Ordered by Base Case FER



Statistical Tests Conducted on FER

- Full sample of 21 replications per case
  - Variance homogeneous (Levene (2,60)=.414 p=.663)
  - One-Way ANOVA significant <.001;
    - SNK, lsd => BC<AT=FE
  - Kruskal-Wallace significant <.001; Median test < .001
  - Pairwise Mann-Whitney => BC<AT; BC<FE; AT=FE
- Reduced – 21 for Base Case, 7 for alts
  - Variance homogeneous (Levene (2,32)=.1.221 p=..308)
  - One-Way ANOVA significant <.001, lsd=> BC<AT=FE
  - Kruskal-Wallace significant =.001; Median test < .001
  - Pairwise Mann-Whitney => BC<AT; BC<FE; AT=FE

Blue Loss ordered by Base Case



Statistical Tests Conducted on Blue Loss

- **Full sample of 21 replications per case**
  - **Variance homogeneous** (Levene (2,60)=1.071 p=.349)
  - **One-Way ANOVA significant <.001;**
    - **lsd, SNK => BC>AT=FE**
  - **Kruskal-Wallace significant <.001; Median test < .001**
  - **Pairwise Mann-Whitney => BC>AT; BC>FE; AT=FE**
- **Reduced – 21 for Base Case, 7 for alts**
  - **Variance homogeneous** (Levene (2,32)=.0.094 p=..910 (suspicious?))
  - **One-Way ANOVA significant <.001;**
    - **SNK, lsd=> BC>AT=FE**
  - **Kruskal-Wallace significant =.001; Median test < .001**
  - **Pairwise Mann-Whitney => BC>AT; BC>FE; AT=FE**

Another Approach? – Method using Mean

- **Take mean, SE, n of the Base Case sample**
- **Sort the Base Case results in ascending order & select 7 replications**
  - **the one nearest the mean, three above and three below the mean value and note their case numbers.**
- **Run the 7 cases in the alternate using the same case number "seeds"**
- **Do a t-test between the mean value of the Base Case and the mean of the alternate sample. Use the SE $[sd/(n)^{1/2}]$ of the Base Case for the comparison.**

Ideas?

- **This presentation is in the pure sense of a clinical case.**
- **Using the Median or the Mean as a marker in CASTFOREM for replications is an idea that seems plausible.**
- **Is this approach logical?**
  - We evaluate at a system of systems level, not one measure.
  - Usually the first measure we are interested in is mission accomplishment, the others are side issues. So for different measures would this reduced population have different base populations, would we end of running more runs?
- **Is another approach better when the intent is a limited number of costly replications to identify potential differences in alternatives?**
  - **For instance, the Median/hinge Gaussian comparison (Velleman & Hoaglin)**
  - **Cioppa's Boot Strap?**

Paul J. Deason, Ph.D.

paul.deason@us.army.mil

505-678-1610 (DSN 258)

Data Mining in Counterterrorism

David Banks
ISDS
Duke University

1. Context

Many entities are exploring the use of data mining tools for help in counterterrorism.

- TIA (DARPA)
- LexisNexis (airline screening)
- Global Information Group (in the Bahamas)
- Terrorist Threat Integration Center (federal)

These efforts raise legal, political, and technical issues.

From the technical standpoint, there are at least three kinds of use that people have in mind:

- Forensic search
- Signature detection, for various prespecified signals
- Prospective data mining for anomalies

2. Forensic Data Mining

This is the least problematic (socially or legally) and the easiest. It is widely, if a bit inefficiently, used by the police:

- The DC snipers
- Background checks for people who purchase guns
- Post-arrest investigation to build cases (e.g., CSI stuff).

- Most forensic data mining problems involve finding matches between known information and large, often decentralized, data.

- Examples include record linkage and biometric identification.

- Also, one sometimes uses multiple search algorithms, and must combine the signal from each.

2.1 Record Matching

Biometric identification tries to rapidly and reliably match physical features. This includes fingerprint matching, retinal scans, and gait analysis.

But the ultimate goal is to use AI to match photos to suspect lists.

DARPA and NIST have a joint project to do automatic photo recognition. The project uses a testbed of images called FERRET. (See Rukhin, 2004, *Chance.*)

The main problem is that the feature extraction systems do not work very well. There are high levels of false positives and false negatives. Shadow and angles are hard.

A data mining strategy I like is:

- Have humans assign a sample of photos impressionistic distances;
- Use multidimensional scaling to embed these photos in a relatively low-dimensional space;
- Use data mining to extract the features of the photo that best predict the low-dimensional metric used by MDS
- Use constrained nonparametric regression to fit the metric.

Looking for matches in large and decentralized databases is much like the classical record-linkage problem (Fellegi and Sunter, 1969).

This methodology has become key in many data mining applications:

- Google search
- Census data quality
- TREC studies

The original method tries to formally match partial or noisy names and addresses.

Text matching builds on linkage, with complex rules for stemming words and using co-present text as covariates.

Winkler (2002) is applying modern data mining to linkage (Bayes nets, SVMs, imputation, etc.)

2.2 Combining Algorithms

A common strategy in looking for matches in a database is to use an algorithm to assign a rank or score to the match between the target (say a photo or a fingerprint) and each record in the database.

A human would then assess the top-ranked matches by hand.

One strategy to improve matching is to use boosting (Freund and Schapire, 1996). This approach successively modifies a weak classifier to produce a much better classification algorithm.

In match finding the classification problem is a little non-standard, but the ideas can be adapted.

The boosting algorithm is:

1. Initialize obs. weights $w_i = 1/n$
2. Do $m=1, \ldots, M$ times:
   a) Fit $G_m(x)$ with weights w_i
   b) Get $e_m = \Sigma\, w_i\, I(\text{error on } x_i) / \Sigma\, w_i$
   c) Compute $\alpha_m = \log[(1 - e_m)/em]$
   d) Use $w_i \exp[\alpha_m\, I(\text{error on } x_i)]$ in place of w_i
3. Use $G(x) = \text{sign}[\Sigma\, \alpha_m\, G_m(x)]$

- The boosting algorithm can be viewed as a weighted sum of M related algorithms.

- Similarly, one can use weighted combinations of unrelated algorithms; this also improves accuracy.

- But there are hard issues on how to weight or combine.

Rukhin (2004) describes issues in combining algorithms for biometric identification.

Using scan statistics, copulas, and partial rankings, he finds a general procedure to combine algorithms to provably improve accuracy for top-rank matches.

One might also try a partitioning approach.

3. Signature Search

Tony Tether at DARPA said that IAO was being used for detecting preset profiles---e.g., Arabic men studying in flight schools.

This is different from forensic search, because the analysts get to pick what they want to find.

DARPA had the TIA/IAO program, the NSA and the DSA have(?) programs, and so forth. The intention is to combine public, federal, and private databases and use statistical techniques to find prespecified risk behaviors.

Contractors have commercialized some programs of this kind, or even taken them off-shore.

The tools used are similar to those needed for forensic datamining. The main distinction is that instead of searching for a match, one has to search for records in a neighborhood of the signature.

This is because terrorists will try to smudge the signal. This could entail false addresses and names, indirect travel patterns, etc.

- The tools used are similar to those needed for forensic datamining. The main distinction is that instead of searching for a match, one has to search for records in a neighborhood of the signature.

- This is because terrorists will try to smudge the signal. This could entail false addresses and names, indirect travel patterns, etc.

- In signature detection problems one probably needs a human in the loop.
- One model is the type of success that El Al screeners have had.
- Data mining can do a lot of preliminary sieving, but humans have domain knowledge that is hard to embed in an AI system.

4. Prospective Mining

This is the Holy Grail in data mining for counterterrorism. One wants to be able to automatically “connect the dots”.

This is extremely hard and perhaps infeasible. The most mature application is in cybersecurity.

Cybersecurity rides two horses:

- Signature detection (CERT, McAfee, spam blockers)
- Anomaly detection.

Anomaly detection tries to identify hack attacks that have not been previously seen. And this has to be done automatically, in order to be sufficiently fast.

DARPA and others have been doing research in anomaly detection for years. The big problem is the false alarm rate. In a high-dimensional space, everything looks funky.

False alarm rates have been a thorn in the side of other federal security programs. The recent NRC study on the use of polygraph data found the error rate was just too high.

Suppose that:

- the false alarm count is f, and the average cost of a false alarm is c.
- the number of missed alarms is m, and the average cost of a missed alarm is d.
- the cost of the program is p.

Then the system should be used if

$$f*c + p < m*d$$

- This decision-theoretic formulation is much too simplistic. One can rank the alarms, do cheap preliminary tests, and use human judgment.

- Nonetheless, this is the right way to assess a system. Any agency that has such a system can easily check whether their program is cost-effective.

5. Conclusions

Data mining has a major role in modern counterterrorism. It is a key methodology in:

- Syndromic surveillance and emergent threats
- Record linkage for background checks
- Cybersecurity

We want to extend the reach of data mining to include:

- Biometric identification
- Automatic dot connection
- Social network discovery
- Prospective classification

But these are hard problems, and some may not be possible. The U.S. government needs to be realistic about what is possible.

Extrapolating Testing for Biological Warfare Agents from the Laboratory to a Field Environment

Charlie E. Holman, Army Evaluation Center
Carl T. Russell, CTR Analytics
Chuck Jennings, EAI Corporation

Field testing using live Biological Warfare Agents (BWA) has long been forbidden in the U.S. However, increasing BWA threats to both troops in the field and to the Homeland has made it imperative for the U.S. to develop systems that can detect and reliably identify BWA threats. Among the systems that the U.S. is developing is a "point" detection system that would be positioned in an area of potential BWA contamination to verify whether or not a BWA is present and to identify it if present. Testing of this system in a laboratory environment with live BWA is possible, but testing with live BWA cannot be done in the field. This paper describes a methodology to extrapolate results from laboratory testing through controlled open-air testing to a field environment. Both the overall methodology and the statistical methodology are discussed. Carefully parameterized logistic regression is the proposed statistical approach, and feasibility results based on previous field testing will be presented. Success of this methodology (if the proposed testing is executed) may be presented at a subsequent ACAS.

Introduction.

The four-service Joint Biological Point Detection System (JBPDS) is among the systems that the United States is developing to deal with the threat posed by Biological Warfare Agents (BWA). The JBPDS is an integrated system designed to automatically detect and identify the presence of BWA aerosols through direct contact with "clouds" potentially containing BWAs. Several versions of the JBPDS are available, including a relatively compact man-portable version that can be pre-positioned in an area potentially subject to BWA attack and a ground-mobile version that can rapidly be deployed to investigate suspicious clouds. The JBPDS provides an audible Alarm together with visual indication of the presence of BWAs displays their identification if any. It can also produce samples for transport to designated laboratories for confirmatory analysis.

As shown in Figure 1, the JBPDS is composed of four main line-replaceable units (LRUs): the Bio Agent Warning Sensor (BAWS), a wetted-wall cyclone collector/concentrator, the Fluidic Transfer System (FTS), and the identifier (Automated Hand-Held Assay, AHHA). An inlet duct on the BAWS LRU provides a pathway for sampling ambient air in close proximity to the instrument. The BAWS particle sensor constantly compares instantaneous measurements to an established background. A fluorescence detector, internal to the BAWS, determines if the sampled air contains aerosolized BWAs. If the BAWS alarms, the collector/concentrator collects and concentrates a sample that is passed by the FTS to the AHHA for identification of the BWA. When the AHHA receives the appropriate signal, a liquid sample from the FTS is automatically injected onto the assay strips. The strips, housed in a carrier, have identification markers that appear when a liquid sample of the BWA is inoculated onto the matching antibody strip. An optical reader of the carrier strips provides a means for identification.

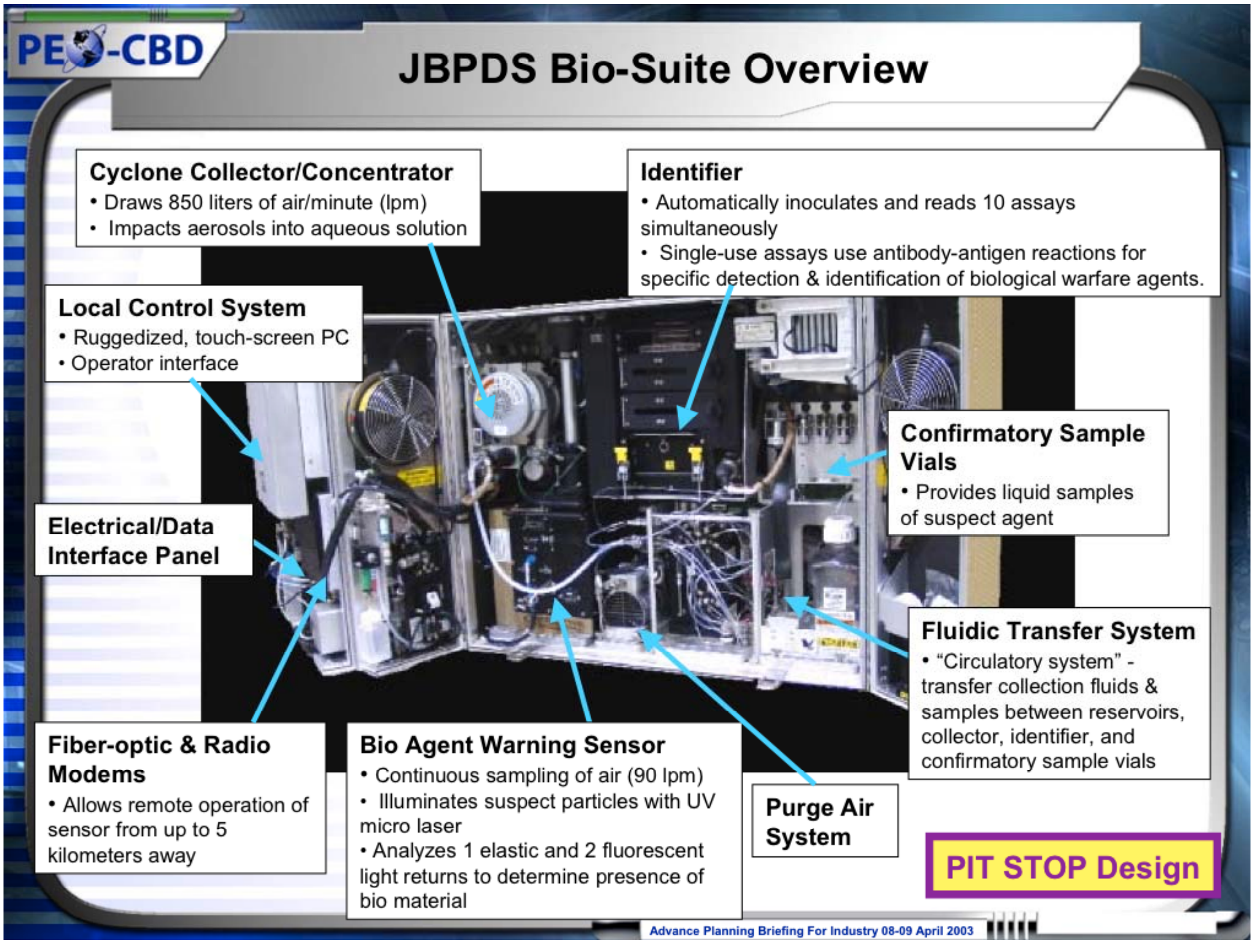


Figure 1. JBPDS Bio-Suite Overview.

The JBPDS has been tested in an enclosed containment chamber, in a controlled outdoor "Ambient Breeze Tunnel" (ABT), and in open field environments. Actual BWAs have only been used in the chamber, however, since outdoor testing with live/active BWAs (or even killed/inactivated BWAs) has long been forbidden in the U.S. In an open field environment (including the ABT), testing is only possible with killed/inactivated agent-like organisms (ALO). and/or live/active (or killed/inactive) biological simulants. However, no systematic study has yet been done to link performance of the JBPDS in the chamber with actual BWAs to performance of the JBPDS in outdoor environments with simulants and killed ALOs. In addition, it has not been possible to test the JBPDS as an integrated system in current containment chambers. There is also an issue concerning how the important cloud concentration factor is measured in the various test environments. Concentration is typically measured in terms of Agent Containing Particles per Liter of Air (ACPLA) for which there are several measurement methods, but the relationships between results for the various methods have not been thoroughly studied.

The Whole System Live Agent Test (WSLAT) is being proposed to address the issues outlined in the previous paragraph. First, WSLAT proposes constructing a containment chamber suitable for testing the smallest man-portable JBPDS configuration as an integrated system. Alternatively, disassembled JBPDS components would be tested simultaneously in existing chambers. For each of the four BWA agent classifications (spore bacteria, vegetative bacteria, viruses, and toxins), WSLAT proposes to test both live/active and killed/inactive BWAs in the chamber along with both live/active and killed/inactive ALOs and live/active and killed/inactive simulants. Then killed/inactive ALOs and both live/active and killed/inactive simulants will be tested outdoors in the ABT and in open field environments. ACPLA measurement will be systematically addressed, but this paper will not cover the ACPLA issue in detail. In addition, the effects of particle size and cloud duration may be investigated.

Analysis Construct and Proof of Principle.

Although no systematic study has yet been done to link performance of the JBPDS in the chamber with actual BWAs to performance of the JBPDS in outdoor environments with live/active simulants and killed ALOs, limited test data are available to investigate whether such linkages are feasible. In particular, the JBPDS has undergone integrated system testing in the ABT and field with simulants, and the BAWS and the assay strips have been tested separately in a containment chamber with both live/active BWAs and live/active simulants. The following analysis construct and proof of principle exploits existing data to show that WSLAT is feasible from an analytic standpoint.

The existing test data for one agent classification were used to develop a logistic regression model for estimating JBPDS Prob[Alarm] and Prob[ID|Alarm] based on "Test" ("Chamber," "ABT," or "Field"), "Agent" ("BWA" or "Sim"), "Particle Size" ("Larger" or "Smaller"), and concentration (actually LogACPLA = Log_{10} ACPLA). The parameters obtained from that model were used to extrapolate chamber results for BWA to ABT and Field in a reasonable manner. Duration of exposure was not considered at this time due to insufficient data across tests. Particle size data was not available for ID data in the chamber, so particle size was not used as a fitting factor for Prob[ID|Alarm]. Examination of results by particle size for ID|Alarm data from ABT and field tests suggest that particle size has little if any influence on ID. Available particle size data relevant to Alarm were not very good, and they also had minimal effect; particle size was used in the model primarily to illustrate how such a factor might be incorporated.

Logistic regression is the standard statistical technique for modeling a discrete response variable as a function of continuous variables or a combination of continuous variables and discrete predictive factors. In the simplest case where there are only two values of the response variable (e.g., "Alarm" and "No Alarm" or "ID" and "No ID") logistic regression fits Prob[Alarm]=$e^{\mathbf{X}\boldsymbol{b}}/(1+e^{\mathbf{X}\boldsymbol{b}})$ (or Prob[ID|Alarm]=$e^{\mathbf{X}\boldsymbol{b}}/(1+e^{\mathbf{X}\boldsymbol{b}})$) where **X** is a matrix of coefficients and ***b*** is a vector of parameters. This is equivalent to fitting the log-odds ratio (ln{Prob[Alarm]/Prob[No Alarm]} or ln{Prob[ID|Alarm]/Prob[No ID|Alarm]}) as a linear model **X*b***.

After much experimentation with available data, the following linear model was fitted using logistic regression to available JBPDS Alarm data for field trials and ABT trials and the chamber testing

$$\ln(\text{Prob[Alarm]}/\text{Prob[No Alarm]}) = \text{intercept} + t(\text{test}) + a(\text{test,agent}) + p(\text{particle size}) + c(\text{test,agent})*(\text{LogACPLA}-m) \quad (1)$$

where the shift parameter m is actually the overall mean of logACPLA. For simplicity, the intercept, test, and agent parameters for each model were grouped together to give an overall "group" parameter g(test,agent) given by

$$g(test,agent) = intercept + t(test) + a(test,agent) - c(test,agent)*m. \qquad (2)$$

Then the reparameterized model was:

$$\ln(Prob[Alarm]/Prob[No\ Alarm]) = g(test,agent) + p(particle\ size) + c(test,agent)*LogACPLA. \qquad (3)$$

Entertaining the notion that the ratio of the BWA slope for the ABT to the BWA slope for the Chamber should be the same as the ratio of the simulant slope for the ABT to the simulant slope for the Chamber (and similarily for the field) gives the constraints:

$$c(ABT,BWA)/c(Chamber,BWA) = c(ABT,Sim)/c(Chamber,Sim) \qquad (4a)$$

and

$$c(Field,BWA)/c(Chamber,BWA) = c(Field,Sim)/c(Chamber,Sim) \qquad (4b)$$

Extrapolating group parameters for ABT and field tests was not so straightforward, but a simple rationale enabled the desired extrapolation. Let c_{50}(Test,Agent) be the concentration at which Prob[Alarm]=0.5 (estimated from the logistic regression). A reasonable constraint for extrapolation of BWA chamber performance to the ABT is that

$$c_{50}(ABT,BWA) - c_{50}(Chamber,BWA) = c_{50}(ABT,Sim) - c_{50}(Chamber,Sim). \qquad (5a)$$

Likewise, a reasonable constraint for extrapolation of BWA chamber performance to the field is that

$$c_{50}(Field,BWA) - c_{50}(Chamber,BWA) = c_{50}(Field,Sim) - c_{50}(Chamber,Sim). \qquad (5b)$$

Since c_{50}(Test,Agent) occurs when Prob[Alarm]=Prob[No Alarm] (i.e., ln(Prob[Alarm]/Prob[No Alarm]) = 0), it follows from formula (3) that (ignoring the effect of particle size since it does not presently depend on test or agent)

$$c_{50}(Test,Agent) = -g(test,agent)/c(test,agent). \qquad (6)$$

Applying constraints (5a) and (5b) together with formulas (4a), (4b), and (6) gives

$$\begin{aligned} g(ABT,BWA) = {} & g(Chamber,BWA)*c(ABT,Sim)/c(Chamber,Sim) \\ & + g(ABT,Sim)*c(Chamber,BWA)/c(Chamber,Sim) \qquad (7a) \\ & - g(Chamber,Sim)*c(Chamber,BWA)*c(ABT,Sim)/\{c(Chamber,Sim)\}^2 \end{aligned}$$

and

$$\begin{aligned} g(Field,BWA) = {} & g(Chamber,BWA)*c(Field,Sim)/c(Chamber,Sim) \\ & + g(Field,Sim)*c(Chamber,BWA)/c(Chamber,Sim) \qquad (7b) \\ & - g(Chamber,Sim)*c(Chamber,BWA)*c(Field,Sim)/\{c(Chamber,Sim)\}^2. \end{aligned}$$

The notion that slopes and relative positions of c_{50} for BWA potentially vary in accordance with equations (4a), (4b), (5a), and (5b) is speculative at this point. However, the charts in Figure 2 (derived by pasting logistic regression output from the SAS JMP statistical package into a Microsoft Excel workbook and performing the calculations described above) indicate reasonable-looking extrapolations from the available test data. For analysis, ACPLA measurements for Chamber testing were scaled similarily to ACPLA from ABT and Field testing. The extrapolations reflect not only the facts that c_{50}(Chamber,BWA) was about half of c_{50}(Chamber,Sim) and that the simulant curves for ABT and field tests were shifted to the right from the chamber test but also the flattening of curves for ABT and Field tests.

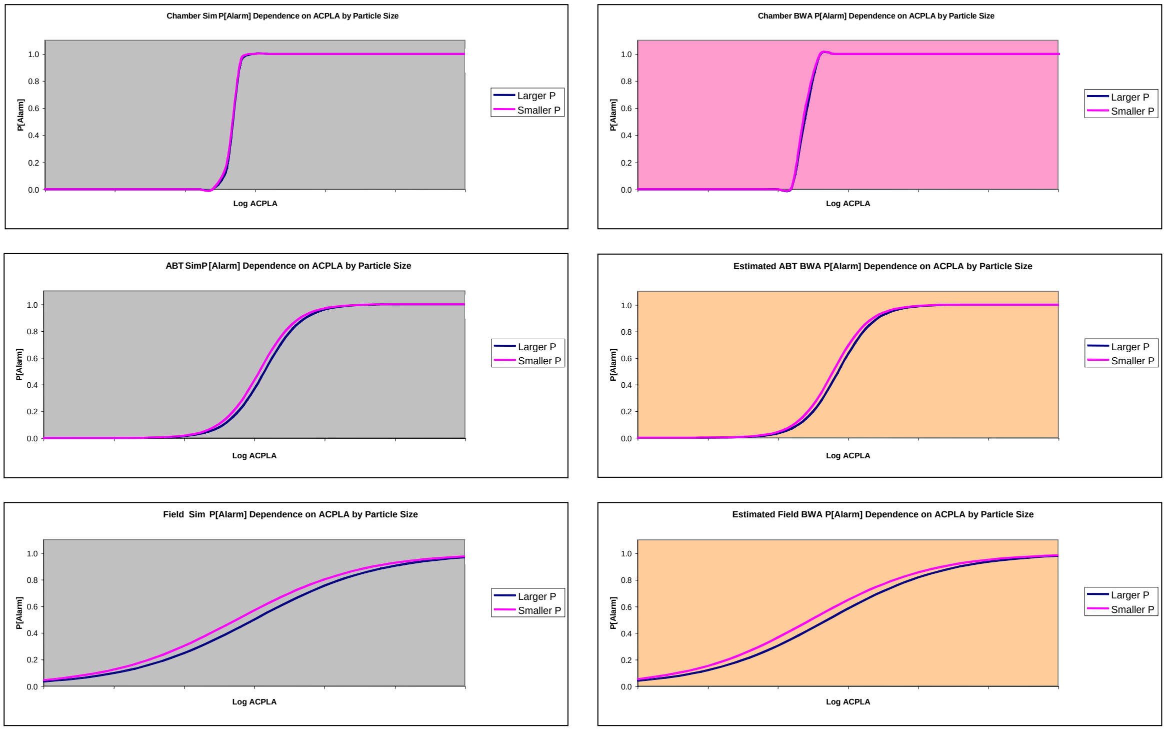


Figure 2. P[Alarm] Dependence on ACPLA by Test Environment and Particle Size
(Grey Background Indicates Simulant, Pink Background Indicates BWA, Orange Background Indicates Extrapolated BWA)

Treatment of ID|Alarm data was very similar to that for the Alarm data. The following linear model (identical to formula (1) for Alarm except that no attempt was made to fit particle size since no particle size data were available for the chamber)

$$\ln(\text{Prob}[\text{ID}|\text{Alarm}]/\text{Prob}[\text{No ID}|\text{Alarm}]) = \text{intercept} + t(\text{test}) + a(\text{test,agent}) + c(\text{test,agent})*(\text{LogACPLA}-m) \quad (8)$$

where as before, the shift parameter m is actually the overall mean of LogACPLA. As with Alarm data, the intercept, test, and agent parameters for each model were grouped together to give an overall "group" parameter g(test,agent) given by

$$g(\text{test,agent})=\text{intercept} + t(\text{test}) + a(\text{test,agent}) - c(\text{test,agent})*m. \quad (9)$$

Then the reparameterized model was:

$$\ln(\text{Prob}[\text{ID}|\text{Alarm}]/\text{Prob}[\text{No ID}|\text{Alarm}]) = g(\text{test,agent}) + c(\text{test,agent})*\text{LogACPLA}. \quad (10)$$

Logistic regression output was again pasted into a Microsoft Excel workbook and equations (4a), (4b), (7a), and (7b) were used to calculate revised and extrapolated parameters. As with Alarm data, the charts from Excel displayed in Figure 3 show that the extrapolations give reasonable results. In chamber ID|Alarm testing, solutions of known concentration were inoculated directly onto the identification media, so that ACPLA was known very precisely and the curve inflections are very sharp. If there had been more uncertainty in chamber ACPLA determination, curves for chamber ID|Alarm data would have been much flatter (as they were in ABT and field tests where ACPLA was determined with more uncertainty). The extrapolations

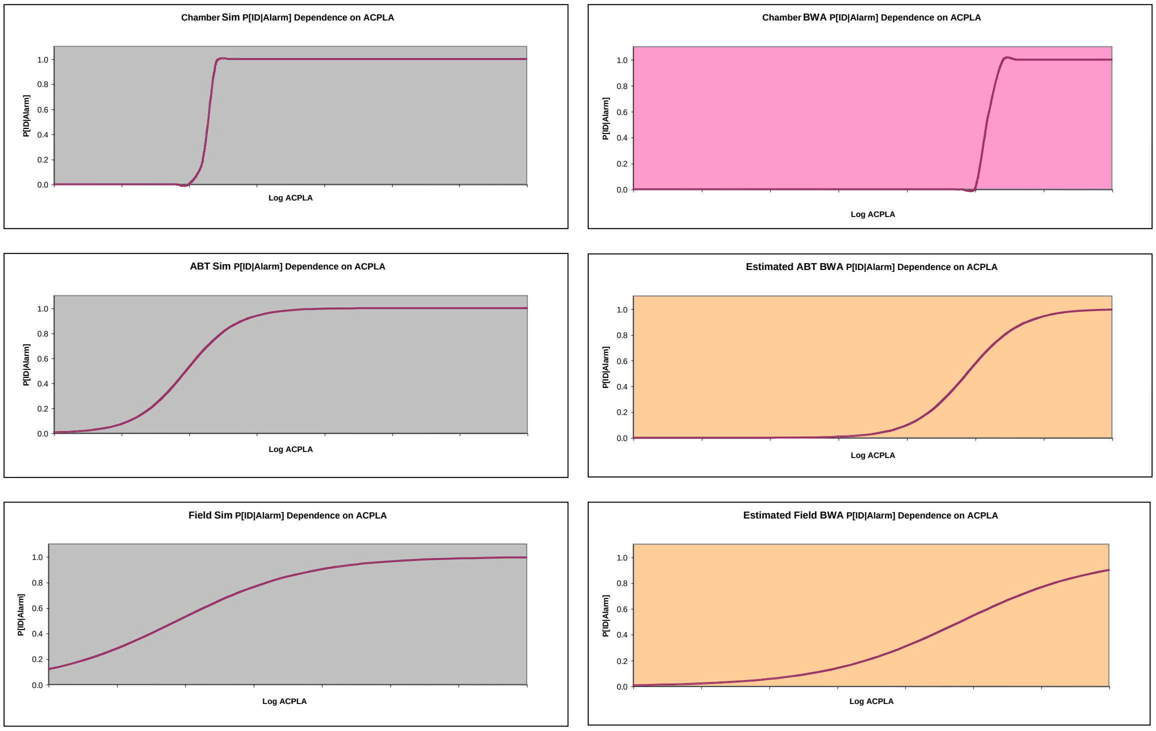


Figure 3. P[ID|Alarm] Dependence on ACPLA by Test Environment
(Grey Background Indicates Simulant, Pink Background Indicates BWA, Orange Background Indicates Extrapolated BWA)

used reasonably reflect flattening in ABT and field tests by shrinking the c(Chamber, agent) parameters. As expected, they also capture the slight shift to the left suggested by simulant data for the ABT and field tests.

Extension to Full WSLAT.

If WSLAT is conducted, it is anticipated that the analysis construct used above will be the starting point for analysis. In addition to possibly having more than one particle size for each Agent, "Shorter" and "Longer" durations are possible for cloud releases, and there will be a "Live/Killed" status factor. In particular, it is anticipated that both live/active and killed/inactive BWAs as well as both live/active and killed/inactive ALOs and simulants will be used in each agent classification (spore bacteria, vegetative bacteria, virus, and toxins). The live/active and killed/inactive BWAs will be released only in WSLAT chambers, but it is anticipated that both live/active and killed/inactive ALOs and simulants will be released in the ABT and the field. This will provide the ability to crosswalk the transformations used to extrapolate current BWA chamber data to ABT and field environments between live/active and killed/inactive simulants, between live/active simulants and killed/inactive ALOs, etc. Examination of these relationships in WSLAT test data may build confidence in the ratio approach or suggest a better approach. It is anticipated that separate fits would be done for each agent classification. Tentative factors and levels for WSLAT are listed in the following table.

Factor	Parameter	Nesting	Levels	Level Labels
Test	t	None	Chamber, ABT, Field	C,A,F
Particle Size	p	None	Larger, Smaller	LP,SP
Duration	d	None	Larger, Smaller	LD,SD
Agent	a	Test	BWA, Sim	B,S
Status	s	Test, Agent	Live/Active, Killed/Inactive	L,K
Concentration	c	Test, Agent	NA (coefficient)	

The initial log-linear model to be entertained for Alarm is:

$$\ln(\text{Prob[Alarm]}/\text{Prob[No Alarm]}) = \text{intercept} + t(\text{test}) + p(\text{particle size}) + d(\text{duration}) + a(\text{test,agent}) + s(\text{test,agent}) + c(\text{test,agent}) * (\text{LogACPLA} - m) \quad (11)$$

and a similar model will be entertained for ID|Alarm. Reparameterization and extrapolation is expected to proceed as it did for currently available data in this analysis construct.

Some Things Economists Know That Just Aren't So*

James R. Thompson
L. Scott Baggett
William C. Wojciechowski
Rice University

This research was supported, in part, by the Army Research Office (Durham) under W911NF-04-1-0354.

It Isn’t Ignorance So Much That Hurts Us. It’s The Things We Know That Just Aren’t So.

Will Rogers

Be most slow to believe what you most want to be true.

Samuel Pepys

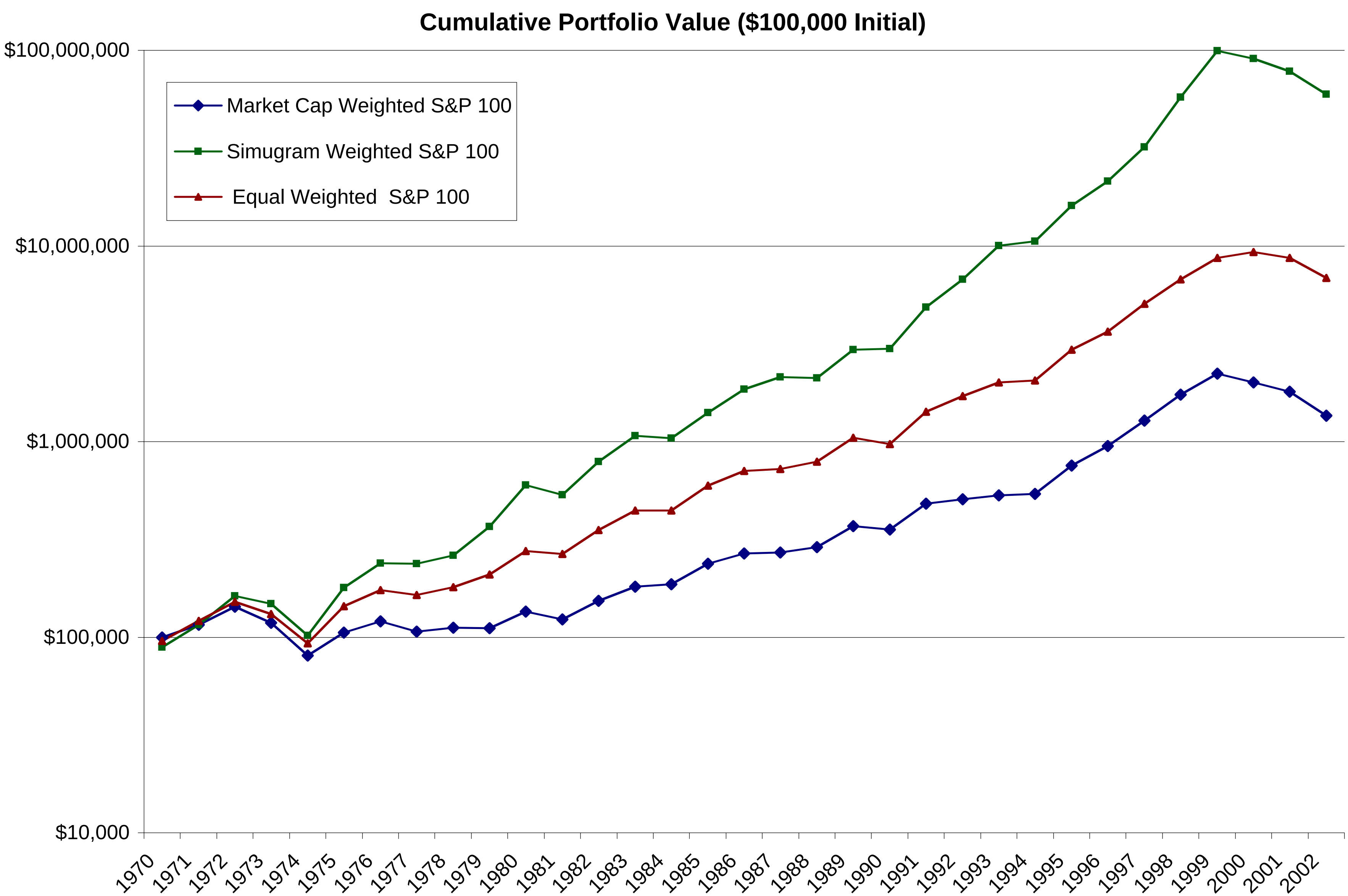
QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Bad News

- Many of the basic models of contemporary computational finance are inconsistent with real world data.
- We do not at present have alternative models to replace these flawed models.

Good News

- In the absence of models there is much that can be done with empirical data based techniques.
- We can, over time, develop models that do work.


Cumulative Portfolio Value ($100,000 Initial)
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
Market Cap Weighted S&P 100
Simugram Weighted S&P 100
Equal Weighted S&P 100
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002

Goals

- To subject some of the accepted models of finance to concordance with real data.
- To come up with practical means for dealing with risk.
- To estimate the long-term aggregate risk of a portfolio strategy, taking into account the history of the securities considered.
- To obtain useful means for portfolio selection. This amounts to solving empirically an ill-posed problem.

Some Lessons From The Bad Old Days

Marxian Memories

Antonio Gramsci: **Indeed in politics the assumption of the law of statistics as an essential law operating of necessity is not only a scientific error, but becomes a practical error of action.**

What is more it favors mental laziness and superficiality …

The situationing of the problem as a search for laws and for constant, regular and uniform lines is connected to a need, conceived in a somewhat puerile and ingenuous way, to resolve in preemptory fashion the practical problem of the predictability of historical events...

Georg Lukács

“If theory does not conform to the facts,then so much the worse for the facts.”

George Orwell

O’Brien to Winston Smith

- “The law of gravity is preposterous. No such law exists. If I think I float and you think I float, then it happens.”
- “If you want a picture of the future, Winston, think of a boot stepping onto a man’s face forever.”

Freedom is the freedom to say that two plus two equal four.
Given that, all else follows.

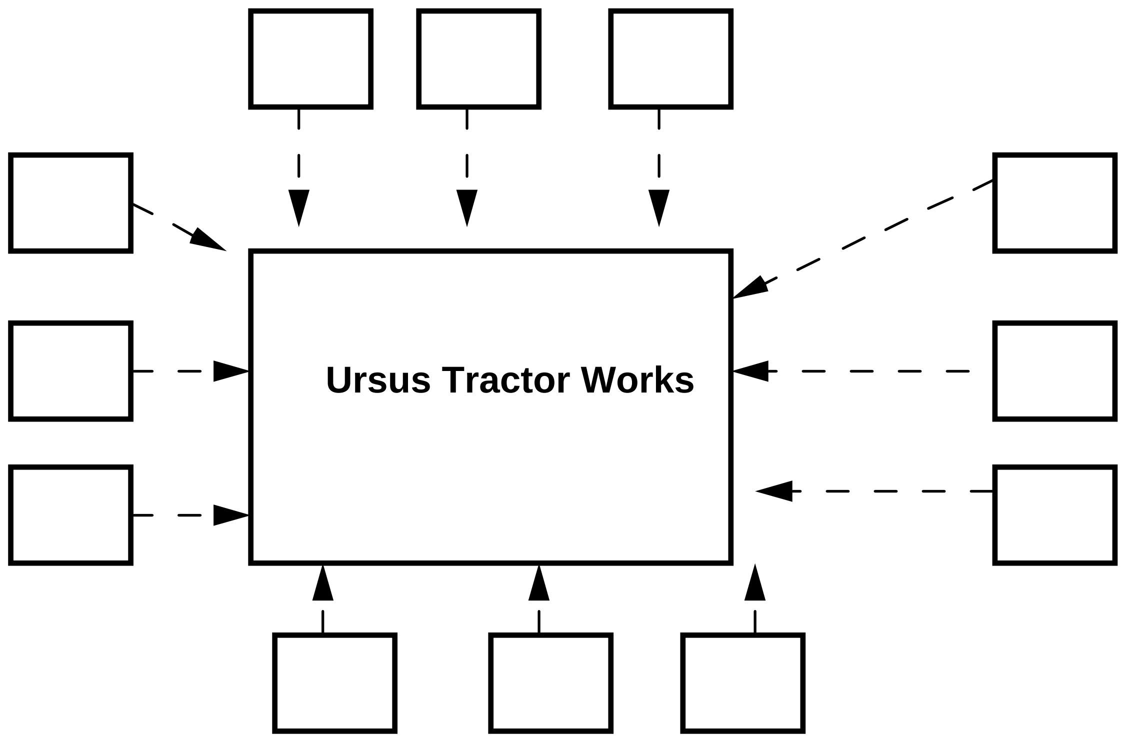
If there is hope, it lies with the proles.

Mikail Ostatny

Fall of Communism June 4, 1989

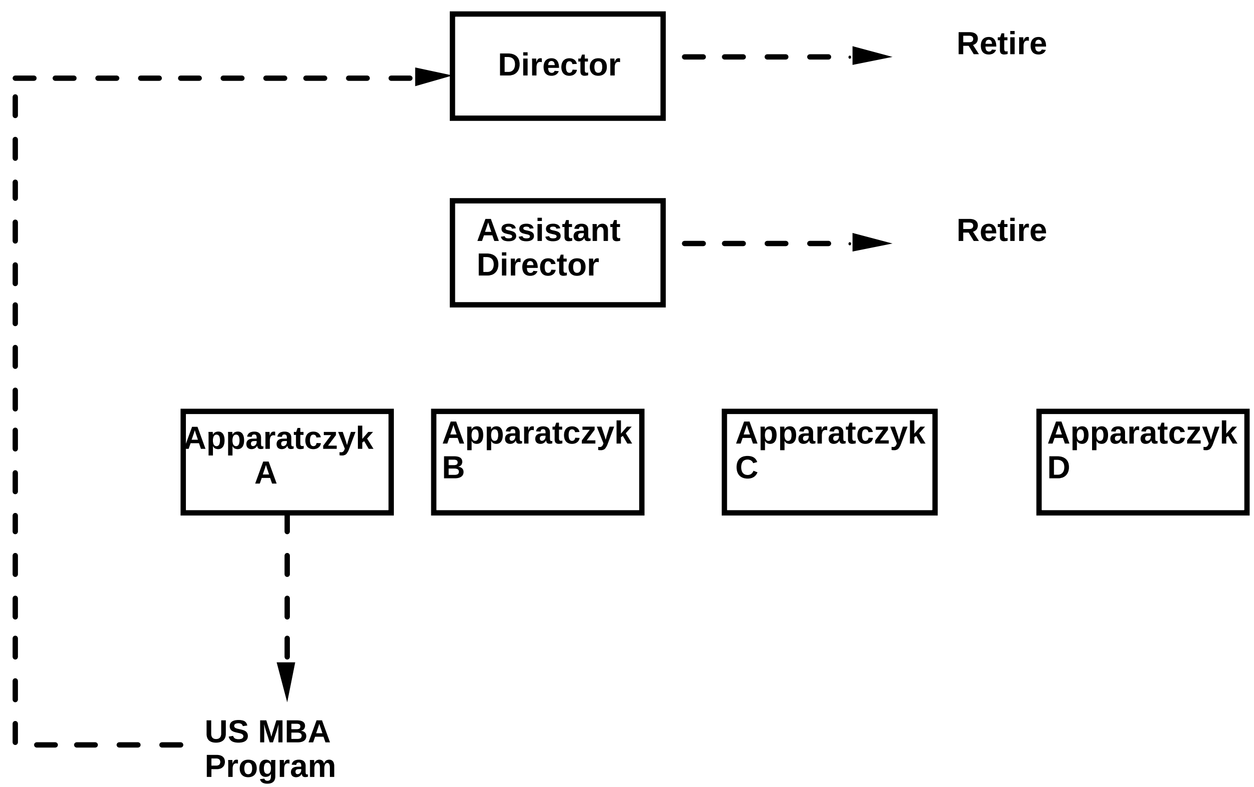
A Poison Pill for Ursus

Marxist Poison Pill (#613)

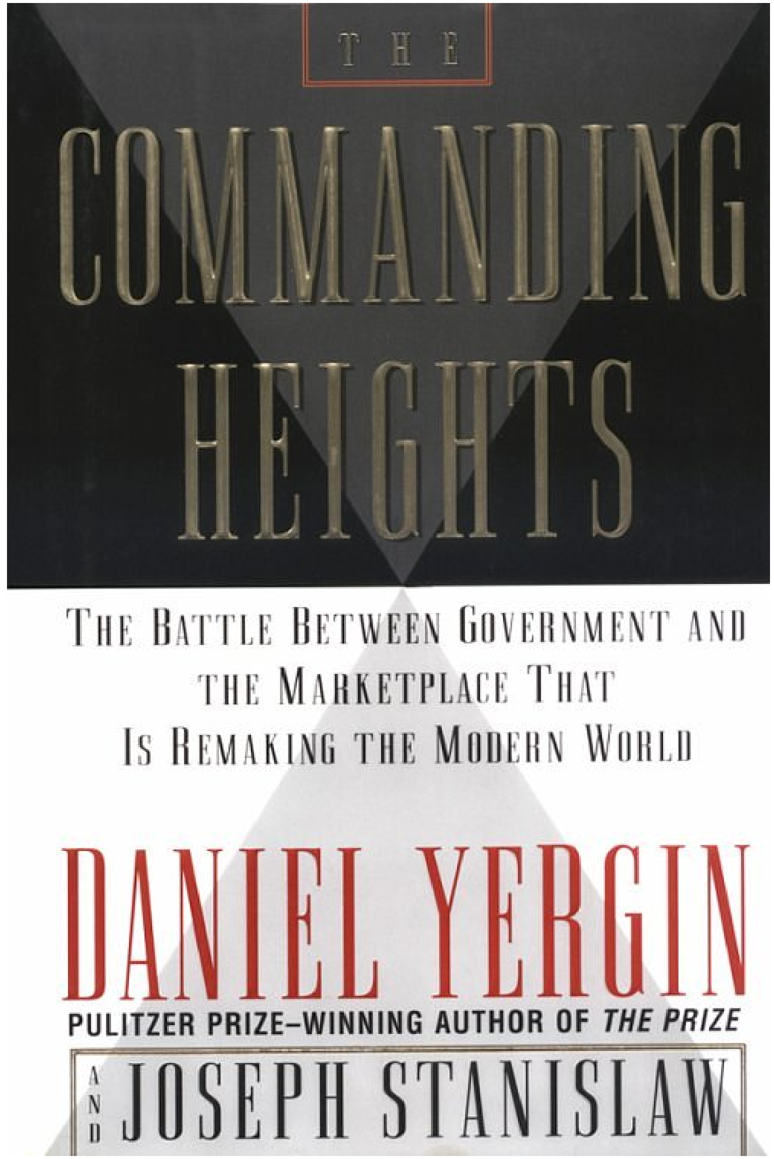


“Help from the USA”

Reorganization of "Elites" Recommended By USA and World Bank "Experts"



Jeffrey Sachs Capitalism Cold Turkey


THE
COMMANDING HEIGHTS
THE BATTLE BETWEEN GOVERNMENT AND THE MARKETPLACE THAT IS REMAKING THE MODERN WORLD
DANIEL YERGIN
PULITZER PRIZE–WINNING AUTHOR OF THE PRIZE
AND JOSEPH STANISLAW

Marxism Makes for Simplicity In Economic Modeling

So Does The Efficient Market Hypothesis

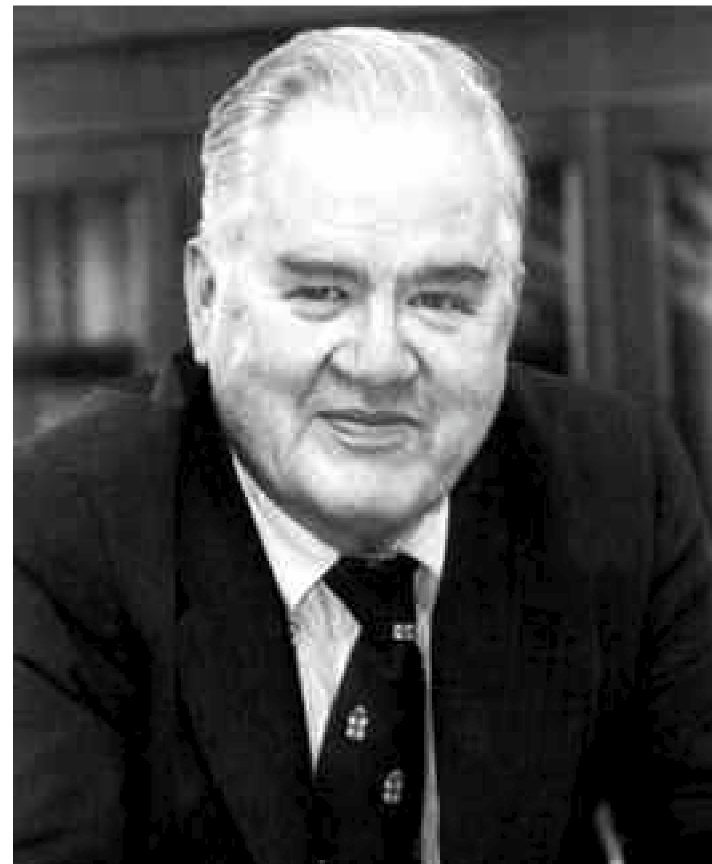
- Martingales abound
- Pricing of options becomes simply another application of the heat equation
- Data analysis is unnecessary, since we know our models are correct, facts notwithstanding

The View In This Paper

The Models Being Flawed, We Need To Turn To The Empiricism of Exploratory Data Analysis

Tukey's Maxim

Far better an approximate answer to the right question, which is often vague,than an exact answer to the wrong question,which can always be made precise.



John Maynard Keynes

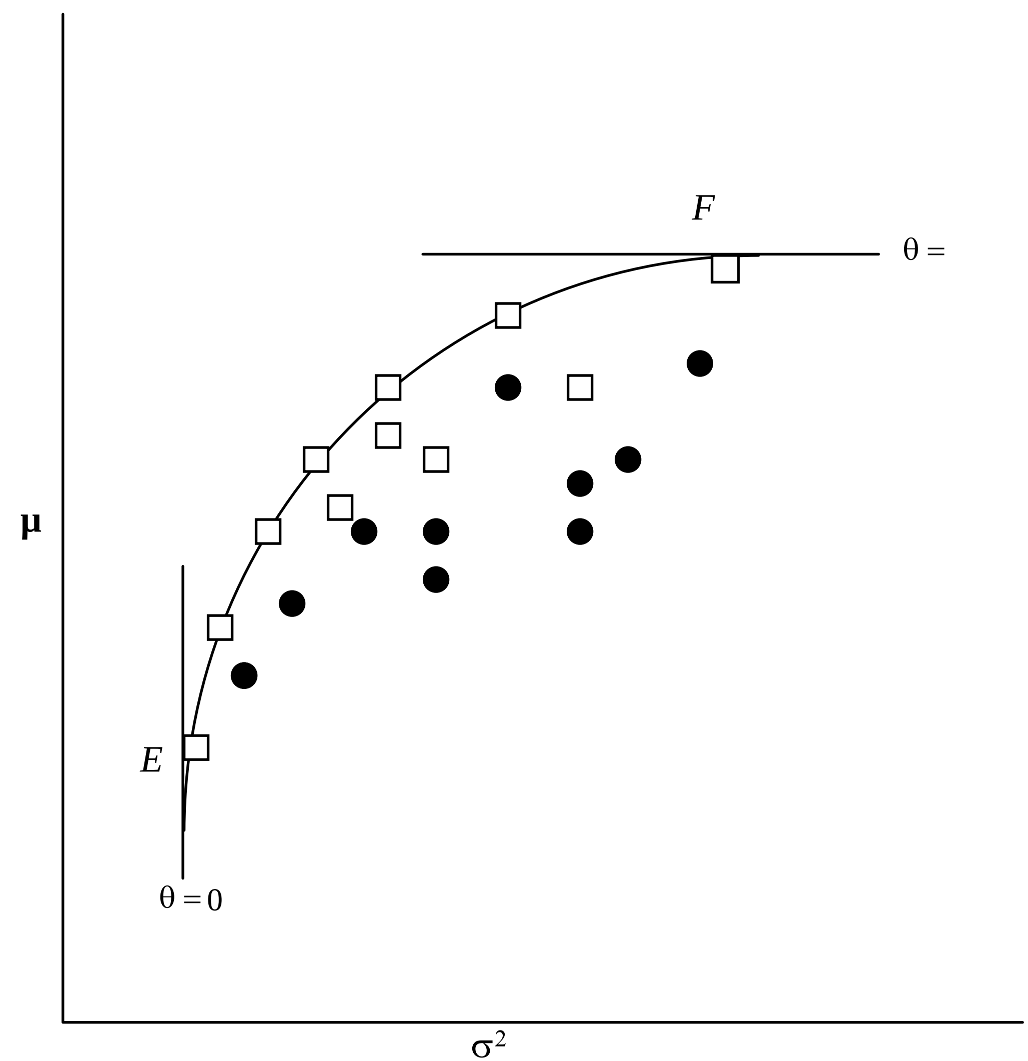
The market can remain irrational longer than you can remain solvent

Some Famous Flawed Models

Harry Markowitz

Fifty years ago, Harry Markowitz posed and solved the following problem:

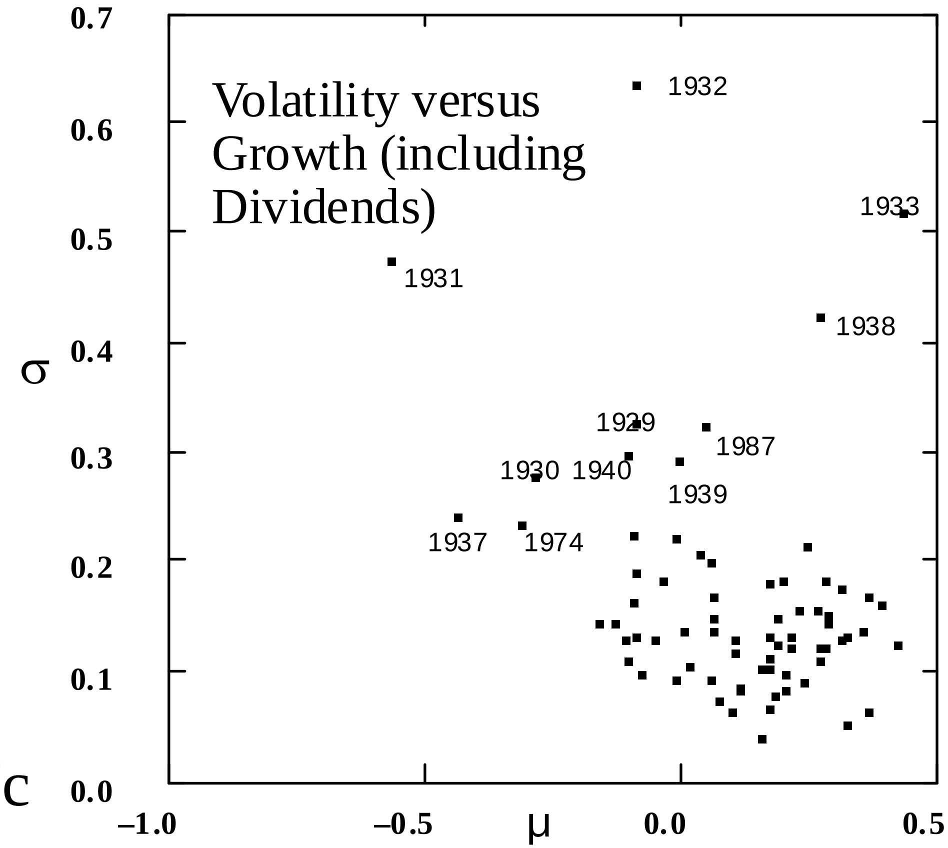
Given a set of n stocks and a capital to be invested of C, what is the allocation of capital which maximizes the expected return, at a future time t, of the portfolio $P(t)$ for an acceptable volatility of the total portfolio $\sigma(t)$?



