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

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

Malcolm S. Taylor
OAO Corporation

Abstract

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

Introduction

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

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

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

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

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

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

Time Series Model

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

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

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

Bootstrap Application

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

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

Figure 1. Serbian Text

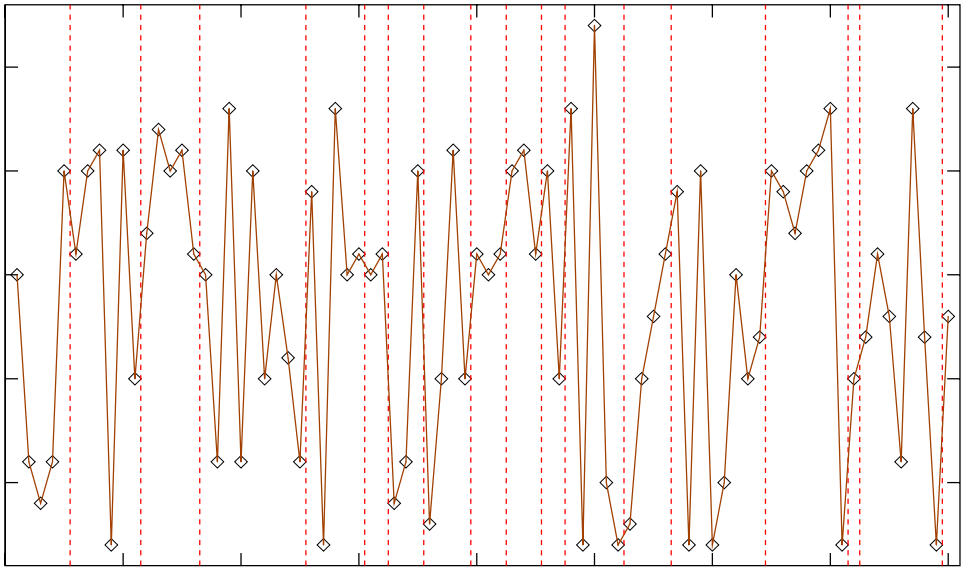


Figure 2. Time Series Representation

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