F
θ =
μ
E
θ = 0
σ2

For this contribution, Markowitz received the Nobel Prize. His result is the foundation of portfolio analysis. However, it is flawed. “Volatility” is the square root of the variance of the value of the portfolio. It is a poor surrogate for risk. The concept of risk is a hard one to grasp. Laurence Siegel, treasurer of the Ford Foundation, defines risk rather forcefully, if imprecisely:

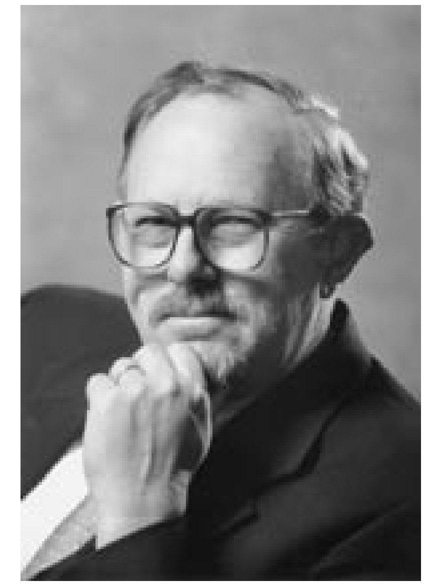
… *risk is the possibility that, in the long run, stock returns will be terrible.*

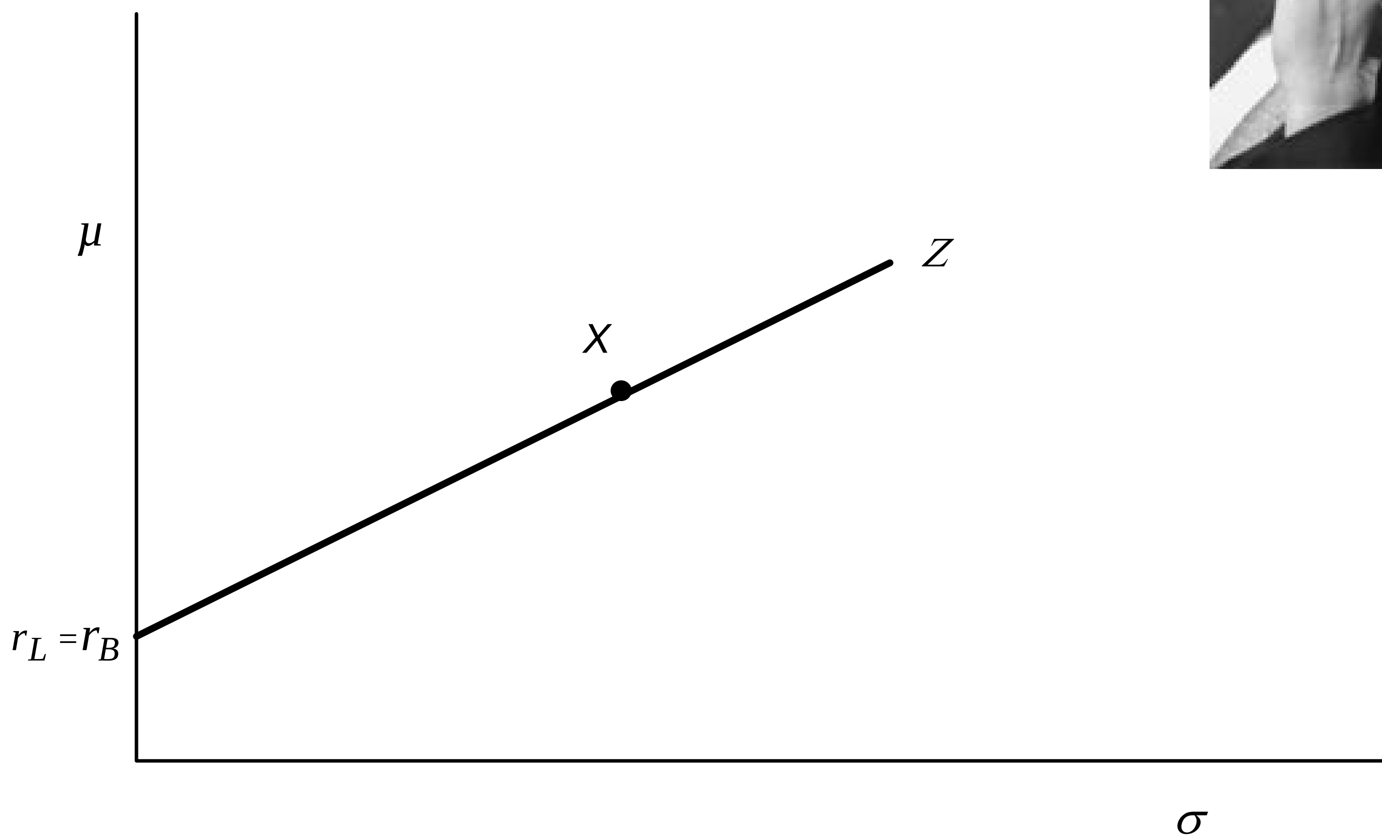


$\rho = -.317c$

Cover of *Models for Investors in Real World Markets*
James R.Thompson, Edward E. Williams & M. Chapman Findlay III

William Sharp (Nobel 1990)
Capital Market Line
CREF
John Bogle Vanguard

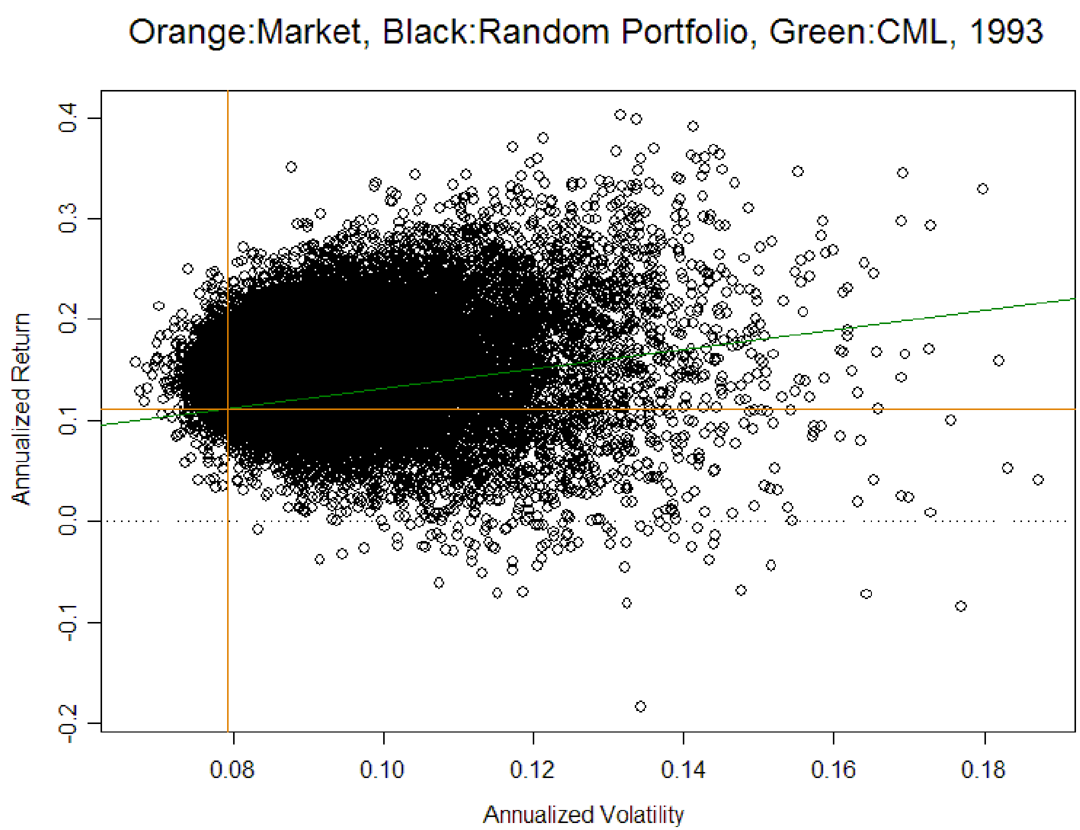




Our Large Scale
Simulation Shows That Since 1960 65% Of The Random Portfolios Beat The Market Cap Weighted Portfolio.

"Market Cap Weighting - Where's the Risk Management?"

William C. Wojciechowski and James R. Thompson (2004)



Orange:Market, Black:Random Portfolio, Green:CML, 1993
Annualized Return
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
0.08
0.10
0.12
0.14
0.16
0.18
Annualized Volatility

Black-Scholes-Merton and Their Amazing Money Machine

Definition: A Call Option is the right (but not the obligation) to buy a security of current price $S(0)$ for strike price X at a future time T.

What is the "fair price" C of such a call option?

Answer: There is no such thing.

Wrong Answer!

How about

$$C = exp(-\mu T)\, E\{Max(0, S(T) - X)\}$$

$$= e^{-\mu T}\left(e^{\mu T} S(0)\Phi\left(\frac{\log(S(0)/X) + (\mu + \sigma^2/2)T}{\sigma\sqrt{T}}\right)\right.$$

$$\left. - X\Phi\left(\frac{\log(S(0)/X) + (\mu - \sigma^2/2)T}{\sigma\sqrt{T}}\right)\right)$$

No. What is wanted is:

$$C_{BS} = e^{-rT}\Big(\; e^{rT} S(0)\Phi\left(\frac{\log(S(0)/X) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}\right)$$

$$-X\Phi\left(\frac{\log(S(0)/X) + (r - \sigma^2/2)T}{\sigma\sqrt{T}}\right) \Big)$$

Transforms a noisy game into a sure thing.

Some Problems with Black-Scholes

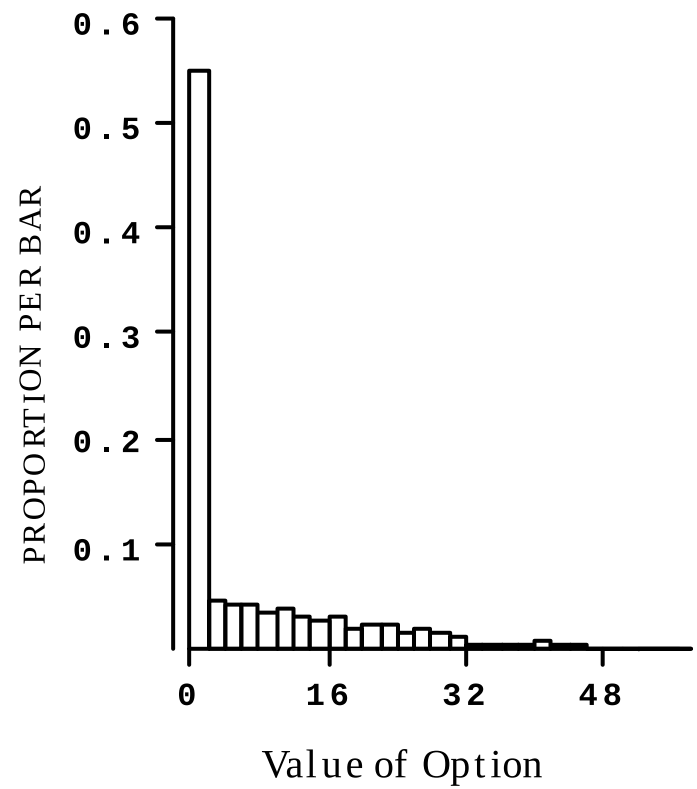
Transaction costs are not really free.
The closer the hedge gets to being riskless, the more frequently one must rebalance (and this results in material transaction costs).

The realistic value of r will be significantly higher than that of a Treasury bill.

Historical records show that the Black-Scholes formula, generally does not give the actual market price of a call option. To correct this imperfection in nature, it is customary for some traders to plug in whatever value is necessary for σ to give the market price for the option.

Moreover, if we look at the same execution time T and two different strike prices, then we generally get two different plug-in estimates for implied volatility.

Looking at expectations and variances does not tell the story.We need to look at the distribution function of the payoffs.



μ=.15; T=0.5; σ=0.20; X=\$108;
S(0) = \$100 E= \$7.23

C=\$3.54 55%of the time V=0.

$$C_{vendor} = e^{-\eta T} \Big(e^{\eta T} S(0) \Phi\left(\frac{\log(S(0)/X) + (\eta + \sigma^2/2)T}{\sigma\sqrt{T}} \right)$$

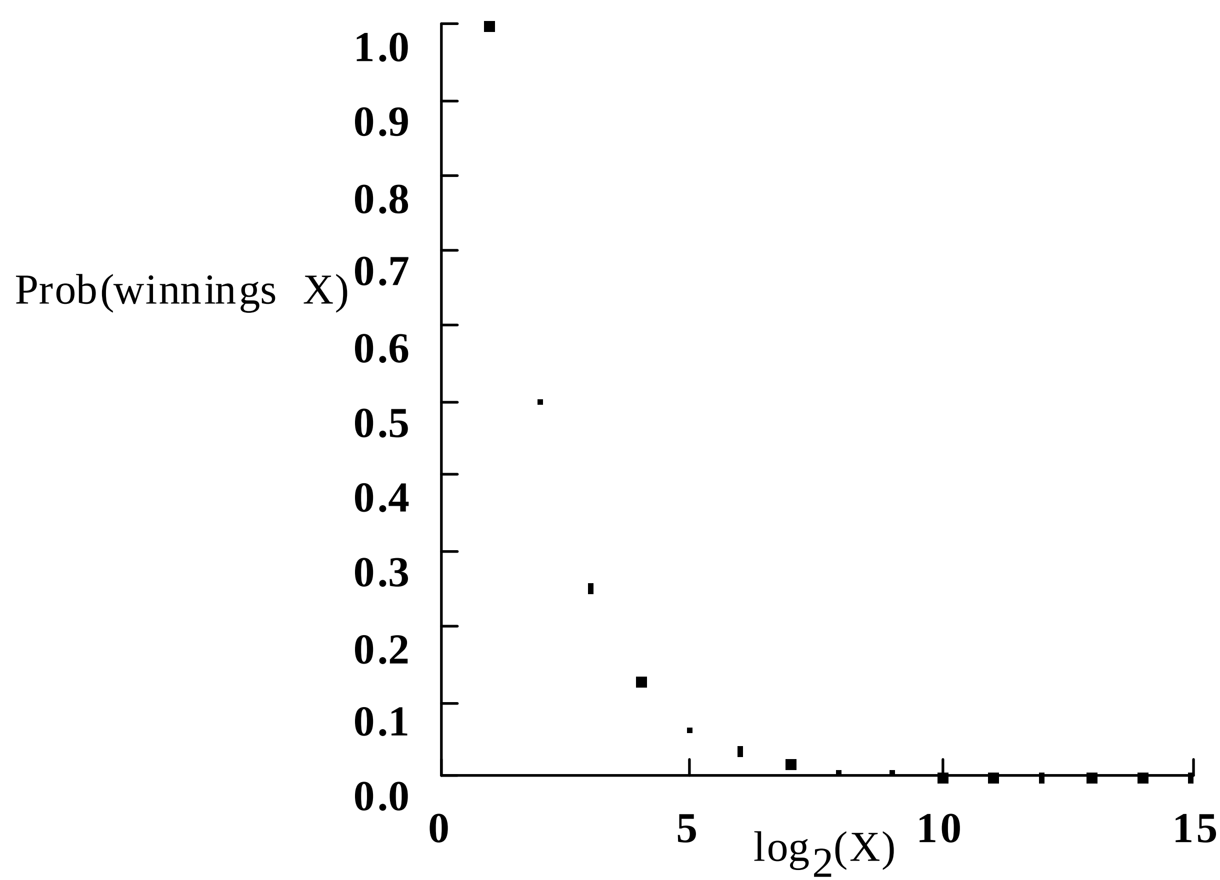
$$- X\Phi\left(\frac{\log(S(0)/X) + (\eta - \sigma^2/2)T}{\sigma\sqrt{T}} \right) \Big)$$

$$C_{buyer} = e^{-\mu T} \Big(e^{\mu T} S(0) \Phi\left(\frac{\log(S(0)/X) + (\mu + \sigma^2/2)T}{\sigma\sqrt{T}} \right)$$

$$- X\Phi\left(\frac{\log(S(0)/X) + (\mu - \sigma^2/2)T}{\sigma\sqrt{T}} \right) \Big)$$

$$\mu > \eta > r$$

There are dangers with maximizing the expectation of payoff.



St. Petersburg Paradox

When Models Fail

In 1998, Alan Greenspan organized a 3.5 billion dollar bailout of the failed LTCM "hedge fund." Long Term Capital Management, like Enron, produced nothing. It simply bought and sold stocks, bonds and derivatives with leveraging aplenty (typically, a "hedge fund" is actually a collection of speculative ventures). It was organized based on the "risk neutral" theories of Black, Scholes and Merton, which theories had been rewarded with the 1997 Nobel Prize in Economics. Indeed, Scholes and Merton were conspicuous advisors (Black was deceased) to LTCM.

Unfortunately, Dr. Greenspan acted like a true believer who, when facts are not in accord with cherished beliefs, fails to use facts to modify theory. He reacted quickly to avert the embarrassment caused by what was supposed to be "a six sigma event." Across America, company chieftains, growing accustomed to cooking their books in order to gain the time necessary for their "risk neutral" approaches to bear fruit, heaved a collective sigh of relief and redoubled their cooking. Indeed, the writing of uncovered options and other dubious business practices expanded after LTCM. Greenspan tried to quell irrational exuberance by raising the prime. This cut off the oxygen to high tech.He then tried to resuscitate the patient by dropping the prime to 1%. Unfortunately, the patient was already dead.

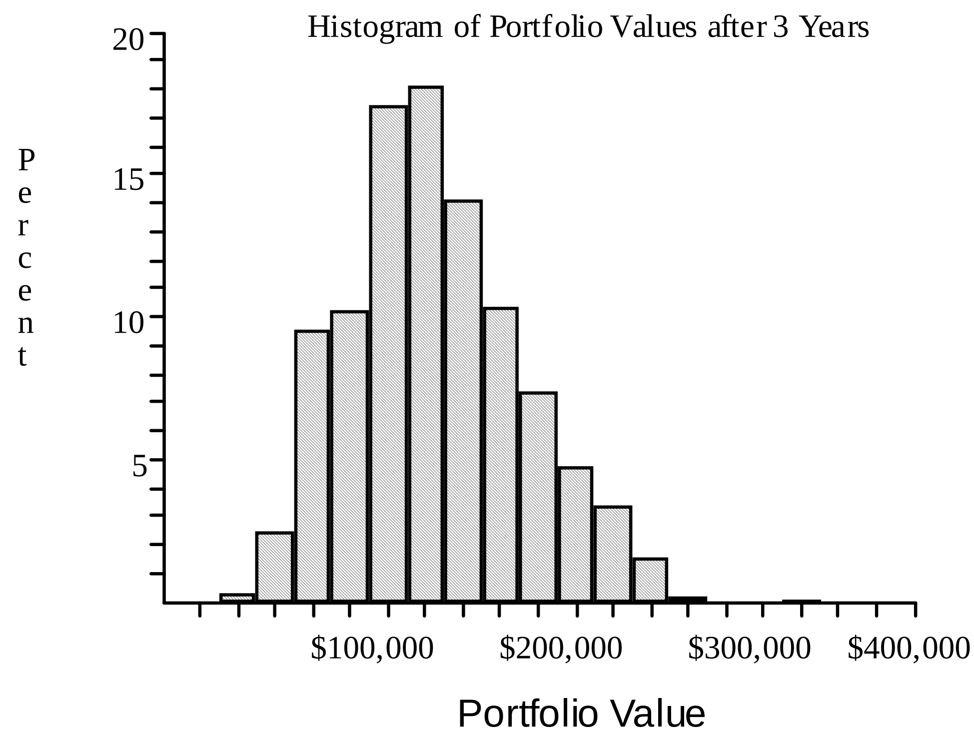
From the standpoint of the dollars involved, the 1998 crash of LTCM (a $3.5 billion dollar bubble) was orders of magnitude less significant than that of the $62 billion Enron debacle in late 2001. The Enron collapse was too large for even Dr. Greenspan to make disappear. Then there is the long list of other companies zapped by belated discovery of their irresponsible accounting practices in 2002 and subsequently. The total wreckage will easily top a hundred times the LTCM figure.

The Simugram™

An Expert System for Forecasting the Probability Distribution of Future Security Prices

The Simugram* Using Ibbotson Index Data

Look at a simugram for an Ibbotson Index Portfolio of initial value $100,00 over a 3 year period, using data using data from 1926-2000.



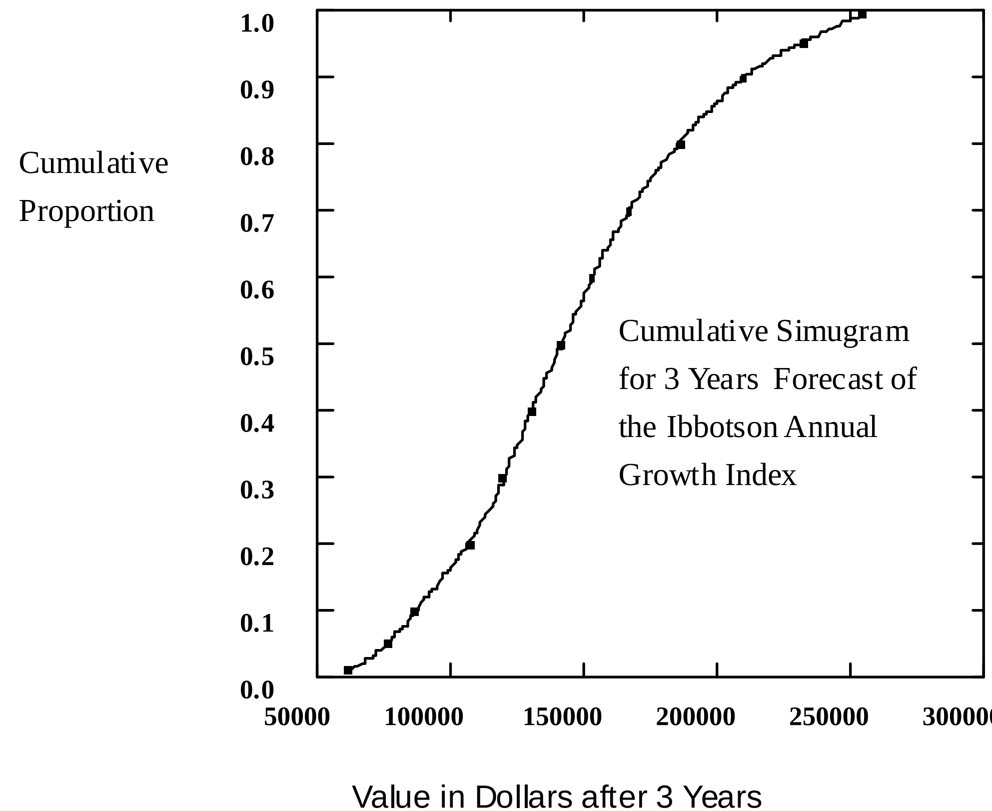
*Copyright and Trademark Granted, Patent Pending

Easier to use is the cumulative simugram* shown below

From this diagram we can note

- that the value of the portfolio is less than $142,000 50% of the time

- and that it is less than $86,000 10 % of the time.

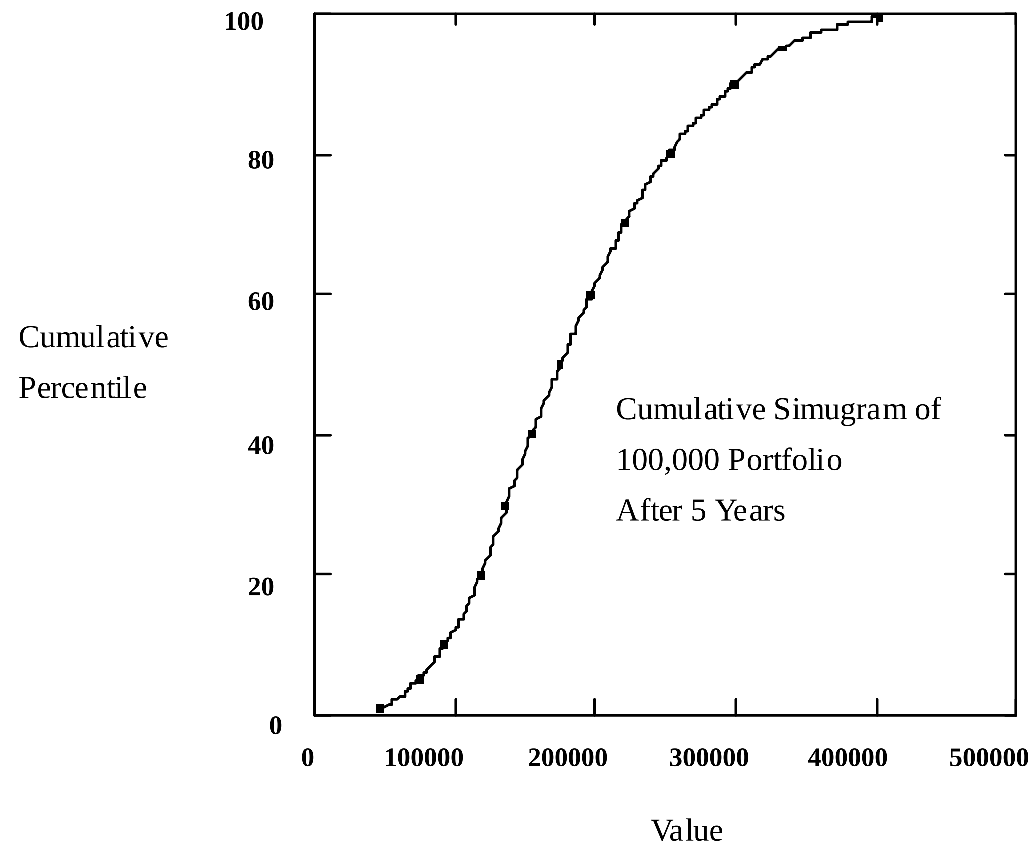


*Copyright and Trademark Granted, Patent Pendin

The mean value of a $100,000 portfolio after five years is $192,676.

The median value is $175,530 (growth rate of .1125).

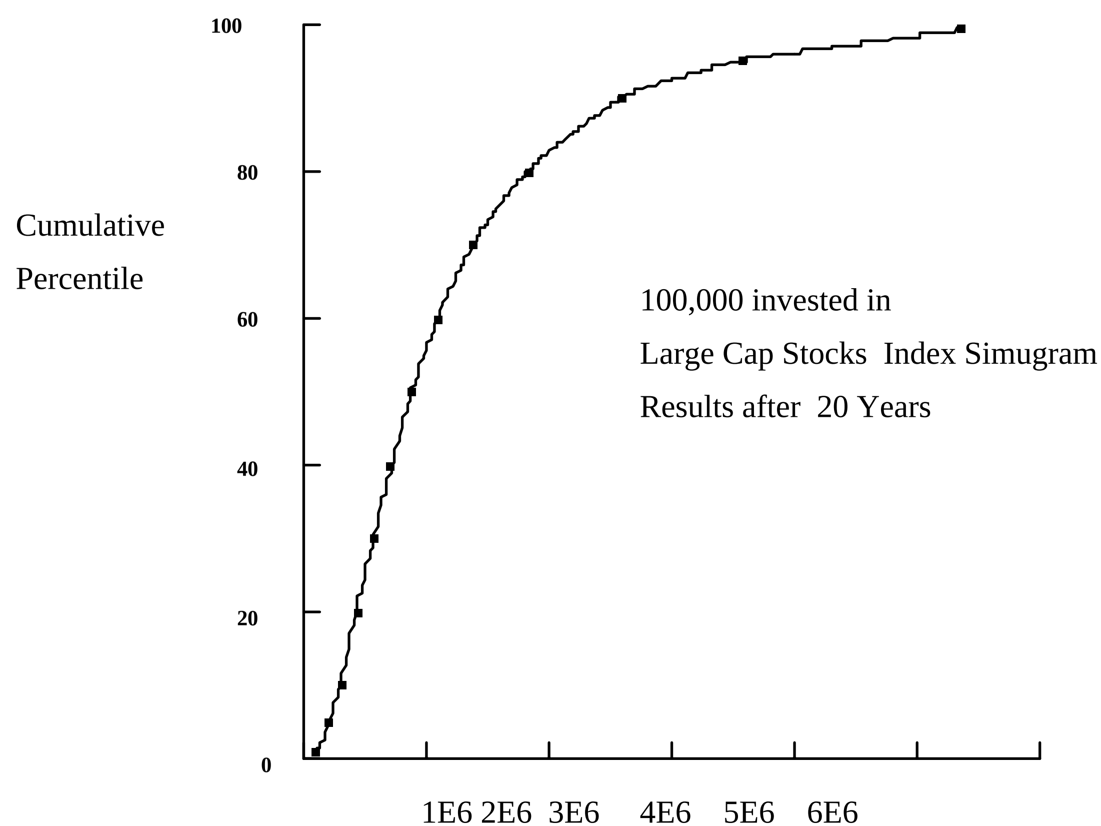
However, the lower ten percentile is $92,747 (growth rate of -.015).



Next, we consider the same scenario except looking 20 years into the future.

The median value is $873,100, an annual increase of 10.8%.

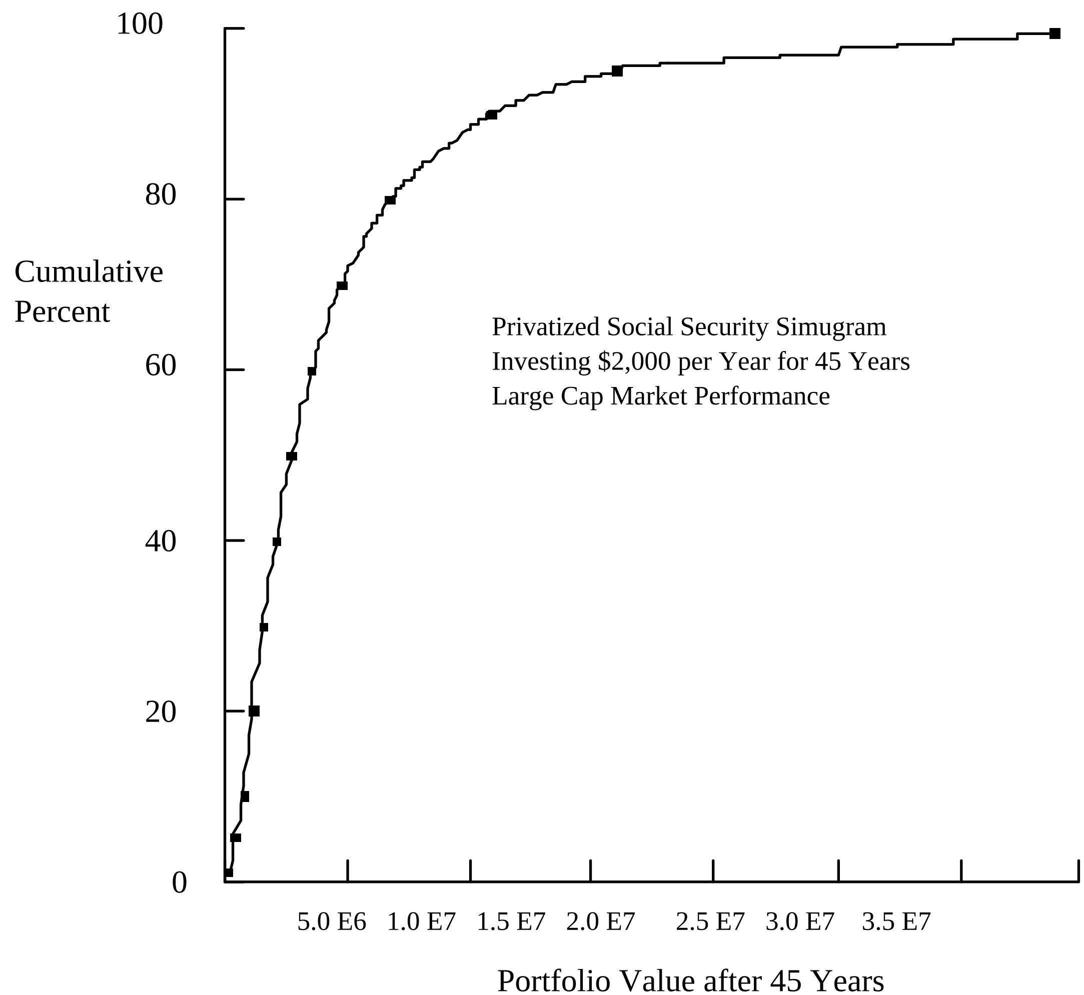
Even the lower ten percentile value of $285,590 represents a growth rate of 5.2%.

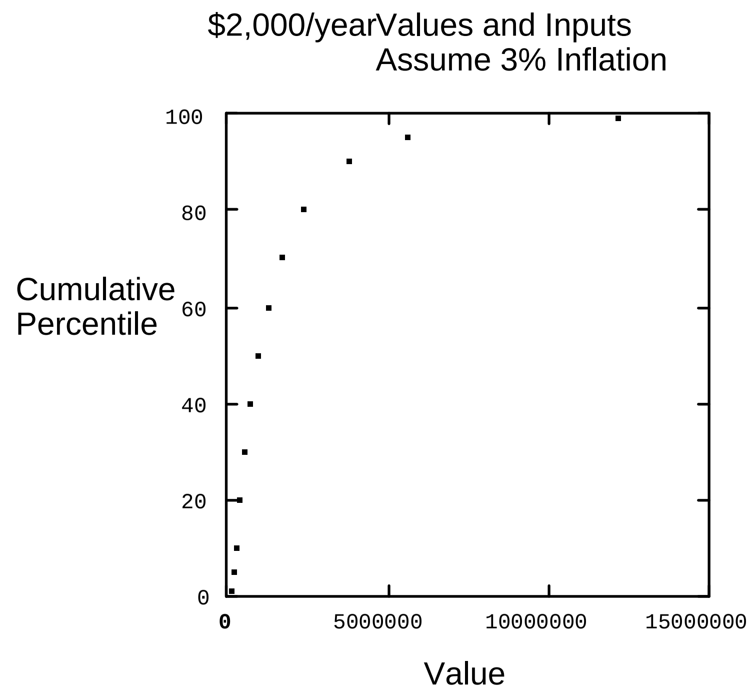


The Stock Market As Casino

- In the case of a casino, over the long haul, the player loses.
- In the case of the US market, over the long haul, the player wins.
- It is rather like a St. Petersburg Trust (see p 45 in Models for Investors in Real World Markets).

Mean = $4,850,000
Median = $2,725,000
Lower Ten Percentile = $695,000





Lower 10 Percentile = $269,000

Median = $993,000

Mean = $1,745,000

A Portfolio Case Study

Next we take the 90 stocks in the S&P 100 that were in business prior to 1991.

Optimal investment allocation maximizing a combination of various simugram percentiles

Constraining each stock to no more than 5% of the portfolio share

Table A.3. Portfolio Allocation from S&P 100. Maximizing One Year 20 Percentile with Max 5% in Any Stock.

id	permno	ticker	$\check{s}$	se	par alloc	npar alloc
1	10104	ORCL	0.44	0.65	0.05	0.05
2	10107	MSFT	0.39	0.48	0.05	0.05
3	10145	HON	0.12	0.44	0.00	0.00
4	10147	EMC	0.48	0.61	0.00	0.01
5	10401	T	0.05	0.40	0.00	0.00
6	10890	UIS	0.36	0.68	0.00	0.01
7	11308	KO	0.07	0.31	0.00	0.00
8	11703	DD	0.05	0.28	0.00	0.00
9	11754	EK	* 0.11	0.35	0.00	0.00
10	11850	XOM	0.12	0.17	0.00	0.00
11	12052	GD	0.19	0.25	0.05	0.05
12	12060	GE	0.23	0.26	0.00	0.00
13	12079	GM	0.08	0.36	0.00	0.00
14	12490	IBM	0.32	0.34	0.05	0.00
15	13100	MAY	0.07	0.29	0.00	0.00
16	13856	PEP	0.13	0.28	0.00	0.00
17	13901	MO	0.13	0.32	0.00	0.00
18	14008	AMGN	0.32	0.39	0.05	0.05
19	14277	SLB	0.12	0.36	0.00	0.00
20	14322	S	0.05	0.36	0.00	0.00
21	15560	RSH	0.27	0.49	0.05	0.05
22	15579	TXN	0.40	0.56	0.00	0.01
23	16424	G	0.09	0.33	0.00	0.00
24	17830	UTX	0.21	0.35	0.00	0.00
25	18163	PG	0.16	0.31	0.04	0.00
26	18382	PHA	0.12	0.30	0.00	0.00
27	18411	SO	0.14	0.24	0.05	0.05
28	18729	CL	0.25	0.33	0.05	0.05
29	19393	BMY	0.20	0.26	0.01	0.03
30	19561	BA	0.05	0.35	0.00	0.00
31	20220	BDK	0.06	0.38	0.00	0.00
32	20626	DOW	0.07	0.30	0.00	0.00
33	21573	IP	0.07	0.36	0.00	0.00
34	21776	EXC	0.17	0.33	0.05	0.05
35	21936	PFE	0.26	0.28	0.05	0.05

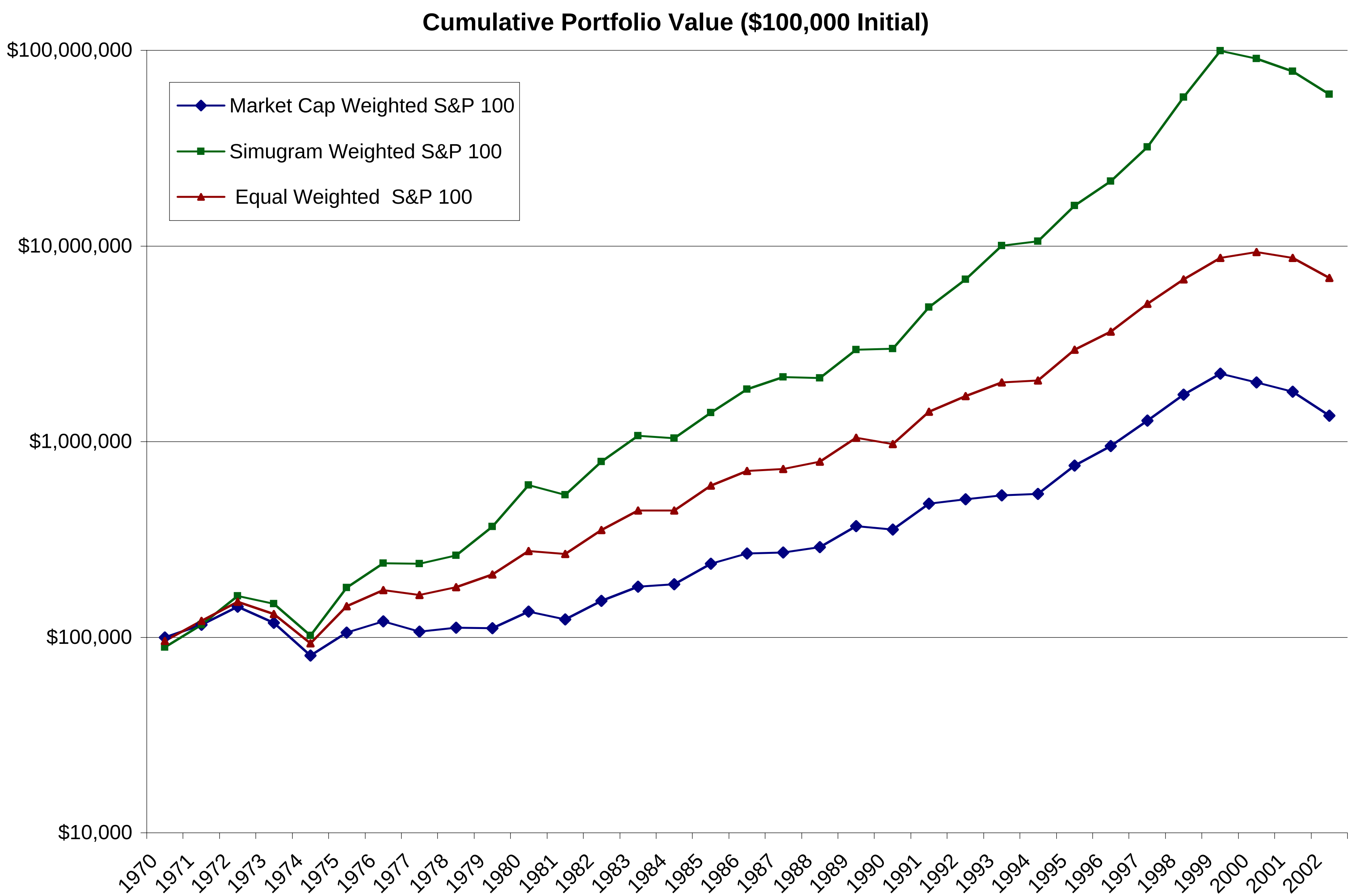
Simugram Systems™
Patent Pending

Table A.3. Portfolio Allocation from S&P 100. Maximizing One-Year 20 Percentile with Max 5% in Any Stock (continued).

id	permno	ticker	„		par alloc	npar alloc
36	22111	JNJ	0.20	0.26	0.00	0.02
37	22592	MMM	0.14	0.25	0.00	0.00
38	22752	MRK	0.17	0.31	0.00	0.00
39	22840	SLE	0.11	0.31	0.00	0.00
40	23077	HNZ	0.06	0.25	0.00	0.00
41	23819	HAL	¡ 0.03	0.47	0.00	0.00
42	24010	ETR	0.11	0.29	0.00	0.01
43	24046	CCU	0.27	0.38	0.00	0.00
44	24109	AEP	0.04	0.22	0.00	0.00
45	24643	AA	0.21	0.37	0.00	0.00
46	24942	R TN	0.02	0.44	0.00	0.00
47	25320	CPB	0.04	0.29	0.00	0.00
48	26112	D AL	0.00	0.33	0.00	0.00
49	26403	DIS	0.05	0.33	0.00	0.00
50	27828	HWP	0.11	0.48	0.00	0.00
51	27887	BAX	0.21	0.24	0.05	0.05
52	27983	XRX	0.04	0.61	0.00	0.00
53	38156	WMB	0.14	0.33	0.00	0.00
54	38703	WF C	0.20	0.31	0.00	0.00
55	39917	WY	0.07	0.32	0.00	0.00
56	40125	CSC	0.17	0.48	0.00	0.00
57	40416	A VP	0.24	0.47	0.00	0.00
58	42024	BCC	0.00	0.33	0.00	0.00
59	43123	A TI	¡ 0.05	0.39	0.00	0.00
60	43449	MCD	0.05	0.26	0.00	0.00
61	45356	TYC	0.37	0.33	0.05	0.05
62	47896	JPM	0.15	0.38	0.00	0.00
63	50227	BNI	0.03	0.27	0.00	0.00
64	51377	NSM	0.35	0.69	0.05	0.05
65	52919	MER	0.31	0.44	0.00	0.01
66	55976	WMT	0.32	0.31	0.05	0.05
67	58640	NT	0.23	0.64	0.00	0.00
68	59176	AXP	0.19	0.31	0.00	0.00
69	59184	BUD	0.20	0.21	0.05	0.05
70	59328	INTC	0.37	0.52	0.00	0.00
71	59408	BA C	0.14	0.34	0.00	0.00
72	60097	MDT	0.28	0.28	0.05	0.05
73	60628	FD X	0.23	0.35	0.00	0.00
74	61065	TO Y	0.06	0.48	0.00	0.00
75	64186	CI	0.20	0.28	0.00	0.00
76	64282	LTD	0.16	0.42	0.00	0.00
77	64311	NSC	¡ 0.02	0.34	0.00	0.00
78	65138	ONE	0.10	0.37	0.00	0.00
79	65875	VZ	0.11	0.29	0.00	0.00
80	66093	SBC	0.12	0.29	0.00	0.00
81	66157	USB	0.12	0.37	0.00	0.00
82	66181	HD	0.33	0.32	0.05	0.05
83	66800	AIG	0.26	0.26	0.05	0.05
84	69032	MWD	0.34	0.47	0.00	0.00
85	70519	C	0.35	0.36	0.03	0.05
86	75034	BHI	0.11	0.41	0.00	0.00
87	75104	VIA	0.21	0.37	0.00	0.00
88	76090	HET	0.09	0.39	0.00	0.00
89	82775	HIG	0.24	0.37	0.01	0.00
90	83332	LU	0.10	0.54	0.00	0.00

Next We Examine Results
Using A Fixed (over 33 years)
Propietary Simugram* Strategy

*Copyright and Trademark Granted, Patent Pending


Cumulative Portfolio Value ($100,000 Initial)
Market Cap Weighted S&P 100
Simugram Weighted S&P 100
Equal Weighted S&P 100
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002

Some Summary Statistics 1970–2002

Type of Fund	Annualized Return	Total Downside Loss
S&P 100 Index	8.2%	118.13%
S&P 100 Equal Weight	13.2%	90.57%
S&P 100 Simugram Weight	20.00%	112.74%.

Conclusions

1. Financial analysis is in a primitive stage of development.

2. We should focus on EDA rather than on simplistic models.

3. Looking at the mean and variance is not enough.

4. Our risk analysis should be of higher dimensionality.

5. The main weapons of the investor are diversification and time.

6. We can construct computer intensive forecasting paradigms which enable us readily and intuitively to consider questions of risk and growth simultaneously.

Smoothed Residual Based Goodness-of-Fit Statistics for Logistic Hierarchical Regression Models

Rodney X. Sturdivant
Department of Mathematics Sciences
United States Military Academy, West Point, New York

David W. Hosmer, Jr.
Department of Biostatistics and Epidemiology
University of Massachusetts – Amherst, Amherst, MA

1 Executive Summary

We extend goodness-of-fit measures used in the standard logistic setting to the hierarchical case. We develop theoretical asymptotic distributions for a number of statistics using residuals at the lowest level. Using simulation studies we examine the performance of statistics extended from the standard logistic regression setting: the Unweighted Sums of Squares (USS), Pearson residual and Hosmer-Lemeshow statistics. Our results suggest such statistics do not offer reasonable performance in the hierarchical logistic model in terms of Type I error rates. We also develop Kernel smoothed versions of the statistics and apply a bias correction method to the USS and Pearson statistics. Our simulations demonstrate satisfactory performance of the Kernel smoothed USS statistic, using Type I error rates, in small sample settings. Finally, we discuss issues of bandwidth selection for using our proposed statistic in practice.

2 Introduction

The logistic regression model is a widely used and accepted method of analyzing data with binary outcome variables. The standard logistic model does not easily address the situation, common in practice, in which the data is clustered or has a natural hierarchy. . For example, in education students are grouped by teachers, schools and districts. In medicine, patients may have the same doctor or use the same clinic or hospital. In recent years, statistical research has led to development of models that explicitly account for the hierarchical nature of the data. Many of these models are now available in commonly used software packages. While use of the models has increased, the development of methods to assess model adequacy and fit has not been commensurate with their popularity.

3 The Hierarchical Logistic Regression Model

The standard logistic approach models the probability that Y takes on the value one, denoted $\pi = \Pr(Y = 1)$. For simplicity, first consider the case where there are two levels in the hierarchy. Further, suppose in this situation there is a single predictor variable. Ignoring the second level, the standard logistic regression model is:

$$Y_{ij} = \pi_{ij} + \varepsilon_{ij},$$

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{ij}, \qquad (1)$$

where $i = 1,\ldots,n_j$ is the subject or level one indicator and $j = 1,\ldots,J$ is the group or level two indicator. The model assumes the distribution of the outcome variable is binomial: $Y_{ij} \sim \mathrm{B}(1,\pi_{ij})$. The standard assumptions about the error structure are then that the errors are independent with moments:

put in text line $\mathrm{E}(\varepsilon_{ij}) = 0$ and $\mathrm{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2 = \pi_{ij}(1-\pi_{ij})$.

.

The hierarchical logistic regression model accounts for the structure of the data by introducing random effects to model (1). In this case, with two levels, we might suppose that either or both coefficients (intercept and slope of the linear logit expression) vary randomly across level two groups. Assuming both are random the hierarchical logistic model is written:

$$\mathrm{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + \beta_{1j}x_{ij}, \qquad (2)$$

with $\beta_{0j} = \beta_0 + \mu_{0j}$, and $\beta_{1j} = \beta_1 + \mu_{1j}$. The random effects are typically assumed to have a normal distribution so that $\mu_{0j} \sim \mathrm{N}(0,\sigma_0^2)$ and $\mu_{1j} \sim \mathrm{N}(0,\sigma_1^2)$. Further, the random effects need not be uncorrelated so we have $\mathrm{Cov}(\mu_{0j},\mu_{1j}) = \sigma_{01}$. Assumptions about ε_{ij} (the level one errors) remain the same as in the standard logistic model.

Substituting the random effects into expression (2) and rearranging terms, the model is:

$$\mathrm{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = (\beta_0 + \beta_1 x_{ij}) + (\mu_{0j} + \mu_{1j}x_{ij}). \qquad (3)$$

In this version of the model, we see a separation of fixed and random components which suggests a general matrix expression for the hierarchical logistic regression model given by:

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\pi} + \boldsymbol{\varepsilon} \\ \mathrm{logit}(\boldsymbol{\pi}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}, \end{aligned} \qquad (4)$$

where $\mathbf{y}$ is an $N \times 1$ vector of the binary outcomes; $\boldsymbol{\pi}$ the vector of probabilities; $\mathbf{X}$ is a design matrix for the fixed effects; and $\boldsymbol{\beta}$ a $p \times 1$ vector a parameters for the fixed portion of the model. The level one errors have mean zero and variance given by the diagonal matrix of binomial variances:

$$\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{W} = \mathrm{diag}[\pi_{ij}(1-\pi_{ij})].$$

Choppy These quantities, then, are the same as in standard logistic models. The quantities added to the model to introduce the random effects are the design matrix for the random effects, $\mathbf{Z}$, and the vector of random parameters, $\boldsymbol{\mu}$. This latter vector has assumed distribution $\boldsymbol{\mu} \sim N(\mathbf{0},\boldsymbol{\Omega})$ with a block diagonal covariance matrix.

Section on estimation

Several methods are available for estimating the parameters of this model. By conditioning on the random effects and then integrating them out, an expression for the maximum likelihood estimates is available. This integral is difficult to evaluate, but recently estimation techniques using numerical integration, such as adaptive Gaussian quadrature, have been implemented in software packages. This method is computational intensive and suffers from instability. In some packages, the ability to handle larger models is lacking.

Don't expand on EM

A second closely related method uses the E-M algorithm to maximize the conditional likelihood function [1]. In this case, the random effects are treated as "missing data" and the algorithm, in the "E" (Expectation) step estimates these parameters by obtaining their conditional (on the data and current estimates of the fixed parameters) expected values. Then, with random parameter estimates in place, the "M" or Maximization step is invoked in which standard generalized least squares (GLS) estimates of the fixed parameters are calculated. The algorithm alternates between E and M steps until some convergence criteria is met. The E-M algorithm also involves heavy computation and is not available in most commercial software packages.

Bayesian methods of estimation have increased in popularity although they have not been implemented in the more popular software packages. Gibbs Sampling [3] and Metropolis-Hastings (M-H) are Markov Chain Monte Carlo simulation techniques [4] typically used to produce parameter estimates under this approach. Again, the techniques involve heavy computation.

The most readily available methods in software packages involve quasi-likelihood estimation [5]. For the logistic hierarchical model the idea is generally to use a Taylor approximation to "linearize" the model. The estimation is then iterative between fixed and random parameters. These procedures suffer from known bias in parameter estimates [6]. However, there are methods to reduce this bias available [7]. Further, the methods are easily implemented and generally converge with less computational effort than other methods. Throughout this study we use the SAS GLMMIX macro which implements a version of quasilikelihood estimation SAS refers to as PL or "pseudo-likelihood" [8].

4 Theoretical Asymptotic Distribution Development of ?

better start of section

Copas [9] proposed the unweighted sum of squares (USS) statistic as a goodness-of-fit measure for the standard logistic model. If consistent number of subscripts y_i is the observed response for the i^{th} subject (here we are only concerned with indexing at level one) and $\hat{\pi}_i$ the model predicted value based on the estimated parameters, the USS statistic is the sum of the squared residuals, $\hat{e}_i = y_i - \hat{\pi}_i$, or:

$$\hat{S} = \hat{\mathbf{e}}^t \hat{\mathbf{e}} = \sum (y_i - \hat{\pi}_i)^2$$

Hosmer et. al. give the asymptotic moments of $\hat{S}$ for the standard logistic case as:

$$\mathrm{E}(\hat{S}) \cong \mathrm{trace}(\mathbf{W})$$

and

$$\mathrm{Var}[\hat{S} - \mathrm{trace}(\mathbf{W})] \cong \mathbf{d'(I - M_1)Wd},$$

where $\mathbf{d}$ is the vector with general element $d_i = 1 - 2\pi_i$, $\mathbf{W}$ is the covariance matrix in standard logistic regression given by $\mathbf{W} = \mathrm{diag}[w_i = \pi_i(1-\pi_i)]$, and $\mathbf{M}_1 = \mathbf{WX(X'WX)}^{-1}\mathbf{X'}$ is the logistic regression version of the "hat" matrix.

Model fit is then assessed forming a standardized version of the statistic for comparison to the standard normal distribution:

$$\frac{\hat{S} - \mathrm{trace}(\hat{\mathbf{W}})}{\sqrt{\widehat{\mathrm{Var}}[\hat{S} - \mathrm{trace}(\hat{\mathbf{W}})]}}.$$

Evans follows a similar procedure to produce a standardized statistic for a logistic 2-level mixed model with random intercept only. Our simulation studies of a version of this statistic in models with random slopes suggest that the theoretical normal distribution under the null hypothesis of a correctly specified model does not hold in smaller samples typically encountered in practice. The statistic itself is inflated due to either large or small observations in the covariance matrix.

le Cessie and van Houwelingen [12] note similar problems with goodness-of-fit measures in certain standard logistic regression settings. They observe a shrinkage effect when considering the approximation for the test statistic. The estimated test statistic may be written as the statistic with true values minus a quantity that is always positive. In order to control the problem in the standard logistic case they use kernel smoothing of Pearson residuals. We similarly develop a USS statistic in the hierarchical logistic model using kernel smoothed residuals.

The smoothed residuals are weighted average of the residuals which controls for issues with extremely large or small values. One can perform kernel smoothing of the residuals in either the "y-space" or "x-space" [10]. In the "x-space" all covariates are used in developing the weights. In the "y-space", the weights are produced using relative distances of the model predicted probabilities of the outcome given by:

$$\hat{\boldsymbol{\pi}} = \begin{pmatrix} \hat{\pi}_1 \\ \vdots \\ \hat{\pi}_n \end{pmatrix}.$$

We use Kernel smoothing of the residuals in the "y-space" in this research. In the standard logistic setting the difference between the two approaches was negligible [10]. The "y-space" smoothing is somewhat simpler and, as demonstrated in the next section, produces reasonable results.

The vector of smoothed residuals is given by:

$$\hat{\mathbf{e}}_{\mathbf{s}} = \boldsymbol{\Lambda}\hat{\mathbf{e}},$$

where $\boldsymbol{\Lambda}$ is the matrix of smoothing weights:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & & \lambda_{1n} \\ & \ddots & \\ \lambda_{n1} & & \lambda_{nn} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \\ \boldsymbol{\lambda}_n \end{bmatrix}.$$

The weights, λ_{ij}, are produced using the kernel density by:

$$\lambda_{ij} = \frac{\mathrm{K}\left(\frac{|\hat{\pi}_i - \hat{\pi}_j|}{h}\right)}{\sum_j \mathrm{K}\left(\frac{|\hat{\pi}_i - \hat{\pi}_j|}{h}\right)}. \tag{5}$$

where $\mathrm{K}(\xi)$ is the Kernel density function and *h* is the bandwidth.

We explore three choices used in other studies for the Kernel density function. The first was the uniform density used in a study of a goodness-of-fit measure in standard logistic regression [12] defined as:

$$\mathrm{K}(\xi) = \begin{cases} 1 \text{ if } |\xi| < 0.5 \\ 0 \text{ otherwise} \end{cases}.$$

A second choice used in standard logistic studies involving smoothing in the "y-space" ([10] and [13]) was the cubic kernel given by:

$$K(\xi) = \begin{cases} 1 - |\xi|^3 \text{ if } |\xi| < 1 \\ 0 \text{ otherwise} \end{cases}.$$

Finally, we tested the Gaussian Kernel density [14] defined:

$$K(\xi) = \tfrac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}\xi^2) \quad .$$

The bandwidth, *h*, controls the number of observations weighted in the case of the uniform and cubic densities. For the Gaussian Kernel, all observations are weighted. However, observations outside of two or three standard deviations of the mean effectively receive zero weight. The bandwidth then determines how many residuals are effectively given zero weight in the Gaussian case.

The choice of Kernel function is considered less critical than that of the bandwidth [15]. There are several methods available (plug-in, cross-validation etc.) for selecting the "optimal" bandwidth. Here we are more concerned with the efficacy of smoothing as an approach. Thus, we examine several bandwidth choices.

Simulations suggest that, using the uniform Kernel in the Pearson statistic for standard logistic models, a bandwidth in which approximately $\sqrt{n}$ of the observations have non-zero weights is best and the weighting too many observations is too conservative [12]. The same criteria worked well with the cubic Kernel in the standard logistic case [10].

Some preliminary work suggested that the use of fewer observations is preferred in the hierarchical setting. Sentence makes no sense ->This is not surprising as the shrinkage effect in the statistic appears even more pronounced. We thus test the bandwidth weighting $\sqrt{n}$ of the residuals for the uniform and cubic kernel for each $\hat{\pi}_i$, as well as smaller bandwidths so that $0.5\sqrt{n}$ or $0.25\sqrt{n}$ of the kernel values were not zero. For the Gaussian kernel, the chosen bandwidth places the selected number of observations within two standard deviations of the mean of the $N(0,1)$ density used in the kernel estimation (somewhat analogous to the other kernels as outside of 2 standard deviations the weights are extremely small in the normal density).

Regardless of the bandwidth criteria, we choose a different bandwidth h_i for each $\hat{\pi}_i$ (in the fashion of Fowlkes, [13]). The weights are then standardized so that they sum to one for each $\hat{\pi}_i$ by dividing by the total weights for the observation as shown in expression (5).

The USS statistic based upon these smoothed residuals is then given by:

$$\hat{S}_s = \sum_{i=1}^{n} \hat{e}_{si}^2 = \hat{\mathbf{e}}_{\mathbf{s}}{}'\hat{\mathbf{e}}_{\mathbf{s}}$$

The distribution (moments?) of this statistic under the null hypothesis that the model is correctly specified is extremely complicated. However, we can produce expressions to approximate the moments of the statistic. We make first approximate the residuals in terms of the level one errors [16]:

$$\hat{\mathbf{e}} \approx (\mathbf{I} - \mathbf{M})\mathbf{e} + \mathbf{g} \ , \qquad (6)$$

where take out of in line equations $\mathbf{M} = \mathbf{WQ}\left[\mathbf{Q}'\mathbf{WQ} + \mathbf{R}\right]^{-1}\mathbf{Q}'$ and $\mathbf{g} = \mathbf{WQ}\left[\mathbf{Q}'\mathbf{WQ} + \mathbf{R}\right]^{-1}\mathbf{R}\boldsymbol{\delta}$. In these expressions, $\mathbf{Q} = \left[\mathbf{X} \quad \mathbf{Z}\right]$ is the design matrix for both fixed and random effects, and $\hat{\boldsymbol{\delta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\mu}} \end{pmatrix}$ the vector of estimated fixed and random effects. The other matrix in the expression involves the estimated random parameter covariances and is defined: $\mathbf{R} = \begin{bmatrix} 0 & 0 \\ 0 & \hat{\boldsymbol{\Omega}}^{-1} \end{bmatrix}$.

Under the null hypothesis of correct model specification, the errors have known moments allowing us to produce the approximate mean and variance for the statistic. We first write the statistic using the approximation of (6):

$$\begin{aligned} \hat{S}_s &= \hat{\mathbf{e}}_{\mathbf{s}}{}'\hat{\mathbf{e}}_{\mathbf{s}} \\ &= \hat{\mathbf{e}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\hat{\mathbf{e}} \\ &\approx [(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + \hat{\mathbf{g}}]'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}[(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + \hat{\mathbf{g}}] \\ &= \mathbf{e}'(\mathbf{I} - \hat{\mathbf{M}})'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + 2\hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + \hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\hat{\mathbf{g}}. \end{aligned}$$

Standard methods to calculate the expected value of a quadratic form (for example, [17]) allow us to express the first moment as:

$$\begin{aligned} \mathrm{E}(\hat{S}_s) &= \mathrm{E}[\mathbf{e}'(\mathbf{I} - \hat{\mathbf{M}})'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + 2\hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + \hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\hat{\mathbf{g}}] \\ &= \mathrm{trace}[(\mathbf{I} - \hat{\mathbf{M}})'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{W}] + \hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\hat{\mathbf{g}} \ . \end{aligned} \tag{7}$$

The variance is expressed as:

$$\begin{aligned} \mathrm{Var}(\hat{S}_s) &= \mathrm{Var}[\mathbf{e}'(\mathbf{I} - \hat{\mathbf{M}})'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + 2\hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + \hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\hat{\mathbf{g}}] \\ &= \mathrm{Var}(\mathbf{e}'\mathbf{A}_4\mathbf{e}) + \mathrm{Var}(\mathbf{b}_4{}'\mathbf{e}) + 2\,\mathrm{Cov}(\mathbf{e}'\mathbf{A}_4\mathbf{e}, \mathbf{b}_4{}'\mathbf{e}) \end{aligned}$$

where no 4: $\mathbf{A}_4 = (\mathbf{I} - \hat{\mathbf{M}})'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})$ and $\mathbf{b}_4{}' = 2\hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})$. To evaluate this expression we use a lesser known result [18] so that the final expression becomes:

$$\begin{aligned} \mathrm{Var}(\hat{S}_s) &= \mathrm{Var}(\mathbf{e}'\mathbf{A}_4\mathbf{e}) + \mathbf{b}_4{}'\hat{\mathbf{W}}\mathbf{b}_4 + 2\,\mathrm{Cov}(\mathbf{e}'\mathbf{A}_4\mathbf{e}, \mathbf{b}_4{}'\mathbf{e}) \\ &= \sum_{i=1}^{n}[a_{4ii}^2 w_i(1 - 6w_i)] + 2\,\mathrm{trace}(\mathbf{A}_4\hat{\mathbf{W}}\mathbf{A}_4\hat{\mathbf{W}}) + \mathbf{b}_4{}'\hat{\mathbf{W}}\mathbf{b}_4 \\ &\quad + 2\sum_i a_{4ii} b_{4i}\pi_i(1 - \pi_i)(1 - 2\pi_i) \ . \end{aligned} \tag{8}$$

The moment expressions are then used to create a standardized statistic:

$$\frac{\hat{S}_s - \mathrm{E}(\hat{S}_s)}{\sqrt{\mathrm{Var}(\hat{S}_s)}}. \tag{9}$$

Under the null hypothesis of correct model fit this statistic has don't use "theory here…reword: theoretical asymptotic standard normal distribution. In order to test model fit, the moments are evaluated using the model estimated quantities where necessary in expressions (7) and (8).

5 Simulation Study Results

The standardized statistic of expression (9) has a dittotheoretical standard normal distribution asymptotically. In a large enough sample, one would expect these statistics to appropriately reject the null hypothesis for a given rejection rate. In a hierarchical model "large enough sample" has two implications. First, the total sample must be large. Further, the number of subjects in each group should also be large. In practice, both conditions may not always be met. Usually the total sample size for hierarchical data is reasonably large, but the cluster sizes might still cause us to question the validity of asymptotic results. We used simulations to examine the performance of the statistics in settings with small sample and cluster sizes likely to occur in practice.

The simulation study consisted of 28 different settings involving four factors: dimension, number of covariates, Intra-class or Intra-cluster Correlation (ICC), and random effects.

The first factor, dimension, involved the number of levels in the hierarchy as well as cluster sizes. Noting that hierarchical models of more than three levels are rare in practice, we defined four levels for this factor:

1A: 2-level model with 20 groups of 20 subjects (400 subjects)
1B: 2-level model with 50 groups of 4 subjects (200 subjects)
1C: 2-level model with 25 groups of 4 subjects (100 subjects)
1D: 3-level model with 10 groups at level three, each with 5 subgroups of 4 subjects (200 subjects)

The second factor, , had 2 levels defined as:

2A: A single continuous covariate at each of level one and level two
2B: A single continuous covariate at level two and 5 covariates at level one (three continuous and two dichotomous)

The ICC was also broken into two levels. We used several measures to calculate the ICC and experimented to determine what values of each constituted high and low ICC values. One of these, ρ_{hl} [19], is sufficient to give an idea of the factor levels:

3A: Moderately low ICC; this corresponds to ρ_{hl} of roughly 0.20 to 0.24.

3B: Moderately high ICC; this corresponds to ρ_{hl} of roughly 0.50 to 0.57.

The final factor involves the number of random effects in the models and was broken into three levels, again based on the most likely scenarios in previous studies:

4A: Random intercept (only in the three-level model).

4B: Random intercept and one random slope for a level one continuous covariate.

4C: Random intercept and two random slopes (level one continuous and dichotomous variables); available for factor 2 level 2B only.

The resulting 28 simulations are shown in Table 1.

Table 1: Four Factor Simulation Study Design

SIMULATION	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
1	1A	2A	3A	4B
2	1A	2A	3B	4B
3	1A	2B	3A	4B
4	1A	2B	3B	4B
5	1A	2B	3A	4C
6	1A	2B	3B	4C
7	1B	2A	3A	4B
8	1B	2A	3B	4B
9	1B	2B	3A	4B
10	1B	2B	3B	4B
11	1B	2B	3A	4C
12	1B	2B	3B	4C
13	1C	2A	3A	4B
14	1C	2A	3B	4B
15	1C	2B	3A	4B
16	1C	2B	3B	4B
17	1C	2B	3A	4C
18	1C	2B	3B	4C
19	1D	2A	3A	4A
20	1D	2A	3A	4B
21	1D	2A	3B	4A
22	1D	2A	3B	4B
23	1D	2B	3A	4A
24	1D	2B	3A	4B
25	1D	2B	3A	4C
26	1D	2B	3B	4A
27	1D	2B	3B	4B
28	1D	2B	3B	4C

We generated 1000 data sets for each of the 28 simulations outlined in the previous section. We then fit the appropriate hierarchical logistic model using the SAS Glimmix macro (PQL estimation). Finally, proposed kernel smoothed USS goodness-of-fit statistic was computed using the model output. In this study, we were concerned with rejection rates for the statistic when the correct model was fit to the data.

We considered, in particular, how often the null hypothesis was rejected at three commonly used significance levels ($\hat{\alpha}$): 0.01, 0.05 and 0.1. A statistic for which the asymptotic distribution continues to hold in the smaller samples rejects at the same rate as $\hat{\alpha}$ in the 1000 simulations. Using 1000 replications in each simulation, approximate

95% confidence intervals are within 0.6%, 1.4% and 1.9% of the respective values of $\hat{\alpha}$ used.

We simulated the statistic for three different choices of kernel density and bandwidth. In most simulations, the cubic Kernel density was best. In general, the three kernels appear similar, for their optimal bandwidth choice. In our study, the cubic might appear better due to having an optimal bandwidth nearest one of the three bandwidths we chose. In practice the density chosen appears to be much less important than the bandwidth.

The three simulated bandwidths ranged from the smallest which weighted the fewest subjects among the three choices (roughly $\frac{1}{4}\sqrt{n}$ subjects). The other two bandwidths weight more of the subjects: roughly $\frac{1}{2}\sqrt{n}$ subjects and $\sqrt{n}$ subjects respectively. After reviewing the results for these three bandwidth choices, the optimal choice appeared to weight fewer subjects than the smallest bandwidth in five of the simulation settings (simulations 7, 12, 13, 20 and 28). In those cases, we ran additional simulations to find the approximately optimal bandwidth choice.

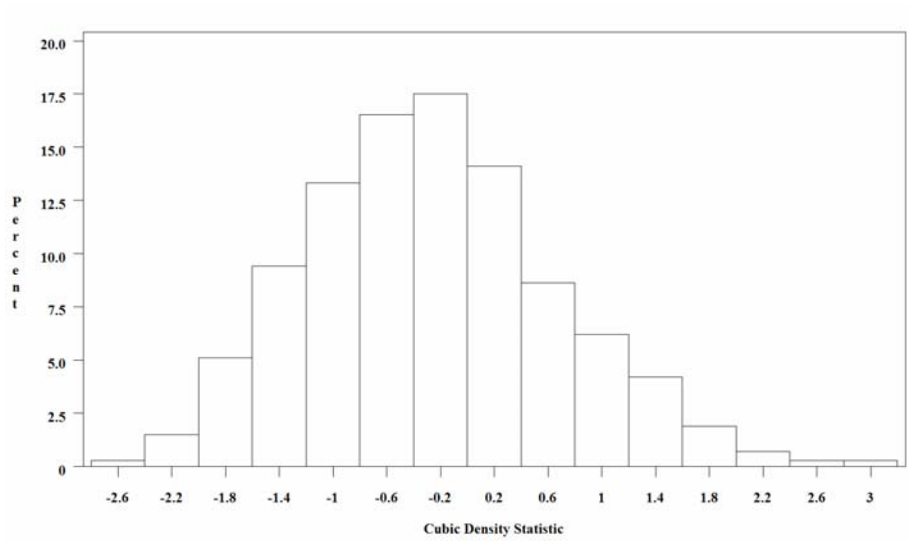
Results for the estimated optimal choice are shown in Table 2 for all 28 simulation settings using the cubic kernel density. In each case, the simulated rejection rate based on 1000 replications is displayed for each of the three significance levels ($\hat{\alpha}$): 0.01, 0.05 and 0.1. Shaded cells in the table are simulation runs in which the 95% confidence interval for the estimated rejection level includes the desired value.

Table 2: USS Kernel Statistic (Cubic Kernel Density Function) Simulation Study Results

	Significance				Significance		
Simulation	**0.01**	**0.05**	**0.1**	**Simulation**	**0.01**	**0.05**	**0.1**
1	0.02	0.062	0.103	**15**	0.011	0.049	0.093
2	0.009	0.032	0.085	**16**	0.006	0.042	0.076
3	0.011	0.047	0.084	**17**	0.014	0.039	0.094
4	0.011	0.039	0.083	**18**	0.013	0.038	0.08
5	0.017	0.062	0.1	**19**	0.016	0.052	0.091
6	0.016	0.042	0.084	**20**	0.035	0.03	0.06
7	0.01	0.04	0.09	**21**	0.011	0.052	0.089
8	0.018	0.055	0.097	**22**	0.014	0.064	0.115
9	0.008	0.049	0.102	**23**	0.007	0.035	0.074
10	0.009	0.048	0.087	**24**	0.014	0.062	0.11
11	0.014	0.051	0.099	**25**	0.017	0.06	0.112
12	0.01	0.04	0.07	**26**	0.012	0.037	0.074
13	0.01	0.04	0.08	**27**	0.016	0.053	0.11
14	0.009	0.031	0.07	**28**	0.01	0.04	0.08

reverse shading – maybe not shade
footnote to table indicating shading
As shown, the statistic rejects appropriately in nearly all simulation runs. The simulations suggest that the USS Kernel statistic is appropriate for use in logistic hierarchical models. The study includes a variety of small sample settings and the use of our theoretical asymptotic distribution performs admirably.

We do note that tests of normality typically reject (in all but five settings) the normal distribution in the simulation runs. However, we believe that this is in part due to the power to detect departures from normality with 1000 replications. Examination of the histogram (Figure 1) and QQ Plot (Figure 2) in a typical simulation setting suggests that the assumption of normality for the standardized statistic holds even in the small sample setting. We note a slight skew in the statistic but, coupled with rejection rates at various



significance levels, believe the statistic is appropriate for use in practice.

Remove figures and verbally describe

Figure 1: Histogram of USS Kernel Smoothed Standardized Statistic Values in Simulation 2 (1000 replications)

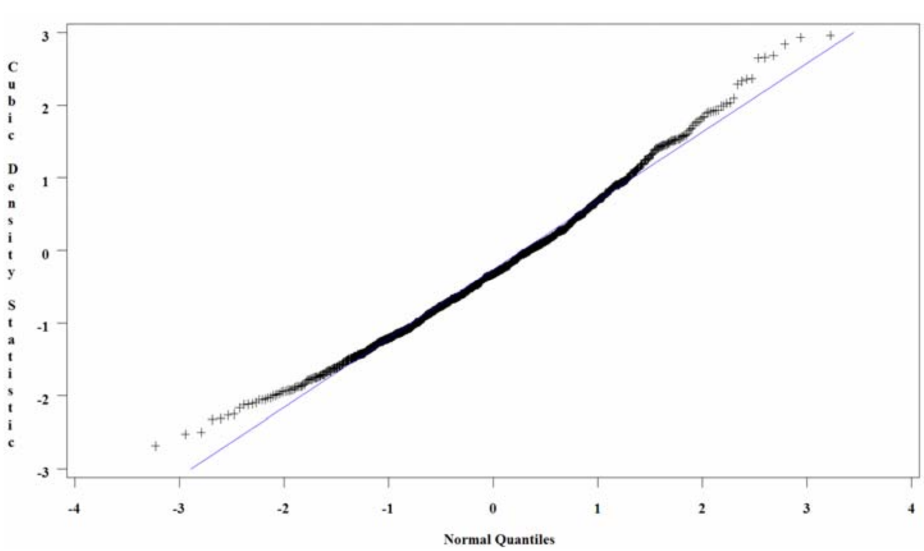


Figure 2: QQ Plot of USS Kernel Smoothed Standardized Statistic Values in Simulation 2 (1000 replications)

6 Discussion

end with what we didn't start with what we did We did not include results for several other versions of goodness-of-fit statistics that were explored in this study. These included a USS statistic using the residuals without smoothing, a statistic using the Pearson residuals (both with and without smoothing) and a version of the Hosmer-Lemeshow statistic. In each case, the theoretical asymptotic distribution did not hold for the small sample settings of our simulation study [16]. We do not recommend these statistics for use in practice.

We do recommend use of the USS Kernel smooth statistic but the choice of bandwidth deserves some discussion. The "optimal" bandwidth choice is not entirely clear and is a subject for further research. Without further study, we offer only a general rule for practice. For reasonably large cluster sizes (20) and number of groups (20) the bandwidth weighting approximately $\frac{1}{2}\sqrt{n}$ of the residuals works well. For smaller cluster or sample sizes, we recommend a smaller bandwidth ($\frac{1}{4}\sqrt{n}$). The scope of our study prevents us from speculating for other data schemes.

Based on our study, these are conservative bandwidth choices; if anything, the USS kernel statistic will reject a bit too often. In fact, we observed that the statistic will generally reject too often when the bandwidth chosen is too large. This suggests that a quick "sensitivity analysis" to bandwidth choice can help. If a larger choice does not reject the analyst can be reasonable certain the selected model is reasonable.

A summary goodness-of-fit statistic is not the only criterion to determine whether a model is acceptable. Rather, it is used to alert the analyst to a potential problem. The possibility of the statistic rejecting too often in isolated cases is not problematic. This result should prompt the model builder to look more closely at the model and data. If no unacceptable problems are discovered upon further research the model will generally still be useful.

The tendency of the statistic to reject too often in some simulation study settings using our recommended bandwidth choices is also somewhat mitigated by the ability to produce significant parameter estimates for those data schemes. We found that the amount of "over rejection" is greater in simulation runs in which one or the other of the random parameters is not "significant". Here, significance is based on the ratio of the estimate to its standard error (to form a z-statistic). In the five settings where are proposed bandwidths would reject too often, one of the random effects is often not significant. In such a situation, the analyst might choose a model excluding the random effect. The kernel smoothed statistic in settings with fewer random effects appears to perform well (in fact, even the unadjusted statistics may be useful in such cases [11]).

References

[1] Guo, G. and Zhao, H. (2000), "Multilevel Modeling for Binary Data", Annual Reviews of Sociology, 26, 441-462.

[2] Bryk, A. and Raudenbush, S. (1992), Hierarchical Linear Models, Sage Publications, Newbury Park.

[3] Zeger, S. and Karim, M. (1991), "Generalized Linear Models with Random Effects; a Gibbs Sampling Approach", Journal of the American Statistical Society, 86, 79-102.

[4] Gilks, W., Richardson, S., and Spiegelhalter, D. (1996), Markov Chain Monte Carlo in Practice, Chapman and Hall, London.

[5] Breslow, N.E. and Clayton, D.G. (1993), "Approximate Inference in Generalized Linear Mixed Models", Journal of the American Statistical Association, 88, 421, 9-25.

[6] Rodriguez, G. and Goldman, N. (1995), "An Assessment of Estimation Procedures for Multilevel Models with Binary Responses", Journal of the Royal Statistical Society A, 158, 1, 73-89.

[7] Goldstein, H. and Rasbash, J. (1996), "Improved Approximations for Multilevel Models with Binary Responses", Journal of the Royal Statistical Society A, 159, 3, 505-513.

[8] Wolfinger, R. and O'Connell, M. (1993), "Generalized Linear Mixed Models: A Pseudo-likelihood Approach", Journal Statistical Computation and Simulation, 48, 233-243.

[9] Copas, J. (1989), "Unweighted Sum of Squares Test for Proportions", Applied Statistics, 38, 1, 71-80.

[10] Hosmer, D., Hosmer, T., le Cessie, S., and Lemeshow, S. (1997), "A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model", Statistics in Medicine, 16, 965-980.

[11] Evans, S. (1998), "Goodness-of-Fit in Two Models for Clustered Binary Data", Ph.D. Dissertation, University of Massachusetts Amherst, Ann Arbor: University Microfilms International.

[12] le Cessie, S., and van Houwelingen, J. (1991), "A Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods", Biometrics, 47, 1267-1282.

[13] Fowlkes, E. (1987), "Some Diagnostics for Binary Logistic Regression via Smoothing Methods", Biometrika, 74, 503-515.

[14] Wand, M. and Jones, M. (1995), Kernel Smoothing, Chapman & Hall/CRC, Boca Raton.

[15] Hardle, W. (1990), Applied Nonparametric Regression, Cambridge University Press, Cambridge.

[16] Sturdivant, R. (2005), "Goodness-of-Fit in Hierarchical Logistic Regression Models", Ph.D. Dissertation, University of Massachusetts Amherst, Ann Arbor: University Microfilms International.

[17] Searle, S. (1982), Matrix Algebra Useful for Statistics, John Wiley and Sons, New York.

[18] Seber, G. (1977), Linear Regression Analysis, John Wiley and Sons, New York.

[19] Hosmer, D. and Lemeshow, S. (2000), Applied Logistic Regression 2nd Edition, John Wiley & Sons, Inc., New York.

Recurrent Event Model and Army Hospitalization

Yuanzhang Li, Timothy E. Powers

Introduction

In many fields of study, research questions involve life course events that can happen repeatedly over time, such as injuries, hospital admissions, risk events, etc. Researchers analyzing such data are often concerned with both the "if" and "when" of event occurrence. Traditional survival analysis methods, however, are most commonly applied to situations in which the outcome of interest that can occur only once for each subject under study. Different models, or at least adaptations of these traditional survival analysis techniques, are needed to examine data involving recurrent events.

One setting in which such models are needed is the study of hospitalizations among new enlistees in the United States Army. Early hospitalizations of enlistees are not only quite costly, but have been shown to be a strong risk factor for another costly problem -- early attrition. Hence, it is of interest to understand the risk factors for early hospitalizations, and multiple hospitalizations, among new enlistees.

A number of multivariate regression models have been proposed for use with recurrent events. [1-5] In this study, we will apply five different recurrent models that have been proposed in the scientific literature to Army enlistee hospitalization data, and to closely related data simulations. We will examine results for consistency across models, and for robustness of each model as alterations are made to event timing, percentage of subjects experiencing events, and sample size.

Methods

Five different recurrent models were compared by using SAS procedures: Poison Process, Counting Process, Conditional A, Conditional B, and Marginal. Each of these models will be used to estimate the influence of several three factors (hospitalization timing, proportion of subjects hospitalized, and sample size) on control variable effect estimates. The control variables include gender, race, age, Armed Forces Qualification Test (AFQT) percentile score, indicators of body weight (underweight or overweight), and an indicator of medical qualification status at the time of application for service (qualified, temporarily disqualified, permanently disqualified).

Details of the models to be used are given below.

The Poison process model

The Poison model is essentially event counting, with the assumption that each event is independent of other events. Under this model, a subject contributes to the risk set for an event as long as the subject is under observation. Further, it is assumed that all time periods of the same length have the same probability of an event occurring. In the current setting, it should be noted that some hospitalization causes (e.g. injuries which leave a subject at greater risk for subsequent injury) might not satisfy the above assumptions.

This model ignores the order of the events, leaving each subject to be at risk for any event as long as they are still under observation. The data structure for the Poison model would consist only of the length of time at risk and the number of events experienced during that time.

The counting process model

The second model is the counting process. This model assumes that each event is independent. A subject contributes to the risk set for an event as long as the subject is under observation at the time the event occurs; the data for a subject with multiple events could be described as data for multiple subjects while each subject has different entry date and is followed until the next event occurs. For example, in the data set we see that the first subject will be at risk for any event occurring between 0 and 80 months and subject two for any events occurring between 0 and 71 months.

This model, thus, ignores the order of the events leaving each subject to be at risk for any event as long as they are still under observation at the time of the new event occurring. This implies that a subject could be at risk for a subsequent event without having experienced the prior events. Since this model ignore the event order, same as the Poison model, for some causes of hospitalization, it might not be fine. But for some causes of hospitalization, which can't be recovered well, such as asthma, mental disease, Schizophrenia, etc, it might be improper to use this model.

The conditional model A

The third model is a conditional model. Using such a model, it assumes that it is not possible for a subject to be at risk for event 2 without having experienced event 1. It means if a subject is at risk for a subsequent event, then it is already having experienced the previous event; In order to contract the data in such an order, a strata variable is designed to indicate the event number. In this model the time interval of a subsequent event starts at the end of the time interval for the previous event.

This model is useful for modeling the full time course of the recurrent event process. In the data set the time intervals are set up exactly the same as in the counting process model with each time interval starting at the time of the previous event occurring. But the difference between this model and the counting process model is that we are using the stratum variable to keep track of the event number; thus, ensuring that it is not possible to be at risk for subsequent events without having experienced the previous events.

The conditional model B

Next model is also a conditional model. This model only differs from the Conditional Model A in the way how the time intervals are structured. For the data, each time interval starts at zero and ends at the length of time until the next event occurs. The result is that the risk sets for each of these conditional models are completely different and the questions that these analysis answer are also very different. This model is very useful for modeling the time between each of the recurring events rather than the full time period of the recurrent event process. In the data set the first subject experiences four time intervals which each start at time zero but end at the length of time until the next event. This model ignores the length of full time period. For the data including a lot of subjects without events, it might overestimate the significance of the factors we studied.

As in Conditional Model A we use the stratum variable to keep track of the event number; thus, ensuring that it is not possible to be at risks for subsequent events without having experienced the previous events.

The marginal model

In the marginal model each event is considered as a separate process. We assume that the time for each event starts at the beginning of follow up time and ends at the time of event occurring or to the end of the follow up time; Each subject is considered to be at risk for all events; The number of events at risks for all subjects are the same, which is the maximum number of events among all subjects. In other word, all subjects in the study contribute follow up times to all possible recurrent events. The marginal model considers each event separately and models all the available data for the specific event. By using the marginal model, the size of the data set is much bigger than that used by other models, especially, if there are a lot of subjects without experienced events.

Table 1 below shows the data structures necessary to implement each of the models described above. In this table, the "Time" variable shows the observation time (in months) used in the various models. The variable "Event" shows the number of events (in this case, hospitalizations) occurring in the given period. Finally the variable "Status" shows indicates the ordering of events (those with multiple events showing Status=1 indicate models that do not consider event order).

Table 1: Data Structure:

Table 1. Structure of Recurrent Event Data for Various Models

Model	Enlistee	Time(month)	Event	Status	Enlistee	Time	Event	Status
Poison Process	1	(0, 80)	4	1	2	(0,71)	2	1
Counting Process	1	(0, 5]	1	1	2	(0, 32]	1	1
		(5, 9]	1	1		(32, 62]	1	1
		(9, 56]	1	1		(62,71]	0	1
		(56, 80]	1	1				
Conditional A	1	(0, 5]	1	1	2	(0, 32]	1	1
		(5, 9]	1	2		(32, 62]	1	2
		(9, 56]	1	3		(62, 71]	0	3
		(56, 80]	1	4				
Conditional B	1	(0, 5]	1	1	2	(0, 32]	1	1
		(0, 4]	1	2		(0, 30]	1	2
		(0, 47]	1	3		(0, 9]	0	3
		(0, 24]	1	4				
Marginal	1	(0, 5]	1	1	2	(0, 32]	1	1
		(0, 9]	1	2		(0, 62]	1	2
		(0, 56]	1	3		(0, 71]	0	3
		(0, 80]	1	4		(0, 71]	0	4

For all models except the Poisson process, we use the proportional means regression model. For each observation

$$M(t) = M_0(t)\, e^{X'\beta}$$

where M(t) is the Mean Cumulative Function (MCF) for the number (or associated cost) of events of interest up to time t; X' is a vector of time invariant covariates; and $M_0(t)$ is a baseline MCF.

For the Poison regression, SAS/genmod procedure will be used. For all other models we will use SAS/PHREG.

Data

In this study we use both true data and simulated data. The true data consists of follow-up on all enlistees who began Army service from 1999 to 2002 for hospitalizations occurring during this same period. The life data is censored at Dec. 31 2002 or the date, when the enlistee left the service.

The simulated data are based on the true data, but three features of the data are altered to determine the influence of three factors (hospitalization timing, proportion of subjects hospitalized, and sample size) on control variable effect estimates.
Each of the generated data sets, described in detail below, was generated, 100 times each:

1. Hospitalization timing:

1.1. Early hospitalization

For each enlistee hospitalized k times, we set the date of first hospitalization at x days from the beginning of service, at 2x days for second hospitalization, …k*x days in the k^{th} time, where k=1, 2, 3. Eight datasets of this type were generated using x=5, 10, 15, 20, 30, 60, 70 and 80.

1.2 Late hospitalization

Similar to above, but the hospitalization date was start counting from the date, when they left the service, or the censor enduing date December 31, 2002. The total eight data sets were generated by selecting x=5, 10, 15, 20, 30, 60, 70 and 80.

2. Proportion of subjects hospitalized

In the true data, about 6% of subjects were hospitalized. Simulated data with different hospitalization proportions were created using a stratified sampling technique to select subsets of hospitalized and non-hospitalized subjects from the true data. The ratios of hospitalized to non-hospitalized subjects selected were set to be 1:19, 1:9, 1:7,1:4 and 1:3 respectively. Thus, the hospitalization rates were about 5%, 10%, 12.5%, 20% and 25% in the 5 generated data sets.

3. Sample size

Fixing the percentage of hospitalized subjects at 20%, we use stratified sampling to select samples of sizes 2000, 4000, 8000 and 10000 from the true data.

Fixing the percentage of hospitalized subjects at 20%, we use stratified sampling to select samples of sizes 2000, 4000, 8000 and 10000 from the true data.

Results

True Data

Table 2 shows the estimated control variable effects from each of the five different recurrent events models considered. It is seen that estimates were quite similar across the different models. The marginal model and Poison model have slightly higher significance levels than the others; while the two conditional models have slightly lower significance.

Assessing the factors themselves as hospitalization predictors, hospitalization rates were significant different by age, gender and AFQT. It is interesting that the presence of an initially disqualifying medical condition (presumably surmounted with an accession medical waiver) did not have a significant effect on likelihood of hospitalization. However, those with a temporarily disqualifying medical condition at the time of application had significant higher hospitalization rates than those who did not have any initial disqualification.

Table 2: Control Factors Related to Hospitalization Likelihood: Influence of Model Selection on Estimated Effects

Parameter	Model	Estimates	Standard Error	P_value
AFQT	Poisson	-0.0030	0.001	0.000
	Counting	-0.0030	0.001	0.000
	Conditional A	-0.0027	0.001	0.001
	Conditional B	-0.0026	0.001	0.002
	Marginal	-0.0037	0.001	<.0001
age	Poisson	0.0210	0.005	<.0001
	Counting	0.0201	0.005	<.0001
	Conditional A	0.0160	0.005	0.001
	Conditional B	0.0186	0.005	0.000
	Marginal	0.0243	0.005	<.0001
Female	Poisson	0.7751	0.032	<.0001
	Counting	0.7769	0.032	<.0001
	Conditional A	0.7213	0.032	<.0001
	Conditional B	0.6865	0.032	<.0001
	Marginal	0.7763	0.032	<.0001
Perm DQ	Poisson	0.0734	0.063	0.245
	Counting	0.0660	0.063	0.296
	Conditional A	0.0640	0.063	0.312
	Conditional B	0.0618	0.063	0.328
	Marginal	0.0890	0.063	0.159
Temp DQ	Poisson	0.1289	0.045	0.005
	Counting	0.1261	0.045	0.006
	Conditional A	0.1049	0.045	0.021
	Conditional B	0.1106	0.045	0.015
	Marginal	0.1485	0.045	0.001
Over Weight	Poisson	0.0241	0.033	0.457
	Counting	0.0242	0.033	0.456
	Conditional A	0.0165	0.033	0.612
	Conditional B	0.0171	0.032	0.599
	Marginal	0.0183	0.033	0.574
Less Weight	Poisson	0.0943	0.045	0.035
	Counting	0.0942	0.045	0.036
	Conditional A	0.0893	0.045	0.046
	Conditional B	0.0847	0.045	0.059
	Marginal	0.0897	0.045	0.045

Simulated Data

Table 3 shows the average z-scores of effect estimates from the various data simulation scenarios allowing for variation in the timing of hospitalizations. (Recall that 100 datasets were created under each scenario -- the results below are the averages of results from these.) These results help to examine the stability of model estimates as the timing of hospitalization events among subjects varies. Results from the Poison model are not shown since they depend only on the number of events that occur, and thus do not change according to the timing of events.

It is seen that, in general, the model estimates remain fairly stable as the timing of events is altered. For example, looking at the Conditional A model, the average z-score for the effect of being female when hospitalizations were set to occur an average of every 5 days, was 7.13. As the average time interval between hospitalizations increased, this coefficient did not change much, stabilizing at 7.30 as the average time between hospitalizations grew to 60 days or more. The estimated effects of the other factors were similarly stable within this model, and indeed within all of the models examined. Not surprisingly, then, the actual coefficient estimates also showed little variation (data not shown).

Perhaps the largest impact of timing on effect estimation is seen in the Conditional A model when comparing the effect of hospitalizations occurring early in service to those occurring late in service. For example, the z-scores for the “Temporary disqualification” variable when the hospitalizations occurred early in service ranged from 0.83-0.98, all far from statistical significance. However, when the hospitalizations occurred late in service, the z-scores ranged from 1.87-1.91, a range considered in some settings to indicate borderline statistical significance. A similar pattern was seen for the age variable under this model as the hospitalizations moved from early in service to late in service.

Finally, there were no dramatic differences in results from the different models. The Conditional A model and the Counting Process model yielded results very similar to one another, and the results from the Conditional B and Marginal models were quite similar to one another. This was true not only of the z-scores, but also of the effect coefficients (data not shown).

Table 3: Average Z-scores of Effect Estimates Relating Predictive Factors to Likelihood of Hospitalization: Applying Several Models to Simulated Data

Model	Hospitalization intervals		Average z-scores of effects						
	Zero Point	Days	Female	age	AFQT	Over DQ	Less Weight	Perm DQ	Temp DQ
Conditional A	Start of service	5	7.13	0.38	-0.10	0.32	1.48	1.80	0.83
		10	7.17	0.42	-0.14	0.31	1.52	1.82	0.88
		15	7.23	0.46	-0.15	0.30	1.53	1.87	0.90
		20	7.24	0.50	-0.16	0.28	1.56	1.90	0.96
		30	7.28	0.55	-0.14	0.29	1.60	1.93	0.98
		60	7.30	0.53	-0.14	0.28	1.59	1.92	0.98
		70	7.30	0.53	-0.14	0.28	1.61	1.91	0.96
		80	7.30	0.54	-0.15	0.27	1.61	1.91	0.96
	End of service	80	8.34	1.76	-1.15	-0.05	1.54	2.51	1.90
		70	8.34	1.76	-1.18	-0.06	1.54	2.50	1.88
		60	8.33	1.74	-1.21	-0.04	1.55	2.50	1.87
		30	8.42	1.78	-1.18	-0.10	1.57	2.50	1.91
		20	8.42	1.78	-1.21	-0.12	1.46	2.54	1.89
		15	8.42	1.74	-1.18	-0.11	1.49	2.53	1.90
		10	8.44	1.78	-1.22	-0.10	1.52	2.58	1.90
		5	8.50	1.75	-1.21	-0.03	1.54	2.59	1.91
Conditional B	Start of service	5	6.71	1.50	-0.83	-0.10	1.47	2.55	1.58
		10	6.69	1.50	-0.82	-0.12	1.48	2.54	1.60
		15	6.74	1.49	-0.80	-0.13	1.49	2.57	1.60
		20	6.76	1.48	-0.80	-0.13	1.48	2.56	1.64
		30	6.83	1.47	-0.79	-0.10	1.50	2.59	1.65
		60	6.86	1.45	-0.80	-0.10	1.48	2.59	1.64
		70	6.85	1.45	-0.80	-0.10	1.50	2.58	1.63
		80	6.86	1.46	-0.81	-0.11	1.50	2.58	1.63
	End of service	80	8.17	1.26	-0.49	0.08	1.51	2.02	1.41
		70	8.17	1.28	-0.50	0.07	1.52	2.03	1.38
		60	8.15	1.26	-0.54	0.07	1.53	2.03	1.37
		30	8.22	1.26	-0.53	0.10	1.52	2.02	1.37
		20	8.20	1.15	-0.55	0.12	1.50	2.00	1.33
		15	8.19	1.06	-0.49	0.15	1.52	1.97	1.34
		10	8.16	1.02	-0.45	0.21	1.54	1.93	1.28
		5	8.14	0.97	-0.38	0.24	1.56	1.89	1.26
Counting Process	Start of service	5	7.33	0.44	-0.09	0.29	1.44	1.78	0.86
		10	7.38	0.51	-0.14	0.26	1.45	1.81	0.92
		15	7.45	0.54	-0.16	0.25	1.47	1.86	0.94
		20	7.48	0.60	-0.17	0.22	1.48	1.90	1.00
		30	7.53	0.66	-0.16	0.21	1.51	1.97	1.01
		60	7.54	0.65	-0.16	0.20	1.50	1.97	1.00
		70	7.53	0.66	-0.16	0.20	1.52	1.96	0.99
		80	7.54	0.67	-0.17	0.20	1.52	1.96	0.98

	End of service	80	8.37	1.81	-1.14	-0.09	1.50	2.51	1.91
		70	8.36	1.83	-1.15	-0.10	1.52	2.52	1.88
		60	8.34	1.82	-1.18	-0.09	1.52	2.52	1.87
		30	8.41	1.83	-1.16	-0.07	1.51	2.50	1.87
		20	8.40	1.81	-1.17	-0.08	1.48	2.52	1.87
		15	8.39	1.78	-1.15	-0.07	1.50	2.51	1.87
		10	8.38	1.78	-1.15	-0.06	1.49	2.52	1.87
		5	8.36	1.78	-1.16	-0.04	1.50	2.53	1.88
Marginal	Start of service	5	7.02	1.53	-0.99	-0.10	1.41	2.58	1.74
		10	7.04	1.53	-0.99	-0.10	1.41	2.59	1.74
		15	7.05	1.53	-1.00	-0.10	1.42	2.59	1.74
		20	7.06	1.53	-1.00	-0.10	1.42	2.59	1.75
		30	7.07	1.53	-1.01	-0.09	1.42	2.60	1.75
		60	7.07	1.53	-1.01	-0.09	1.42	2.60	1.75
		70	7.07	1.53	-1.01	-0.09	1.42	2.60	1.75
		80	7.07	1.53	-1.01	-0.09	1.42	2.60	1.75
	End of service	80	7.85	1.98	-1.26	-0.05	1.37	2.36	1.74
		70	7.85	1.98	-1.26	-0.05	1.37	2.36	1.74
		60	7.85	1.98	-1.26	-0.05	1.37	2.36	1.74
		30	7.84	1.96	-1.25	-0.04	1.37	2.36	1.73
		20	7.81	1.93	-1.22	-0.01	1.40	2.35	1.74
		15	7.81	1.93	-1.22	-0.02	1.40	2.35	1.73
		10	7.80	1.93	-1.23	-0.02	1.40	2.36	1.74
		5	7.80	1.93	-1.23	-0.01	1.40	2.36	1.74

The Hospitalization Rate Effect

Tables 4 and 5 show the average effects and z-scores of effect estimates from the various data simulation scenarios allowing for variation in the percentage of subjects experiencing hospitalization. (Recall again that 100 datasets were created under each scenario -- the results below are the averages of results from these.)

Table 4. Average Effect Estimates for Predictive Factors: Applying Several Models to Simulated Data with Different Percentage of Hospitalization.

Model	Percent of subjects hospitalized	Sex	Age	AFQT	Overwt	Underwt	Perm DQ	Temp DQ
Poisson	0.05	0.726	0.019	-0.003	0.023	0.130	0.235	0.170
	0.10	0.698	0.015	-0.003	0.030	0.099	0.141	0.157
	0.15	0.652	0.017	-0.003	0.026	0.112	0.129	0.137
	0.20	0.618	0.017	-0.002	0.022	0.088	0.106	0.118
Conditional A	0.05	0.700	0.015	-0.003	0.014	0.140	0.187	0.148
	0.10	0.671	0.013	-0.003	0.013	0.104	0.124	0.130
	0.15	0.644	0.014	-0.002	0.016	0.101	0.127	0.123
	0.20	0.608	0.015	-0.002	0.019	0.089	0.089	0.104
Conditional B	0.05	0.664	0.017	-0.003	0.014	0.136	0.188	0.154
	0.10	0.638	0.015	-0.002	0.014	0.100	0.119	0.136
	0.15	0.610	0.015	-0.002	0.016	0.097	0.124	0.126
	0.20	0.574	0.016	-0.002	0.020	0.085	0.084	0.108
Counting	0.05	0.739	0.018	-0.003	0.021	0.147	0.211	0.176
	0.10	0.701	0.016	-0.003	0.021	0.106	0.130	0.151
	0.15	0.662	0.016	-0.002	0.021	0.103	0.133	0.137
	0.20	0.616	0.017	-0.002	0.026	0.089	0.091	0.115
Marginal	0.05	0.750	0.020	-0.004	0.014	0.148	0.243	0.205
	0.10	0.719	0.018	-0.003	0.016	0.103	0.158	0.181
	0.15	0.687	0.019	-0.003	0.015	0.099	0.164	0.167
	0.20	0.647	0.020	-0.003	0.021	0.083	0.119	0.145

It is seen that the average effect coefficients and z-scores for the various control factors are quite similar regardless of the percentages of subjects hospitalized. This is true both within and across the models considered. Sex is highly statistically significant in all models and across all percentages of hospitalization, while age, AFQT score, and the two medical qualification status variables are near statistical significance in all models and hospitalization percentages.

It is worth noting that the selected hospitalized and non-hospitalized samples have the same distribution by the control factors (sex, age, etc.) as the true data. Accordingly, the estimated effects of these factors should be similar to those from the modeling of the true data. This was, indeed, generally the case (data not shown).

Table 5: Average Z-scores of Effect Estimates Relating Predictive Factors to Likelihood of Hospitalization: Applying Several Models to Simulated Data with Different Percentage of Hospitalization.

Model	Percent of subjects hospitalized	Sex	Age	AFQT	Overwt	Underwt	Perm DQ	Temp DQ
Poisson	0.05	11.88	2.08	-2.02	0.37	1.54	2.14	2.00
	0.10	13.23	1.85	-1.96	0.55	1.35	1.43	2.14
	0.15	13.87	2.43	-2.04	0.53	1.71	1.47	2.09
	0.20	14.48	2.59	-2.07	0.51	1.46	1.32	1.96
Conditional A	0.05	11.73	1.62	-1.87	0.21	1.68	1.72	1.76
	0.10	13.07	1.65	-1.79	0.25	1.42	1.27	1.78
	0.15	14.12	1.95	-1.91	0.33	1.56	1.47	1.88
	0.20	14.72	2.28	-2.16	0.44	1.51	1.12	1.75
Conditional B	0.05	11.11	1.82	-1.85	0.22	1.68	1.78	1.85
	0.10	12.43	1.84	-1.73	0.25	1.39	1.26	1.87
	0.15	13.40	2.17	-1.84	0.32	1.52	1.48	1.94
	0.20	13.94	2.48	-2.06	0.46	1.47	1.09	1.82
Counting	0.05	11.13	1.75	-1.83	0.29	1.60	1.70	1.82
	0.10	12.46	1.78	-1.71	0.34	1.33	1.21	1.82
	0.15	13.46	2.09	-1.80	0.40	1.47	1.42	1.90
	0.20	14.06	2.38	-2.04	0.54	1.43	1.06	1.78
Marginal	0.05	10.74	1.87	-2.15	0.19	1.53	1.88	2.01
	0.10	11.95	1.92	-2.08	0.24	1.22	1.38	2.05
	0.15	12.84	2.28	-2.24	0.27	1.33	1.62	2.13
	0.20	13.34	2.58	-2.53	0.41	1.22	1.28	2.03

Table 6 shows the standard deviations of the estimated demographic effects in simulations with different percentages of subjects being hospitalized. In general, we would like to use the method with less standard deviation of estimates, if the estimates are unbiased. Overall, the standard deviations from both conditional models are similar to one another, and they are smaller than those from the counting process and marginal models.

Table 6: Standard Errors of Effect Estimates: Applying Several Models to Simulated Data with Different Percentage of Hospitalization

Model	Percent of subjects hospitalized	Sex	Age	AFQT	Overwt	Underwt	Perm DQ	Temp DQ
Poisson	0.05	0.046	0.010	0.002	0.058	0.065	0.089	0.072
	0.10	0.035	0.006	0.001	0.034	0.041	0.051	0.056
	0.15	0.028	0.005	0.001	0.028	0.040	0.042	0.045
	0.20	0.022	0.004	0.001	0.026	0.030	0.030	0.029
Conditional A	0.05	0.048	0.008	0.001	0.046	0.059	0.073	0.063
	0.10	0.032	0.007	0.001	0.035	0.043	0.061	0.045
	0.15	0.028	0.006	0.001	0.028	0.034	0.047	0.040
	0.20	0.023	0.004	0.001	0.020	0.029	0.043	0.033
Conditional B	0.05	0.047	0.008	0.001	0.044	0.058	0.069	0.061
	0.10	0.032	0.007	0.001	0.034	0.042	0.057	0.045
	0.15	0.028	0.006	0.001	0.027	0.034	0.043	0.039
	0.20	0.022	0.004	0.001	0.020	0.028	0.042	0.033
Counting	0.05	0.054	0.009	0.001	0.050	0.067	0.081	0.072
	0.10	0.035	0.007	0.001	0.039	0.048	0.065	0.050
	0.15	0.031	0.006	0.001	0.031	0.037	0.051	0.043
	0.20	0.025	0.005	0.001	0.022	0.031	0.047	0.037
Marginal	0.05	0.056	0.009	0.001	0.051	0.070	0.085	0.074
	0.10	0.038	0.008	0.001	0.040	0.050	0.066	0.053
	0.15	0.033	0.006	0.001	0.032	0.041	0.052	0.046
	0.20	0.027	0.005	0.001	0.024	0.033	0.049	0.040

Comparing Tables 4 and 6, it can be seen that for the sex variable, the standard errors are much less than their respective mean estimated coefficients for all five models and all 4 different hospitalization rates. This means the estimation of this effect is robust across the models, and shows that gender is a significant predictor of hospitalization.

For the other demographic factors, the mean coefficients are comparable in size to their respective standard errors. This means that the estimated coefficient may not be so reliable, and we should not draw conclusions based on only one selected model.

The standard errors of effect estimates are decreasing as the hospitalization rate is increasing, as would be expected. In particular, when the hospitalization rate is about 20%, the effects estimates for age and AFQT are quite reliable across models. When the hospitalization rate is about 5%, the standard errors are quite large compared to the mean estimated coefficients. Model selection and interpretation should be done more carefully at such lower levels of hospitalization percentages.

Sample Size Effect

The results in Table 7 show effect estimates from simulated data of various sample sizes. The selected numbers of hospitalized individuals were 500, 1000, 1500 and 2000 respectively, and four times as many controls were selected for each of these cases. Hence the hospitalization ratio was 20% overall for each data. As in the previous analyses, the distributions of hospitalized and non-hospitalized subjects by the control variables (sex, age, etc.) in this simulation were the same as in the true Army hospitalization data.

The results indicate that sample size has little effect on the effect estimates regardless of which model is used. As expected, however, larger sample size results in reduced standard error and thus greater statistical significance (data not shown).

Table 7: The demographic factor effect s by sample size

Model	Sample Size	Sex	Age	AFQT	Overwt	Underwt	Perm DQ	Temp DQ
Poisson	2,000	0.546	0.017	-0.003	0.022	0.170	0.293	0.203
	4,000	0.569	0.013	-0.002	0.015	0.122	0.125	0.156
	8,000	0.604	0.020	-0.002	0.020	0.108	0.123	0.115
	10,000	0.618	0.016	-0.002	0.032	0.100	0.110	0.125
Conditional A	2,000	0.559	0.014	-0.003	0.017	0.174	0.266	0.190
	4,000	0.566	0.011	-0.002	0.008	0.121	0.120	0.147
	8,000	0.601	0.018	-0.002	0.016	0.104	0.120	0.112
	10,000	0.611	0.014	-0.002	0.026	0.097	0.103	0.113
Conditional B	2,000	0.523	0.016	-0.003	0.016	0.164	0.270	0.190
	4,000	0.536	0.012	-0.002	0.009	0.118	0.113	0.150
	8,000	0.568	0.019	-0.002	0.015	0.103	0.114	0.109
	10,000	0.579	0.015	-0.002	0.027	0.093	0.101	0.116
Counting	2,000	0.548	0.016	-0.003	0.021	0.170	0.283	0.202
	4,000	0.572	0.012	-0.002	0.015	0.122	0.118	0.155
	8,000	0.607	0.020	-0.002	0.020	0.107	0.118	0.115
	10,000	0.620	0.016	-0.002	0.032	0.100	0.104	0.123
Marginal	2,000	0.580	0.021	-0.004	0.014	0.167	0.348	0.245
	4,000	0.599	0.016	-0.003	0.013	0.121	0.142	0.192
	8,000	0.640	0.023	-0.003	0.016	0.106	0.152	0.135
	10,000	0.651	0.018	-0.003	0.030	0.094	0.142	0.157

Conclusions

Five different model types were used to estimate the effects of several demographic factors on likelihood of hospitalization among new Army enlistees. When applying these models to actual data over a three-year period, the results from these models were quite similar. This result is somewhat comforting in that the major conclusions to be drawn from such modeling would therefore not be dependent on which model was used.

Results from simulated data, generated by making judicious changes to the true data, indicated that these models, and their similarity in results to one another, were fairly robust. In particular, it was found that changes in the timing of hospitalizations, the percentage of subjects hospitalized, and the sample size did not appreciably alter the harmony of findings within or across models.

Nonetheless, some observations were of note from the data simulations:

- After the Poisson (which does not depend the timing of events) the marginal model was the most robust with respect to hospitalization event timing.
- The Conditional A model was more sensitive to events occurring later in service than those occurring earlier.
- There was considerable similarity in results between the Conditional A and counting process models, and between the Conditional B and Marginal process models.

- Estimation of demographic effects was relatively stable as the percentage of subjects hospitalized increased.
- The Conditional models showed generally less variation in results, and lower significance levels of demographic effects, than the other models.
- The estimated effects of several demographic factors on hospitalization likelihood were not so reliable when the percentage of subjects hospitalized was low. In such a case, results from several models should be considered, and conclusions must be drawn carefully.

Modeling of recurrent events while accounting for the timing of those events requires extension of traditional survival analysis techniques. The results of this study indicate that any of the five models presented in this paper are adequate choices for the modeling of hospitalization data among new Army enlistees, and perhaps in other settings as well.

David W., Jr. Hosmer, Stanley Lemeshow, Applied survival analysis: Regression modeling of time to event data, 1999, Wiley-Interscience;
Fleming, T. R. and Harrington, D. P. (1991) Counting Processes and survival analysis, New York Willey

Lawless, J. F. and Nadeau, C. (1995), Some simple robust methods for the analysis of recurrent events, Technometrics, 37

Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000), Semiparametric regression for the mean and rate functions of recurrent events, J. R. Statisc. Soc. B., 62,

Gordon Johnston andYing So, Analysis of data from recurrent events, SUGI 28, SAS Institute Inc. Cary, North Carolina.

Evaluation of Advanced Aerospace Aluminum Alloys For Armor and Structural Applications

Reference: ARL-TR-3185

John F. Chinella

U.S. Army Research Laboratory,
AMSRD-ARL-WM-MD
Aberdeen Proving Ground, MD 21005

UNCLASSIFIED- UNLIMITED DISTRIBUTION

Introduction

Solutions to High Fuel and Material Costs

Materials, Processing, and Performance Improvements

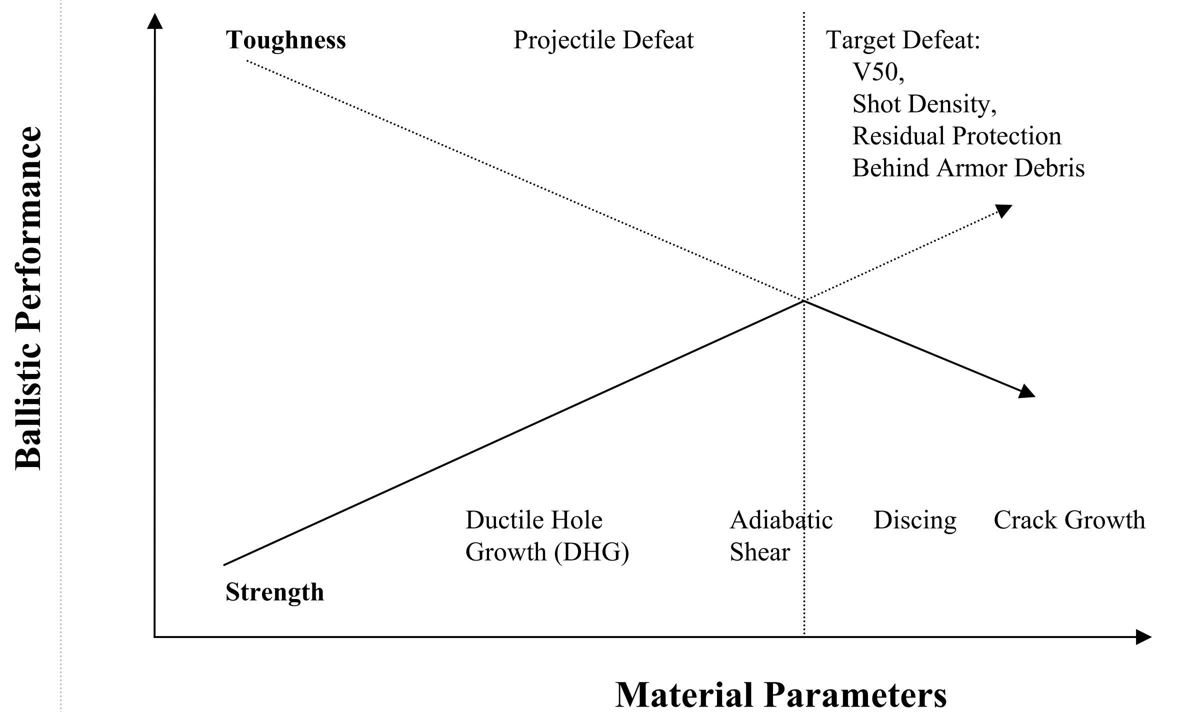
- **Durability**
- **Specific-Strength and Toughness**
- **Low Material and Operating Costs**
- **Isotropic Mechanical Properties**
- **Weldable Al-Cu-Li Alloys**

Lithium as Alloy-Element

- **Non-Toxic**
- **3% Density-<u>Reduction</u> / Weight-%**
- **6% Elastic Modulus <u>Increase</u> / Weight-%**

Introduction

Aluminum Armor:
Material Parameters and
Ballistic Performance of Small-Caliber Projectiles



- Plastic Flow, and Failure Modes Determine Performance
- At High-Strength Levels, Localization of Plastic Flow and Fracture Decrease Ballistic Performance

Introduction

- **Al-Li Alloys versus Al-Armor 7039**
  - **- Literature: Investigations Controversial-**
    - **a. 2090-T8: No significant improvement, either at 0° or 30°;**
    - **b. 2090-T8: Improvement claimed with 0° impact obliquity, but not oblique; Al-Li material parameters claimed significant**
    - **c. 2090-T8 & 2049 Improvement claimed at 30° obliquity, effect improves with strength.**
- **Quantitatively,**
  What are the V50 mean and variance of 7039 Al-armor performance?
  Do the experimental materials provide significant V50 improvements?
  What are the improvements?

- **Do the material parameters or failure modes enhance ballistic protection?**

Experimental Aluminum Alloy Target Materials

- **C47A, C458** Al-Cu-Li-Zn (C458 → 2099)
- **7055** Al-Zn-Cu-Mg
- **2195** Al-Cu-Li-Mg-Ag

Experimental Results

Experimental Mechanical Properties In Tension with Range Variation - Comparison to Reference Alloys

Property	Experimental Alloys and Tempers				Reference Data:Al Armor and Al-Li		
	C47A -T8	C458 -T861	2195 -T8	7055 -T7751	7039 -T64	2519 -T87	2090 -T8
Young's Modulus (GPa)	76.8 (0.6)	77.7 (0.4)	75.9 (0.6)	70.6 (0.6)	70	72	79
0.2% Yield Strength (MPa)	437 (4.2)	525 (5.2)	592 (1.8)	602 (1.9)	400	423	490
Ultimate Strength (MPa)	469 (1.5)	558 (2.7)	627 (0.9)	632 (2.1)	458	465	550
Elongation (%)	13.2 (1.7)	9.3 (1.4)	9.6 (2.7)	14.5 (0.7)			
Reduction of Area (%)	44.3 (0)	23.5 (7.1)	22.2 (9.1)	40.7 (1.1)			

Experimental Results: average of 3 specimens strained to failure, except C458 = 6 specimens

Experimental Results

Physical Properties And Specific Strength

Property	Experimental Alloy				Al Armor			Ti Armor
	C47A -T8	C458 -T861	2195 -T8	7055 -T7751	5083	7039	2519	6Al-4V
Density (ρ), (g/cm^3)	2.642	2.633	2.709	2.866	2.66	2.73	2.82	4.43
Hardness (HRB)	75.0	80.4	88.3	92.1	—	—	—	—
Young's Modulus (GPa)	76.8	77.7	75.9	70.6	70.3	69.6	72.4	110
Specific Modulus $(E/\rho)/10^{-8}$ (cm)	2.96	3.01	2.86	2.51	2.69	2.60	2.62	2.53
Specific Strength $(YS/\rho)/10^{-6}$ (cm)	1.69	2.03	2.23	2.14	1.20	1.45	1.53	1.75–2.10

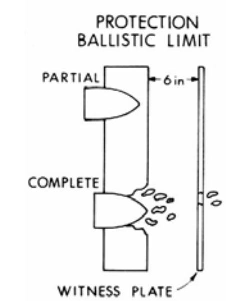
- **Al-Cu-Li Alloys Have the Highest Specific Modulus and Strength**

Experimental Results

Ballistic V50 Test: Procedure, Results, Criteria

Projectile		Target Characteristics			Penetration Result Intervals		V50 Prot. Limit Averaged	
Type	Impact Obliquity (°)	Alloy	Thickness (t) (mm)	AD (kg/m^2)	Target Impacts (No.)	Distribution of All P, C, Penetrations	Shots (No.)	(S) Spread (m/s)
0.50-cal. FSP	0	C47A	19.10	50.46	8 UM	3P ≤ S ≤ 5C	2	1.2
	0	2195	19.13	51.83	5 UM	3P≤ S ≤ 2C	2	2.4
	0	C458	19.91	52.43	8 M	3P≤ S ≤ 3C	4	15.8
20-mm FSP	0	7055	31.72	90.95	5 UM	3P≤ S ≤ 2C	2	5.8
	0	2195	40.16	108.8	6 UM	4P≤ S ≤ 2C	2	17.7
0.30-cal. APM2	0	C47A	19.02	50.26	10 UM	7P≤ S ≤ 3C	2	2.7
	0	2195	19.11	51.76	4 M	1P≤ S ≤ 1C	4	15.5
	0	C458	19.91	52.43	6 UM	3P≤ S ≤ 3C	2	11.0
	0	7055	18.95	54.34	4 UM	2P≤ S ≤ 2C	2	4.9
	0	2195	31.76	86.02	9 UM	4P≤ S ≤ 5C	2	4.3
	0	7055	31.75	91.03	11 UM	8P≤ S ≤ 3C	2	3.7
	0	2195	40.18	108.9	11 UM	7P≤ S ≤ 4C	2	13.4
	30	C47A	19.08	50.40	10 M	3P≤ S ≤ 3C	6	23.2
	30	2195	19.11	51.76	12 M	4P≤ S ≤ 3C	4	20.4
	30	C458	19.94	52.50	9 M	4P≤ S ≤ 3C	4	17.4
	45	C458	19.90	52.40	12 UM	9P ≤ S ≤3C	2	7.9
0.50-cal. APM2	0	2195	39.65	107.4	9 UM	5P ≤ S ≤ 4C	4	7.9

Test Criteria

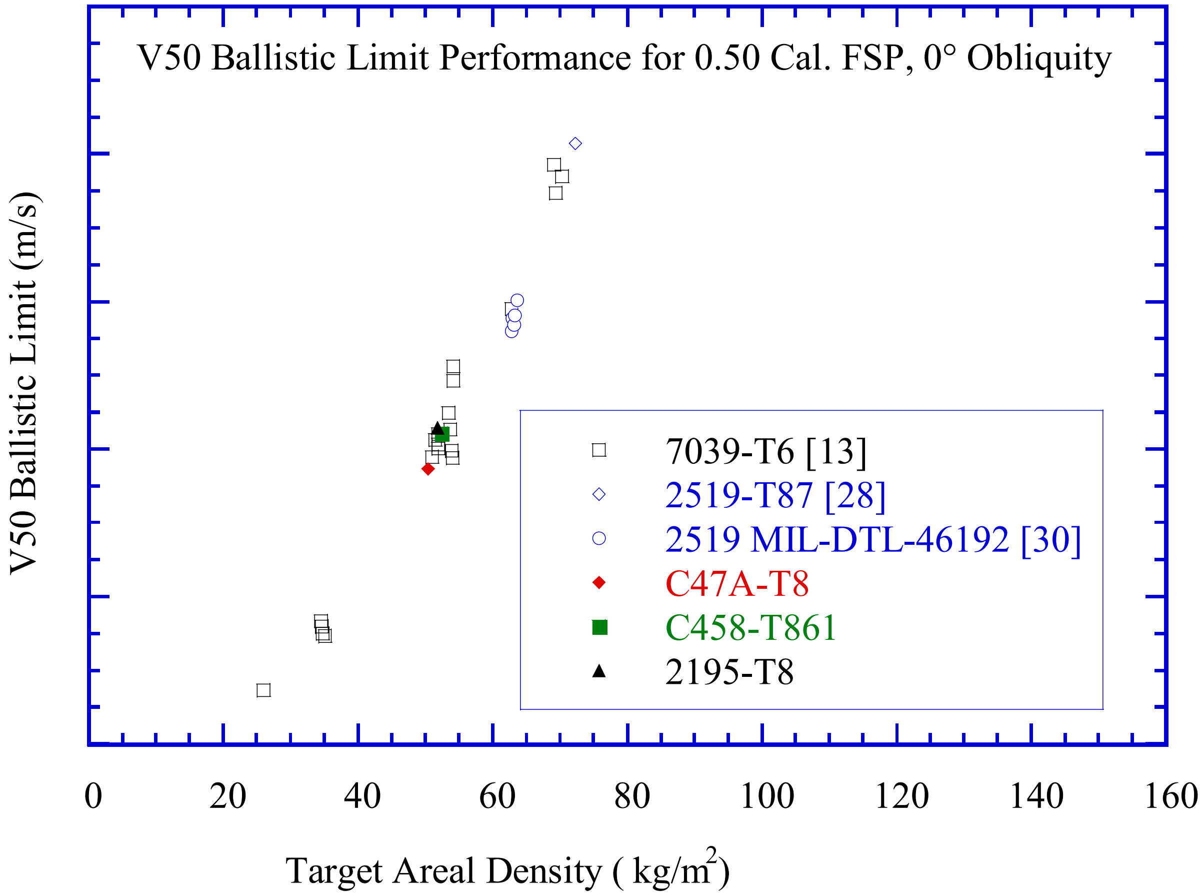


UM = UnMixed, all P or all C results
M = Mixed, includes P & C results
P = Partial Penetration
C = Complete Penetration
S = spread, range of velocities of P and C results used in V50 estimate

ACAS, 20-22 October, 2004

Results - Discussion

0.50 cal. FSP, 0° Obliquity

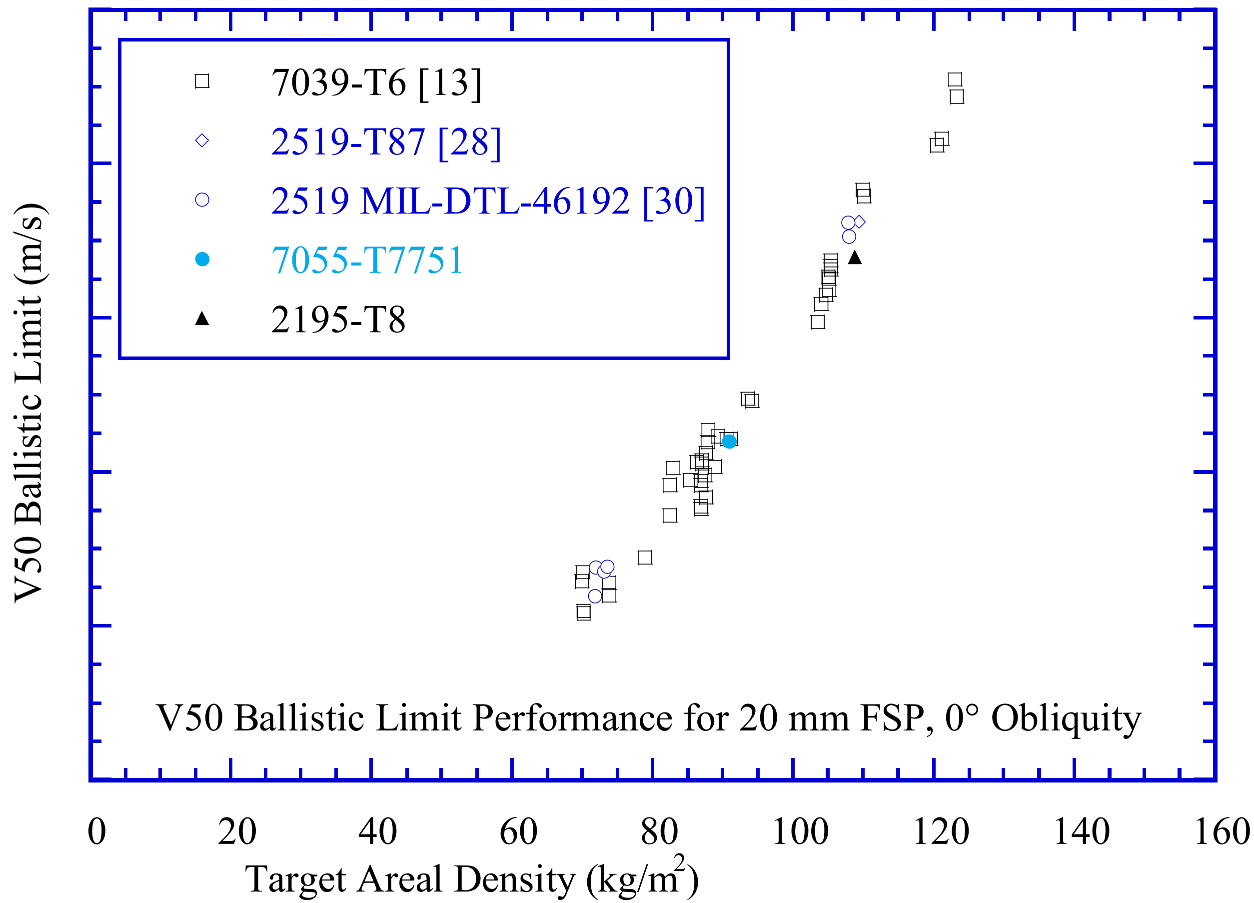


Scales of V50 and AD axes:
FSP = 0.30 cal. APM2
= 0.50 cal. APM2(V50)

Experimental Alloys with 7039 and 2519 Al armor V50 Performance

Results - Discussion

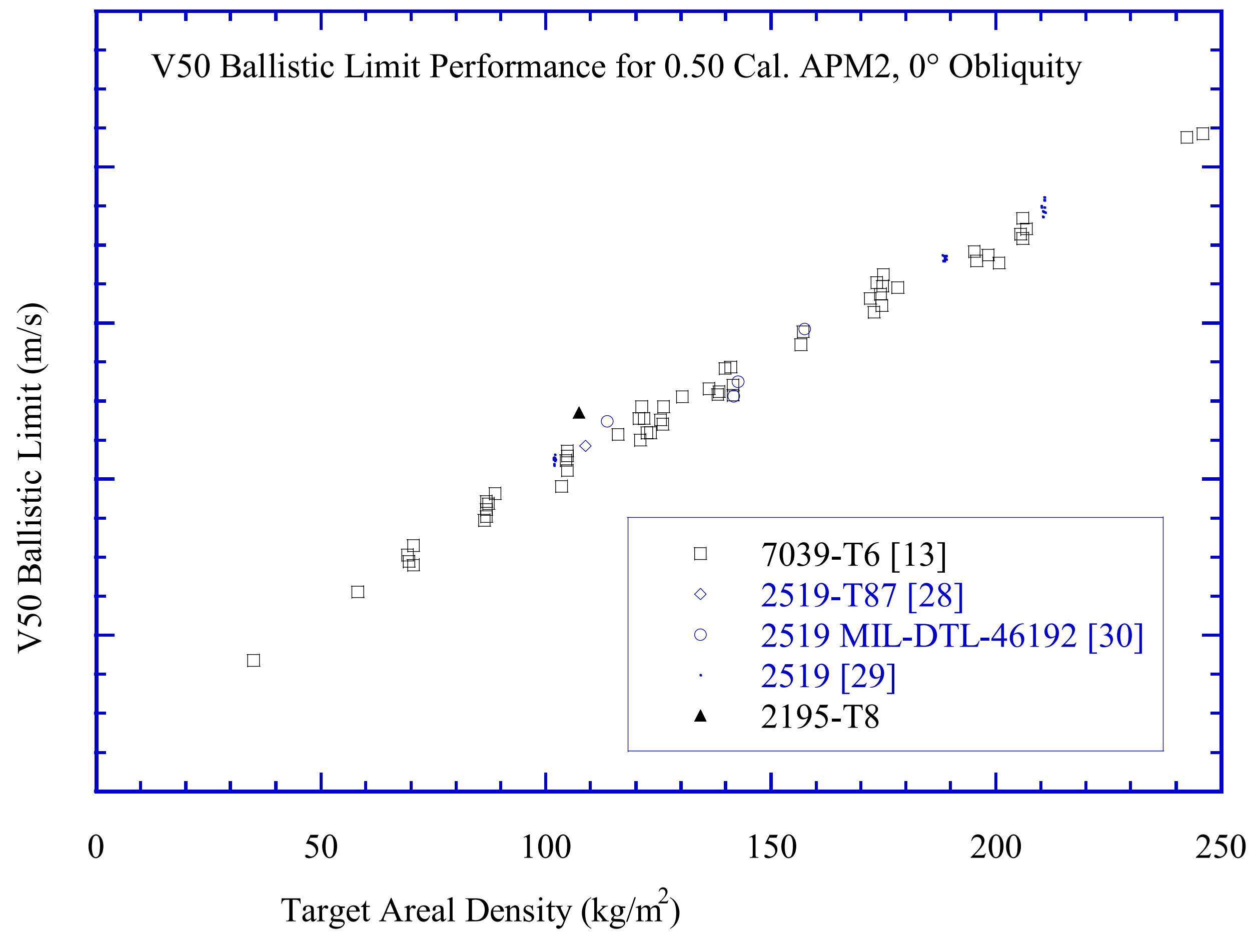
20 mm FSP, 0° Obliquity



Experimental Alloys with 7039 and 2519 Al armor V50 Performance

Results - Discussion

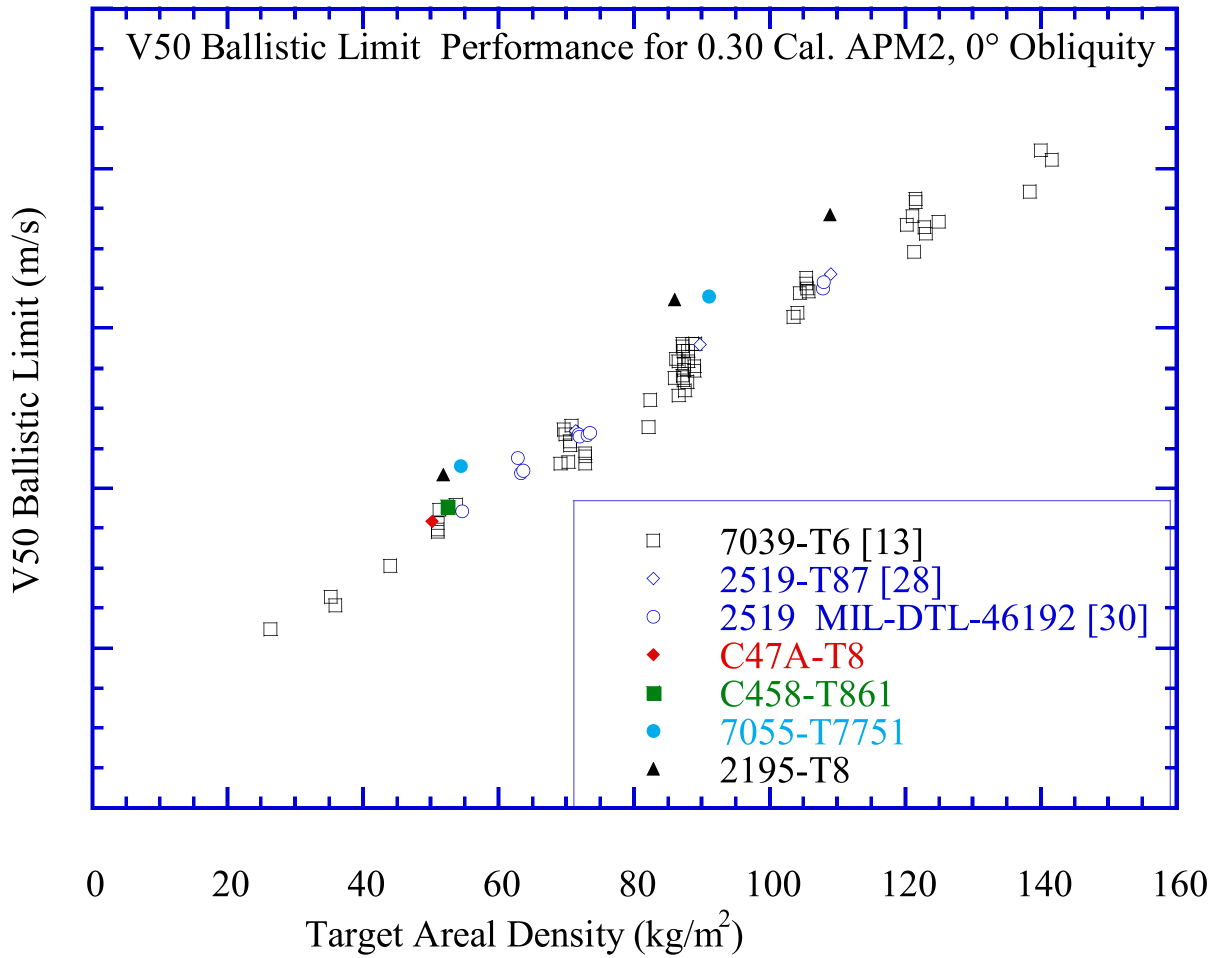
0.50 cal. APM2, 0° Obliquity



Experimental Alloys with 7039 and 2519 Al armor V50 Performance

Results - Discussion

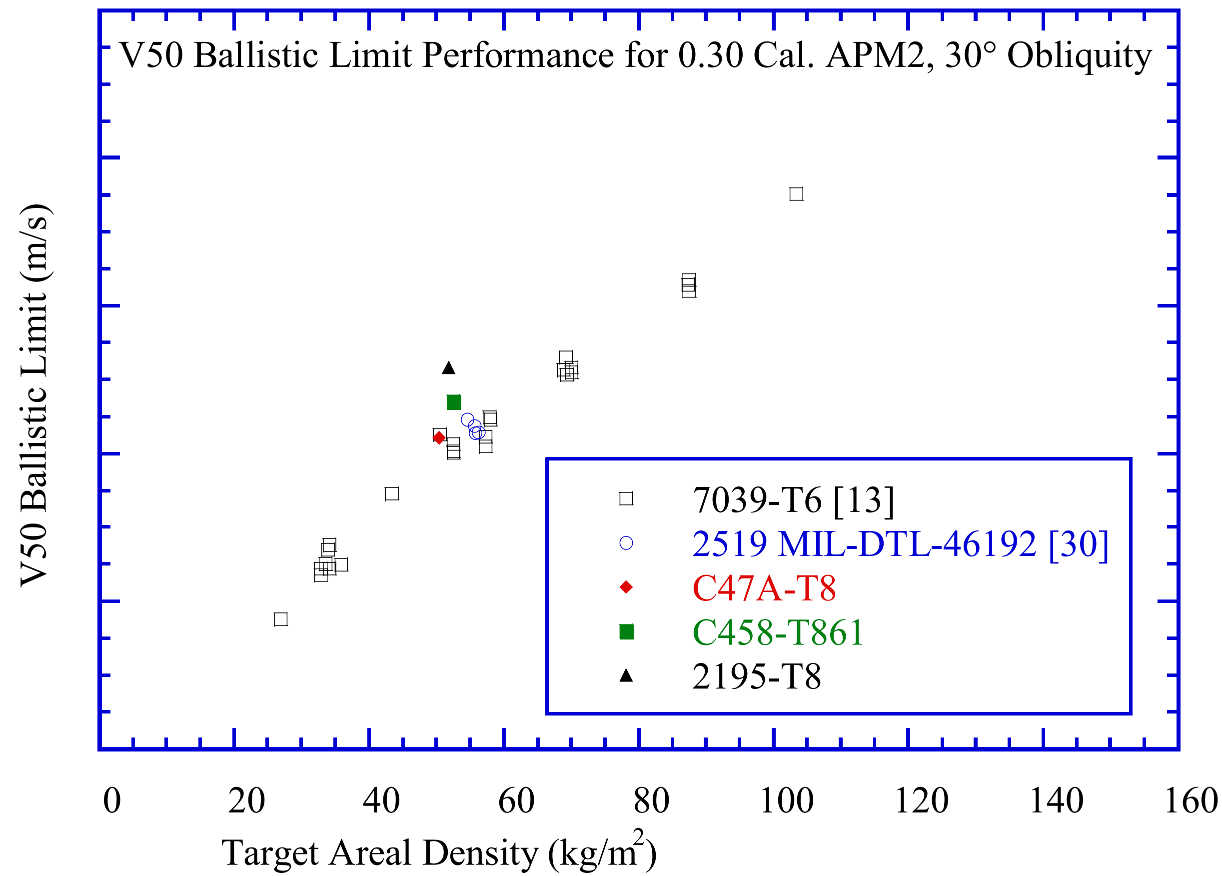
0.30 cal. APM2, 0° Obliquity



Experimental Alloys with 7039 and 2519 Al armor V50 Performance

Results - Discussion

0.30 cal. APM2, 30° Obliquity

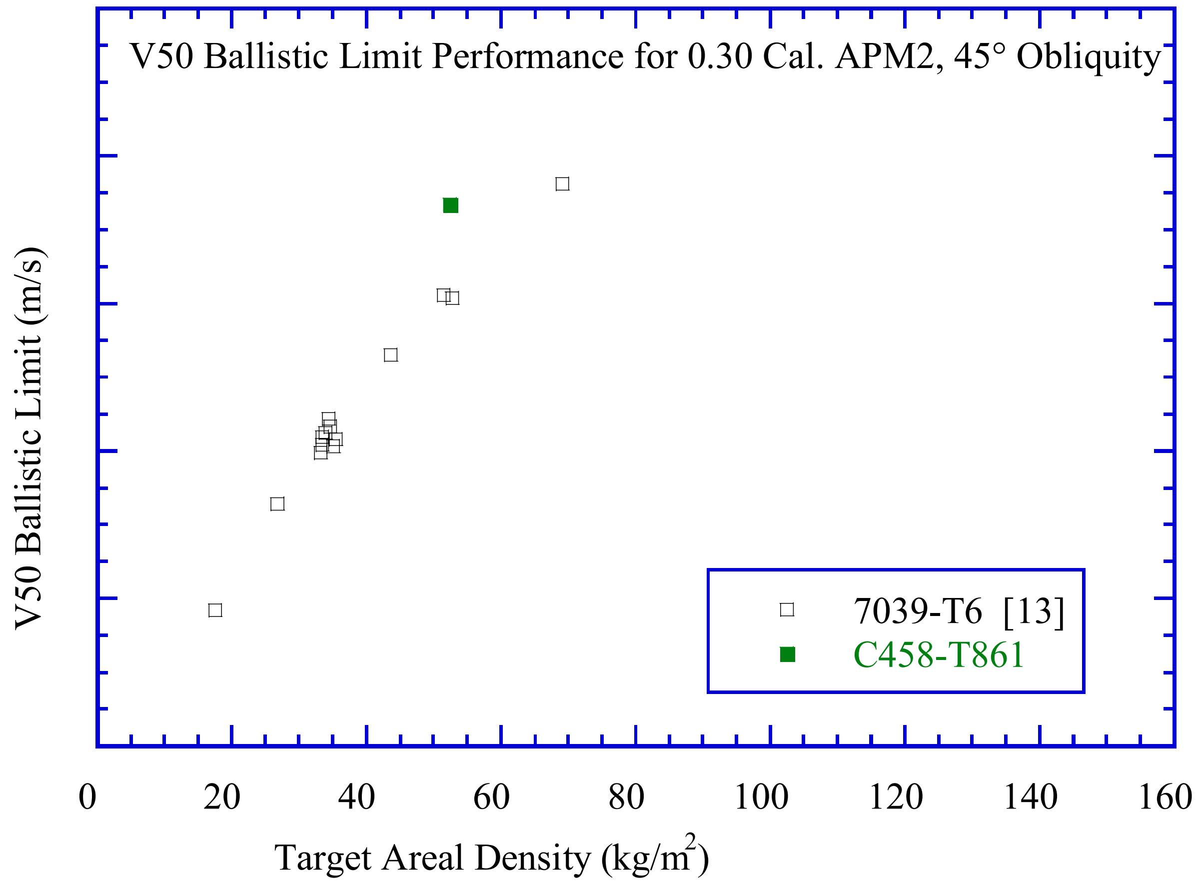


Experimental Alloys with 7039 and 2519 Al armor V50 Performance

- Oblique-Impact V50 Performance of Al-Li Alloys > 7039

Results - Discussion

0.30 cal. APM2, 45° Obliquity

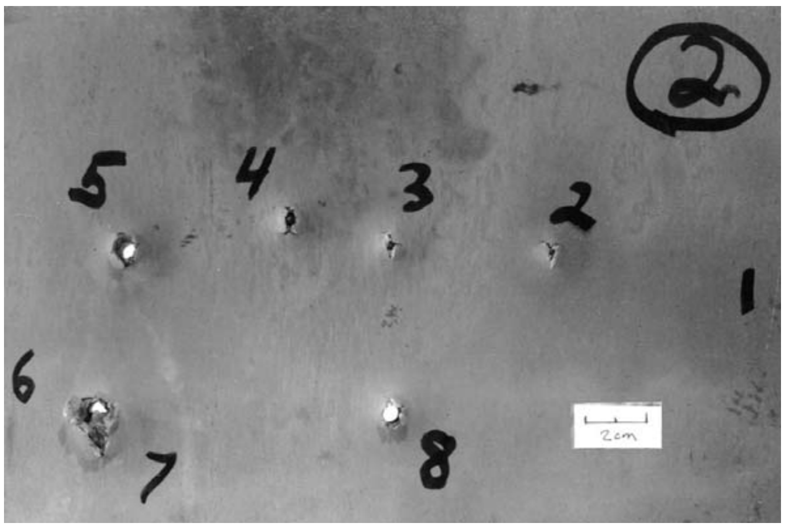


Experimental Alloys with 7039 and 2519 Al armor V50 Performance

- Oblique-Impact V50 Performance of Al-Li Alloys > 7039

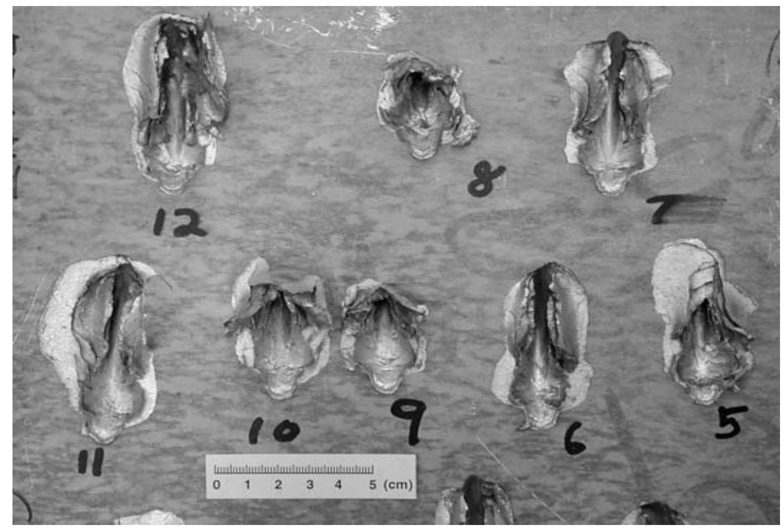
Results and Discussion

0.30 cal. APM2 Penetration Modes



C458-T861: 0° Obliquity

- Damage-tolerant
- High shot-density
- Multiple impacts



C458-T861: 45° Obliquity

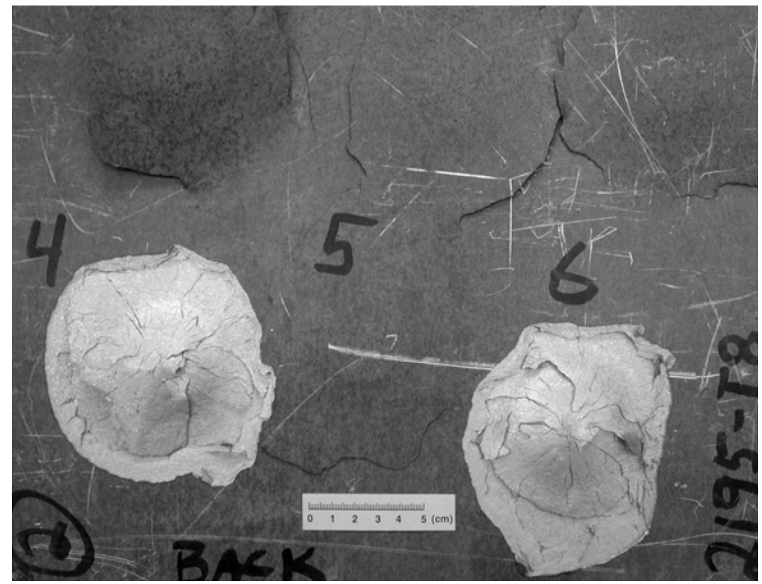
V50 45° = 1.62 x V50 0°

Results and Discussion

FSP Penetration Mode

C47A-T8 versus 0.50 cal. FSP

- Damage-tolerant
- High shot-density
- Multiple impacts



2195-T8 versus 20 mm FSP

- Ductile Hole Growth
  + Discing Failure

Results and Discussion

Linear Regression in Matrix Notation

$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ — Model with k order or regression variables, $p = k + 1$ regression coefficients (parameters)

$\beta j, = 0, 1,2,\ldots k$, $\mathbf{Y}$ = n x 1 vector of observed values, $\mathbf{X}$ = n x p matrix of levels of regressor variable(s), β = p x 1 vector of regression coefficients, ε = n x 1 vector of random errors

$(\mathbf{X'X})\hat{\beta} = \mathbf{X'Y}$ — Normal equations

$\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ — Parameter estimates

$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ — Fitted values (mean-estimates) corresponding to observed values, $\mathbf{Y}$

$= \mathbf{X(X'X)}^{-1}\mathbf{X'Y}$ — "Hat matrix" maps vector of observed values into vector of fitted values

$= \mathbf{HY}$

$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I-H})\mathbf{Y}$ — Residuals, vector of residual errors **e**

$\mathbf{Var}(\hat{\beta}) = (\mathbf{X'X})^{-1}\,\sigma^2$ — Variance of $\hat{\beta}$

$\mathbf{Var}(\hat{\mathbf{Y}}) = \mathbf{X}[\mathbf{Var}\,\hat{\beta}\,]\mathbf{X'}$ — Variance of all mean estimates, $\hat{\mathbf{Y}}$

$= \mathbf{X}(\mathbf{X'X})^{-1}\,\mathbf{X'}\,\sigma^2 = \mathbf{H}\,\sigma^2$

$\mathbf{Var}(\hat{\mathbf{Y}}_{\text{pred}}) = (\mathbf{I} + \mathbf{H})\;\sigma^2$ — Variance of prediction, single point future

$\mathbf{Var(e)} = (\mathbf{I} - \mathbf{H})\;\sigma^2$ — Variance of the residual errors

$\mathbf{Y'Y} - n\overline{\mathbf{Y}}^2 = (\hat{\beta}'\mathbf{X'Y} - n\overline{\mathbf{Y}}^2) + (\mathbf{Y'Y} - \hat{\beta}'\mathbf{X'Y})$ — Total sum of squares (corrected) partitioned into errors due to regression + residual, $S_{yy} = SS_R + SS_E$

Results and Discussion

Regression in Matrix Notation

$$R_a^2 = 1 - \frac{SS_E / n-p}{S_{yy} / n-1}$$ Coefficient of determination

$$= 1 - \frac{n-1}{n-p}(1-R^2)$$ Adjusted coefficient of determination

$$\frac{S_{yy}}{n-1} = \frac{SS_R}{k} + \frac{SS_E}{n-p}$$ Mean sums of squares partitioned, corrected, $k = p - 1$, k = variables n = observations, p = coefficients

$$MS_E = \frac{SS_E}{n-(k+1)} = \frac{SS_E}{n-p}$$ the residual mean square s^2, a model-dependent estimate of variance σ^2, where $SSE = \sum_{i=1}^{n} e_i^2 = \mathbf{Y'Y} - \hat{\boldsymbol{\beta}}'\mathbf{X'Y}$

$$T = \frac{\hat{Y}_o - \mu_{Y|1, x_{01}, x_{02}, \ldots, x_{0k}}}{s\sqrt{\mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x}_o}}$$ statistic for construction of 100(1-α)% confidence intervals on the mean (predicted) response $\mu_Y = | 1, x_{01}, x_{o2}, \ldots x_{ok}$ where s^2 is an estimate of the variance σ^2, $\mathbf{x}_o$ or $\mathbf{x'}_0 =$

$[1, x_{01}, x_{02}, \ldots x_{0k}]$ is the condition vector, and where the statistic is T distribution probability determined by degrees of freedom $\upsilon = n - p = n - k - 1$

$$s(\hat{Y}_0) = \sqrt{\mathbf{x}_0'(\mathbf{X'X})^{-1}\mathbf{x}_0 s^2}$$ Conditional standard error on the mean estimate (response)

$$s(\hat{Y}_0 - Y_0) = \sqrt{(\mathbf{1} + \mathbf{x}_0'(\mathbf{X'X})^{-1}\mathbf{x}_0)\, s^2}$$ Conditional standard error for a single point future prediction on an observed response

[1] Walpole, R. E.; Myers, R. H.; Meyers, 1972.
[2] Montgomery, D. C.; Peck, 1992.
[3] Rawlings, J. O. 1988.
[4] Gerald, C. F. 1980.

Discussion

Summarize & Predict V50 Performance of 7039-T6 Armor,

Mean Estimates and Confidence Intervals,

Independent Variable = Areal Density(AD) the Target Weight / Unit Area

$$V50_{i\ 7039\text{-}T6} = \beta_0 + \beta_1 AD_i + \beta_2 AD_i^2 + [\beta_3 AD_i^3],\quad \text{(Mathematica v2.2)}$$

Polynomial Least Square (LS) Regression, Coefficients and Statistics

Projectile		Polynomial Regression, 7039-T6 V50 Ballistic Performance f(AD)								
Type	Obl. (°)	Obs. (n)	R^2_a	s (m/s)	Coefficient Estimates B_0	B_1	B_2	B_3	t Dist. t (1-0.025, ν)	d.f. (ν)
0.50-cal. FSP	0	23	0.99170	32.36	XXX.xxx	-XX.xxxx	0.xxxxxx	–0.00xxxxxx	2.093	19
20-mm FSP	0	44	0.98110	25.76	XXX.xx	-XX.xxx	0.xxxxxx	–0.00xxxxxx	2.021	40
0.30-cal. APM2	0	58	0.98090	19.69	XXX.xxx	X.xxxx	0.0xxxxx	–0.000xxxxx	2.005	54
	30	26	0.99100	14.09	–XX.xxxx	XX.xxxx	–0.xxxxx	0.00xxxxxx	2.074	22
	45	14	0.99087	13.22	XXX.xxx	XX.xxxx	–0.0xxxx	—	2.201	11
0.50-cal. APM2	0	55	0.99180	13.53	XXX.xxx	X.xxxxx	–0.00xx	0.0000xxxx	2.008	51

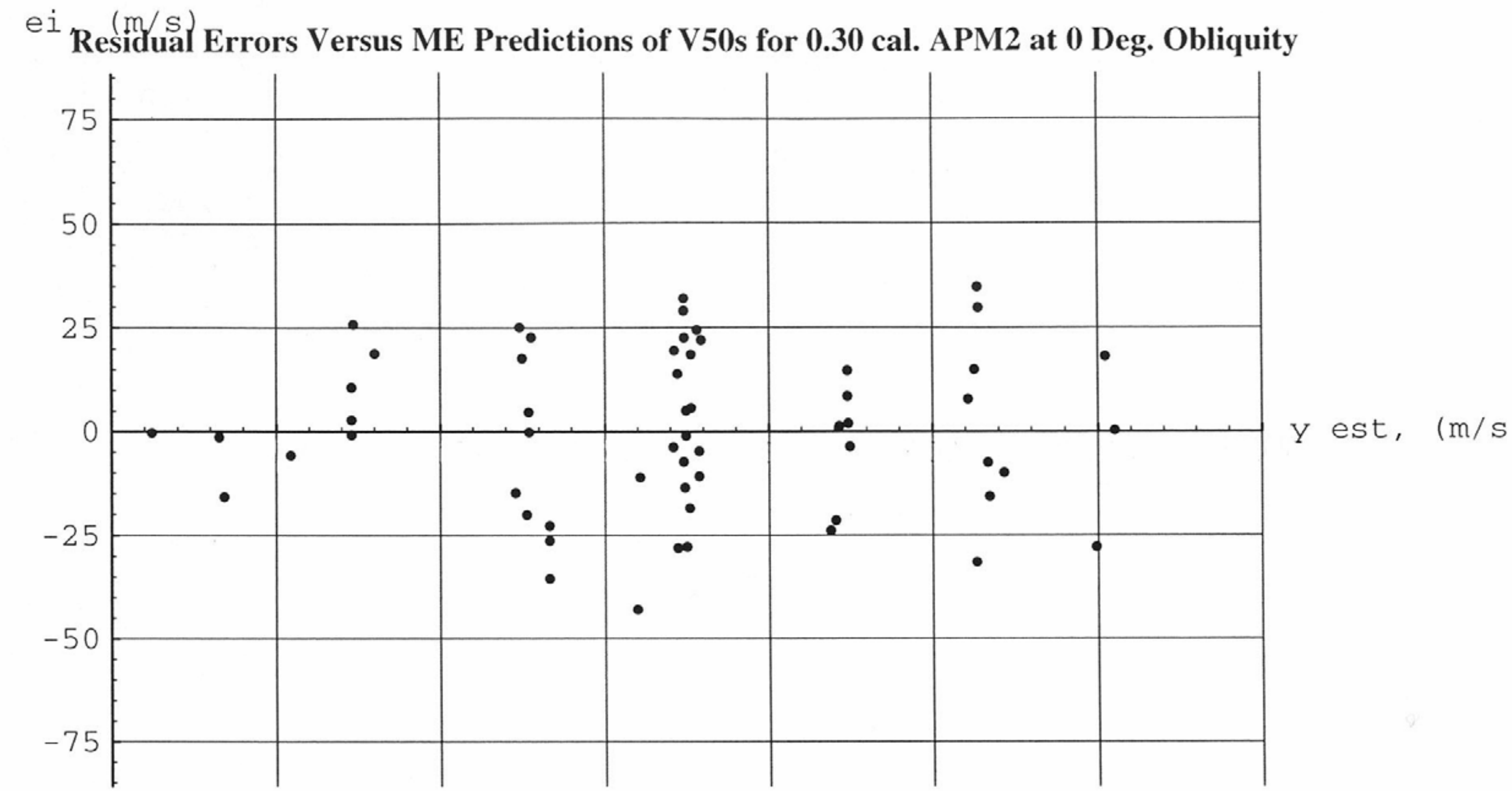
Notes: s = standard error

t = critical values of t-distribution for determination of a two-tailed 95% confidence interval

ν = degrees of freedom

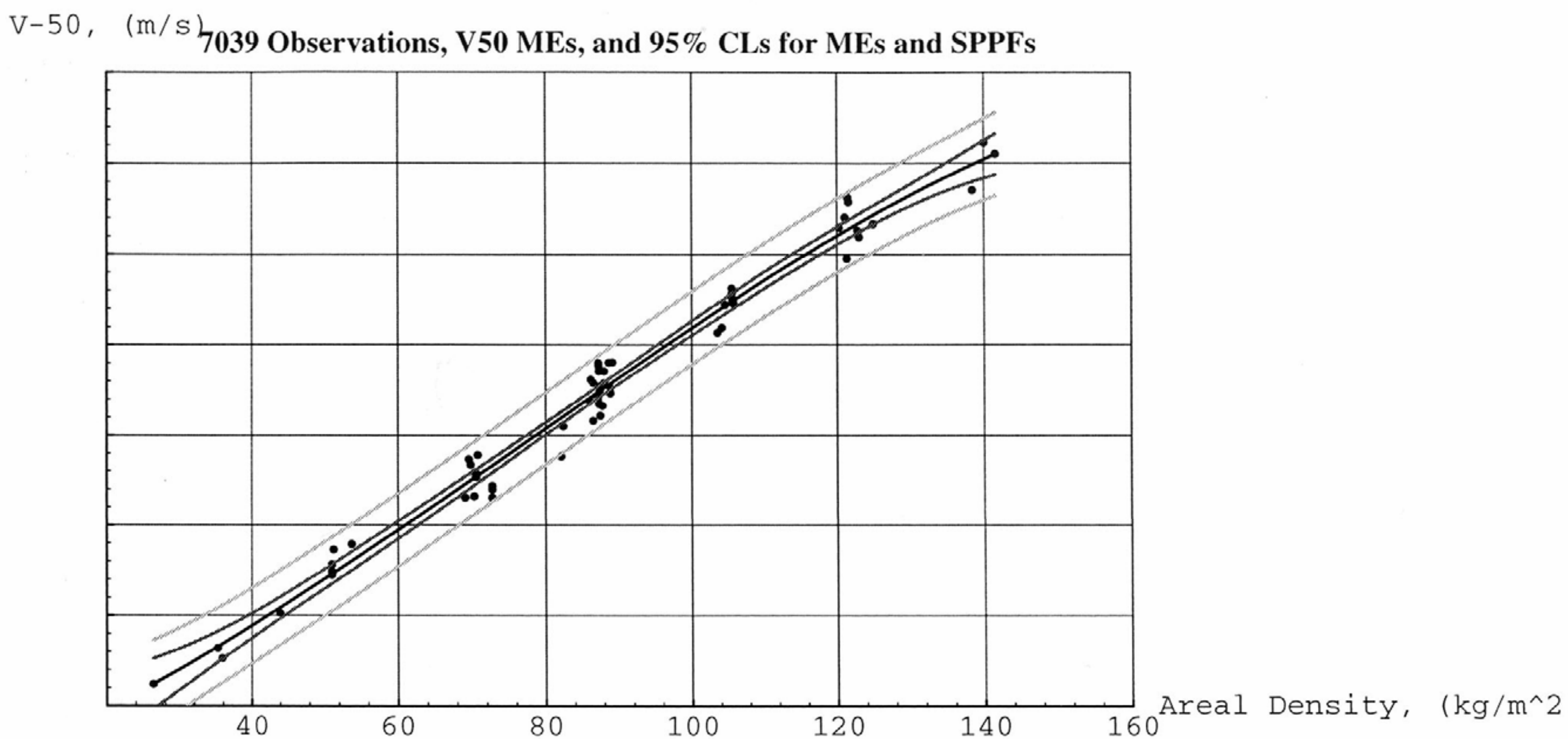
R^2_a = adjusted coefficient of determination

Discussion



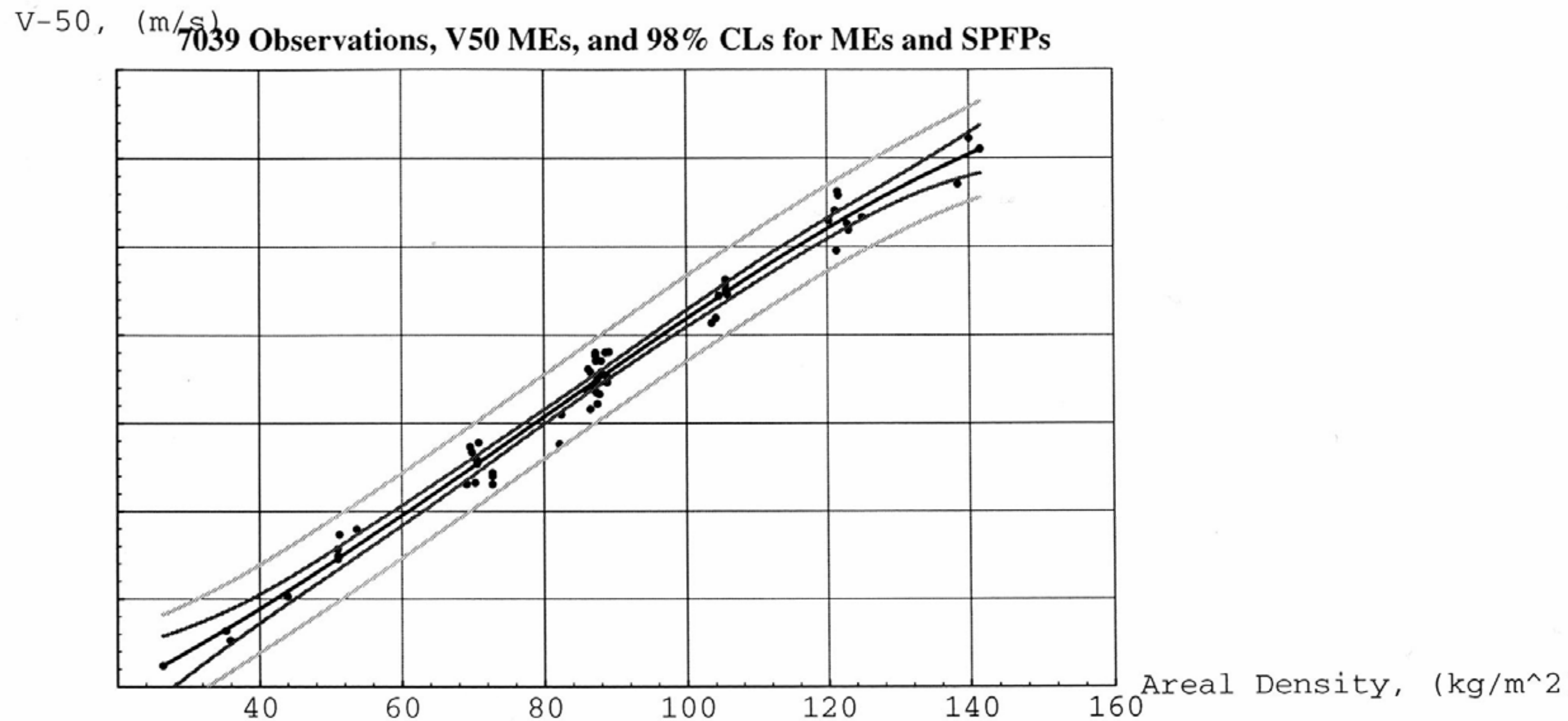
Plot of residuals versus fitted regression values of 7039-T6 V50s

Discussion



7039-T6 V50 observations, n = 58, and 95% CLs for mean estimates (MEs), and single point future predictions (SPFPs)

Discussion

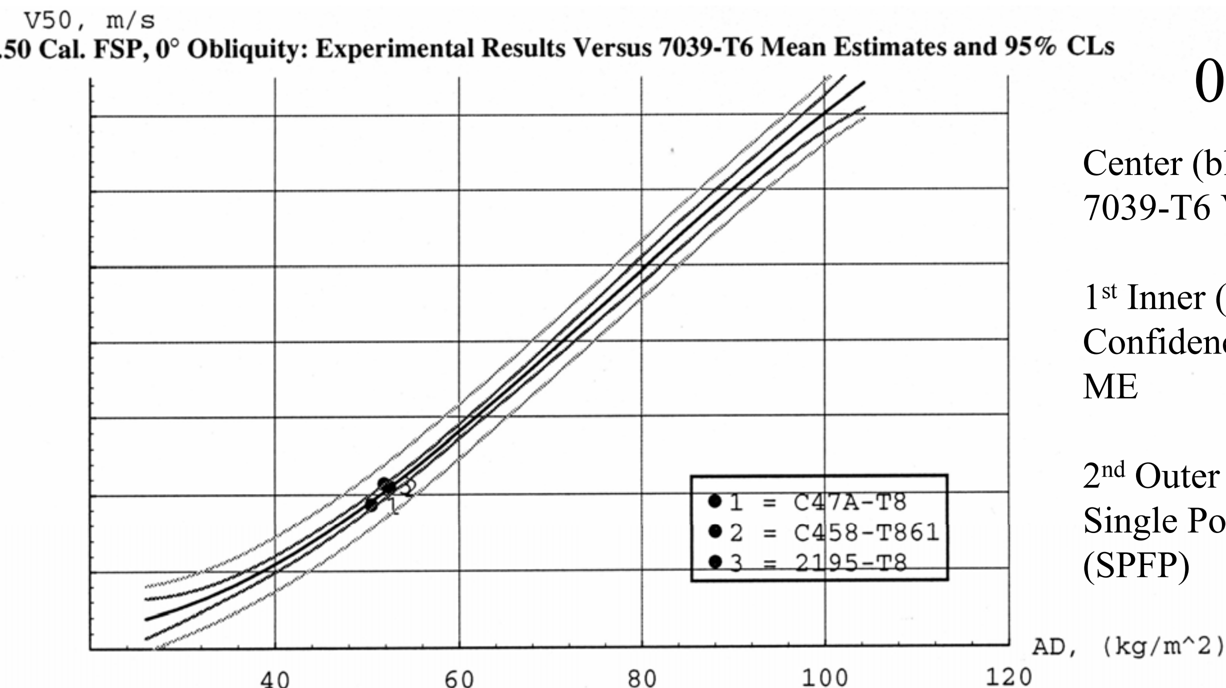


7039-T6 V50 observations, n = 58, and 98% CLs for mean estimates (MEs), and single point future predictions (SPFPs)

Discussion

Comparisons to 7039-T6 V50 Performance: LS Regression

0.50 Cal. FSP, 0° Obliquity: Experimental Results Versus 7039-T6 Mean Estimates and 95% CLs



0.50 cal. FSP, 0°

Center (black) line = Predicted Mean 7039-T6 V50 Estimate, (ME)

1st Inner (gray) lines = 7039-T6 ME Confidence Interval Limits = 95% (CL) ME

2nd Outer (gray) lines = 7039-T6 CL Single Point Future Prediction = 95% CL (SPFP)

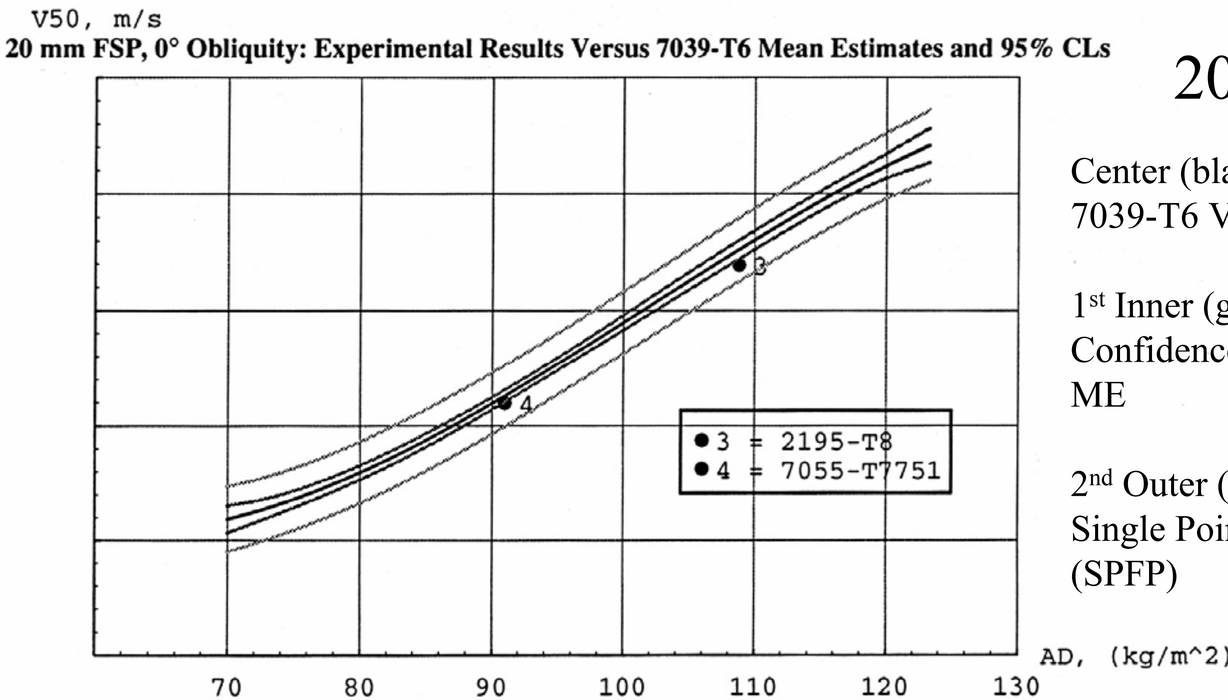
0.50 cal. FSP

Exp. Results Vs 7039-T6 Mean V50 Estimates and Confidence Interval Limits of the Mean and Single Point Future Prediction

Alloy (i)	AD (kg/m2)	V50-VME7039 (m/s)	Conditional s, mean m/s)	Conditional s, pred (m/s)	t	ME 95% CL (± m/s)	SPFP 95% CL (± m/s)
1	50.46	-8.489	9.238	33.65	-0.2523	19.33	70.44
3	51.83	21.83	9.032	33.6	0.6499	18.9	70.32
2	52.43	3.622	8.947	33.57	0.1079	18.73	70.27

Discussion

Comparisons to 7039-T6 V50 Performance: LS Regression



20 mm FSP, 0°

Center (black) line = Predicted Mean 7039-T6 V50 Estimate, (ME)

1st Inner (gray) lines = 7039-T6 ME Confidence Interval Limits = 95% (CL) ME

2nd Outer (gray) lines = 7039-T6 CL Single Point Future Prediction = 95% CL (SPFP)

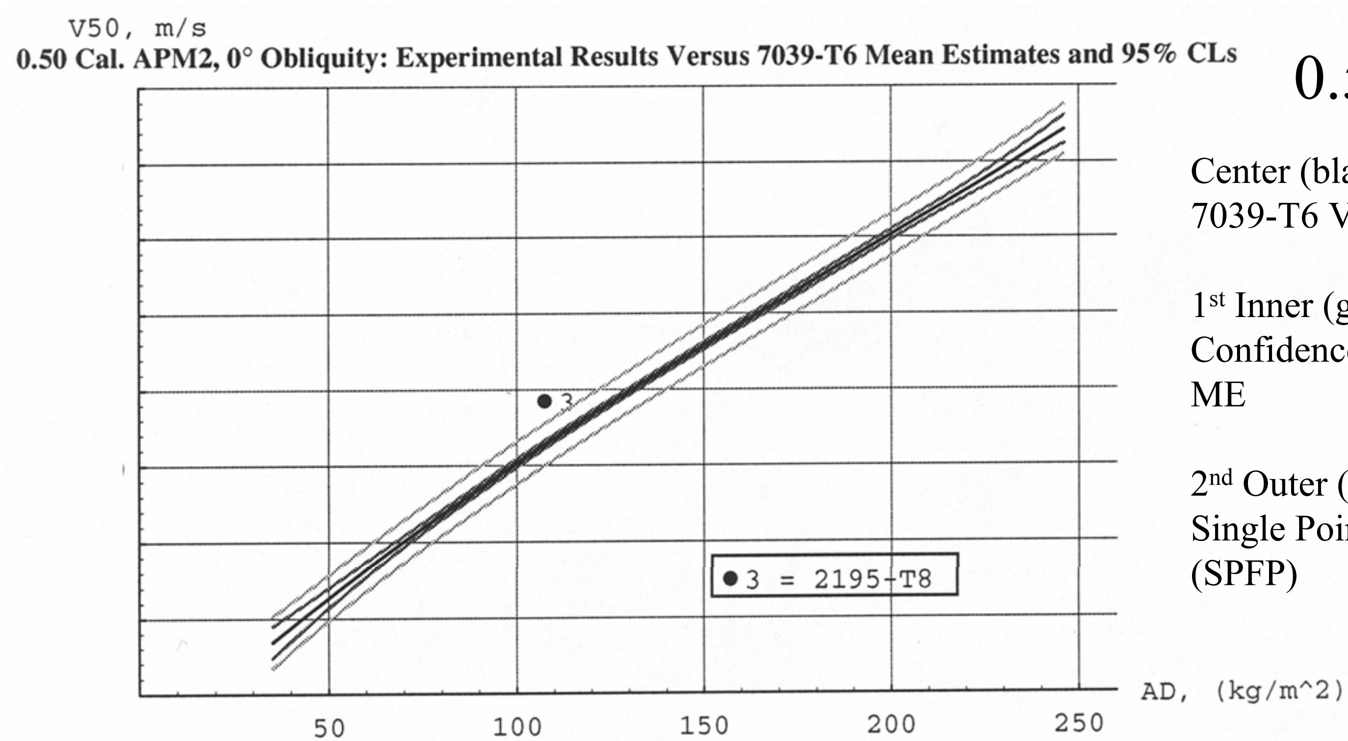
20 mm FSP

Exp. Results Vs 7039-T6 Mean V50 Estimates and Confidence Interval Limits of the Mean and Single Point Future Prediction

Alloy (i)	AD (kg/m2)	V50-VME7039 (m/s)	Conditional s, mean (m/s)	Conditional s, pred (m/s)	t	ME 95% CL (± m/s)	SPFP 95% CL (± m/s)
4.	90.95	-12.39	5.186	26.28	-0.4715	10.48	53.11
3.	108.8	-26.13	7.944	26.96	-0.9691	16.06	54.49

Discussion

Comparisons to 7039-T6 V50 Performance: LS Regression



0.50 cal. APM2, 0°

Center (black) line = Predicted Mean 7039-T6 V50 Estimate, (ME)

1st Inner (gray) lines = 7039-T6 ME Confidence Interval Limits = 95% (CL) ME

2nd Outer (gray) lines = 7039-T6 CL Single Point Future Prediction = 95% CL (SPFP)

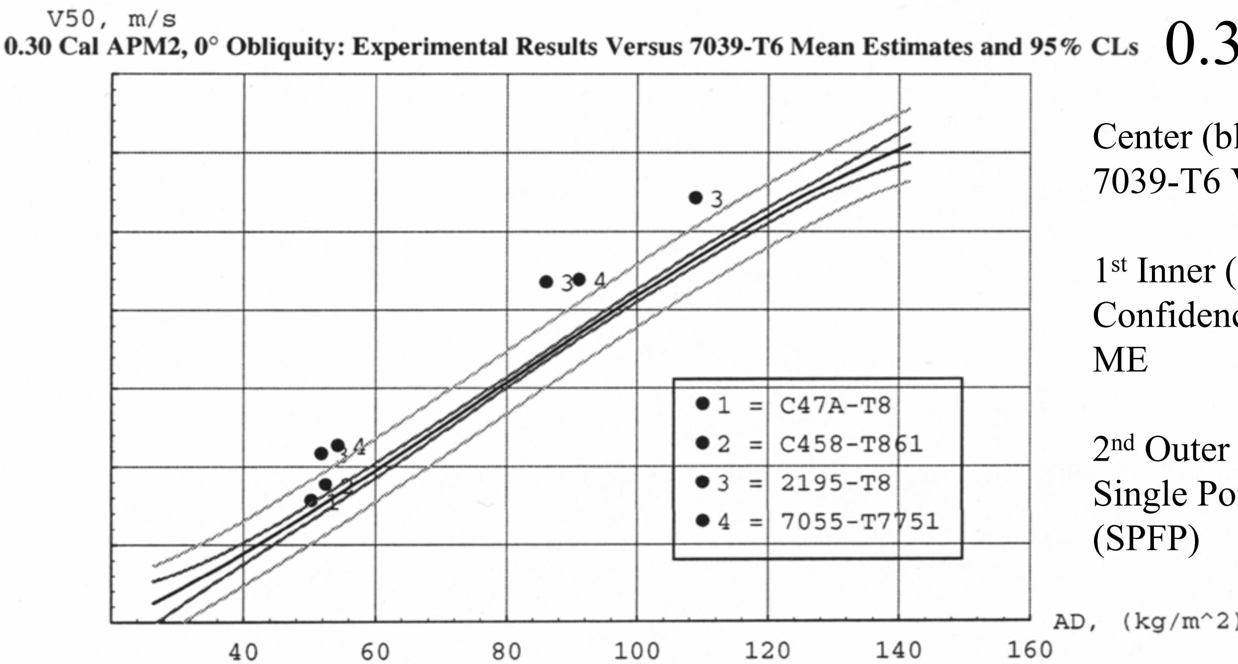
50 cal. APM2

Exp. Results Vs 7039-T6 Mean V50 Estimates and Confidence Interval Limits of the Mean and Single Point Future Prediction

Alloy (i)	AD (kg/m2)	V50-VME7039 (m/s)	Conditional s, mean (m/s)	Conditional s, pred (m/s)	t	ME 95% CL (± m/s)	SPFP 95% CL ± m/s)
3.	107.4	57.76	2.7	13.8	4.186	5.42	27.7

Discussion

Comparisons to 7039-T6 V50 Performance: LS Regression



0.30 cal. APM2, 0°

Center (black) line = Predicted Mean 7039-T6 V50 Estimate, (ME)

1st Inner (gray) lines = 7039-T6 ME Confidence Interval Limits = 95% (CL) ME

2nd Outer (gray) lines = 7039-T6 CL Single Point Future Prediction = 95% CL (SPFP)

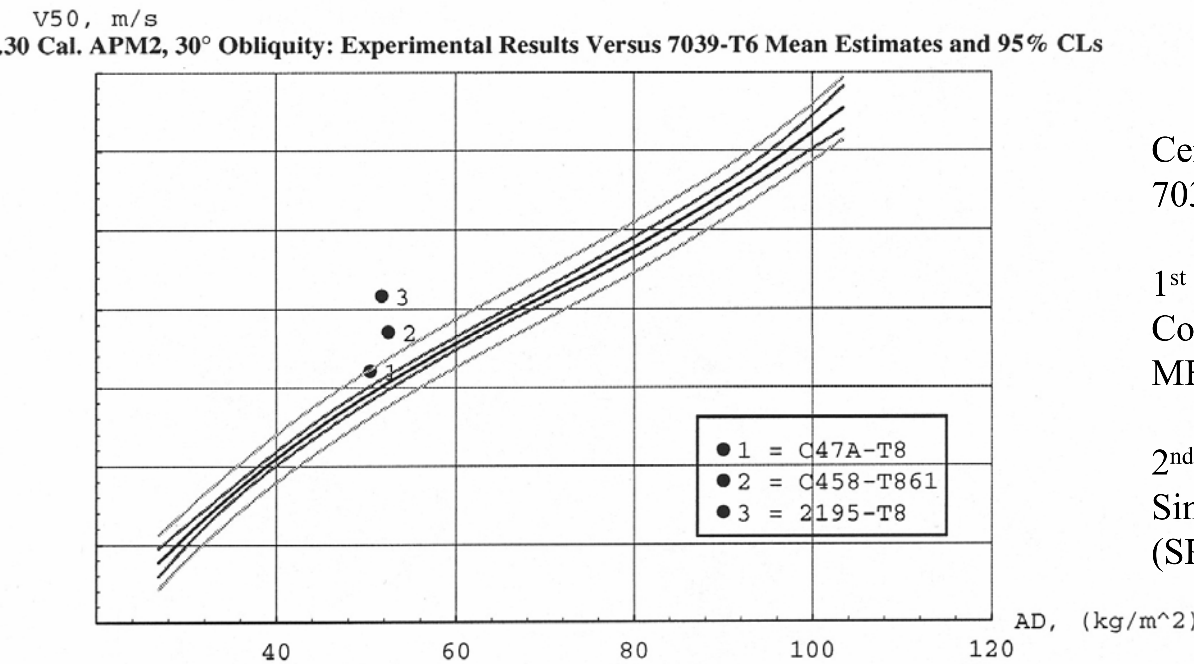
0.30 cal. APM2, 0° Obliquity
Exp. Results Vs 7039-T6 Mean V50 Estimates and Confidence Interval Limits of the Mean and Single Point Future Prediction

Alloy (i)	AD (kg/m2)	V50-VME7039 (m/s)	Conditional s, mean (m/s)	Conditional s, pred (m/s)	t	ME 95% CL (± m/s)	SPFP 95% CL (± m/s)
1	50.26	14.63	5.346	20.4	0.7173	10.72	40.9
3	51.76	67.41	5.264	20.38	3.308	10.55	40.86
2	52.43	23.61	5.231	20.37	1.159	10.49	40.84
4	54.34	63.88	5.146	20.35	3.139	10.32	40.8
3	86.02	94.82	3.253	19.95	4.752	6.522	40.01
4	91.03	70.06	3.385	19.98	3.507	6.787	40.05
3	108.90	77.14	4.556	20.21	3.818	9.134	40.51

Discussion

Comparisons to 7039-T6 V50 Performance: LS Regression

0.30 Cal. APM2, 30° Obliquity: Experimental Results Versus 7039-T6 Mean Estimates and 95% CLs



0.30 cal. APM2, 30°

Center (black) line = Predicted Mean 7039-T6 V50 Estimate, (ME)

1st Inner (gray) lines = 7039-T6 ME Confidence Interval Limits = 95% (CL) ME

2nd Outer (gray) lines = 7039-T6 CL Single Point Future Prediction = 95% CL (SPFP)

0.30 cal APM2, 30 deg

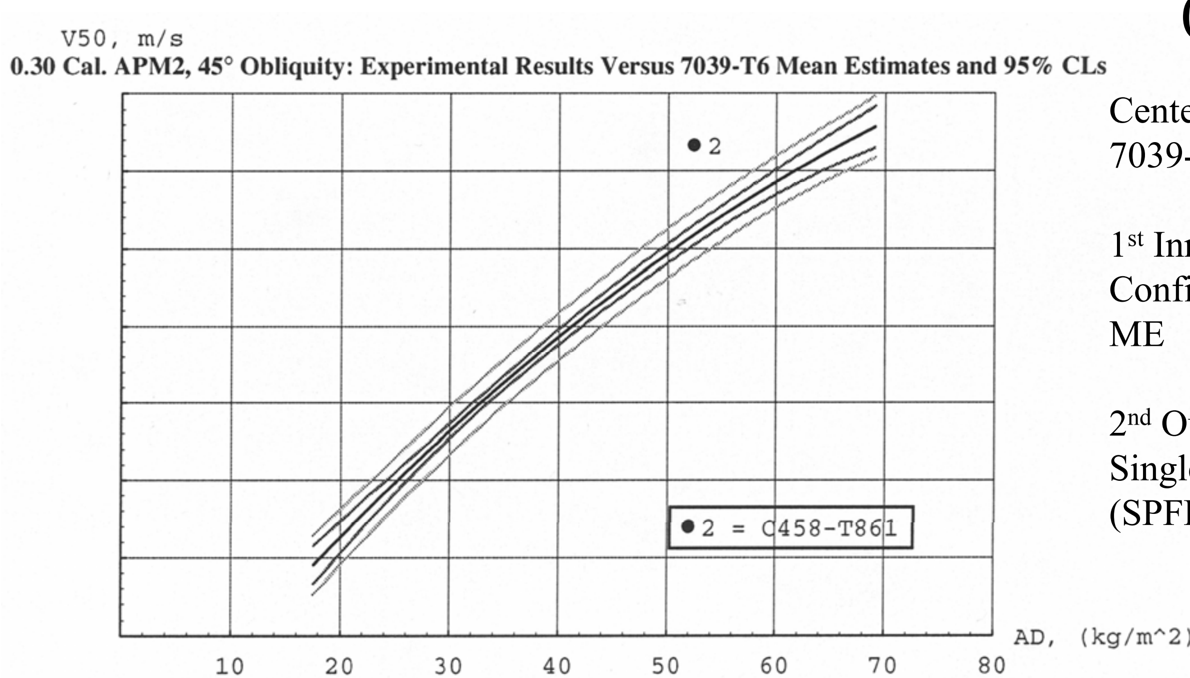
Exp. Results Vs 7039-T6 Mean V50 Estimates and Confidence Interval Limits of the Mean and Single Point Future Prediction

Alloy (i)	AD (kg/m2)	V50-VME7039 (m/s)	Conditional s, mean (m/s)	Conditional s, pred (m/s)	t	ME95% CL (± m/s)	SPFP 95% CL (± m/s)
1	50.4	29.01	4.582	14.82	1.958	9.502	30.73
3	51.76	115.3	4.506	14.79	7.792	9.345	30.68
2	52.5	64.12	4.457	14.78	4.339	9.244	30.65

Discussion

Comparisons to 7039-T6 V50 Performance: LS Regression

0.30 cal. APM2, 45°



Center (black) line = Predicted Mean 7039-T6 V50 Estimate, (ME)

1st Inner (gray) lines = 7039-T6 ME Confidence Interval Limits = 95% (CL) ME

2nd Outer (gray) lines = 7039-T6 CL Single Point Future Prediction = 95% CL (SPFP)

0.30 cal. APM2, 45 deg
Exp. Results Vs 7039-T6 Mean V50 Estimates and Confidence Interval Limits of the Mean and Single Point Future Prediction

Alloy	AD	V50-VME7039	Conditional s, mean	Conditional s, pred	t	ME 95% CL	SPFP 95% CL
(i)	(kg/m2)	(m/s)	(m/s)	(m/s)		(± m/s)	(± m/s)
2	52.4	116.1	5.762	14.42	8.047	12.68	31.74

Discussion

Oblique-V50 Results:

Regression-Estimate Comparisons to V50s of 0°-ADs Up-Scaled by LOS Thickness / Target Thickness (secant θ of Target Obliquity)

Proj. Obl. θ	Experimental Alloys: Oblique Results and Predicted 0° Obliquity V50s					
	Alloy	Thick. (mm)	AD, θ (kg/m^2)	Equiv. LOS at θ = 0° Thick. (mm)	Equiv. LOS at θ = 0° AD (kg/m^2)	Improvement, $V50_\theta$–LOS 0°V50 (m/s)
30	C47A	19.08	50.40	22.03	58.20	21.3
30	2195	19.11	51.76	22.07	59.77	55.8
30	C458	19.94	52.50	23.02	60.62	48.9
45	C458	19.90	52.40	28.14	74.10	235.9

Proj. Obl. θ	7039-T6: Oblique and 0° Regression Predicted MEs, Reference Data		
	AD, θ (kg/m^2)	Equiv. LOS at θ = 0° AD (kg/m^2)	Improvement, $V50_\theta$–LOS 0°V50 (m/s)
30	50.40	58.20	7.0
30	51.76	59.77	7.9
30	52.50	60.62	8.4
45	52.40	74.10	143.4

Proj. Obl. θ	2519: Oblique Point Estimates and Simple Linear 0° Obliquity MEs, Reference Data			
	AD, θ° PE		Equiv. LOS at θ = 0° AD	Improvement, $V50_\theta$–LOS 0°V50
30	55.62		64.24	–31.7

Notes: For 7039, LOS V50s were estimated from LOS ADs by the 0.30-cal. APM2, 0° regression equation

For experimental alloys, LOS V50s were estimated at 0° obliquity with LOS ADs and the 0.30-cal. APM2, 0° regression equation with b_0 = coefficient adjusted upward 14.63 for C47A, 23.61 for C458, and 67.41 for 2195 to fit V50 improvements of the 19- to 20-mm nominal thickness targets.

Discussion

Experimental Oblique-Impact V50 vs. (LOS t / t) Scaled, 0° V50

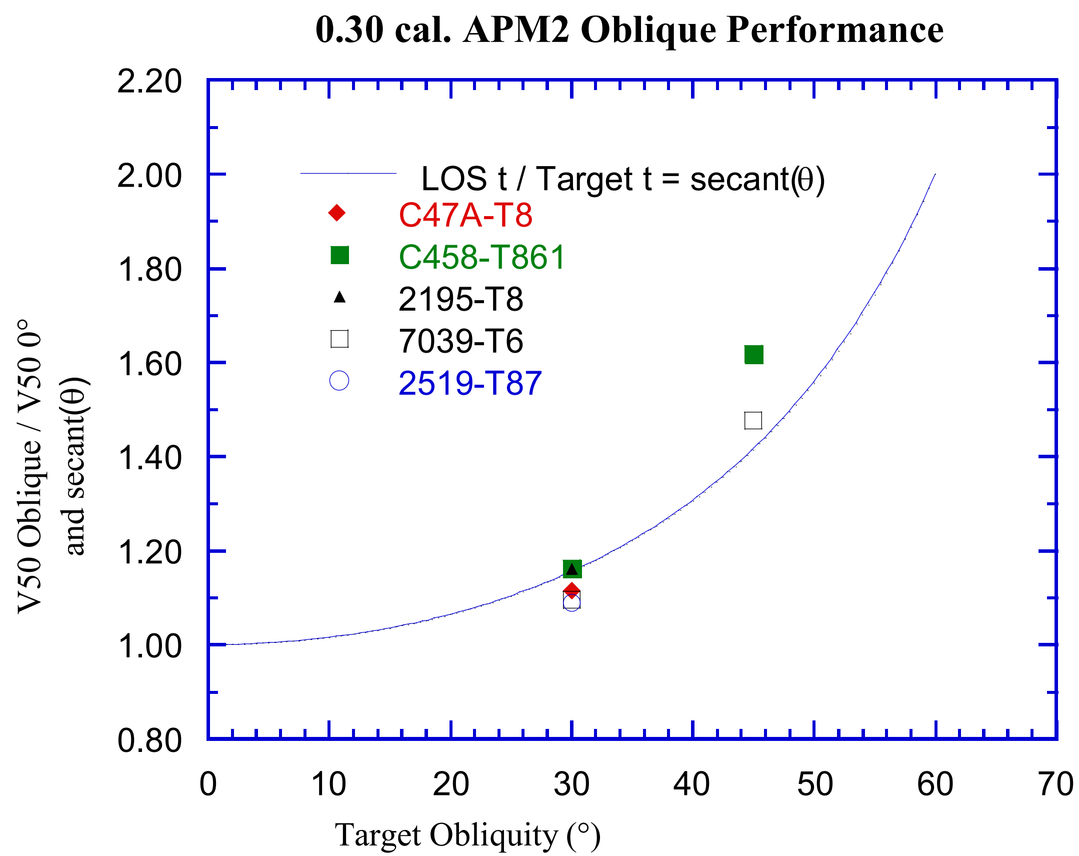
Experimental Results, 0.30-cal. APM2 (Figure 5c–d)						V50, θ Oblique vs. θ = 0° Sec(θ) Scaled Exp.		
Obliquity θ Deg. 0 and θ°	Exper. Alloy	θ = 0° Exp. Data		θ = 30 or 45° Exp. Data		sec θ	sec x 0°	Excess
		Thick. (mm)	AD (kg/m^2)	Thick. (mm)	AD (kg/m^2)	LOSt/ target t	AD (kg/m^2)	V50θ–V500°secθ (m/s)
0 and 30°	C47A	19.02	50.26	19.08	50.40	1.155	58.20	–21.7
0 and 30°	2195	19.11	51.76	19.11	51.76	1.155	59.77	4.0
0 and 30°	C458	19.91	52.43	19.94	52.50	1.155	60.62	4.4
0 and 45°	C458	19.91	52.43	19.90	52.40	1.414	74.10	117.6
Regression of 7039-T6 vs. 0.30-cal.						V50 ME, θ Oblique vs. θ = 0° Sec(θ) Scaled ME.		
Obliquity θ Deg. 0 and θ°	AD Point	θ = 0° Regression		θ = 30, 45° Obl. Regr.		sec θ	sec x 0°	Excess
			AD (kg/m^2)		AD (kg/m^2)	LOSt/ target t	AD (kg/m^2)	V50θ–V500°secθ (m/s)
0 and 30°	C47A		50.26		50.40	1.155	58.20	–33.82
0 and 30°	2195		51.76		51.76	1.155	59.77	–33.45
0 and 30°	C458		52.43		52.40	1.155	60.62	–32.41
0 and 45°	C458		52.43		52.40	1.414	74.10	34.87
Simple Linear and Point Estimates, 2519 vs. 0.30-cal. APM2 Ref. Data						V50 PE, θ Oblique vs. θ = 0° Sec(θ) Scaled SLE		
θ Deg. 0 and θ°	AD Point	θ = 0° SLE Mean		θ = 30° (4 PE Mean)		sec θ	sec x 0°	Excess
			AD (kg/m^2)		AD (kg/m^2)		AD (kg/m^2)	V50θ–V500°secθ (m/s)
0 and 30°	2519		55.62		55.62	1.155	64.23	–37.8
2519 vs. 0.50-cal. APM2, Ref. Data						V50, θ Oblique vs. θ = 0° Sec(θ) Scaled		
θ Deg. 0 and θ°	AD Point	θ = 0° Ref. Data		θ = 45° Ref. Data			sec x 0°	Excess
			AD (kg/m^2)		AD (kg/m^2)	sec θ	AD (kg/m^2)	V50θ–V500°secθ (m/s)
0 and 45°	2519		108.8		109.3	1.414	154.5	24.4

Notes:

ME = LS Regression Mean Estimate, SLE = Straight Line Estimate), PE = (4) Point Estimate

Discussion

Experimental Oblique-Impact V50 vs. V50 0° and (LOS t / Target t)



- 2195-T8 and C458-T861 V50 performance and mass efficiency improve with target obliquity, C4585 V50 Oblique 45°/ V50 0° = 1.62 versus sec(θ) =1.414

Conclusions

- Al-Li Alloys: Best Specific Strengths
- Protection versus FSP, 0°, approximate 7039-T6 Average
- Superior Protection Vs AP Projectile:
  - -V50 Performance Improves with Target Strength
  - -Al-Li V50 & M_e Performance Improves w Oblique Impact
    - * Exceeds 7039 or 2519 performance
    - * Material Parameters and Failure Mode of 2195 and C458 Enhance Protection versus AP projectile
    - *Reduced sensitivity failure mode to increased V
- Damage-Tolerant, High Shot Density,

  Multiple Impact Capabilty

Interval Estimates for Probabilities of Non-Perforation Using a Generalized Pivotal Quantity

David W. Webb
U. S. Army Research Laboratory
Aberdeen Proving Ground, MD 21005

Abstract

A generalized pivotal quantity is developed that yields confidence intervals for the cumulative distribution function (CDF) at a specific value when the underlying distribution is assumed to be normal. This problem is similar to the development of a tolerance interval, and, unsurprisingly, its solution involves the non-central t distribution. Generalized confidence bands for a normal CDF follow easily. Military applications include vulnerability and lethality assessment, for example, interval estimation for the probability of non-perforation against homogeneous armored targets.

Introduction

An engineer at the U.S. Army Research Laboratory at Aberdeen Proving Ground approached the author in the spring of 2004 requesting help in estimating the probability that a projectile would not penetrate beyond the thickness of an armor plate. That is, the client wanted an estimate for the probability of non-perforation. This estimate was to be obtained from sample data of depths of penetration into homogeneous armor of "infinite" thickness.

To model this phenomena, one could let X be defined as the penetration depth of a random projectile and assume that X is a normally distributed random variable with mean μ and variance σ^2. If x_0 is the thickness of the plate for which the probability of non-perforation estimate is desired, then the client seeks an estimate for $P(X < x_0)$, or equivalently $\Phi\left(\frac{x_0 - \mu}{\sigma}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

Using the sample mean and the sample standard deviation from the observations, the plug-in estimate, $\Phi\left(\frac{x_0 - \bar{x}}{s}\right)$, serves as an adequate point estimate for the probability of non-perforation. Of course, point estimates yield no information on the error of estimation. What we'd prefer to report is a confidence interval for $\Phi\left(\frac{x_0 - \mu}{\sigma}\right)$ so that one can say, e.g., "… with 95% confidence, the probability of non-perforation is between some lower confidence limit (LCL) and upper confidence limit (UCL)." Interval estimation of $\Phi\left(\frac{x_0-\mu}{\sigma}\right)$ is not a new problem – see Owen & Hua (1977), Odeh & Owen (1980), Hahn & Meeker (1990), or Patel & Read (1996). However, in this paper we will examine this problem from the perspective of generalized confidence intervals, a technique that allows one to obtain confidence intervals for *any* function of μ and σ^2.

Classical and Generalized Confidence Intervals

In the ensuing discussion, it is imperative that we make the notational distinction between a random variable and its observed value. An observable random variable (either scalar or vector) will always be denoted by a capital letter, for example, $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$; its observed value will be denoted using small case, for example, $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$.

In recalling the classical definition of confidence interval, let $\vec{X}=\{X_1, X_2, \ldots, X_n\}$ be a random sample of size *n*, and θ be a scalar parameter of interest. It should be noted that θ may actually be a function of other parameters. If there exists statistics $L(\vec{X})$ and $U(\vec{X})$ whose distributions are free of any unknown parameters, and that satisfy $1-\alpha=P(L(\vec{X})<\theta<U(\vec{X}))$, then a $(1-\alpha)100\%$ confidence interval for θ is given by $(L(\vec{x}), U(\vec{x}))$.

One technique that is used to construct a classical confidence interval for a parameter θ is the pivotal method. It is comprised of the following four steps: 1) Obtain a pivot, that is, a random quantity whose distribution does not depend upon any unknown parameters. 2) Write a simple probability statement that bounds the pivot. 3) Invert this into a probability statement that bounds θ. 4) If the bounds for θ do not depend on any unknown parameters, one can use them to obtain a confidence interval for θ.

We note that a pivot is a function of:

a) the random data (typically a set of sufficient statistics),
b) the unknown parameter of interest, and
c) perhaps other nuisance parameters.

As long as the cumulative distribution function associated with the observations is continuous, a pivot will exist (Mood, Graybill & Boes, 1974). However, for many problems, no pivot exists which can be inverted to form a confidence interval. In many cases this is due to the presence of nuisance parameters. In such instances, approximate methods, or some other technique for constructing a confidence interval is needed.

One relatively new technique involves the use of generalized pivots (Weerahandi, 1993). A generalized pivot is a function of the same three arguments as a traditional pivot *plus* the observed data. A generalized pivot can be written as $R(\vec{X},\vec{x},\theta,\eta)$, reminding us of its four arguments; and it must satisfying the following conditions:

a) the distribution of $R(\vec{X},\vec{x},\theta,\eta)$ is free of any unknown parameters, and
b) the observed value $r=R(\vec{x},\vec{x},\theta,\eta)$ is free of the nuisance parameter η.

Furthermore, we say that $R(\bar{X}, \bar{x}, \theta, \eta)$ is a <u>generalized pivot for θ</u> if $r = \theta$. The percentiles of a generalized pivot for θ yield generalized confidence limits/bounds for θ.

<u>Constructing Generalized Pivots</u>

Since their introduction, generalized confidence intervals have been utilized on a limited basis, in part because of the lack of a clear method for constructing generalized pivots. Even in his seminal paper Weerahandi wrote "The problem of finding an appropriate pivotal quantity is a nontrivial task", and "… the construction of pivotals requires some intuition." Recently, a seven-step algorithm has been proposed (Iyer & Patterson, unpub.) that elucidates how generalized pivots and confidence intervals can be obtained. Their algorithm is given here, along with its implementation towards an interval estimate for $\theta = \Phi\left(\frac{x_0 - \mu}{\sigma}\right)$, the probability of non-perforation.

Step 1: Find a set of independent, sufficient statistics for the sample.

Assuming a normal population, the sample mean, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and the sample variance, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ are sufficient statistics.

Step 2: From these, find a same-sized set of statistics whose distributions are independent of the unknown parameters.

We have $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Step 3: Solve for the unknown parameters in terms of the statistics in Step 2.

With some simple algebraic manipulation, one obtains $\mu = \overline{X} - ZS\sqrt{\frac{n-1}{nV}}$ and $\sigma = S\sqrt{\frac{n-1}{V}}$.

Step 4: Substitute the expressions for the unknown parameters in Step 3 into θ.

Starting with the parameter of interest, θ, we make substitutions for μ and σ, then simplify:

$$\theta = \Phi\left(\frac{x_0 - \mu}{\sigma}\right) = \Phi\left(\frac{x_0 - \left(\overline{X} - ZS\sqrt{\frac{n-1}{nV}}\right)}{S\sqrt{\frac{n-1}{V}}}\right) = \Phi\left(\frac{x_0 - \overline{X}}{S\sqrt{\frac{n-1}{V}}} + \frac{Z}{\sqrt{n}}\right).$$

Step 5: Substitue the (random) sufficient statistics with their observed values.

In the previous expression for θ, the sufficient statistics appear in the first addend. After substituting the observed values, we have the random variable $\Phi\left(\frac{x_0-\bar{x}}{s\sqrt{\frac{n-1}{V}}}+\frac{Z}{\sqrt{n}}\right)$. Notice that because of the substitution, this expression is no longer equated to θ. But more importantly, note that this random variable has a distribution that is independent of either μ or σ^2.

Step 6: Substitute the remaining random terms with their sufficient-statistic based equivalents. Finally, this is the generalized pivot for θ.

Denoting this generalized pivot by *R*, one obtains $R=\Phi\left(\frac{x_0-\bar{x}}{s\sqrt{n-1}}\sqrt{\frac{(n-1)S^2}{\sigma^2}}+\frac{1}{\sqrt{n}}\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)$. *R* meets the definition of a generalized pivot for θ since it has the same parameter-free distribution as $\Phi\left(\frac{x_0-\bar{x}}{s\sqrt{\frac{n-1}{V}}}+\frac{Z}{\sqrt{n}}\right)$, and its observed value is

$$r=\Phi\left(\frac{x_0-\bar{x}}{s\sqrt{n-1}}\sqrt{\frac{(n-1)s^2}{\sigma^2}}+\frac{1}{\sqrt{n}}\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)=\Phi\left(\frac{x_0-\mu}{\sigma}\right)=\theta .$$

Step 7: The percentiles of the generalized pivot form generalized confidence limits (or confidence bounds) for θ.

In general, these percentiles may be obtained through Monte-Carlo simulation. For example, we know that $R=\Phi\left(\frac{x_0-\bar{x}}{s\sqrt{n-1}}\sqrt{\frac{(n-1)S^2}{\sigma^2}}+\frac{1}{\sqrt{n}}\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)$ has the same distribution as $\Phi\left(\frac{x_0-\bar{x}}{s\sqrt{\frac{n-1}{V}}}+\frac{Z}{\sqrt{n}}\right)$. Therefore, one could generate, and order a "large" sample of values of *R*, e.g., $R_{(1)}, R_{(2)}, \ldots, R_{(10000)}$. Then a 95% two-tailed generalized confidence interval is $\left(\frac{R_{(250)}+R_{(251)}}{2}, \frac{R_{(9750)}+R_{(9751)}}{2}\right)$.

Equating the Generalized and Classical Confidence Intervals for Probability of Non-Perforation

In some cases, the percentiles of the generalized pivot may be expressed in closed form. When a conventional confidence interval exists (for example, a confidence interval for the mean of a normal random variable), the generalized confidence interval will reduce to it. In the following exercise, we will show how the generalized confidence interval for $\theta = \Phi\left(\frac{x_0 - \mu}{\sigma}\right)$ is equivalent to the interval derived by Owen & Hua.

Consider a $(1-\alpha)100\%$ lower confidence bound for $\theta = \Phi\left(\frac{x_0 - \mu}{\sigma}\right)$. It is defined in Step 7 to be that value, B_L, for which $1-\alpha = P(B_L \le R)$. Recalling the distributional equivalent of R discussed in Step 6, we have.

$$1-\alpha = P\left(B_L \le \Phi\left(\frac{x_0 - \bar{x}}{s\sqrt{\frac{n-1}{V}}} + \frac{Z}{\sqrt{n}} \right) \right).$$

Algebraically manipulating this equation, one obtains

$$1-\alpha = P\left(\frac{\bar{x} - x_0}{s/\sqrt{n}} \le \frac{Z - \sqrt{n}\Phi^{-1}(B_L)}{\sqrt{V/(n-1)}} \right)$$

Notice that the right side of the inequality is a non-central t random variable with non-centrality parameter $-\sqrt{n}\Phi^{-1}(B_L)$ and $n-1$ degrees of freedom. However, a non-central t random variable with non-centrality parameter $-\sqrt{n}\Phi^{-1}(B_L)$ is the mirror image of a non-central t random variable with non-centrality parameter $\sqrt{n}\Phi^{-1}(B_L)$ (Johnson and Kotz, 1970). Therefore,

$$1-\alpha = P\left(T_{n-1,\sqrt{n}\Phi^{-1}(B_L)} \le \frac{x_0 - \bar{x}}{s/\sqrt{n}} \right).$$

But this probability is equivalent to the cumulative distribution function of a non-central t random variable with non-centrality parameter $\sqrt{n}\Phi^{-1}(B_L)$ and $n-1$ degrees of freedom, evaluated at $\frac{x_0 - \bar{x}}{s/\sqrt{n}}$. This is the same result as proven by Owen and Hua. The value of the non-centrality parameter that satisfies this probability statement (and in turn produces the desired lower confidence bound, B_L) must be solved using numerical methods, e.g., the bisection method.

Application

Fifteen projectiles ($n=15$) are fired into armor plates to record the depth of penetration. The results in ascending order are given in the table below. Find an estimate for the probability of non-perforation if the production armor is to be 60 units thick. (Note: these data are not actual, but have been contrived for illustrative purposes.)

Sample Data

29.4	46.1	47.5	52.7	57.6
34.6	46.2	50.9	55.9	60.8
43.3	46.4	52.7	56.4	69.5

The sample mean is $\bar{x}=50$ and sample standard deviation is $s=10$. Therefore, the plug-in point estimate is $P(X<60)=\Phi\left(\frac{60-\bar{x}}{s}\right)=\Phi\left(\frac{60-50}{10}\right)=\Phi(1)=.841$.

A 95% generalized confidence interval is (B_L, B_U) where B_L satisfies $.975=P\left(T_{14,\sqrt{15}\Phi^{-1}(B_L)} \le \frac{60-50}{10/\sqrt{15}}\right)=P\left(T_{14,\sqrt{15}\Phi^{-1}(B_L)} \le \sqrt{15}\right)$ and B_U satisfies $.025=P\left(T_{14,\sqrt{15}\Phi^{-1}(B_L)} \le \sqrt{15}\right)$. Using the bisection method in MATLAB®, we obtain $B_L=.642$ and $B_U=.947$.

By allowing the thickness of the armor to vary, one can generate confidence bands for the normal cumulative distribution function (see Table 1). Perhaps of more importance to the armor design engineer would be a plot that shows the relationship between armor thickness and a one-sided lower confidence bound for the probability of non-perforation (see Table 2). Such a chart would allow the engineer to choose that thickness which offers the desired level of protection against perforation by an enemy projectile.

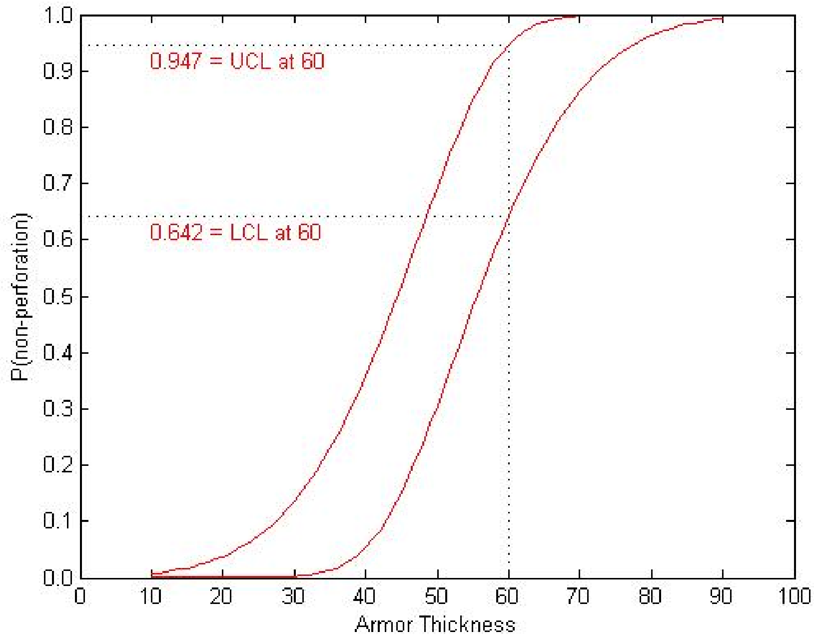


Figure 1. 95% confidence bands for the probability of non-perforation.

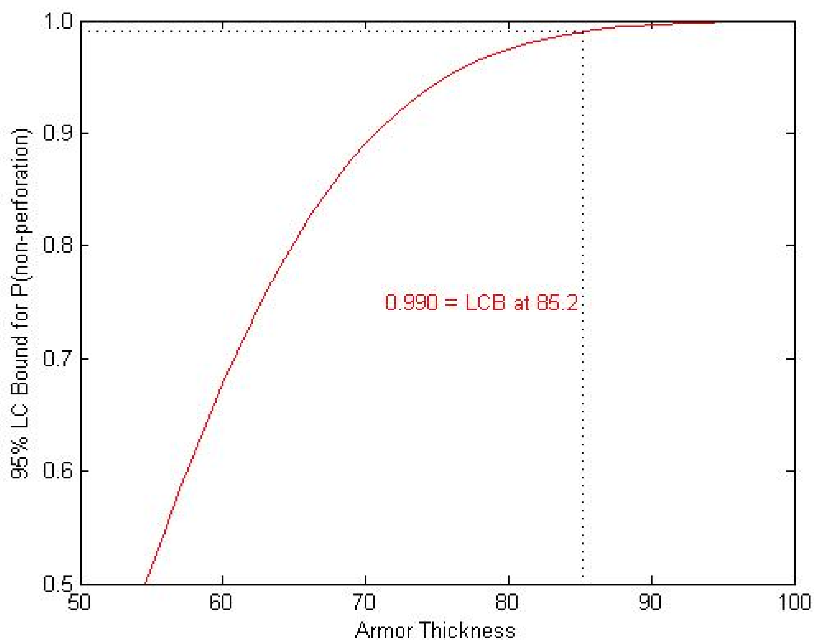


Figure 2. Armor thickness versus 95% lower confidence bound for the probability of non-perforation. For example, a thickness of 85.2 yields a high level (at least 99%) of protection against perforation.

Concluding Remarks

The beauty of this theory is that a generalized confidence interval can be constructed for any function of the normal parameters. In particular, if the parameter of interest is some function of the normal parameters, $\theta = g(\mu, \sigma)$, then it can be shown that a generalized pivot for θ is $R = g\left(\bar{x} - (\bar{X} - \mu)\frac{s}{S}, \frac{s\sigma}{S} \right)$. For example, Weerahandi (2004) discusses estimation of $\theta = \dfrac{\mu + \sigma}{\mu^2 + \sigma^2}$; and Iyer and Patterson (unpub.) derive interval estimates for $\theta = P(a < X < b) = \Phi\left(\dfrac{b - \mu}{\sigma}\right) - \Phi\left(\dfrac{a - \mu}{\sigma}\right)$.

On the surface, it may appear that generalized confidence intervals hold great potential for solving complex estimation problems. However, it is important to understand that the actual coverage probability of a generalized confidence interval may not necessarily equal $1 - \alpha$ when the percentiles of the pivot have no closed-form solution. The actual coverage probability may be influenced by nuisance parameters. A detailed simulation study is necessary to evaluate the (approximate) coverage probability, and hence the quality of the generalized confidence interval.

References

Hahn G. J., and Meeker, W. Q. (1991). *Statistical Intervals: A Guide for Practitioners*, New York: John Wiley & Sons.

Iyer, H. K. and Patterson, P. D. (unpublished). A recipe for constructing generalized pivotal quantities and generalized confidence intervals.

Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions – 2*, New York: Houghton Mifflin.

Mood, A., Graybill, F., and Boes, D. (1974). *Introduction to the Theory of Statistics*, 3rd ed., New York: McGraw-Hill.

Odeh, R. E. and Owen, D. B. (1980), *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*, New York: Marcel Dekker.

Owen, D. B., and Hua, T. A. (1977). Tables of confidence limits on the tail area of the normal distribution". *Communications in Statistics – Simulation and Computation B*, 6, 285-311.

Patel, J. K., and Read, C. B. (1996). *Handbook of the Normal Distribution,* 2nd ed., New York: Marcel Dekker.

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association*, **88**, 899-905.

Weerahandi, S. (2004). *Generalized Inference in Repeated Measures*, New York: John Wiley & Sons.

Statistical Tests for Bullet Lead Comparisons

NRC Report to the FBI

Karen Kafadar

University of Colorado at Denver

kk@math.cudenver.edu

`http://math.cudenver.edu/~kk`

Clifford H. Spiegelman, Texas A&M

Acknowledgements:

Bullet Lead Committee Members; M. Cohen, NRC

OUTLINE

1. Introduction and Background
   Charge to *Committee on Scientific Assessment of Bullet Lead Elemental Composition Comparison*
2. Bullet Manufacturing Process
3. FBI tests on individual elements
4. Available data sets
5. False Positive Probabilities
6. Probative impact of CABL
7. Alternative tests for matching bullets
8. Committee Recommendations; Further work

1. Introduction

Background

- Crime → evidence → bullets
- Gun recovered: match striations on bullet and gun barrell (separate NRC committee)
- *No gun*: **Compositional Analysis of Bullet Lead** (CABL)
- "Working hypothesis": chemical concentration of lead used to make a "batch" of bullets provides a "unique signature" ⇒ "equal" concentrations of elements in bullet from Crime Scene (CS) and bullets from Potential Suspect (PS) may indicate "guilt"

- Local police dept sends CS, PS bullets sent to FBI lab
- FBI measures (in triplicate) concentrations of 7 elements
- Reports "analytically indistinguishable concentrations" between CS and PS bullets if "mean $\pm$ 2·SD intervals overlap for *all* 7 elements" (**2-SD-overlap**)
- FBI also uses "**range overlap**" and "**chaining**"
- FBI court testimony when requested

Charge to the Committee

1. Is analytical procedure (ICP-OES) sound, best available? Choice of elements, use of isotopes?
2. **"Are the statistical tests used to compare two samples appropriate?"**
3. **"Can known variations introduced in manufacturing process be used to model specimen groupings and provide improved comparison criteria?"**
4. Interpretation issues (probative value): "What are the appropriate statements that can be made to assist the requester in interpreting the results of compositional bullet lead comparison? Can significance statements be modified to include effects of such factors as the analytical technique, manufacturing process, ...

2. Manufacturing Process

- Most bullets made from lead in recycled batteries (5%)
- Process involves removal of impurities (Cu, Se, Zn, S, ...) by cooling, heating, crystalizing, addition of chemicals (e.g., charcoal, Zn, Sb for hardness)
- Formed into blocks ("pigs," "ingots," "slugs")
- Extruded into wire of dimension depending upon caliber
- Cut into pieces for bullets; poured into bins
- Bullets sent to cartridge manufacturer
- Mixed in bins; placed in boxes
- Boxes shipped to customer (many to 1 store in small town, or many to many stores in large city, or ...)

'Homogeneity' of bullets within a 'source'

- "melt": some oxidation of elements, but likely insignificant
- pig, ingot, billet: some segregation of solutes during solidification
- wires, slugs, bullets: "uniformity" along length of wire (Randich et al. 2002; Koons and Grant 2002)
- CIVL = "compositionally indistinguishable volume of lead"
- All discussion refers to chemical composition of a CIVL, which could yield 12,000–35 million bullets, depending upon manufacturing consistency, volume, bullet size

3. FBI Tests on Individual elements

2 bullets, 3 measurements of 7 elements on each bullet
(As, Sb, Sn, Bi, Cu, Ag, Cd)

Measurement = average of triplicates of one element:
Each bullet [bullet fragment] has three measurements

Crime Scene (CS) bullet: $\mathbf{X_i} = (X_{i1}, ..., X_{i7})'$
$\bar{X}$ and s_X = vector of means, SDs

Potential Suspect (PS) bullet: $\mathbf{Y_i} = (Y_{i1}, ..., Y_{i7})'$
$\bar{Y}$ and s_Y = vector of means, SDs

FBI presented three procedures for assessing "match":

- 2-SD-overlap
- range overlap
- "chaining"

Committee reformulated question as follows:

Do they come from populations (CIVLs) with same mean concentration on each of the 7 elements?

Three issues:

1. Components of variance
   - σ_e = measurement variation
   - σ_w = variation between bullets **within** batch
   - σ_b = variation between bullets **between** batches
2. False positives:
   P{claim 'match' | bullet mean concentrations differ by δ}
3. Sensitivity (and specificity):
   - P{bullet mean concentrations $< \delta$ | test claims "match"}
   - P{bullet mean concentrations $> \delta$ | claims "no match"}

2-SD-overlap test: Claim "match" if "2-SD-intervals overlap":

$$\bar{X} + 2s_X > \bar{Y} - 2s_Y \text{ or } \bar{Y} + 2s_Y > \bar{X} - 2s_X$$
$$\text{i.e., } |\bar{X} - \bar{Y}| < 2(s_X + s_Y) \text{ on each element}$$

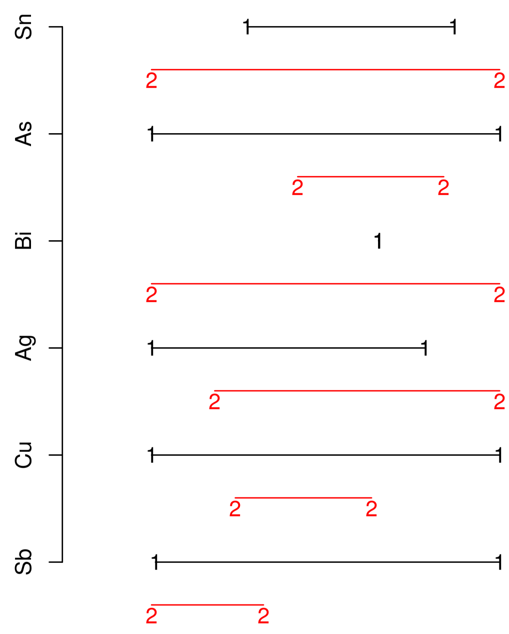
1. σ better estimated by s_p =pooled SD from many bullets: comparable error in measuring concentrations in CS, PS bullets (close in time, same sets of standards, etc)
2. Chemical concentrations *lognormal*, not normal: $\log(X) \sim \mathrm{N}(\mu_X, \sigma^2)$, $\log(Y) \sim \mathrm{N}(\mu_Y, \sigma^2)$
3. $\mathrm{SD}(X) \sim \mathrm{SD}(\log(X))$ if $\sigma/\mu < 0.05$
4. $\mathrm{SD}(\bar{X} - \bar{Y}) \approx \sigma\sqrt{2/3}$, not 2σ
5. FBI allowance of $\approx 4\sigma$ is too wide

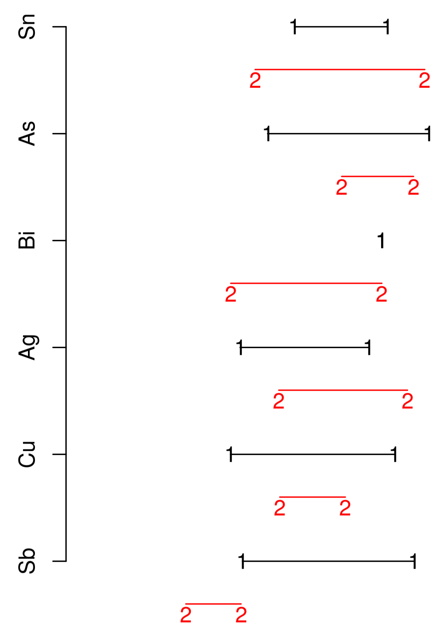
Federal bullet F001

	icpSb	icpCu	icpAg	icpBi	icpAs	icpSn
a	29276	285	64	16	1415	1842
b	29506	275	74	16	1480	1838
c	29000	283	66	16	1404	1790
mean	29260.7	281.0	68.0	16	1433.0	1823.3
SD	253.4	5.3	5.3	0	41.1	28.9
mean-2SD	28754.0	270.4	57.4	16	1350.8	1765.5
mean+2SD	29767.4	291.6	78.6	16	1515.2	1881.2
minimum	29000	275	64	16	1404	1790
maximum	29506	285	74	16	1480	1842

Federal bullet F002

	icpSb	icpCu	icpAg	icpBi	icpAs	icpSn
a	28996	278	76	16	1473	1863
b	28833	279	67	16	1439	1797
c	28893	282	77	15	1451	1768
mean	28907.3	279.7	73.3	15.7	1454.3	1809.3
SD	82.4	2.1	5.5	0.6	17.2	48.7
mean-2SD	28742.5	275.5	62.3	14.5	1419.8	1712.0
mean+2SD	29072.2	283.8	84.4	16.8	1488.8	1906.7
minimum	28833	278	67	15	1439	1768
maximum	28996	282	77	16	1473	1863


2−SD overlap
Sn
As
Bi
Ag
Cu
Sb
'Analytically indistinguishable'


Range overlap
Sn
As
Bi
Ag
Cu
Sb
All elements analytically
indistinguishable except Sb

5. FPP = False Positive (Match) Probability (1 element)

Mean concentrations differ by δ, measurement error SD σ_e:

$$\mathrm{P}\{(\bar{x}+\bar{y}) < 2(s_x+s_y) \mid |\mu_x-\mu_y| = \delta\ \}$$
$$\mathrm{P}\{(\bar{x}+\bar{y}) < 2(s_x+s_y) \mid |\mu_x-\mu_y| = \delta\ \}$$
$$= \mathrm{P}\{(\bar{x}+\bar{y})/(s_p\sqrt{2/3}) < 2(s_x+s_y)/(s_p\sqrt{2/3}) \mid |\mu_x-\mu_y| = \delta\ \}$$

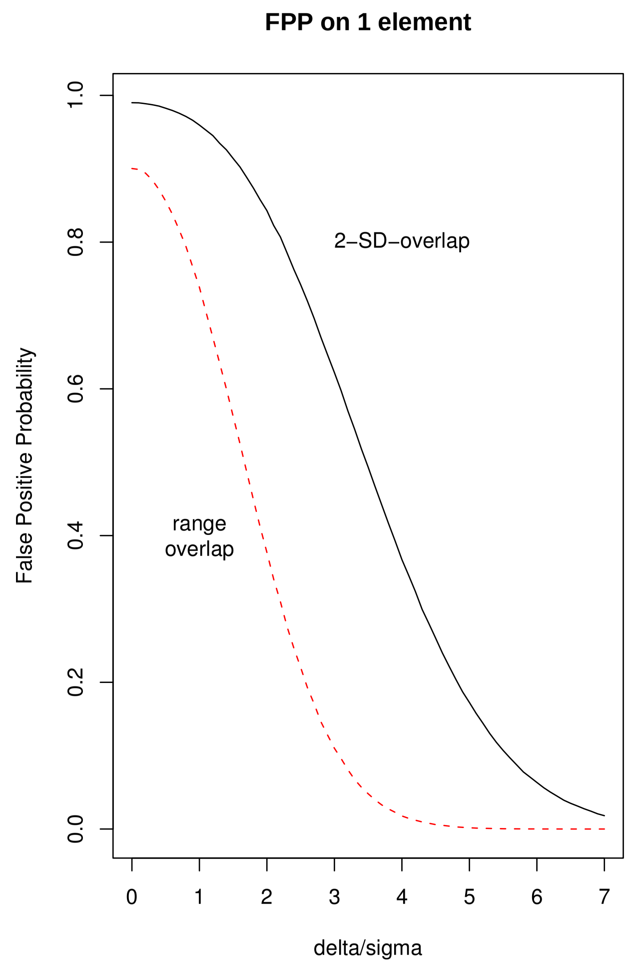
FPP is a function of only δ/σ_e:

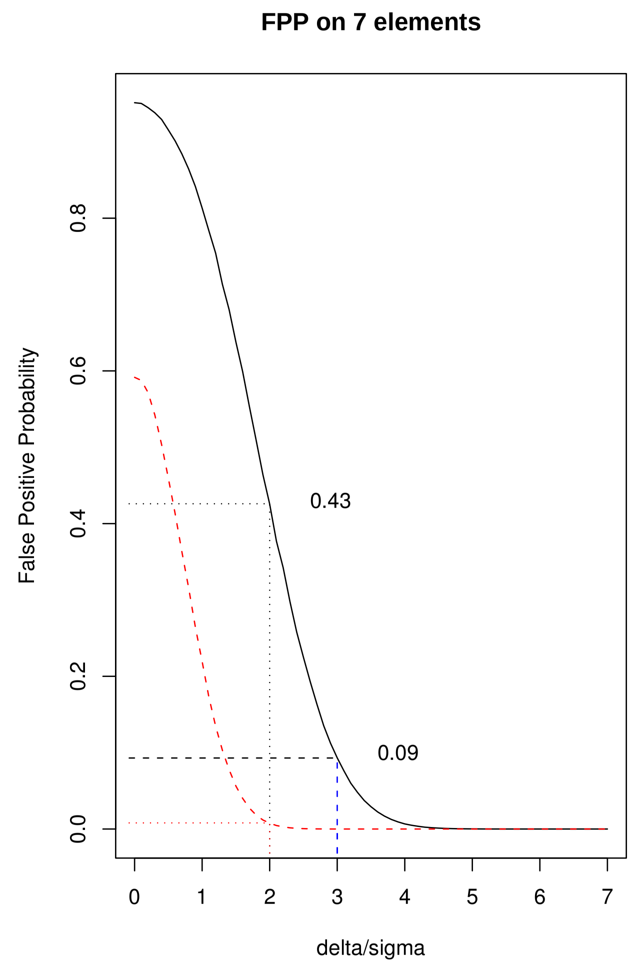
$\mathrm{E}(s_x) = \mathrm{E}(s_y) = 0.8812\sigma$, $\mathrm{E}(s_p) \approx \sigma$ (many d.f. in s_p):

$$\mathrm{P}\{(\bar{x}+\bar{y})/(s_p\sqrt{2/3}) < 4.317 \mid |\mu_x-\mu_y| = \delta\ \}$$

[rough approximation: $\mathrm{E}(\mathrm{P}\{t < r.v.\}) \neq \mathrm{P}\{t < E(r.v.)\}$]

- Distribution of $\frac{\bar{X}-\bar{Y}}{s_x+s_y} = \frac{N(\delta,2\sigma_e^2)}{\sqrt{\chi_2^2}+\sqrt{\chi_2^2}}$ is theoretically possible but messy
- Easier: 100,000 simulations
- Plot $FPP(\delta)$ for given levels of σ_e
- FPP = function of δ/σ, $\sigma = \sqrt{\sigma_e^2 + \sigma_b^2 + \sigma_w^2}$
- Large $\delta \Rightarrow$ small probability of false match
- Small δ ($\delta < 3\sigma_e$) $\Rightarrow$ high probability
- Smaller FPP for range overlap (E(range of 3) $\approx 1.6\sigma_e$)


FPP on 1 element
False Positive Probability
delta/sigma
2–SD–overlap
range
overlap
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7


FPP on 7 elements
False Positive Probability
delta/sigma
0.43
0.09
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6
7

FPP for 2-SD-overlap, independent measurement errors

σ	$\delta = 0$	1	2	3	4	5	6	7
0.5	0.931	0.298	0.001	0.000	0.000	0.000	0.000	0.000
1.0	0.931	0.749	0.298	0.036	0.001	0.000	0.000	0.000
1.5	0.931	0.849	0.612	0.303	0.084	0.013	0.001	0.000
2.0	0.931	0.883	0.747	0.535	0.302	0.125	0.036	0.007
2.5	0.931	0.903	0.817	0.669	0.487	0.302	0.151	0.062
3.0	0.931	0.911	0.850	0.748	0.615	0.450	0.298	0.175

FPP for 2-SD-overlap, Federal correlation matrix

σ	$\delta = 0$	1	2	3	4	5	6	7
0.5	0.950	0.426	0.007	0.000	0.000	0.000	0.000	0.000
1.0	0.950	0.813	0.426	0.093	0.007	0.000	0.000	0.000
1.5	0.950	0.884	0.713	0.426	0.163	0.048	0.007	0.001
2.0	0.950	0.916	0.813	0.638	0.426	0.225	0.093	0.029
2.5	0.950	0.929	0.864	0.754	0.599	0.426	0.258	0.135
3.0	0.950	0.938	0.884	0.813	0.713	0.553	0.426	0.298

4. Data Sets Available to Committee

- '800-bullet'
- '1837-bullet' (subset 854 bullets)
- Tables 1+2 in Randich et al. (2002)
- Table 3 in Koons and Grant (2002)
- Others in published articles (but not analyzed here)

800-bullet data set

- Peele, Havekost, Peters, Riley, Halberstam, Koons (1991), "Comparison of Bullets Using the Elemental Composition of the Lead Component," *Proceedings of the International Symposium on the Forensic Aspects of Trace Evidence*
- 4 manufacturers (CCI, Federal, Remington, Winchester)
- 4 boxes per manufacturer
- 50 bullets per box
- Replicates a, b, c per bullet

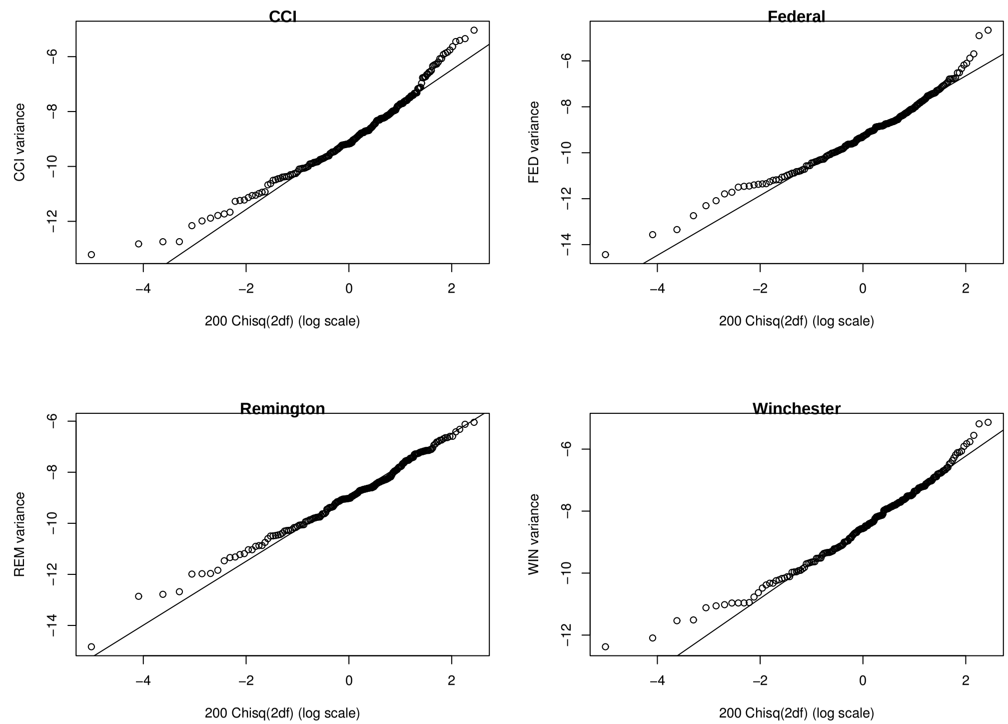
What this data set can provide:

- Rough idea of distributions of concentrations across bullets
- measurement error variance σ_e^2
- correlations between errors in measuring two different elements on same bullet (only Federal data measured 6 elements by ICP-OES; none measured Cadmium)

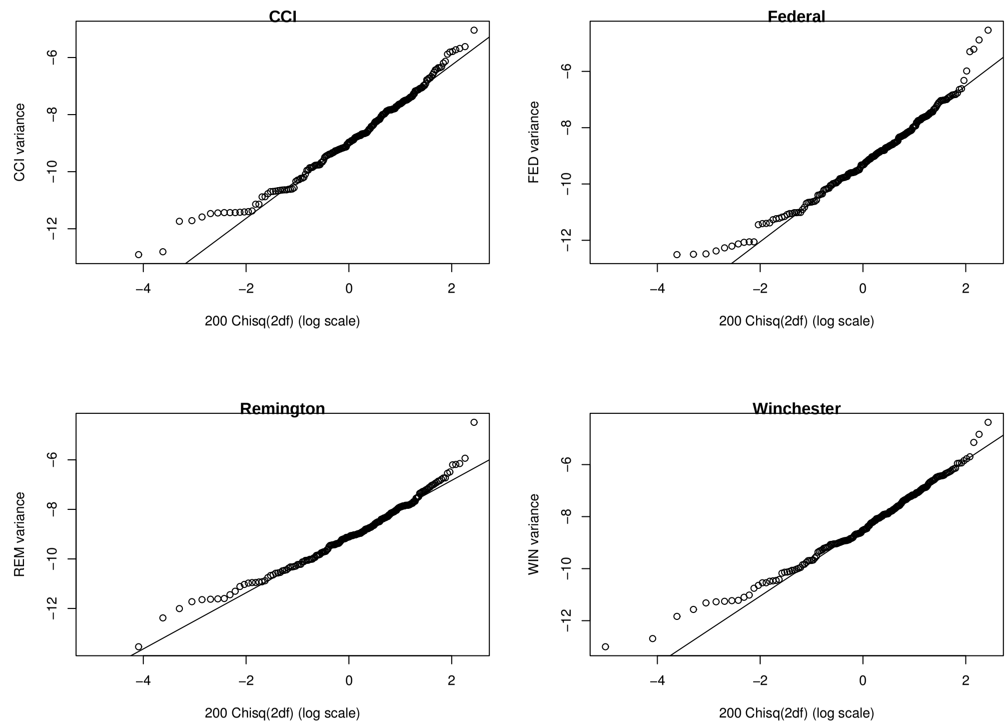
What this data set not provide:

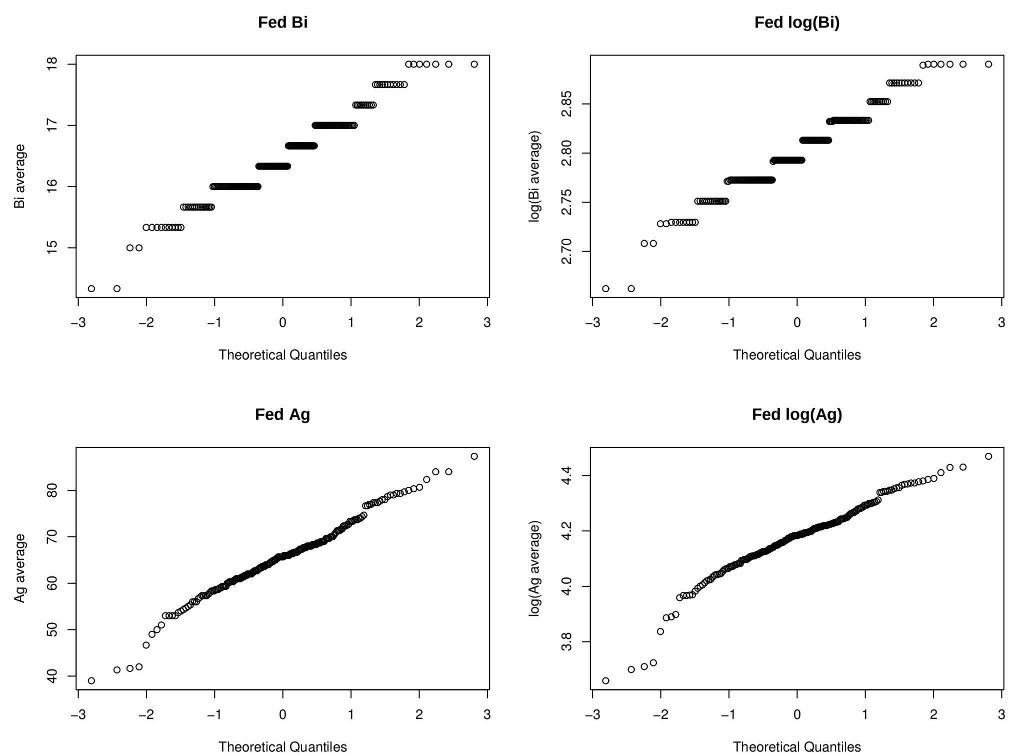
- Does not provide estimates of within-batch homogeneity σ_w
- Does not provide estimates of between-batch variability σ_b – unless one believes that the "batches" defined by one of thousands of possible clustering algorithms are "homogeneous" (e.g., ISU Tech Report, FBI "chaining").

Distributions of within−bullet variances on log(ICP−Sb)
(y−axis: 200 within−bullet variances, log scale)

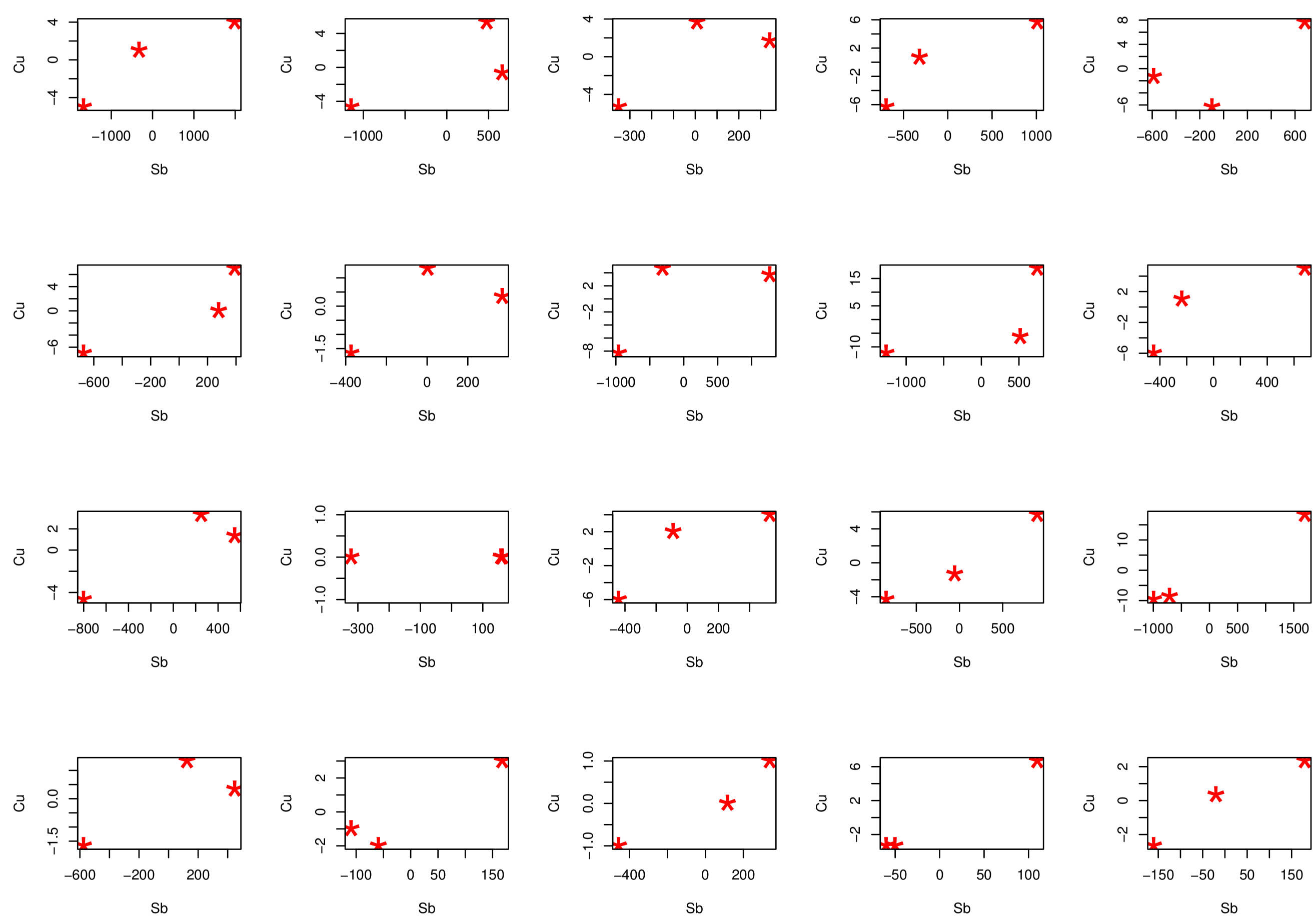


Distributions of within–bullet variances on log(ICP–Cu)
(y–axis: 200 within–bullet variances, log scale)




Fed Bi
Bi average
Theoretical Quantiles
Fed log(Bi)
log(Bi average)
Theoretical Quantiles
Fed Ag
Ag average
Theoretical Quantiles
Fed log(Ag)
log(Ag average)
Theoretical Quantiles

Sb, Cu on 20 CCI bullets



Sample correlation matrix: Federal bullets

	As	Sb	Sn	Bi	Cu	Ag	(Cd)
As	1.000	0.320	0.222	0.236	0.420	0.215	0.000
Sb	0.320	1.000	0.390	0.304	0.635	0.242	0.000
Sn	0.222	0.390	1.000	0.163	0.440	0.154	0.000
Bi	0.236	0.304	0.163	1.000	0.240	0.179	0.000
Cu	0.420	0.635	0.440	0.240	1.000	0.251	0.000
Ag	0.215	0.242	0.154	0.179	0.251	1.000	0.000
(Cd)	0.000	0.000	0.000	0.000	0.000	0.000	1.000

Consequence:

$$\mathrm{P}\{7\ |t|\ \text{statistics} < K_\alpha(\nu, n)\} \neq \alpha^7\ (\text{maybe } \alpha^5)$$

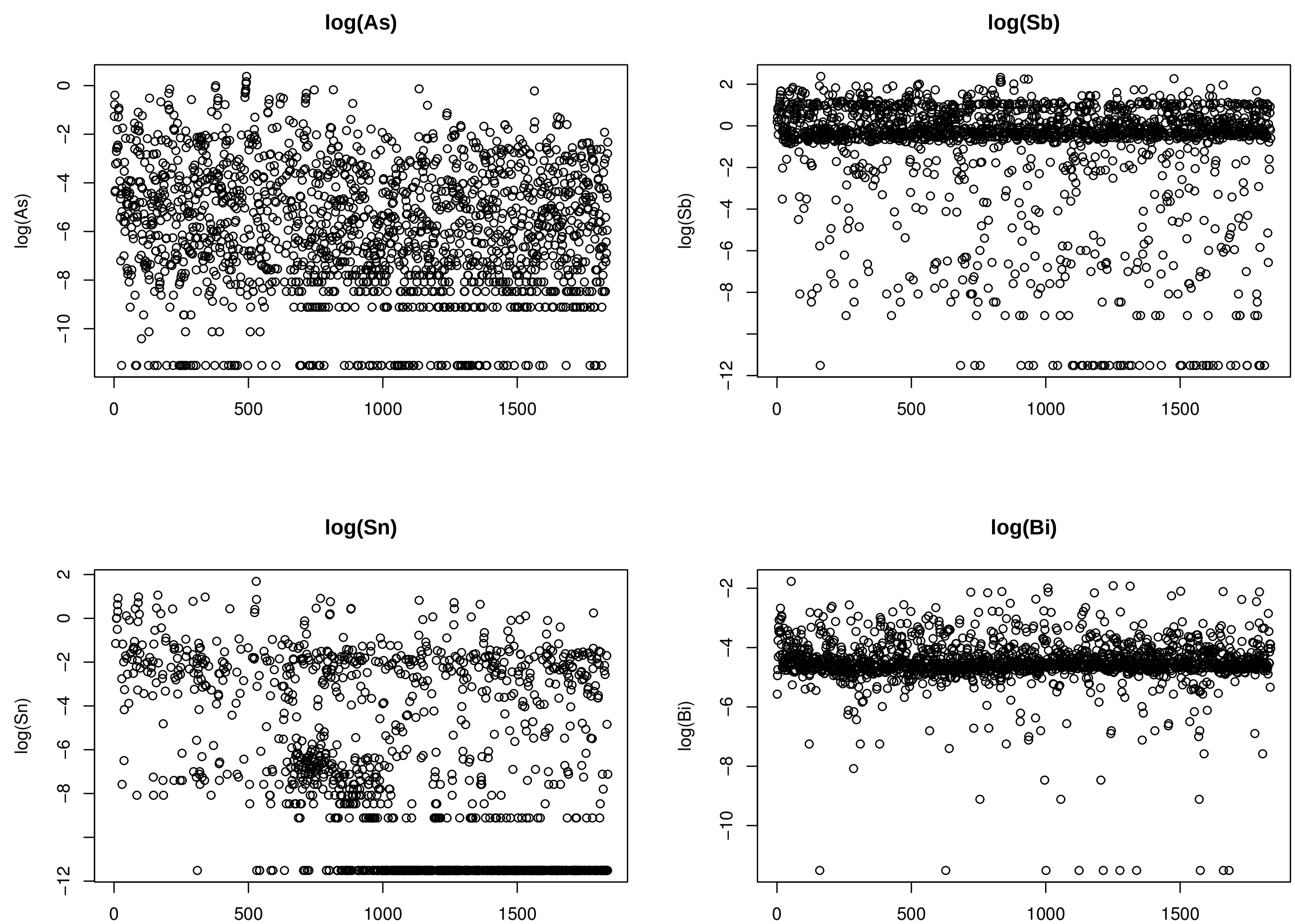
1837-bullet data set

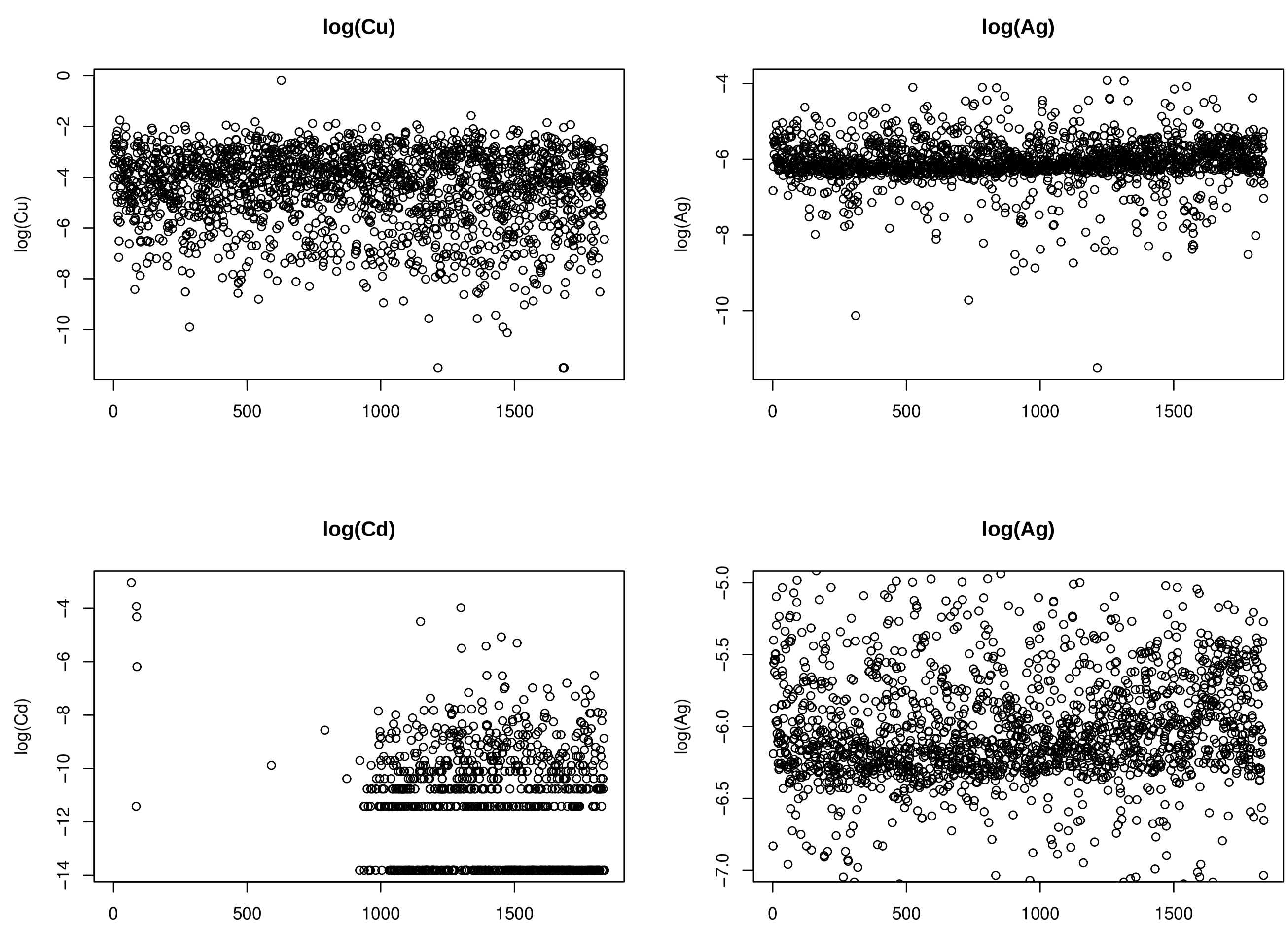
- Part of the complete data log containing chemical analyses on 71,000+ bullets
- FBI "selected" 1837 bullets that were believed to be "different"
- 1837-bullet set = FBI's attempt at different "melts"
- Only 854 of 1837 had all 7 elements (1997 or later)

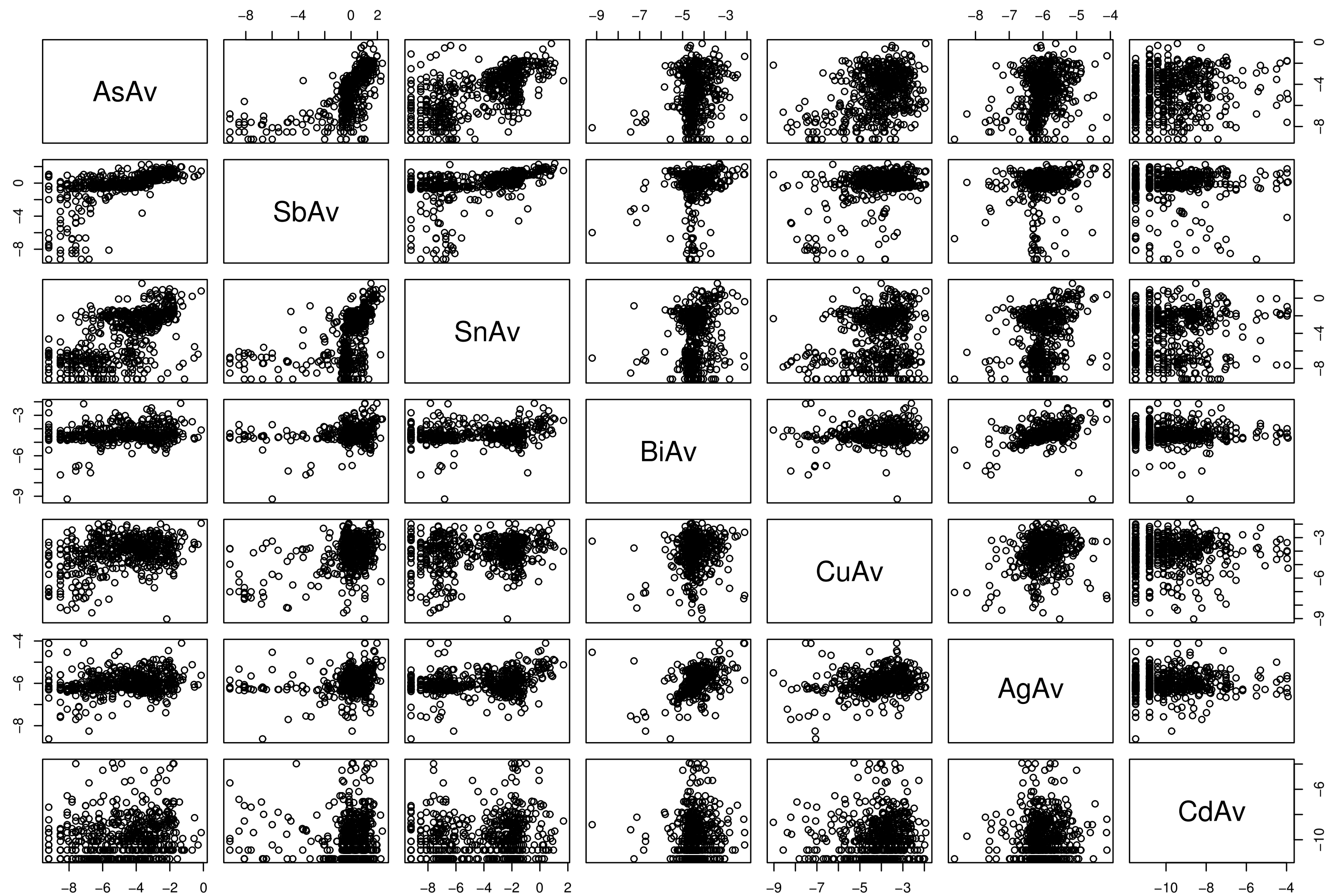
FBI Notes on 1837-bullet data set

"To assure independence of samples, the number of samples in the full database was reduced by removing multiple bullets from a given known source in each case. To do this, evidentiary submissions were considered one case at a time. For each case, one specimen from each combination of bullet caliber, style, and nominal alloy class was **selected** *and that data was placed into the test sample set. In instances where two or more bullets in a case had the same nominal alloy class, one sample was randomly selected from those containing the maximum number of elements measured. . . . The test set in this study, therefore, should represent an unbiased sample in the sense that each known production source of lead is represented by only one randomly selected specimen."* (Notes on 1837-bullet dataset)

- FBI used it to estimate FPP=False Positive Probability: 693 2-SD-overlap "matches" among 1,686,366 comparisons $\Rightarrow$ "about 1 in 2500"
- Committee: This FPP (1 in 2500) is not valid (useless)
- 1837-bullet data set is not a random sample
- See Cochran, Mosteller, Tukey (1954), "Principles of Sampling" (*JASA*)
- Does provide some information on distributions of concentrations across unspecified collection of bullets


log(As)
log(Sb)
log(Sn)
log(Bi)


log(Cu)
log(Cu)
0
−2
−4
−6
−8
−10
0
500
1000
1500
log(Ag)
log(Ag)
−4
−6
−8
−10
0
500
1000
1500
log(Cd)
log(Cd)
−4
−6
−8
−10
−12
−14
0
500
1000
1500
log(Ag)
log(Ag)
−5.0
−5.5
−6.0
−6.5
−7.0
0
500
1000
1500

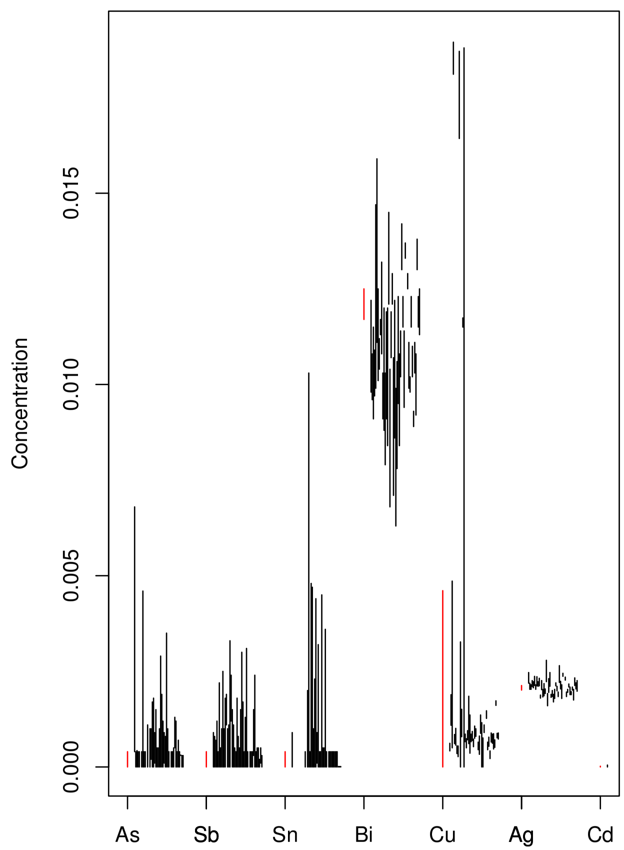


545 bullets from 1837–bullet data set with 7 measureable elements

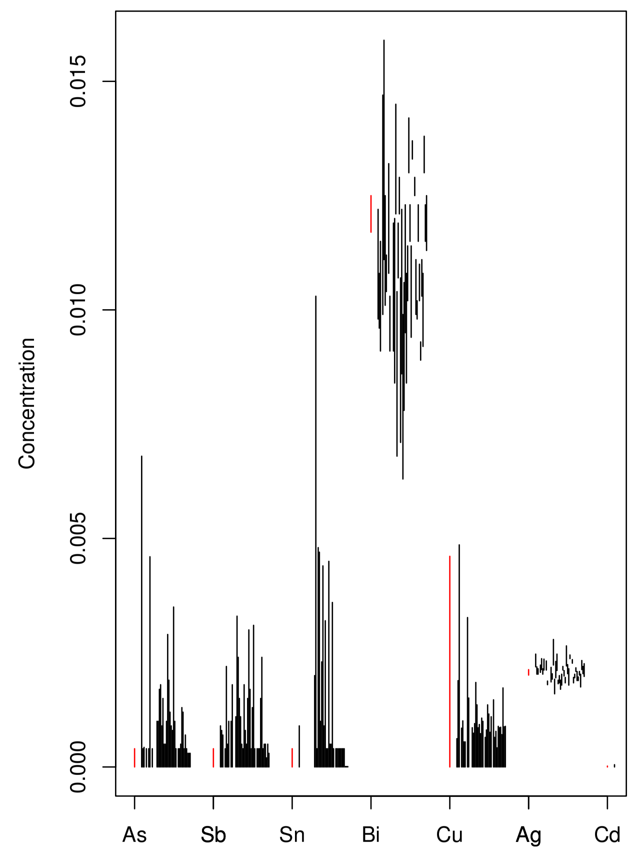
Chaining (FBI's version of "clustering"):

"The mean element concentrations of the first and second specimens in the known material population are compared based upon twice the measurement uncertainties from their replicate analysis. If the uncertainties overlap in all elements, they are placed into a composition group; otherwise they are placed into separate groups. The next specimen is then compared to the first two specimens, and so on, in the same manner until all of the specimens in the known population are placed into compositional groups." (Peters, C.A.: *Comparative Elemental Analysis of Firearms Projectile Lead By ICP-OES*, FBI Laboratory Chemistry Unit. Issue date: October 11, 2002.)

Resulting "compositional group" could be very diverse:



41 matched bullets to #1044



37 matched bullets to #1044

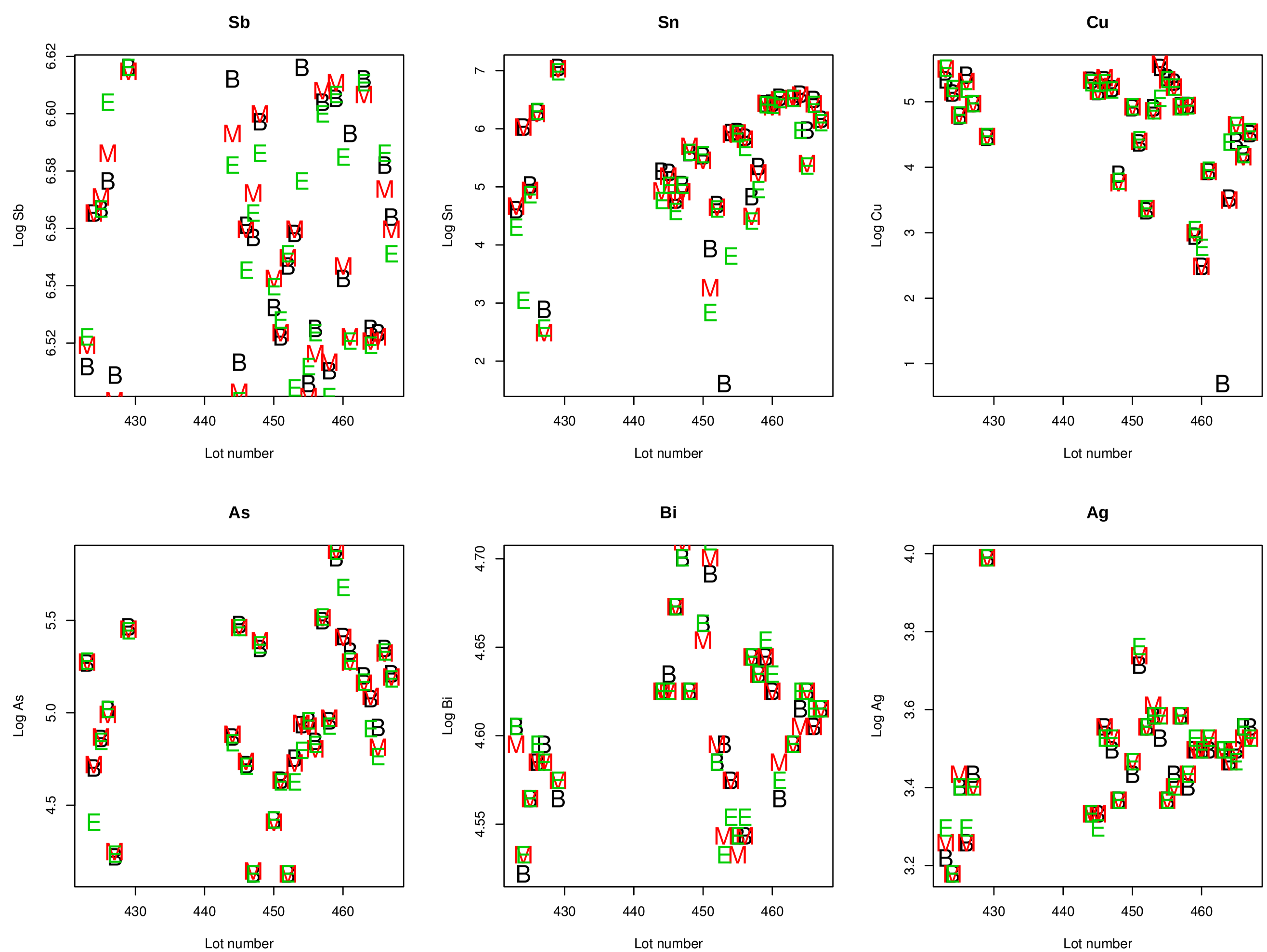
Do we need to measure all 7 elements?

Principal components analysis on 854 bullets on which all 7 elements were measured:

- Total variance (sum of eigenvalues of $X'X$ matrix): 136.944
- Percent of total variance explained with:
  - 3 elements: 83.6% (Sb, Sn, Cd)
  - 4 elements: 96.1% (Sb, Sn, Cd, As)
  - 5 elements: 97.5% (Sb, Sn, Cd, As, Cu)
  - 6 elements: 98.5% (Sb, Sn, Cd, As, Cu, Ag)
- Conclusion: Little gained by adding Bi+Ag, but little cost; need to confirm with complete 71,000+ bullet data set.

Randich et al. (2002) data set

- Erik Randich, Wayne Duerfeldt, Wade McLendon, William Tobin (2002), "A metallurgical review of the interpretation of bullet lead compositional analysis," *Forensic Science International* 127: 174–191.
- 28 lots of nominal 0.7wt.% alloy, manufactured Jan'99–Mar'00
- 3 samples per lot (Beginning, Middle, End of pour)
- Six elements (all but CD), Tables 1–2
- One reported value per sample, no standard errors
- Limited information of σ_b (between lots) compared to σ_w (within lots)
- Compare $\hat{\sigma}_w$ for these lots to FBI's σ_e


Sb
Log Sb
Lot number
Sn
Log Sn
Lot number
Cu
Log Cu
Lot number
As
Log As
Lot number
Bi
Log Bi
Lot number
Ag
Log Ag
Lot number

Variation among "B", "M", "E" consistent with FBI measurement error variation, random ordering

Comparing within-bullet, within-lot variances

	NAA-As	ICP-Sb	ICP-Cu	ICP-Bi	ICP-Ag
between lots:					
Randich	4981.e-04	40.96e-04	17890e-04	60.62e-04	438.5e-04
within bullet:					
800-bullet set	26.32e-04	4.28e-04	4.73e-04	18.25e-04	20.88e-04
within lot:					
Randich et al.	31.32e-04	3.28e-04	8.33e-04	0.72e-04	3.01e-04
within lot to within bullet	1.2	0.8	1.8	0.04	0.14

Koons and Grant (2002) data set

- R.D. Koons, D.M. Grant (2002), "Compositional variation in bullet lead manufacture," *J. Forensic Sci.* 47: 950–958
- Homogeneity within a "melt"
- 2 smelters (+ 1 ammunition manufacturer), Sep'97–Aug'99
- Lead poured into disks (7.4cm diameter; 1.6cm thick)
- Subdivided into 3 or 5 vertical layers, 3 wedges
- Reported mean & SD, triplicate measurements
- Limited information of σ_b compared to σ_w
- Most CVs $<$ 2%, except with low concentrations
- Measurement error $\approx$ 3–5%

Pour date	AsAv	SbAv	SnAv	BiAv	CuAv	AgAv
09.06.97	61.5	0.9	25.0	0.4	1.4	1.5
08.13.99	10.3	0.6	28.6	0.5	1.0	1.1
01.04.99	29.4	0.6	9.1	0.3	0.8	0.9
04.09.98	46.2	1.2	20.0	1.0	2.3	1.3
11.29.98	NA	0.9	76.9	0.8	0.9	1.4
10.07.97	81.8	2.1	1.4	1.9	3.4	2.8
10.30.98	1.3	0.8	14.3	0.4	1.3	1.3
11.22.97	0.8	0.0	100.0	0.1	1.4	0.8
07.03.99	0.8	0.7	1.7	0.7	0.9	0.6
08.12.99	0.8	0.6	20.0	0.9	0.8	0.8
11.04.97	0.2	0.3	100.0	0.4	0.3	0.4
10.21.97	11.1	0.8	0.8	1.2	1.7	0.8
01.02.98	62.5	0.6	100.0	1.0	1.0	0.8
02.22.99	2.2	1.7	33.3	1.4	2.7	0.2
02.01.99	0.9	0.5	100.0	0.7	0.7	0.8

Conclusions:

- σ_w (homogeneity within "CIVL") $\approx$ or $< \sigma_e$ (FBI measurement error)
- Concentration distributions $\approx$ lognormal
- Measurement error SDs can be pooled
- Measurement errors are not uncorrelated
- No reliable information on "within-batch" vs "between-batch" variances
- No "honest" data set from which to compute false positive probability of match using FBI 2-SD or range overlap methods, so resort to simulation

False Positive Probability:

Given the difference between mean concentrations of the CIVLs from which the bullets were manufactured (δ) and the measurement error variation (σ_e), what is P$\{$match $\mid \delta, \sigma_e\}$?

The practical issue:

Given that the 2-SD-overlap test claims "match", what is the probability that the two bullets really did come from the same or different sources?

"Likelihood" that two bullets came from CIVLs whose mean difference is no more than δ:

$$\text{Prob}\{|\mu_x - \mu_y| < \delta | \text{ 2-SD-overlap "match" }\}$$
$$\text{Prob}\{|\mu_x - \mu_y| > \delta | \text{ 2-SD-overlap "match" }\}$$

Practical issue (cont'd):

- Bayes rule $\Rightarrow$ need to know Prob$\{\delta\}$; i.e., typical δ's (distribution of distances between bullets, SD $= \sigma_b$).
- Depends upon caliber, manufacturer, geographical location, ... None of the available data sets provides reliable, unbiased information.
- We don't know how often δ may be BIG or small. It depends on how "different" the sources are.

What we **can** say:

Scenario 1: Two bullets.
No tests on them, just two bullets. What is the probability they came from same source? (tiny)

Scenario 2: Two bullets.
We measure them; "2-SD-overlap" says "match".
What is the probability they came from same source? (probably higher than in scenario 1, whose bullets were never measured)

Probability that two bullets came from the same CIVL is increased by a finding that they are analytically indistinguishable, versus no evidence of match status.

6. Probative impact of matching evidence

From the *FBI Handbook of Forensic Sciences 36* (rev 1999):

"Differences in the concentrations of manufacturer-controlled elements and uncontrolled trace elements provide a means of differentiating among the lead of manufacturers, among the leads in individual manufacturer's production lines, and among specific batches of lead in the same production line of a manufacturer."

Accordingly, FBI testimony has included statements such as:

- "Could have come from the same box"
- "Could have come from the same box or a box manufactured on the same day"
- "Were consistent with their having come from the same box of ammunition"

- "Probably came from the same box"
- "Must have come from the same box or another box that would have been made by the same company on the same day"
- "had come from the same batch of ammunition: they had been made by the same manufacturer on the same day and at the same hour"
- "likely originated from the same manufacturer's source (melt) of lead"
- "The specimens within a composition group are analytically indistinguishable. Therefore, they originated from the same manufacturer's source (melt) of lead."

(*NAS Report, pp. 91–92*)

Such testimony grossly overstates probative impact of “matches”; can say only that two bullets that “match” may have come from sources with the same chemical composition — NOT that they came from the same box, wire, ingot, source, or melt.

Finding: The available evidence do not support any statement that a crime bullet came from, or is likely to have come from, a particular box of ammunition, and references to “boxes” of ammunition in any form are seriously misleading under Federal Rule of Evidence 403. Testimony that the crime bullet came from the defendant’s box or from a box manufactured at the same time is also objectionable because it may be understood as implying a substantial probability that the bullet came from defendant’s box.

7. Alternative analyses

(a) Equivalence tests using a series of t statistics:

$$|\bar{X} - \bar{Y}| < K s_{pool} \cdot \sqrt{2/3}$$

- $\bar{X}$, $\bar{Y}$ are the means of the **logarithms** of the three measurements on the CS and PS bullets
- s_{pool} is the root mean square of the standard deviations on the logs of the three measurements from **many** bullets
- Measurement SDs should be monitored (control chart; e.g., Vardeman and Jobe 1999)
- Equivalence test hypotheses:
  H_0: means differ by more than δ_0 (no match)
  H_1: means differ by less than δ_0 ("match")

- K (critical point) depends upon: chosen FPP (α), $n =$ sample size (here, 3 per bullet), $\nu =$ degrees of freedom in s_p, "limit" of "equivalence" δ_0, power at a value $\delta_1 < \delta_0$
- For large ν, K can be solved from:

$$\Phi(K - \delta_0/(s_p\sqrt{2/3})) - \Phi(-K - \delta_0/(s_p\sqrt{2/3})) = \alpha$$

$\Phi(\cdot) =$ cumulative standard Gaussian distribution

For small FPPs (< 0.05) and small δ_0 ($< 2\sigma_e$), $K < 2$.

Some values of K for $\alpha = 0.25$, n=3 ($\alpha^5 \approx 0.001$):

Values of δ_0/σ_e

	0.25	0.33	0.50	1	1.5	2	3
df= 3	0.3577	0.3703	0.4109	0.6814	1.1924	1.7741	2.9155
20	0.3363	0.3482	0.3866	0.6481	1.1690	1.7730	2.9816
50	0.3348	0.3466	0.3848	0.6458	1.1676	1.7742	2.9922
100	0.3344	0.3462	0.3843	0.6450	1.1672	1.7746	2.9960

(mean difference/pooled SD) $< K \cdot \sqrt{2/3}$

Gaussian probabilities of false match, false non-match

(b) Multivariate Equivalence Hotelling's T^2 test:

$$H_0 : |\mu_x - \mu_y| \geq \delta_{\mathbf{0}}$$

$$H_1 : |\mu_x - \mu_y| < \delta_{\mathbf{0}}$$

Assume:

$$\mathbf{X_i} \sim N_7(\mu_x, \Sigma), \quad \mathbf{Y_i} \sim N_7(\mu_y, \Sigma)$$

Test: Reject if

$$(\bar{\mathbf{X}} - \bar{\mathbf{Y}})'\hat{\Sigma}^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}}) < \frac{7\nu}{(\nu - 6)} F_{7,n-7}(ncp) \text{ (noncentral F)}$$

Hotelling's T^2 and 7 t tests:

- $\mathrm{P}\{T^2 < F_{crit} \mid |\mu_x - \mu_y| \geq \delta_0 \cdot \mathbf{1}\} = \alpha$
- t-statistics:
  $\mathrm{P}\{\ |t_j| < K \mid |\mu_x - \mu_y| > \delta_0\ \} = \alpha_j \equiv \alpha$
  (given α, K comes from noncentral Student's t distribution, or Gaussian approximation above)
- $\mathrm{P}\{$ all t-statistics $< K \mid |\mu_x - \mu_y| > \delta_0\ \} \neq \alpha^7$ (correlation)
- Compare:
  $p_1 = \mathrm{P}\{\text{indt } t\text{-statistics} < K \mid \delta_0\} = \alpha^7$
  $p_2 = \mathrm{P}\{\text{corr } t\text{-statistics} < K \mid \delta_0\} = \alpha^?$
  $\Rightarrow\ ? = 7 - [\log(p_1) - \log(p_2)]/\log(\alpha)$

Simulation (100,000 trials):

K	Values of δ_0					
	0.0	0.5	1.0	1.5	2.0	2.5
0.000	5.27	5.18				
0.645	5.42	5.15	4.77			
1.167	5.68	5.44	5.27	5.23	5.20	
2.000	6.55	5.91	5.58	5.38	5.06	5.00

Overall FPP $\approx \alpha^5$

(c) Empirical Mahalanobis distances and sampling:
Alicia L. Carriquiry, Michael Daniels, Hal S. Stern, "Statistical treatment of class evidence: Trace element concentrations in bullet lead," Iowa State Technical Report, May 4, 2000 (Table 4 corrected April 19, 2002)

- Noted many of same difficulties with using CABL
- Generate empirical distributions of Mahalanobis distances for bullet pairs in same ("within"), different ("between") batches
- Very sensible strategy, **if** manufacturers would cooperate
- Slight trends in concentrations over time (Ag)

8. Committee Recommendations

- Discontinue discussion of "boxes" in court testimony
- Monitor measurement SDs using standard SQC charts
- Replace "2-SD-overlap" with "successive equivalence t-tests" using per-element α of about $(target\,\alpha)^{1/5}$ or equivalence Hotelling's T^2 using pooled covariance matrix
- Plan to analyze the 71,000+ file of FBI-measured bullets
- **Arrange for a well designed and executed experiment to estimate σ_b, σ_w.** These values will depend on bullet manufacturer, type, caliber, geographic locale.

9. Further work

- How robust is Hotelling's T^2 to misspecified Σ?
- Is Σ easily estimated by non-statisticians (monitoring pooled variances and covariances)?
- Will non-statisticians invert 7×7 covariance matrices? (in Excel?)
- "Chaining"?
- What would the 71,000+ bullet data set show?
- Issues of testing multiple CS/PS bullets

Perturbation Theory and Mixture Models: Application to Particle Physics

ACAS
21 Oct 2004

Cyrus Taylor
Dept. of Physics
Case Western Reserve University
cct@case.edu

(work joint with C. Loader and R. Pilla)

Outline

- Review of score + formula for asymptotic distribution – do we need additional parameters to describe the data?
- Applications to particle physics
  - Ex: search for new particle resonances
  - Ex: energy spectrum of highest energy cosmic rays

Mixture models, score statistic and its asymptotic distribution

- Mixture models:

$$p(x;\eta;\theta) = (1 - \eta)f(x) + \eta \; (x;\theta)$$

- Score statistic

$$S^{?}(x;\theta) := \frac{S(x;\theta)}{\sqrt{n\,C(\theta;\theta^{0})}}$$

$$= \frac{1}{\sqrt{n\,C(\theta;\theta^{0})}} \sum_{i=1}^{n} \frac{(x_i;\theta)}{f(x_i)} - 1$$

- where $C(\theta;\theta^{0}) = \int \frac{f \quad (x;\theta)g^{2}}{f(x)} dx - 1$

Asymptotic distribution

$$\Pr \text{ˆ} \sup_{ò2Ê} Z(ò) \text{ õ } c \text{ˆ}$$

$$\text{ù} \; \frac{\hat{ô}_0}{A_{d+1}} \Pr(ÿ^2_{d+1} \text{ õ } c^2) + \frac{l_0}{2A_d} \Pr(ÿ^2_d \text{ õ } c^2)$$

- $\ddot{o}_0$ described by d-dimensional volume of manifold expressible through covariance function

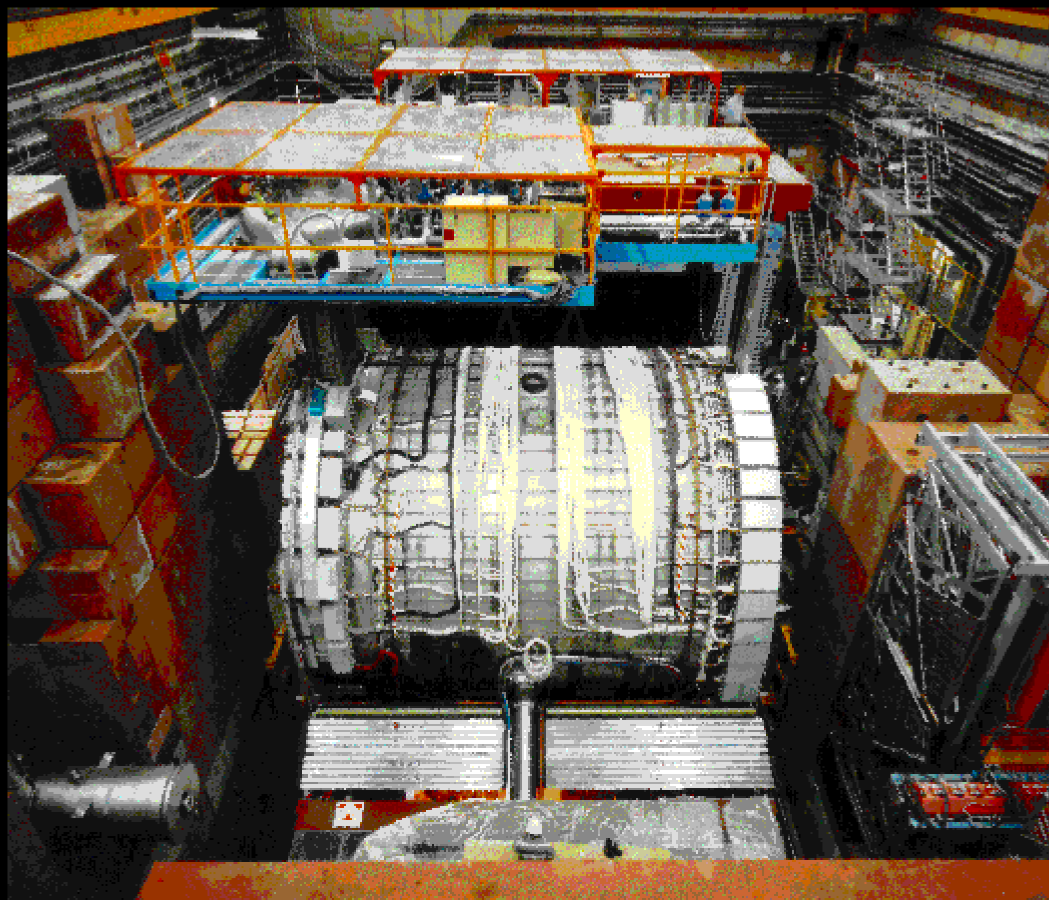
Ex: search for new resonances

- What are physicists searching for?
- Why are we searching for it?
- How do we search for it?

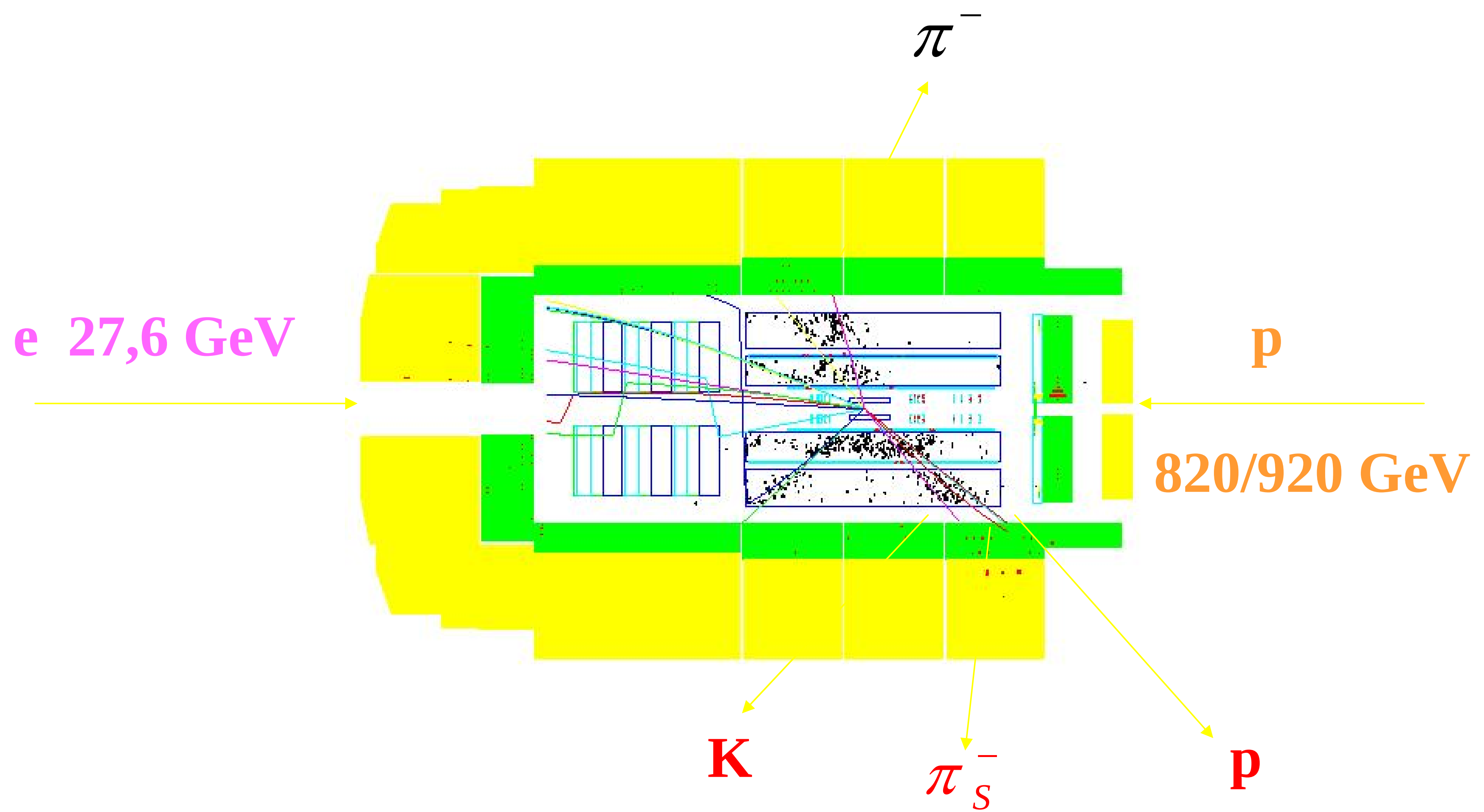
Ex: Pentaquarks

- QCD – nobel prize
- Q q-bar, 3 q states
- Pentaquark discovered
- Charmed pentaquarks
- H1 claims discovery; Zeus doesn't see it

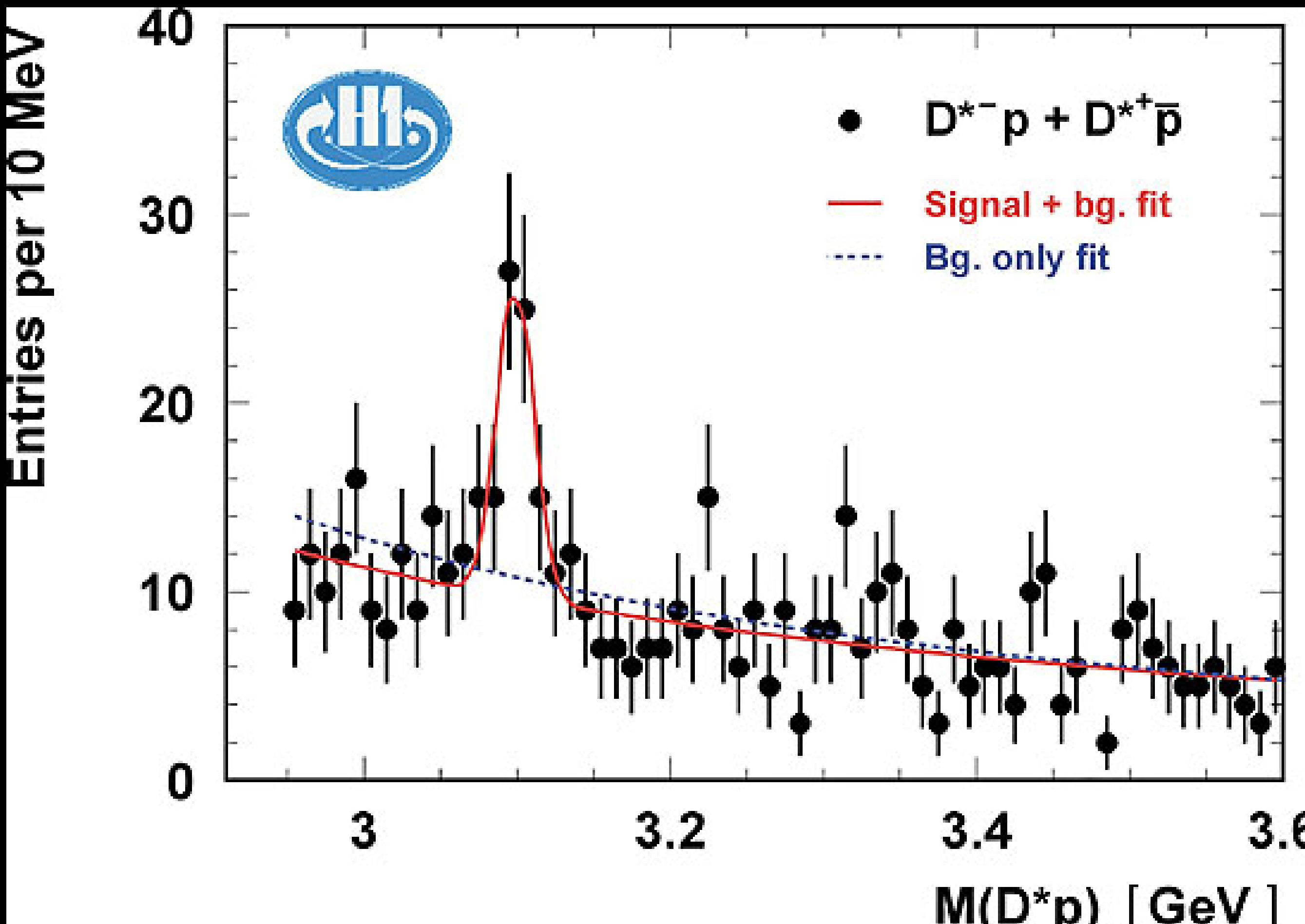
H1 detector

- 

Pentaquark in H1 setup



Ex: pentaquarks



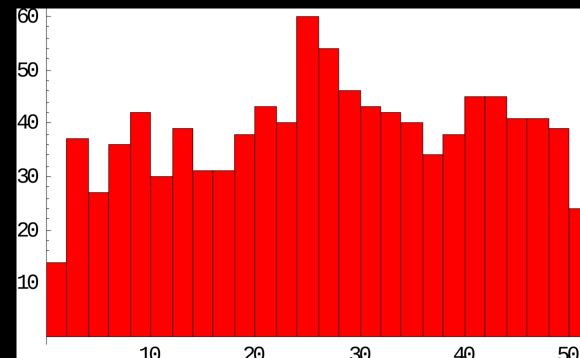
Mixture models in particle physics

- Background : power law
- Perturbation (resonance): Breit-Wigner (Cauchy):

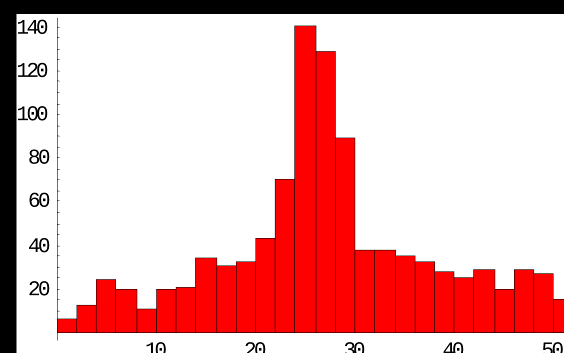
$$(E; E_0) = \text{f È=(2ù)gf}(E \text{ à } E_0)^2 + (\text{È=2})^2\text{g}^{\text{à}1}$$

Application of score analysis (MC)

- 10% mixture

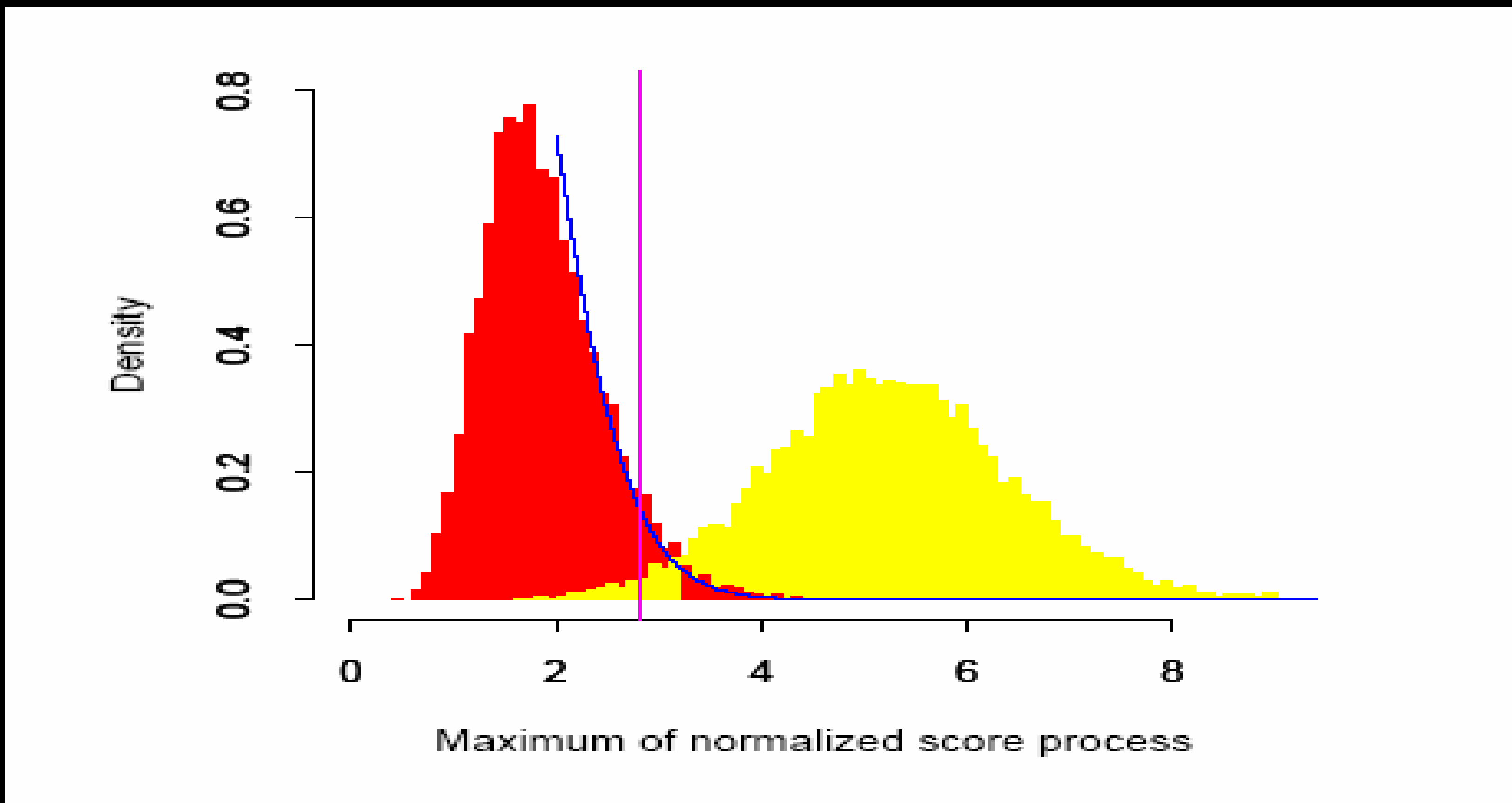


- 50% mixture



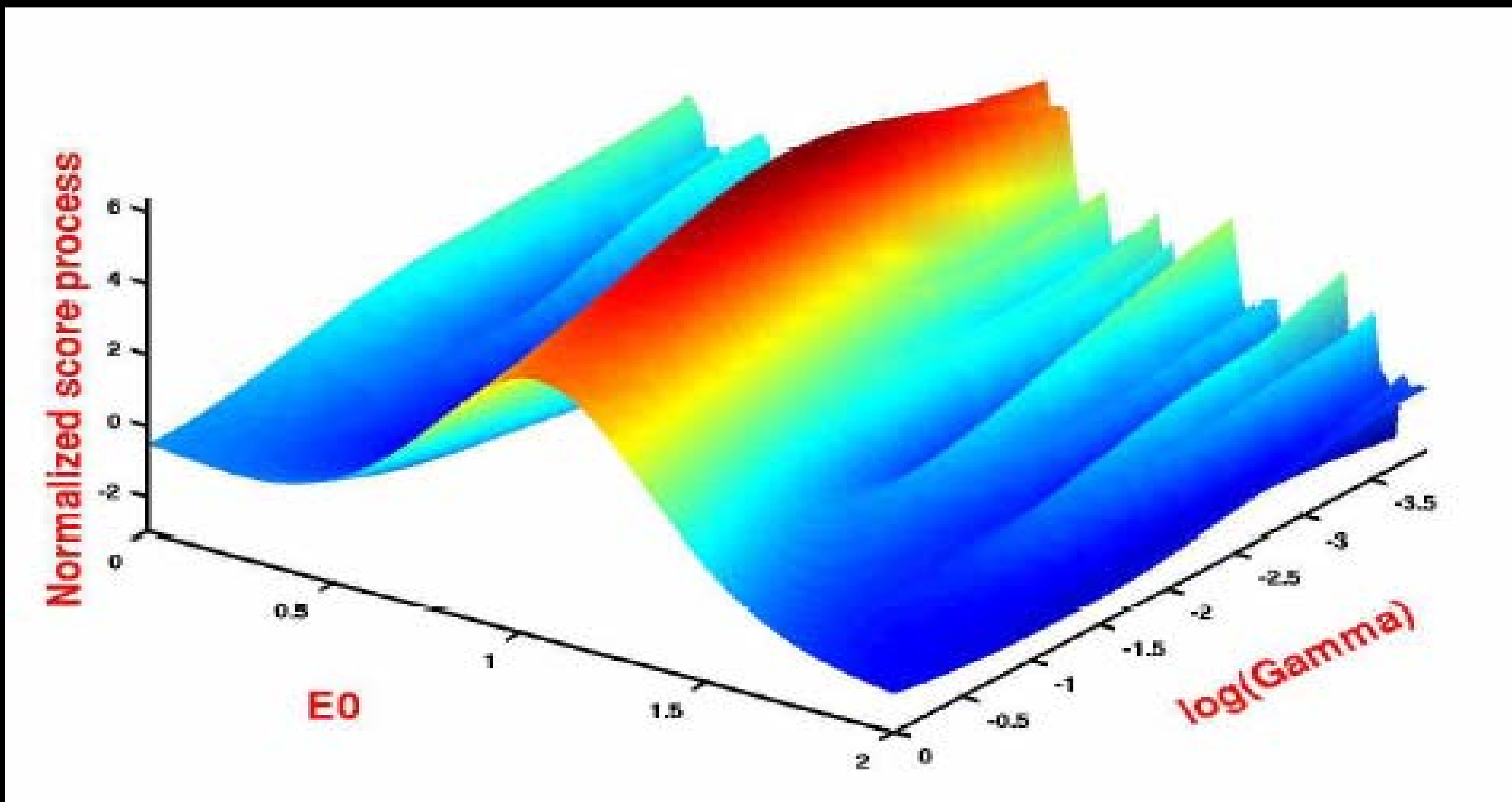
- Model parameters:

T: ñ = 0:1

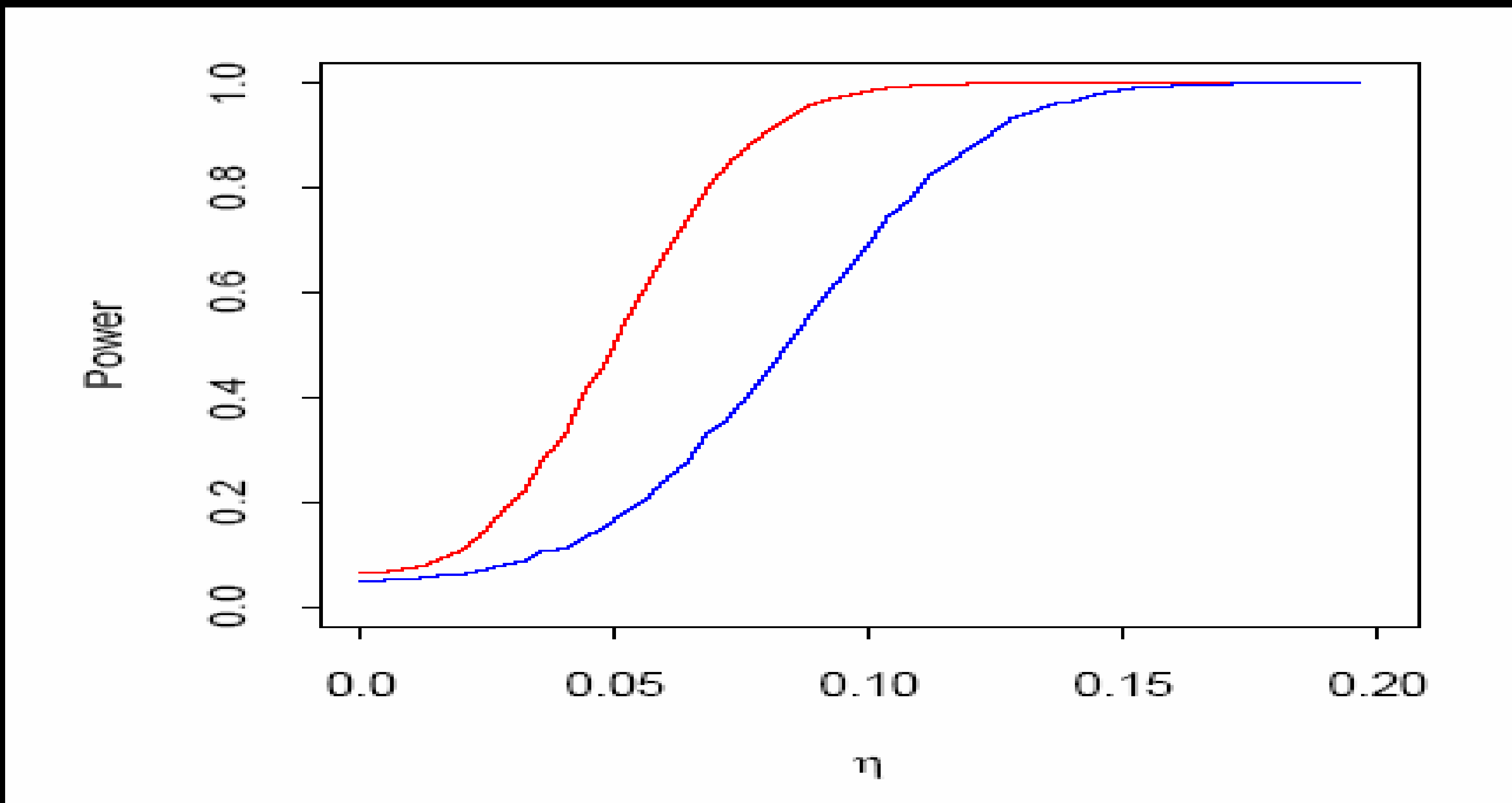


Density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
Maximum of normalized score process

Surface of normalized score process



Power of ÿ ^{2} and normalized score

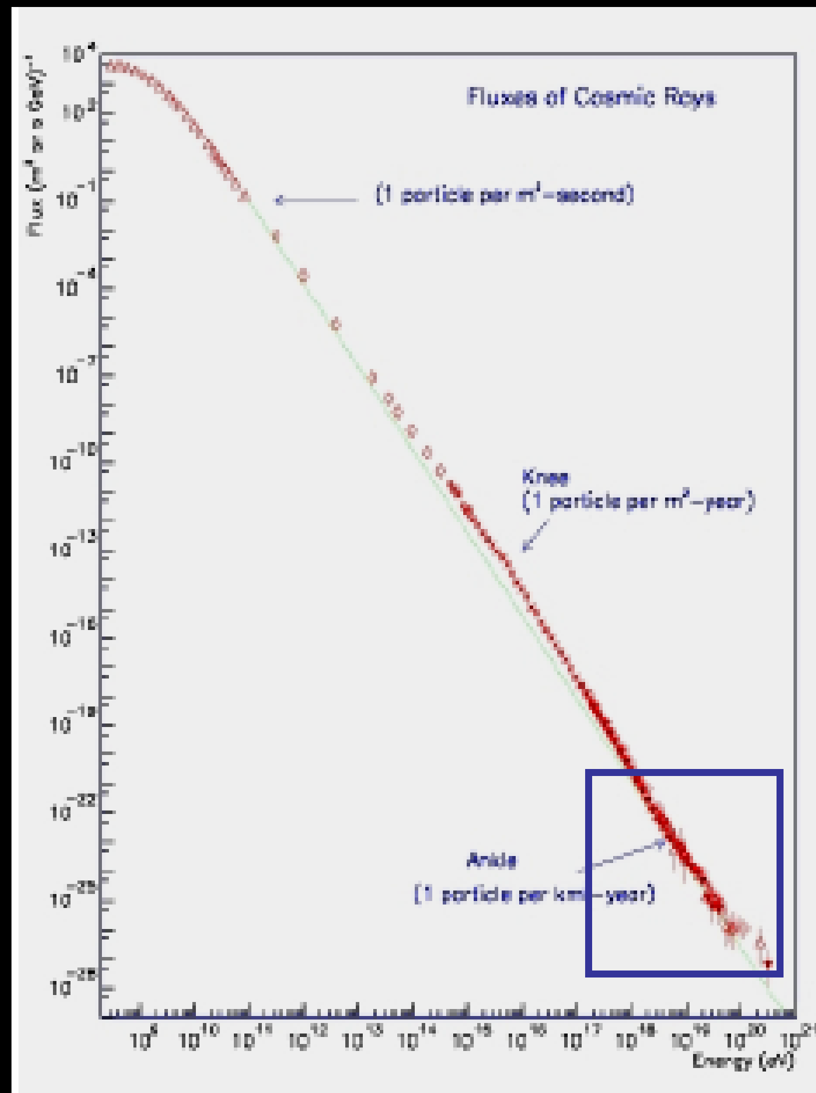


Next: LHC

- Search for Higgs
- Search for SUSY
- Search for the unexpected

Cosmic Rays

- spectrum



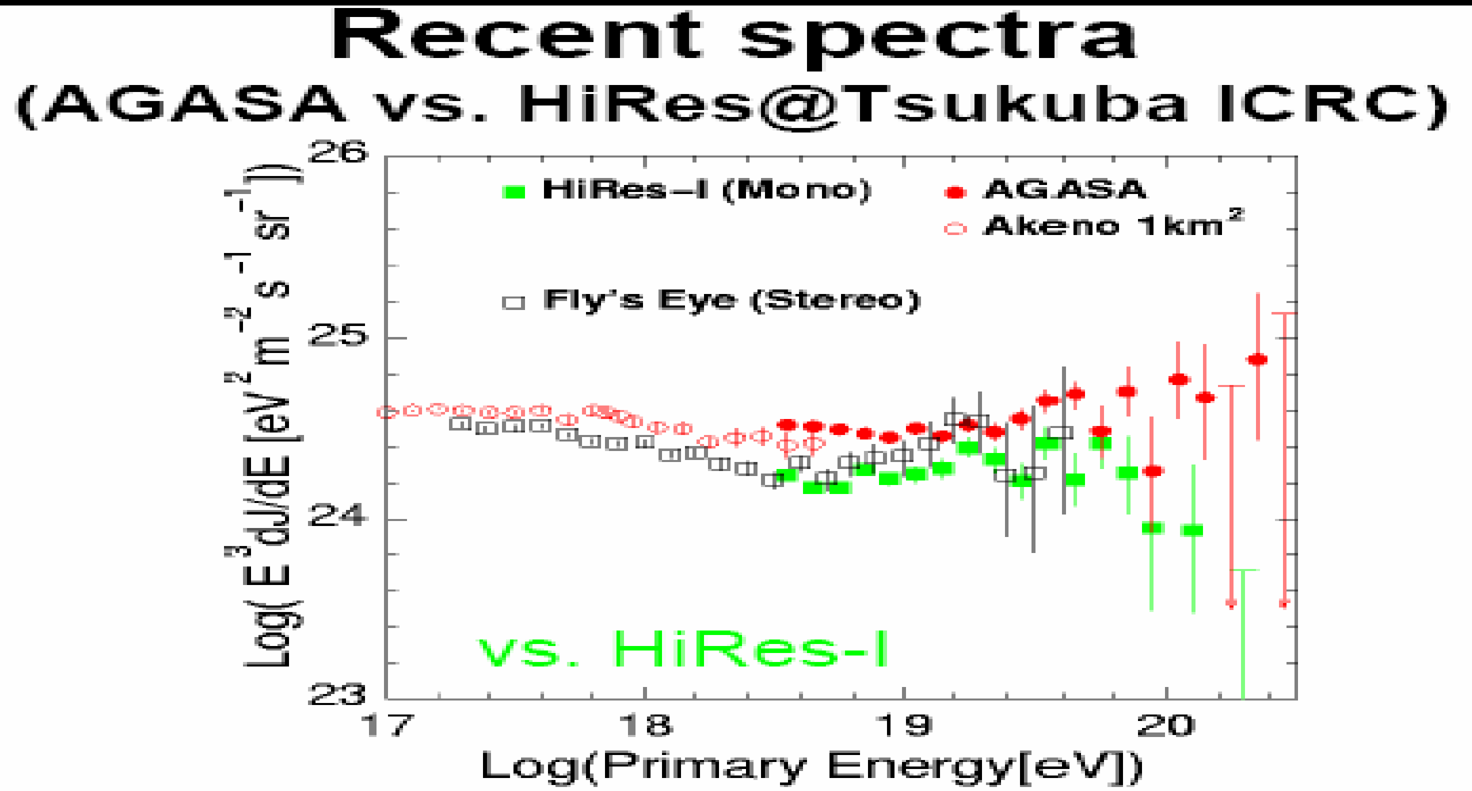
AGASA

Akeno Giant Air Shower Array

- **Detectors**
  - **111 surface detectors (2.2m^2)**
    - 5cm thinck scintillator
    - Optical fibre cable to observatory
      - Delay time monitored @100ps accuracy
    - Location suveyed: $\Delta x,y$=0.1m; Δz=0.3m
  - **27 muon detectors (2.8–10m^2)**
    - Fe / concrete absorber +proportional counters
    - E_{th}>0.5GeV

- **Operation**
  - Started in February 1990 up to now ~95% live ratio
  - **We will shut down at the end of this year…;_;**

High End of spectrum

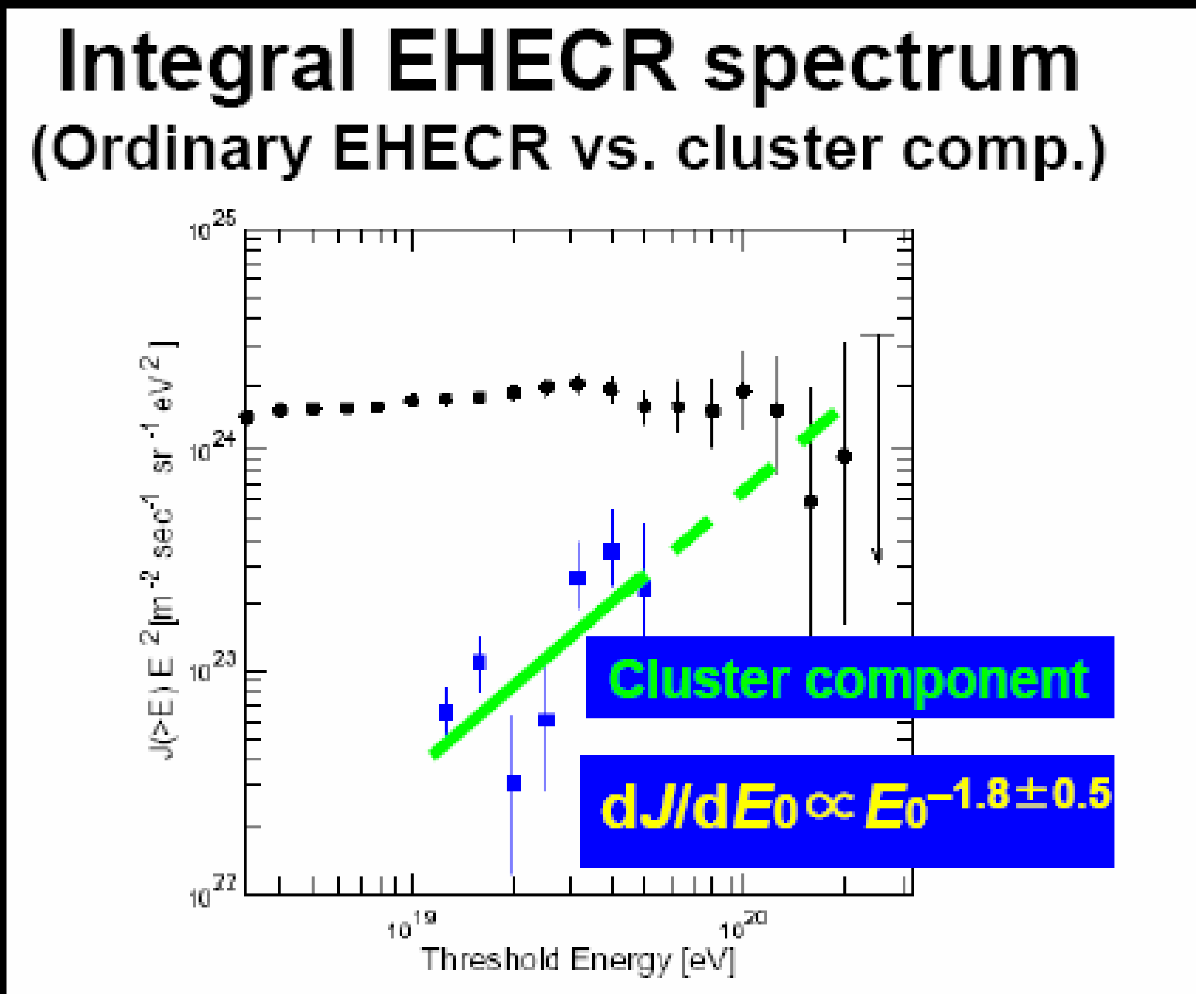


- ~2.5 sigma discrepancy between AGASA & HiRes

Greissen cutoff

- Shouldn’t be any very high energy cosmic rays – interactions with microwave background radiation
- Where are they coming from? New sources?
- Auger project: $50 million air shower array

Auger…New component?

- 

Score sensitivity for power-law mix

$$p_{õ}(x) = \frac{1}{B_{õ}} x^{à õ}$$

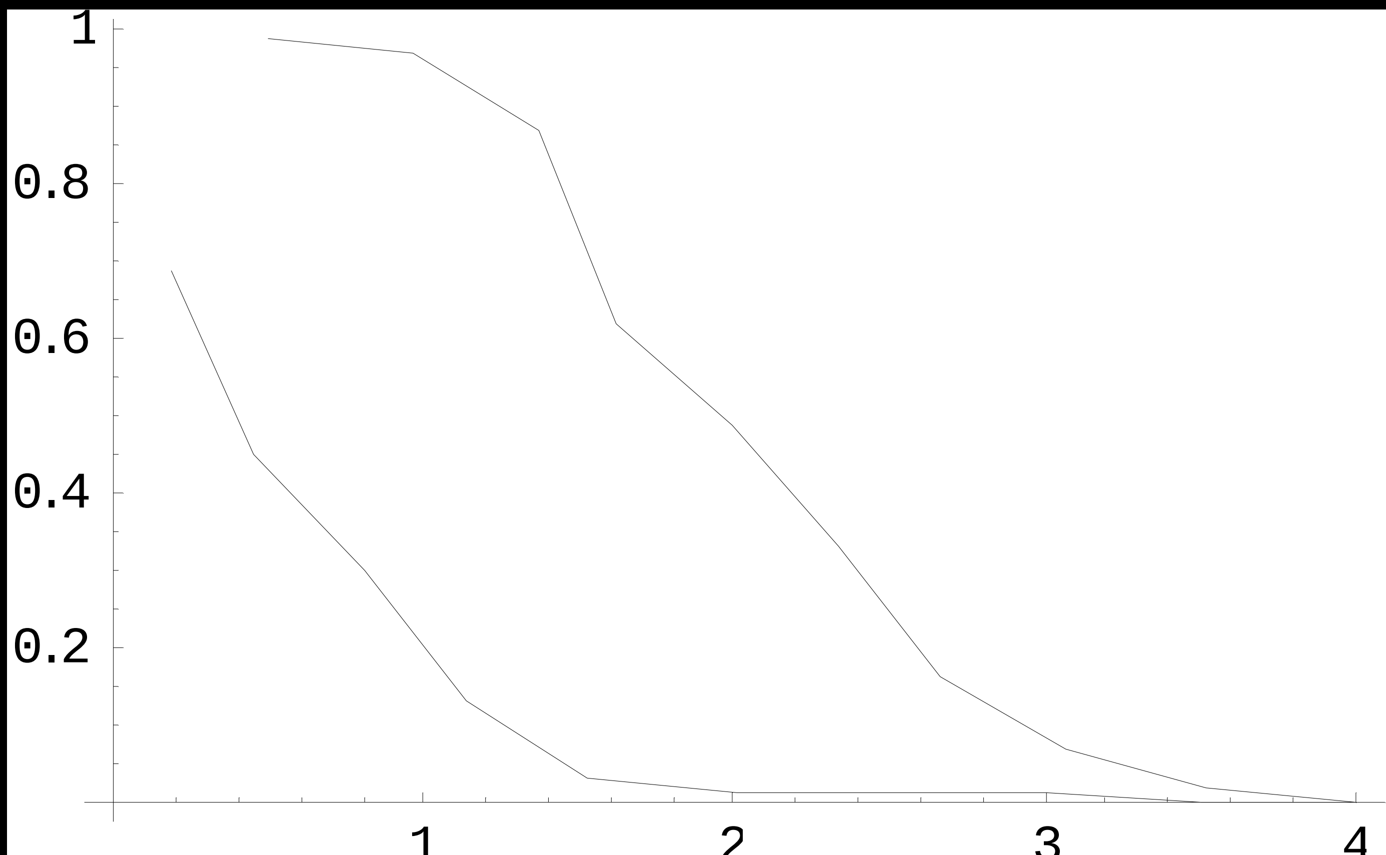
$$B_{õ} = \int_a^b x^{à õ}\, dx$$

$$H_0 = p_{ë}(x)$$

$$H_1 = (1 à ñ)\, p_{ë}(x) + ñ\, p_{ì}(x);\quad ì < ë$$

- Ex: slightly softer (~2.7 added in at 20% level); cut off at region where disagreement begins

Sensitivity   ñ = 0:2; N = 1000



Conclusions

- Score test statistic + asymptotic distribution represents a powerful new tool for the search for new physics in high energy physics and particle astrophysics

Reducing Simulation Runs for Future Combat System Key Performance Parameter Analysis

Lieutenant Colonel Thomas M. Cioppa, Ph.D.
US Army Training and Doctrine Command Analysis Center
PO Box 8695
Monterey, CA 93943-0695

Abstract: The Future Combat System (FCS) Key Performance Parameter (KPP) simulation runs executed in the stochastic Combined Arms and Support Task Force Evaluation Model (CASTFOREM) simulation can exceed 40 hours per replication. Historically, each combination of input factor levels requires 21 replications. These replications, when coupled with the large number of combinations of different input factors, can exceed manpower and computer resources. A methodology to minimize the number of replications required is proposed. Applying normality theory to this issue can be misleading since the output measures may not be normally distributed. The proposed methodology incorporates two techniques to determine the minimum number of replications required and reduce the required number of simulation runs. The first technique examines the output measure using the coefficient of variation, which is defined for the output measure as its standard deviation divided by the mean. The second technique is to use bootstrapping for the executed simulation runs. A bootstrapping sample is obtained by randomly sampling from the original data points. Bootstrap confidence intervals can be constructed to compare various alternatives. These techniques were robustly applied to recent FCS KPP CASTFOREM simulation runs and showed substantial merit.

I. INTRODUCTION

The time to execute a simulation is always of concern and interest to the analyst. The valuable and limited resource of time is best applied to ensuring the simulation setting is accurate, data inputs are valid, and sufficient time is available to analyze the simulation output to provide results to decision makers (Law and Kelton, 2000). The time problem is compounded in military simulations where scenario establishment is time-consuming. Recently, the Future Combat System (FCS) Key Performance Parameter (KPP) analysis required extensive simulation support.

The simulation used at the US Army Training and Doctrine Command Analysis Center (TRAC) at White Sands is the Combined Arms and Support Task Force Evaluation Model (CASTFOREM). CASTFOREM is used to evaluate weapon systems and unit tactics, brigade and below by simulating intense battle conditions at battalion and brigade levels. It models a range of operations to include ammunition resupply, aviation, close combat; combat service support; C3, countermobility, logistics, engineering, mine warfare, fire support, intelligence and electronic warfare, mobility; survivability, and air defense. The time to execute one replication of one scenario exceeded 40 hours.

Previous experience with CASTFOREM showed that after 21 replications of most scenarios, the variance of the measure of effectiveness (MOE) had stabilized (Cherolis, 1992). The problem faced in the FCS KPP study was that even 21 replications of a single scenario could exceed 35 days. An alternative methodology for reducing the required number of replications was needed.

II. BACKGROUND AND NOTATION

Cherolis' (1992) thesis is based on the assumptions that the replications are independent and produce a sequence of independent, identically distributed random variables X_1, X_2, X_3, ..., X_n . The Central Limit Theorem is used to derive confidence intervals and hypothesis tests. When n is sufficiently large, the distribution of the random variable is

$$\frac{\overline{X}}{s_n / \sqrt{n}} ,$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and n is the sample size. This distribution is approximately normally distributed, where $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $s_n = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$. Thus, for sufficiently large n, an approximate $100 \times (1 - \alpha)\%$ confidence interval for the population mean, μ , is given by

$$\overline{X} \pm t_{n-1,1-\alpha/2}\sqrt{\frac{s_n^{\ 2}}{n}}\ ,$$

where $0 < \alpha < 1$ and $t_{n-1,1-\alpha/2}$ is the upper $1 - a/2$ critical point for the t distribution on $n - 1$ degrees of freedom.

Although the 21 replications is commonly used, it is quite possible that more than 21 replications (thus more than 35 days) may be required per scenario or alternative depending on the precision required. A justifiable methodology that could reduce the number of replications required was needed.

III. COEFFICIENT OF VARIATION AND BOOTSTRAP

The coefficient of variation (CV) is defined as the standard deviation divided by mean. It is a statistical measure of the deviation of a variable from its mean. There are no units associated with this measure. A smaller value is better and implies less variability. The data does not have to be normally distributed. A data set with a higher CV will have a larger confidence interval than a data set with a smaller CV. The one drawback of the CV is a measure is that MOE data must be quantitative and positive. This is not considered a significant problem since the majority of MOE data from constructive simulations satisfies these two requirements. From different fields and experience, a CV less than 0.20 indicate a reasonable amount of variability.

The bootstrap method (Efron and Tibshirani, 1994) is commonly used to resample from sparse data sets. A bootstrap sample $x^* = (x^*_1, x^*_2, x^*_3, \ldots, x^*_n)$ is obtained by randomly sampling n times, with replacement, from the original data points $x_1, x_2, x_3, \ldots, x_n$. The corresponding measure of interest (e.g., mean or median) is taken. For example, assume we have seven data points of (3, 9, 8, 5, 6, 1, 10) and its mean is 6. One bootstrap sample of these seven data points might be (6, 6, 1, 8, 1, 8, 10) and its mean is 5.714. A total of 1000 bootstrap samples are done. Above is only one example of the 1000 bootstrap samples. This procedure is done rapidly (within seconds) using a computer. A bias-corrected and accelerated bootstrap confidence interval (BCa) is calculated (via computer) from the 1000 samples and can be used to compare alternatives.

The general algorithm is described. A minimum of five replications is conducted on the simulation. The CV is calculated. If the CV is less than or equal to 0.20, the five replications are bootstrapped and the BCa obtained. If after five replications, the CV is greater than 0.20, another replication is done and a new CV calculated. This procedure is terminated when the CV is less than or equal to 0.20. Note that the CV makes no assumption of normality.

IV. APPLICATION TO DATA SETS

The TRAC element at Monterey previously gained insights on MOE data characteristics from TRAC-White Sands Night Vision and Electronic Surveillance Directorate Search and Targeting Acquisition Modeling Project. In most of the MOE examined, the CV was under 0.20 in most all cases by the time the tenth replication was analyzed.

There were 36 MOE's initially identified for the FCS KPP analysis conducted by TRAC-White Sands. Each of the MOE had four alternative force structures. Thus, a total of 144 data sets existed to determine the soundness of using the CV and bootstrap. There were 11 replications per alternative.

The mean was calculated for the 11 replications. The CV, test for normality, and mean and median 90% BCa were calculated for the first five replications, then for the first six replications, then for the first seven replications, then for the first eight replications, then for the first nine replications, then for the first ten replications, and finally for all 11 replications.

As an example, the first data set included the 11 data points of (279, 287, 356, 297, 302, 291, 294, 288, 286, 352, 306). The sample mean of these 11 replications is 303.5. The sample 90% confidence interval (using parametric statistics) of the 11 replications is (289.2, 317.7), but note the data is non-normal. The true population mean is unknown which is typical in military analysis. Bootstrap samples of 1000 were taken for each of the number of replications. Note that a different 1000 bootstrap samples can yield slightly different numbers, but these differences are negligible. Furthermore, the BCa has a slightly wider confidence interval, but it does not require normality assumptions.

Table 1 shows the results of the CV's and BCa's for this data set. For five replications, it is found that the data is not normally distributed (Kolmogorov-Smirnov test for normality). The CV was 0.087 which indicates little variability in the MOE. The mean 90% BCa is (287.8, 330.4) compared to the 90% confidence interval of the 11 replications of (289.2, 317.7). This indicates that the BCa is a reasonable approximation even with over a 50% decrement in replications required. As the number of replications examined increases, the CV remains relatively constant. Furthermore, the mean 90% BCa experiences some fluctuations, but also is relatively consistent.

Number of Replications	Normal	CV	Mean 90% BCa	Median 90% BCa
5	No	0.087	(287.8, 330.4)	(279, 302)
6	No	0.087	(289.2, 327.2)	(283, 302)
7	No	0.087	(290.7, 324.5)	(279, 297)
8	No	0.087	(290, 318.1)	(287, 297)
9	No	0.087	(289.2, 316.1)	(286, 294)
10	No	0.086	(292.7, 322.8)	(287, 299.5)
11	No	0.086	(293.1, 319.8)	(287, 302)

Table 1. CV's and Confidence Intervals for Representative MOE Data.

Table 1 was for one MOE for one of the four alternatives. For this specific MOE, the other three alternatives were examined. When a CV < .20 was achieved, the procedure was terminated at that number of replications. For this particular MOE, after five replications, each alternative had a CV < .20. Table 2 provides the CV's and confidence intervals for the remaining three alternatives. Note that each of these alternatives also had 11 replications.

	True Mean	True Median	Replications	Normal	CV	Mean 90% BCa	Median 90% BCa
Alternative 2	316.3	314	5	Yes	0.131	(276.6, 318.6)	(258, 316)
Alternative 3	349.9	361	5	Yes	0.066	(337.2, 361.6)	(326, 364)
Alternative 4	235.4	241	5	Yes	0.109	(218.2, 264)	(200, 261)

Table 2. CV and Confidence Intervals for Remaining Three Alternatives for Representative MOE Data.

An important consideration for the analyst is how to present the information to senior decision makers. TRAC-White Sands commonly employs box plots to compare alternatives using the Sheffe, Tukey, Bonferroni, or Fisher's least significant differences approach. One drawback of these approaches is that each assumes normality and equal variances among the alternatives. If the normality assumption is not valid, the non-parametric Kruskal-Wallis test can be used to compare all alternatives and then be followed with the Wilcoxan test to compare a particular pair of treatments. If equal variances are not assumed, then the Games-Howell test and one of Tamhane's tests are recommended.

The proposed alternative emphasizes the BCa and does not rely on normality or equal variance assumptions. The mean 90% BCa's can be displayed and indicate the magnitude and range of the alternatives. Figure 1 illustrates this approach and uses the data from Tables 1 and 2. Thus, if "bigger is better," then Alt 3 > Alt 1 = Alt 2 > Alt 4. This is obtained by examining if significant overlap of the green bands occurs between alternatives. Since there is no overlap between Alternative 3 and the remaining alternatives, Alternative 3, having the highest band, is determined to be best. Although Alternative 1 and Alternative 2 do not coincide exactly (although substantial overlap exists in their BCa's), there is sufficient visual evidence for the senior decision maker to suggest that there is not a remarkable difference between the two. Furthermore, insight is gained that Alt 3 has the least variability.

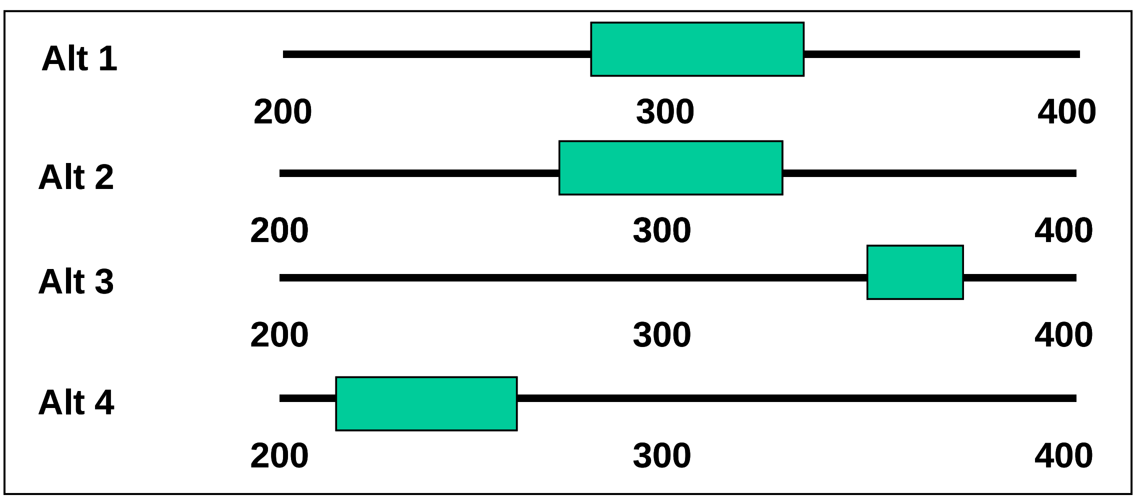


Figure 1. Comparing the Four Alternatives for the Representative MOE Data.

This methodology was executed for each of the remaining 35 MOE's. These results can be presented upon request. The purpose of this work was not to check the work of TRAC-White Sands, but to determine the merits and implementation insights of the CV and bootstrap methods.

V. SUMMARY

Although both the mean and median can be used, the mean appears to be sufficient. There was not a significant difference when comparing alternatives on whether the data was normally distributed or not (assessed using the Kolmogorov-Smirnov Test). The CV was less than .10 in almost 80% of the 144 data sets after five replications. When the bootstrap procedure was done on these five replications, the resulting mean 90% BCa included the 11-replication sample mean in all cases. The CV was less than .20 in over 86% of the 144 data sets after eight replications. When the bootstrap procedure was done on these eight replications, the resulting mean 90% BCa included the 11-replication sample mean in all cases. Approximately 14% of the CV values were greater than .20 (with some greater than .70), but after eight replication, the resulting mean 90% BCa included the 11-replication sample mean in all cases.

If the CV is high for a particular MOE in one alternative, it was found that it is high for all of the alternatives for that MOE. The CV does not significantly change from 5-11 replications. For example, from our MOE example for alternative 1, the CV after five replications was .087 and after 11 replications, the CV was .086. The magnitude of the MOE value does not effect the CV (unitless). For example, the MOE example for Alternative 1 had values ranging from 279 to 352 and had a CV of .087. Another MOE we examined had values ranging from .79 to .831 and had a CV of .017.

If there are available resources, then there is nothing that substitutes for the actual data obtained from executing the simulation. The CV value (especially when paired with a "picture" of the data) appears to be a good measure to determine how many replications are required and does not require normality assumptions. If the FCS KPP simulation runs do require significant resources (mainly time), the bootstrap appears to offer good results after five replications when compared to the 11 replications. Finally, the 90% mean BCa is an excellent analytical and visual tool to show where differences between alternatives exist.

REFERENCES

Cherolis, George. *A Sequential Stopping Rule for Reducing Production Times During CASTFOREM Studies*. Master's Thesis, New Mexico State University, 1992.

Efron, Bradley and Tibshirani, Robert J. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1994.

Law, Averil M. and Kelton, W. David. *Simulation Modeling and Analysis, 3rd Edition*. New York: McGraw Hill Publishers, 2000.

Reducing Simulation Replications for Future Combat System Analysis

TRADOC ANALYSIS CENTER

22 October 2004

Purpose

- Describe an alternative method of determining, and possibly reducing, the simulation replications required to determine differences among alternatives.

Agenda

- Coefficient of variation
- Bootstrapping
- Methodology description
- Application of methodology to FCS data sets
- Insights and summary

Background

- "Accepted" standard is to execute 21 replications in CASTFOREM.
- Due to the long run times of CASTFOREM for the FCS KPP analysis, TRAC-WSMR Deputy Director requested investigation of replications required and comparison of alternatives.
- Subsequent research by TRAC-Monterey, in conjunction with TRAC-WSMR, resulted in:
  – Coefficient of variation (CV) as a tool to determine replications required.
  – Bootstrap to reduce replications.
  – Comparing alternatives using bootstrap confidence intervals.

Coefficient of Variation

- Defined as the standard deviation divided by mean.
- A statistical measure of the deviation of a variable from its mean.
- No units associated with this measure.
- A smaller value is better and implies less variability.
- The data does not have to be normally distributed.
- A data set with a higher CV will have a larger confidence interval than a data set with a smaller CV.

Bootstrap

- **A bootstrap sample $x^* = (x^*_1, x^*_2, x^*_3, \ldots, x^*_n)$ is obtained by randomly sampling n times, with replacement, from the original data points $x_1, x_2, x_3, \ldots, x_n$. The corresponding measure of interest (e.g., mean or median) is taken.**
- **For example, assume we have seven data points of (3, 9, 8, 5, 6, 1, 10) and its mean is 6. One bootstrap sample of these seven data points might be (6, 6, 1, 8, 1, 8, 10) and its mean is 5.714.**
- **A total of 1000 bootstrap samples are done. Above is only one example. This procedure is done rapidly (within seconds) using a computer.**
- **A bias-corrected and accelerated bootstrap confidence interval (BCa) is calculated (via computer) and can be used to compare alternatives.**

Data Sets

- TRAC-Monterey previously gained insights on MOE data characteristics from WSMR's NVESD STAMP effort.
- Requested and received from WSMR FCS MOE data consisting of 36 MOE each with four alternatives (11 replications per MOE and alternative combination).
- Thus, we were provided 144 data sets to determine the potential of the CV and bootstrap.

Methodology

- This methodology was used for each of the 144 data sets.
- The mean and median were calculated for all 11 replications.
- The CV, test for normality, mean 90% BCa, and median 90% BCa were calculated for the first five replications, then for the first six replications, then for the first seven replications,…, then for all 11 replications.
- Insights for applicable CV measures, effect of non-normal data, and assessing alternatives with the BCa were gained.

Methodology Example

- The first data set included the 11 data points of (279, 287, 356, 297, 302, 291, 294, 288, 286, 352, 306).
- The true mean of these 11 replications is 303.5 and true median is 294. The true 90% confidence interval (using parametric statistics) of the 11 replications is (289.2, 317.7), but note the data is non-normal.
- 1000 bootstrap samples were taken for each of the number of replications.
- Note that a different 1000 bootstrap samples can yield slightly different numbers, but these differences are negligible.
- Note the BCa has a slightly wider confidence interval, but it does not require normality assumptions.

Number of Replications	Normal	CV	Mean 90% BCa	Median 90% BCa
5	No	0.087	(287.8, 330.4)	(279, 302)
6	No	0.087	(289.2, 327.2)	(283, 302)
7	No	0.087	(290.7, 324.5)	(279, 297)
8	No	0.087	(290, 318.1)	(287, 297)
9	No	0.087	(289.2, 316.1)	(286, 294)
10	No	0.086	(292.7, 322.8)	(287, 299.5)
11	No	0.086	(293.1, 319.8)	(287, 302)

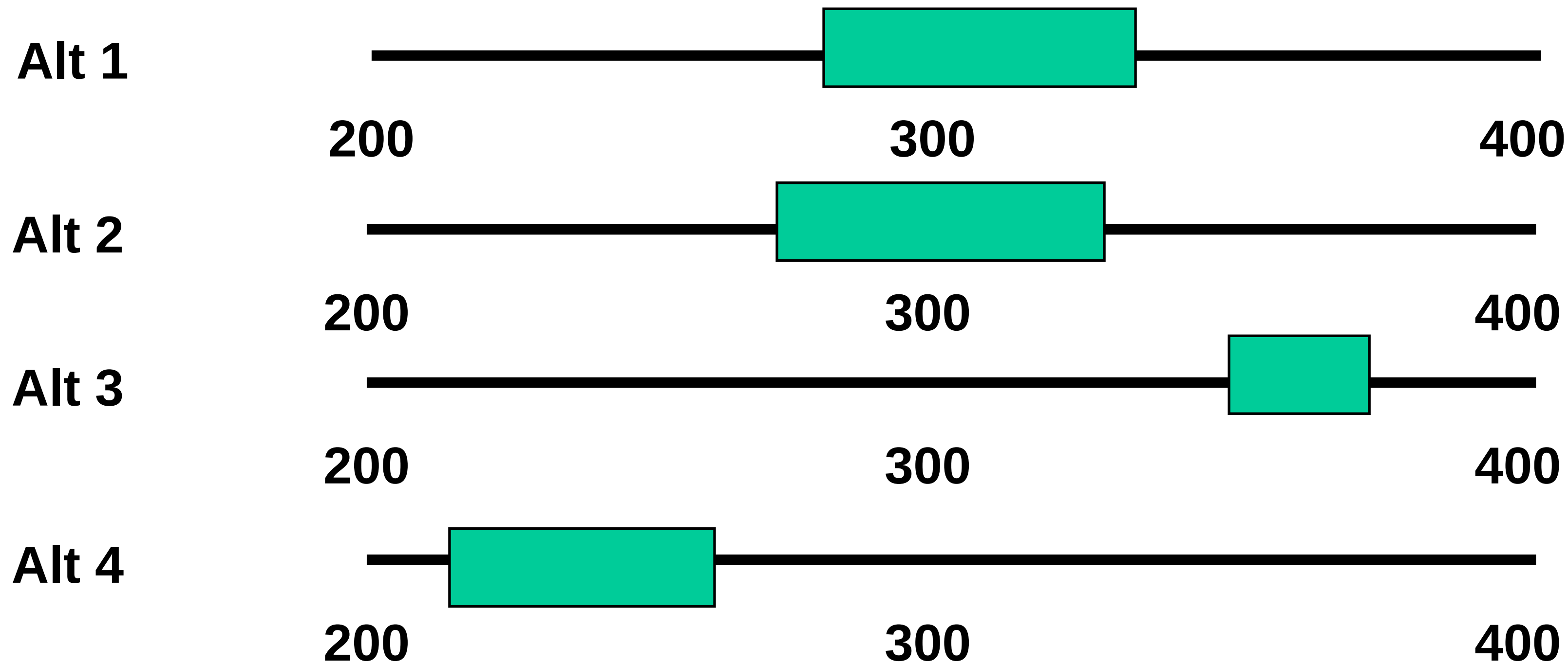
Methodology Example (continued)

- The previous slide was for one MOE on one of the four alternatives. For this specific MOE, the other three alternatives were examined.
- When a CV < .20 was achieved, the procedure was terminated at that number of replications.
- For this particular MOE, after five replications, each alternative had a CV < .20.
- "Regardless of the cost per replication, we (Law & Kelton) recommend always making at least three to five replications of a stochastic simulation to assess the variability of the X_j's."

	True Mean	True Median	Replications	Normal	CV	Mean 90% BCa	Median 90% BCa
Alternative 2	316.3	314	5	Yes	0.131	(276.6, 318.6)	(258, 316)
Alternative 3	349.9	361	5	Yes	0.066	(337.2, 361.6)	(326, 364)
Alternative 4	235.4	241	5	Yes	0.109	(218.2, 264)	(200, 261)

Comparing Alternatives

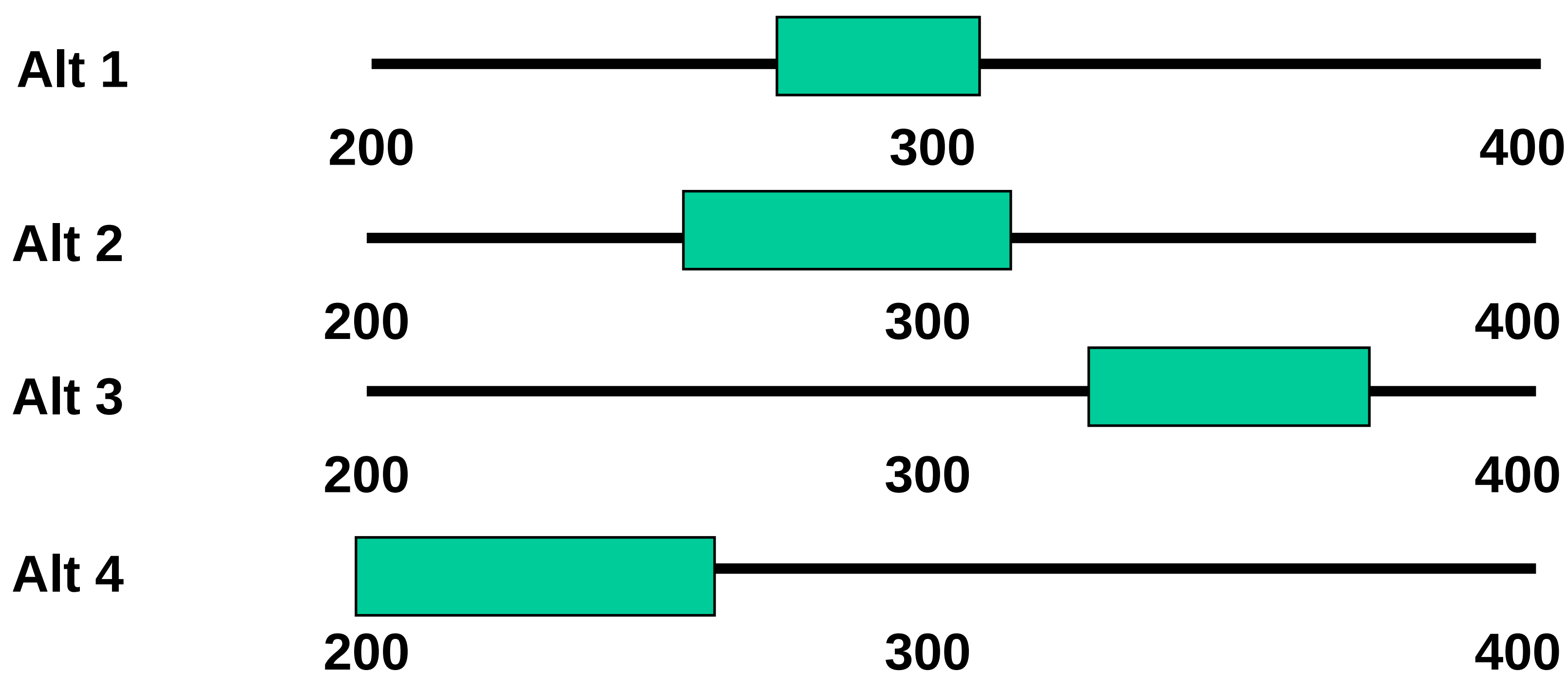
- Use the mean 90% BCa (green rectangles) to compare the MOE for the four alternatives. Experience shows that a 90% BCa provides a robust confidence interval.



- If “bigger is better,” then Alt 3 > Alt 1 = Alt 2 > Alt 4. This is obtained by examining if significant overlap of the green bands occurs between alternatives. Also, insight gained that Alt 3 has the least variability.

Comparing Alternatives (continued)

- Similarly, use the median 90% BCa (green rectangles) to compare the MOE for the four alternatives.



- Similar to the mean, Alt 3 > Alt 1 = Alt 2 > Alt 4 by examining if significant overlap of the green bands occurs between alternatives.

Analysis of Remaining MOE’s

- This procedure was executed for each of the remaining 35 MOE’s.
- These results can be presented upon request.
- The purpose of this work was not to check the work of TRAC-WSMR, but to determine the merits and implementation insights of the CV and bootstrap methods.

Methodology Insights

- **Although both the mean and median can be used, the mean appears to be sufficient. There was not a significant difference when comparing alternatives on whether the data was normally distributed or not (assessed using the Kolmogorov-Smirnov Test).**
- **The CV was less than .10 in almost 80% of the 144 data sets after five replications. When the bootstrap procedure was done on these five replications, the resulting mean 90% BCa included the true mean in all cases.**
- **The CV was less than .20 in over 86% of the 144 data sets after eight replications. When the bootstrap procedure was done on these eight replications, the resulting mean 90% BCa included the true mean in all cases.**

Methodology Insights (continued)

- Approximately 14% of the CV values were greater than .20 (some greater than .70), but after eight replications, the resulting mean 90% BCa included the true mean in all cases.
- If the CV is high for a particular MOE in one alternative, it is high for all of the alternatives for that MOE.
- The CV does not significantly change from 5-11 replications. For example, from our MOE example for alternative 1, the CV after five replications was .087 and after 11 replications, the CV was .086.
- The magnitude of the MOE value does not effect the CV (unitless). For example, the MOE example for Alternative 1 had values ranging from 279 to 352 and had a CV of .087. Another MOE we examined had values ranging from .79 to .831 and had a CV of .017.

Summary

- If you have the available resources, then there is nothing that substitutes for the actual data obtained from executing the simulation.
- The CV value (especially when paired with a “picture” of the data) appears to be a good measure to determine how many replications are required and does not require normality assumptions.
- If the FCS KPP simulation runs do require significant resources (mainly time), the bootstrap appears to offer good results after five replications when compared to the 11 replications.
- The 90% mean BCa is an excellent analytical and visual tool to show where differences between alternatives exist.

Research Directions in Adaptive Mixtures and Model-Based Clustering

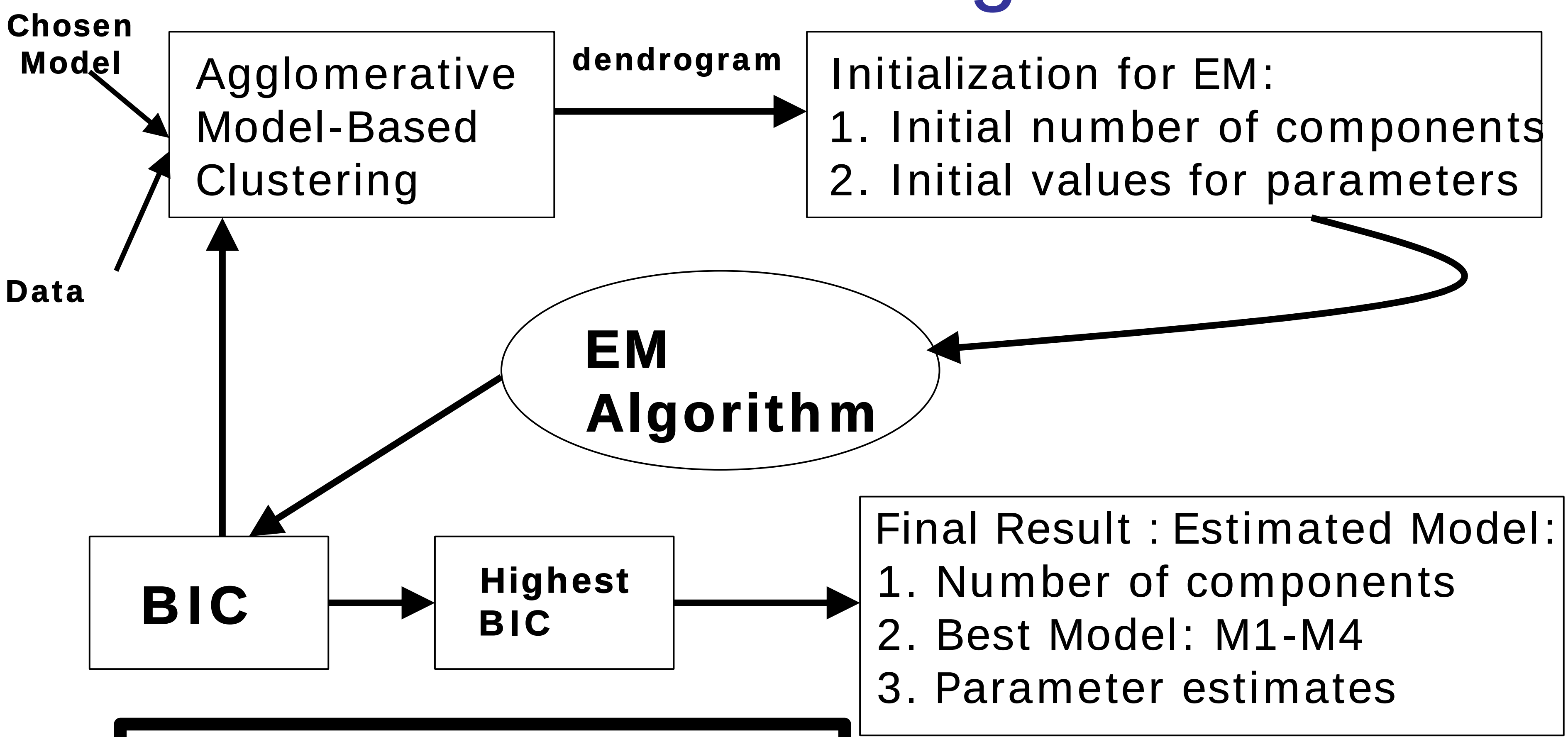
Wendy L. Martinez
Office of Naval Research
Jeffrey L. Solka
NSWCDD/GMU

ACAS 2004

Outline

- Model-based Clustering (MBC).
  - Mixture models and the EM algorithm.
  - The agglomerative step.
  - The model types.
- Adaptive Mixtures Density Estimation
- Their Synthesis
  - Initialization for MB agglomerative clustering
  - MB Adaptive Mixtures Density Estimation
- Preliminary Results.

Model-Based Clustering



Standard MBC performs hierarchical clustering starting with the full dataset, this is computationally intensive.

MODEL-BASED CLUSTERING

- This technique takes a density function approach.
- Uses finite mixture densities as models for cluster analysis.
- Each component density characterizes a cluster.

FINITE MIXTURES REVIEW

$$\hat{f}(x) = \sum_{i=1}^{g} \pi_i f_i(x,\theta)$$

$$f_i(x,\theta) = N(\mu_i, \Sigma_i)$$

- Model the density as a sum of *g* weighted densities.
- Expectation-maximization method used to estimate parameters.
- Must assume distribution for components - usually normal distribution.
- Each component characterizes a cluster.

EXPECTATION-MAXIMIZATION (EM) METHOD

- Method for building or estimating the model.
- Solution of likelihood functions requires iterative procedure.
- E Step - Expectation:
  - Find probability that observations belong to each component density - the posteriors (τ_{ij}'s).
- M Step - Maximization:
  - Update all parameters based on posteriors (π_i, μ_i, Σ_i).

EXPECTATION-MAXIMIZATION (EM) METHOD

- Issues:
  - Can converge to a local optimum.
  - Can diverge.
  - *Requires initial guess at the parameters of the component densities.*
    - *Requires initial guess at the weights (or priors).*
    - *Need an estimate of the number of components.*
  - Requires an assumed distribution for the component densities.
- *Model-based clustering addresses these issues.*

AGGLOMERATIVE MBC

- Regular agglomerative clustering:
  - Each point is in a cluster.
  - Two closest clusters are merged at each step.
  - Closeness is determined by distance and linkage.
- Agglomerative model-based clustering:
  - At each step, two clusters are merged such that the likelihood for the given model is maximized.
- We propose using Adaptive Mixtures to initialize MB agglomerative clustering.

MODEL-BASED CLUSTERING

- Best model is chosen using the Bayesian Information Criterion (m_M is # parameters, L_M is loglikelihood):

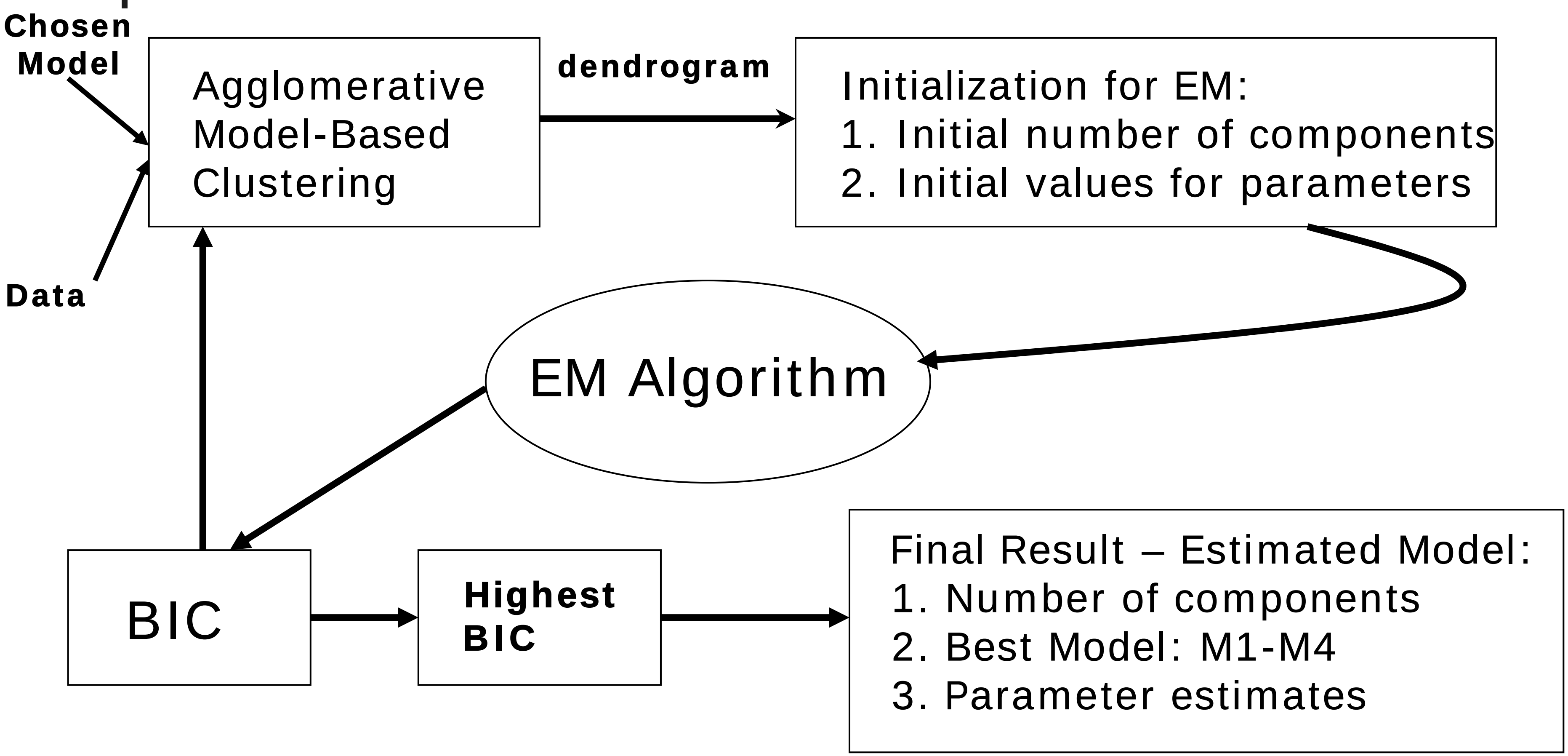
$$BIC \equiv 2L_M(\mathbf{x}, \hat{\boldsymbol{\theta}}) - m_M \log(n)$$

- The four models are (***more models are possible***):
  - Spherical/equal (M1): $\boldsymbol{\Sigma}_K = \sigma^2 \mathbf{I}$
  - Spherical/unequal (M2): $\boldsymbol{\Sigma}_K = \sigma_K^2 \mathbf{I}$
  - Ellipsoidal/equal (M3): $\boldsymbol{\Sigma}_K = \boldsymbol{\Sigma}$
  - Ellipsoidal/unequal (unconstrained) (M4): $\boldsymbol{\Sigma}_K = \boldsymbol{\Sigma}_K$

MODEL-BASED CLUSTERING in a Nutshell

1. Apply the unconstrained agglomerative MBC procedure.
2. Choose number of clusters/densities, g.
3. Choose model: M1 – M4.
4. Find the partition given by step 1 for the specified g.
5. Using this partition, find the weights, means and covariances for each term, based on the model in step 3.
6. Using the chosen g (step 2) and the initial values (step 5), apply the EM algorithm.
7. Calculate the BIC for this value of g and M.
8. Go to step 3 to choose another value of M and repeat.
9. Go to sep 2 and choose another model g and repeat.

MODEL-BASED CLUSTERING



ADAPTIVE MIXTURES DENSITY ESTIMATION (AMDE)

- Priebe and Marchette; 1990s.
- Hybrid of Kernel Estimator and Mixture Model.
- Number of Terms Driven by the Data.
- L1 Consistent.

AMDE ALGORITHM

1 - Given a New Observation.

2 - Update Existing Model Using the Recursive EM.

or

3 - Add a New Term to “Explain” This Data Point.

Recursive EM Update Equations

$$\hat{\tau}_{n+1}^{(i)} = \frac{\pi_n^{(i)} \hat{f}^{(i)}(\vec{x}_{n+1};\hat{\theta}_n)}{\sum_{t=1}^{g} \pi_n^t \hat{f}^{(t)}(\vec{x}_{n+1};\hat{\theta}_n)}$$

$$\hat{\pi}_{n+1}^{(i)} = \hat{\pi}_n^{(i)} + \frac{1}{n}(\hat{\tau}_{n+1}^{(i)} - \hat{\pi}_n^{(i)})$$

$$\hat{\mu}_{n+1}^{(i)} = \hat{\mu}_n^{(i)} + \frac{\hat{\tau}_{n+1}^{(i)}}{n\hat{\pi}_n^{(i)}}[(\vec{x}_{n+1} - \hat{\mu}_n^{(i)})A - \hat{\Sigma}_n^{(i)}]$$

$$A = (\vec{x}_{n+1} - \hat{\mu}_n^{(i)})^T$$

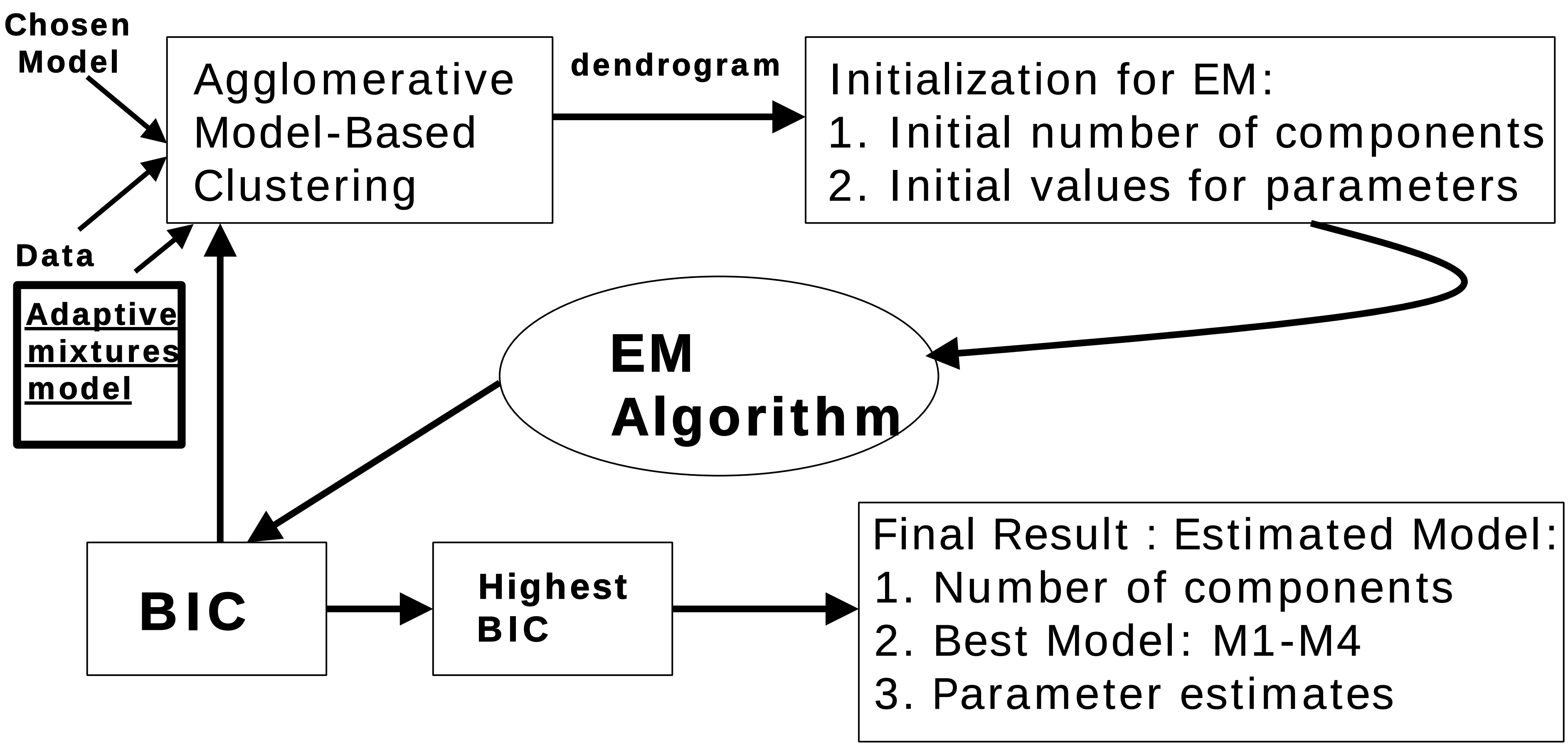
Similarly for Σ

CREATE RULE - AMDE

- Test the Mahalanobis distance from current data point to each mixture term in the existing model.

- Add in a new term when this distance exceeds a certain “create threshold”
  - Location given by current data point.
  - Covariance given by weighted average of the existing covariances.
  - Mixing coefficient set to 1/n.

MBC with an MADE Start



MBC With AMDE Smart Start

1. Form an adaptive mixtures model of the dataset. (Set create threshold in order to guarantee an over determined model.)
2. Partition the data based on the AMDE model using τ_{ij}. (Note some of the original AMDE mixture terms “die” due to insufficient support.)
3. Utilize this partition as a start to the usual MBC procedure. (Instead of starting with as many terms as points we start with approximately log(n) number of points.)

Other Possibilities

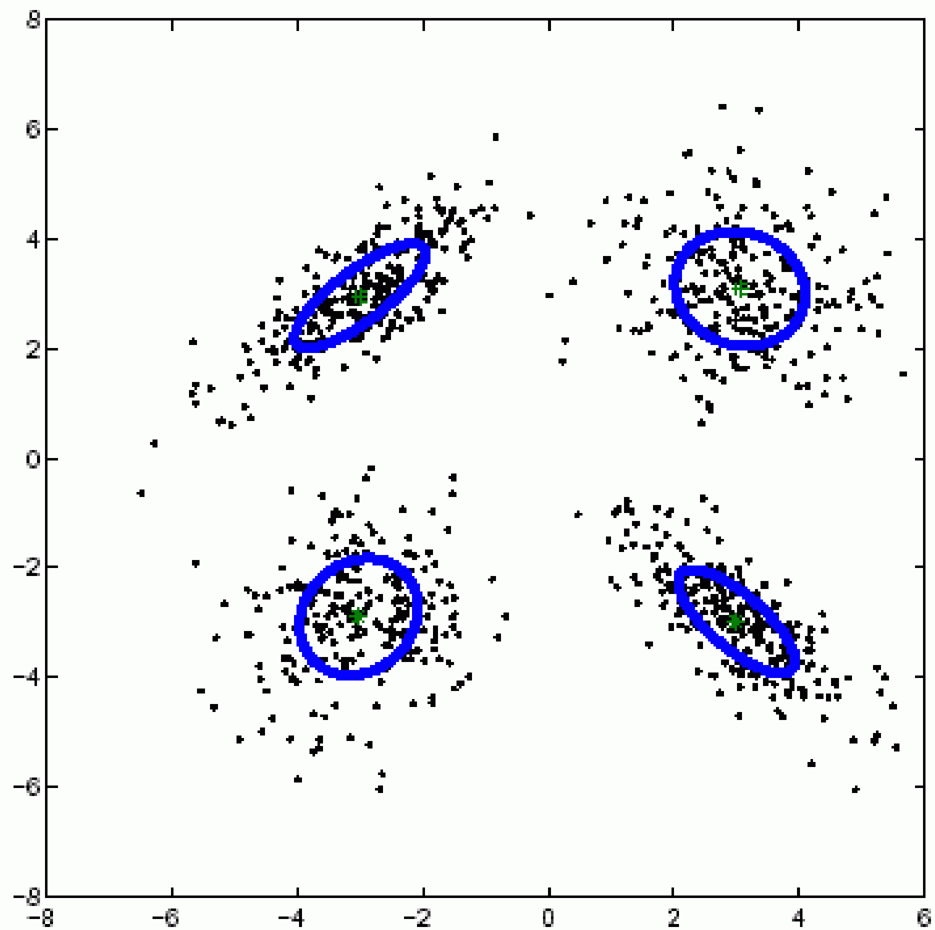
- Other types of initialization:
  - Posse (JCGS) used initial partitions based on minimal spanning tree.
  - K-means

- Benefits of AMDE initialization:
  - Do not have to specify number of clusters as in k-means.
  - Methods like k-means impose a certain structure.
  - In most cases, initial clusters are not singletons.

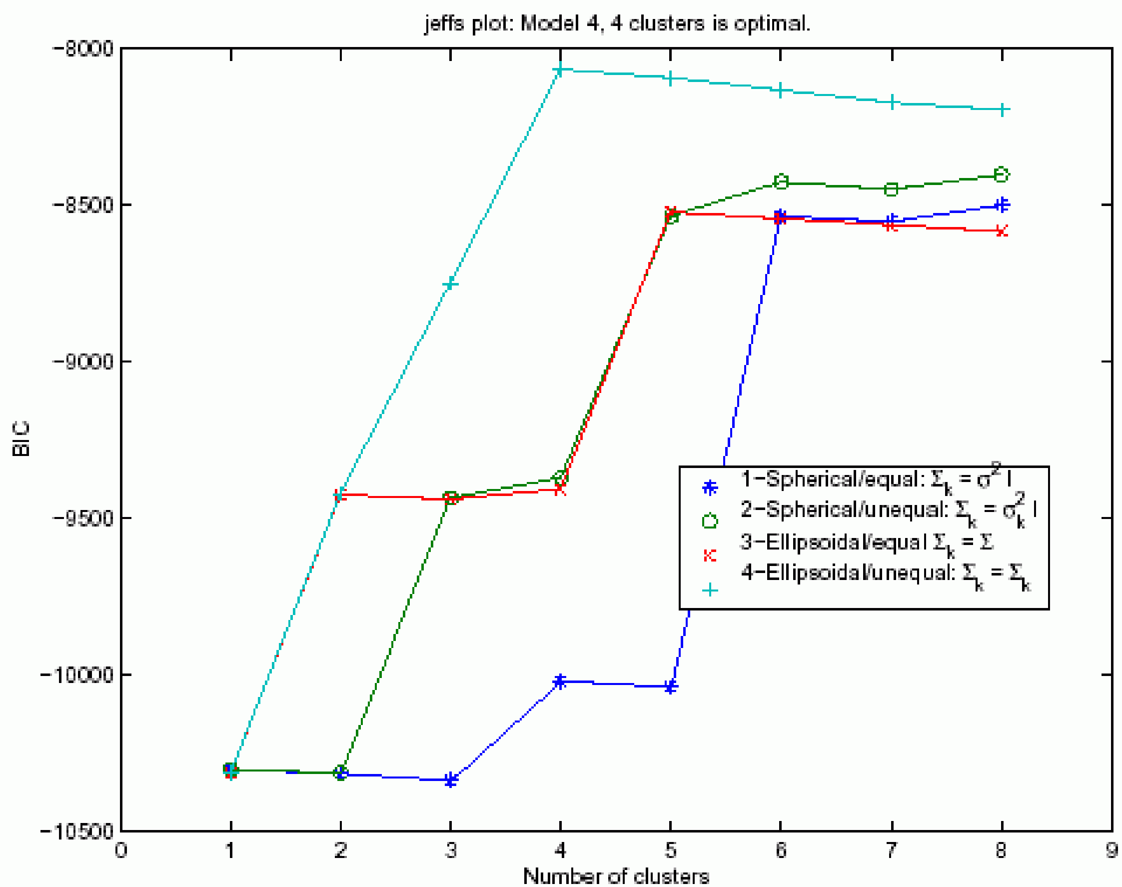
Why Do This?

- Computational tradeoff of the AMDE procedure vs. the agglomerative procedure on the full dataset.
  - Advantages as the size of the dataset grows.
  - Non-singleton clusters
  - Save on storage
- AMDE is data order dependent.
  - Multiple mixture models/clustering can be obtained by merely reordering the dataset.
  - Could get a distribution of models (number of clusters/BICs)

4 Term Test Case



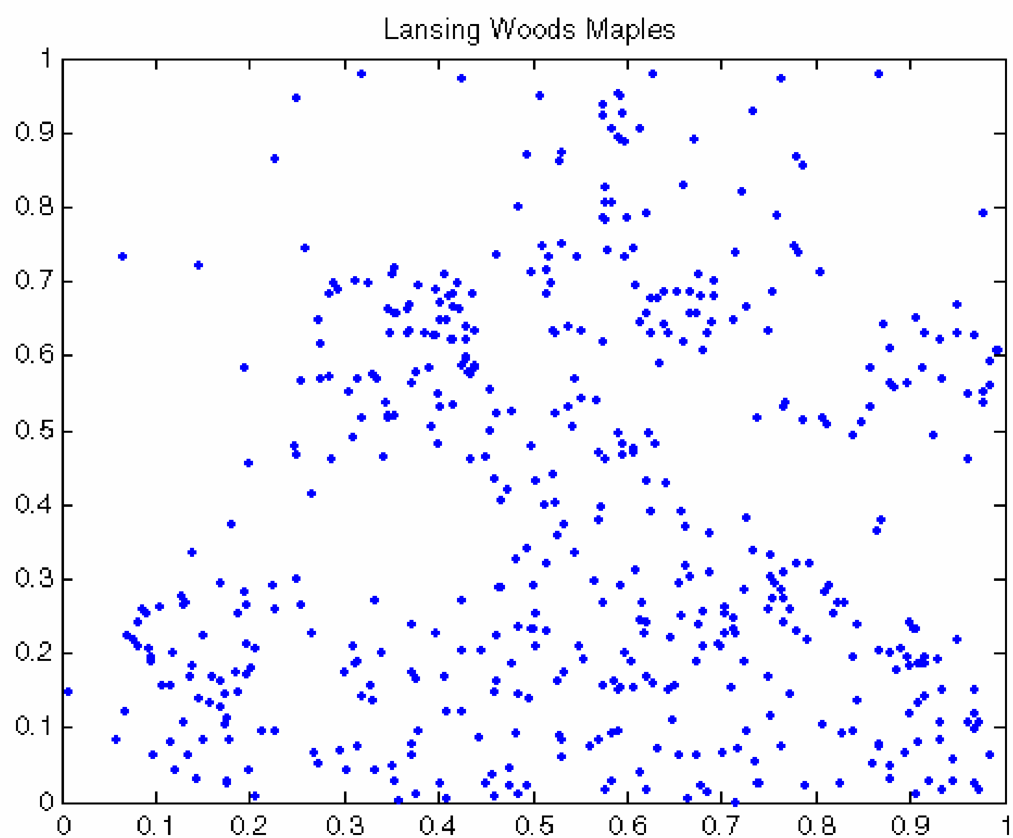
4 Term BIC Curves



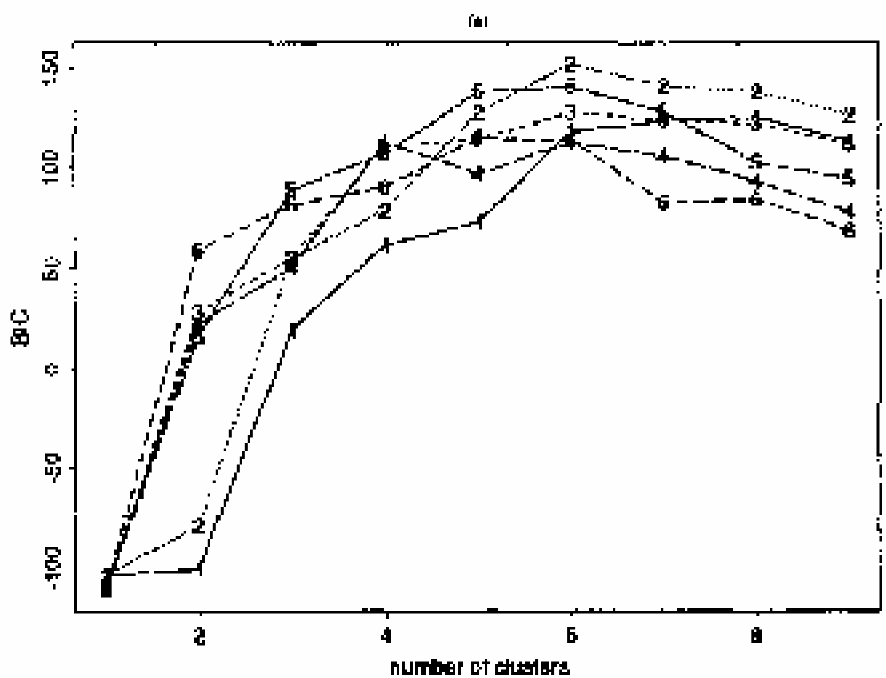
Experiment – Real Data

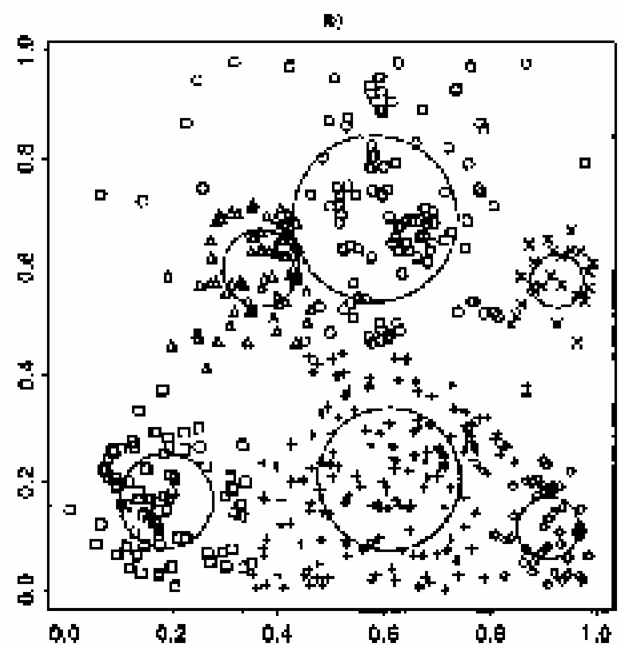
- Model-based clustering was applied to Lansing Woods maples.
- Ran 20 trials with AMDE initialization.
- Re-ordered data each time.
- Maximum BIC model is 6 component non-uniform spherical mixture.
- This is model 2:
  - Covariances are diagonal – equal variances.
  - Covariances are not equal across terms.

The Raw Data

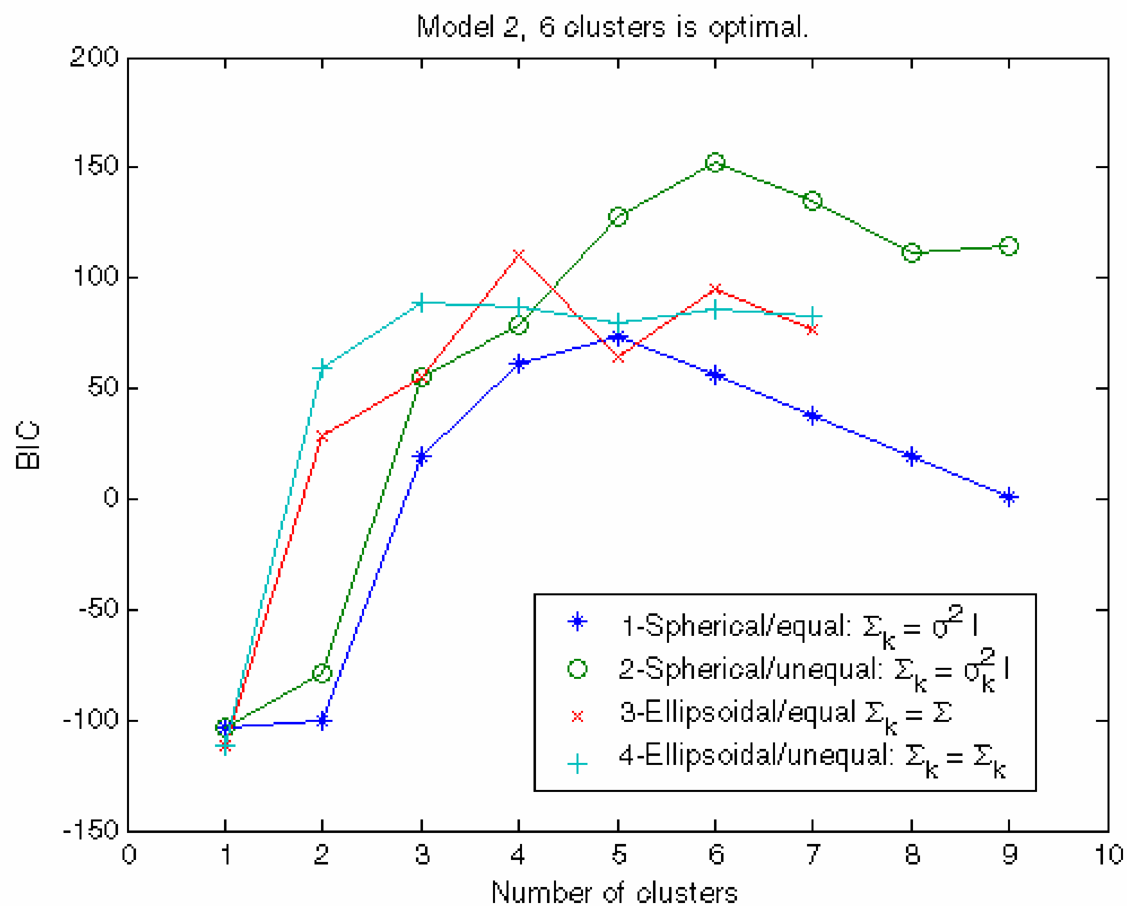


Original Configuration

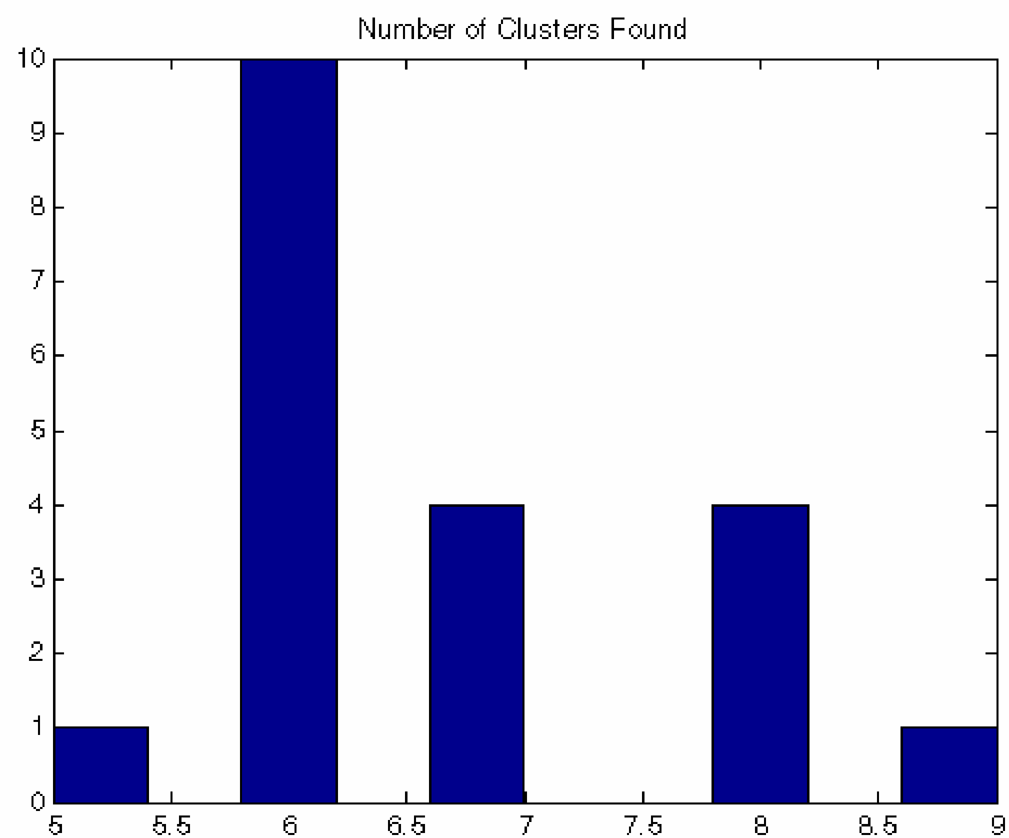




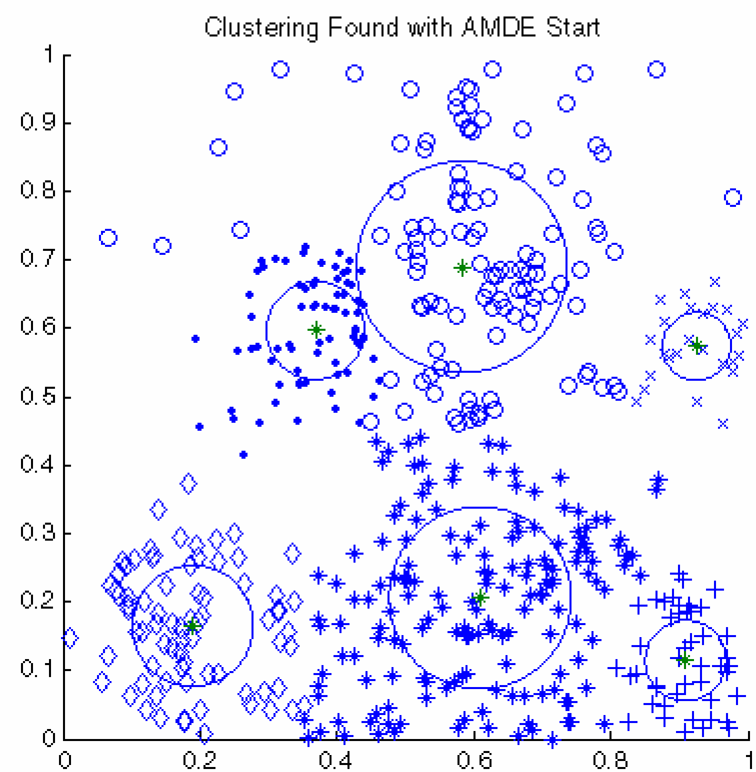
BICS for Best Trial



Number of Clusters



Configuration with AMDE



Conclusion

- Discussed an initialization procedure for the model-based agglomerative clustering.
- Showed applications to synthetic and real data.
- Possible advantages:
  - Savings in storage.
  - Possibly find other solutions – greedy algorithm
- Formulation of Model-Based AMDE.
- Use of MB-agglomerative clustering as a way of pruning terms.

Techniques for Sample Size

Russ Lenth
University of Iowa
russell-lenth@uiowa.edu

Software is available at
http://www.stat.uiowa.edu/~rlenth/Power/

Army Conference on Applied Statistics
October 22, 2004

The University of Iowa

Power basics

Power function

Given a test T of a parameter θ (scalar or vector)

$$\pi(\theta, n, \alpha, \phi) = \Pr(T \text{ is "significant"} \mid \theta, n, \alpha, \phi)$$

where α is the significance level, n is the sample size, and ϕ represents other parameters (e.g., σ^2)

Power basics

Sample-size determination

$$n(\theta^*) = \min\{n : \pi(\theta^*, n, \alpha, \phi) \geq \pi_0\}$$

with θ^* set at a clinically [scientifically] important value of θ.
Typically, people choose $\pi_0 = .8$, $\alpha = .05$.

Two-sample *t* test of significance

- $\theta = \mu_T - \mu_C$, treatment vs. control
- $n = \{n_T, n_C\}$ (often, constrain $n_T = n_C = n$)
- $\phi = \{\sigma_T, \sigma_C\}$ (often, constrain $\sigma_T = \sigma_C = \sigma$)
- Test statistic for $H_0^s : \mu_T = \mu_C$ vs. $H_1^s : \mu_T \neq \mu_C$:

$$U_s = \hat{\theta}/\widehat{SE}(\hat{\theta}) \sim t'(\nu, \theta/SE(\hat{\theta}))$$

with d.f. ν (may be estimated, approximate)

- Power function:

$$\pi_s(\theta, n_T, n_C, \alpha, \sigma_T, \sigma_C) = \Pr(U_s < -t_{\alpha/2,\nu}) + \Pr(U_s > t_{\alpha/2,\nu})$$

THE UNIVERSITY OF IOWA

Two-sample *t* test of equivalence

- Same sampling situation
- Let τ be a threshold for "smallness"
- Test statistic for $H_0^e : |\mu_T - \mu_C| \geq \tau$ vs. $H_1^e : |\mu_T - \mu_C| < \tau$:

$$U_e = \frac{\min\{\hat{\theta} + \tau, \tau - \hat{\theta}\}}{\widehat{SE}(\hat{\theta})} = \frac{\tau - |\hat{\theta}|}{\widehat{SE}(\hat{\theta})}$$

- Power function:

$$\pi_e(\theta, n_T, n_C, \alpha, \sigma_T, \sigma_C) = \Pr(U_e > t_{\alpha,\nu})$$

- This is equivalent to two one-sided t tests of size α; combined test is conservative.

The University of Iowa

Practical example

Strength of two materials

- Goals
  - Want ability to detect a 15% difference ($\theta^* = \log_e 1.15 = 0.14$)
  - A difference of less than 15% is negligible ($\tau = \log_e 1.15 = .14$)
  - Tests with $\alpha = .05$, power goal of $\pi_0 = 80$
- Pilot data on $Y = \log_e$ strength: $\sigma \approx .20$ independent of mean.
- Using GUI. . .
  - Sample size for each test
  - Graphs
  - Budget-based calculations

THE UNIVERSITY OF IOWA

Just the FAQs

Most e-mail questions I get center on two issues

- Sample size for a “medium” effect (per J. Cohen books)
- Retrospective (observed) power

I have some opinions about these. . .

Retrospective power

Compute power based on. . .

- Observed effect size
- Observed SD(s)
- Same sample size and significance level

Rationale: If result is nonsignificant, is it because. . .

- Effect size is too small? ← high retrospective power
- Sample size is too small? ← low retrospective power

The University of Iowa

Retrospective power

Compute power based on. . .

- Observed effect size
- Observed SD(s)
- Same sample size and significance level

Rationale: If result is nonsignificant, is it because. . .

- Effect size is too small?
- Sample size is too small?
- Answer: The power is *always* small in this case (duh!)

Retrospective power—another approach

Given. . .

- Observed effect size
- Observed SD(s)
- Same sample size and significance level

Then the outcome of the test is also known

- Recall that power = $\Pr\{\text{Reject } H_0\}$
- GUI example

THE UNIVERSITY OF IOWA

T-shirt sizes for effects

Power of a two-sample *t* test depends on $d = |\mu_1 - \mu_2| / \sigma$

- Small: $d = .15$
- Medium: $d = .25$
- Large: $d = .40$

T-shirt sizes for effects

Power of a two-sample *t* test depends on $d = |\mu_1 - \mu_2| / \sigma$

- ✲ Small: $d = .15$
- ✲ Medium: $d = .25$
- ✲ Large: $d = .40$

Q. Who is the T-shirt supposed to fit?

- ✲ Human? $\sigma = 1$, say Medium : $|\mu_1 - \mu_2| = .25$ in

T-shirt sizes for effects

Power of a two-sample *t* test depends on $d = |\mu_1 - \mu_2| \,/\, \sigma$

- ✲ Small: $d = .15$
- ✲ Medium: $d = .25$
- ✲ Large: $d = .40$

Q. Who is the T-shirt supposed to fit?

- ✲ Human? $\sigma = 1$ Medium : $|\mu_1 - \mu_2| = .25$ in
- ✲ Hippo? $\sigma = 25$ Medium : $|\mu_1 - \mu_2| = 6.25$ in

T-shirt sizes for effects

Power of a two-sample *t* test depends on $d = |\mu_1 - \mu_2| / \sigma$

- ✲ Small: $d = .15$
- ✲ Medium: $d = .25$
- ✲ Large: $d = .40$

Q. Who is the T-shirt supposed to fit?

- ✲ Human? $\sigma = 1$ $\text{Medium} : |\mu_1 - \mu_2| = .25$ in
- ✲ Hippo? $\sigma = 25$ $\text{Medium} : |\mu_1 - \mu_2| = 6.25$ in
- ✲ Mouse? $\sigma = .04$ $\text{Medium} : |\mu_1 - \mu_2| = .01$ in

A definitive study

. . . is not based on generic criteria

- Specify effect size on the actual scale of measurement, based on well-considered scientific goals
- Need to know σ, approximately *
- If you can't do these things, reaching the bottom line is a matter of luck
- * Except possibly in cases where σ defines population norms

Planning a pilot study

One simple approach

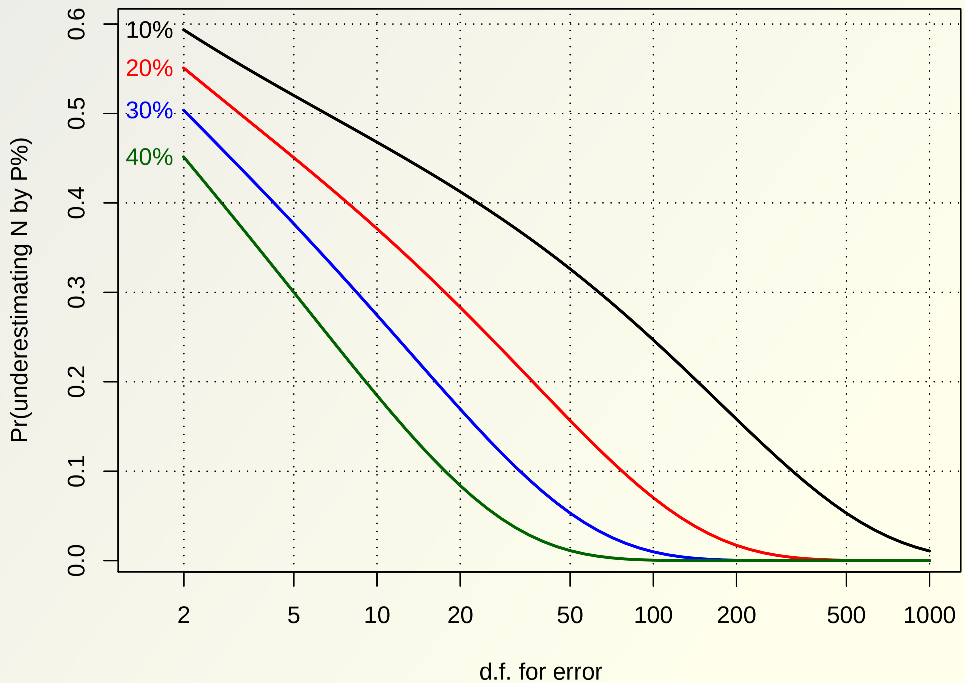
- Control the probability of under-estimating N by a specified percentage P
- Percentage by which N is underestimated = percentage by which σ^2 is underestimated
- In normal case,

$$\begin{aligned} \Pr(S^2 \le (1-P)\sigma^2) &= \Pr(\nu S^2/\sigma^2 \le (1-P)\nu) \\ &= \Pr(Q \le (1-P)\nu) \end{aligned}$$

where S^2 has ν d.f. and $Q \sim \chi^2_\nu$

THE UNIVERSITY OF IOWA

Chart for planning (or fudging)



THE UNIVERSITY OF IOWA

Example: Semiconductor experiment

Structure

- Response measure: Oxide thickness of silicon wafers
- Three whole-wafer treatments
- n lots of three wafers each: one wafer per treatment
- Three sites per wafer

Target effect sizes (for power .80, .05 sig. level)

- Difference of ± 10 between two treatments
- Difference of ± 5 between two site means
- Difference of ± 15 between two treatment*site means

Available data

(From R package `nlme`) 4 lots of 3 wafers each from each of two sources; 3 sites/wafer

source × site

|

LOT

|

WAFER

```
Error: Lot
          Df Sum Sq Mean Sq F value Pr(>F)
Source     1 1830.1  1830.1  1.5261 0.2629
Residuals  6 7195.2  1199.2

Error: Lot:Wafer
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals 16 1922.67  120.17

Error: Within
            Df Sum Sq Mean Sq F value Pr(>F)
Site         2  15.44    7.72  0.6416 0.5313
Source:Site  2  58.33   29.17  2.4234 0.1004
Residuals   44 529.56   12.04
```

SD estimates

- SD(LOT) $\approx \sqrt{(1200 - 120)/9} \approx 11.0$ (not really needed)
- SD(WAFER in LOT) $\approx \sqrt{(120 - 12)/3} = 6.0 \rightarrow$ SD(LOT×treat)
- SD(ERROR) $\approx \sqrt{12} \approx 3.5$

Summary

- Power/sample size is technically messy
- Often have multiple objectives
- Sometimes need to re-define goals
- A flexible user interface can help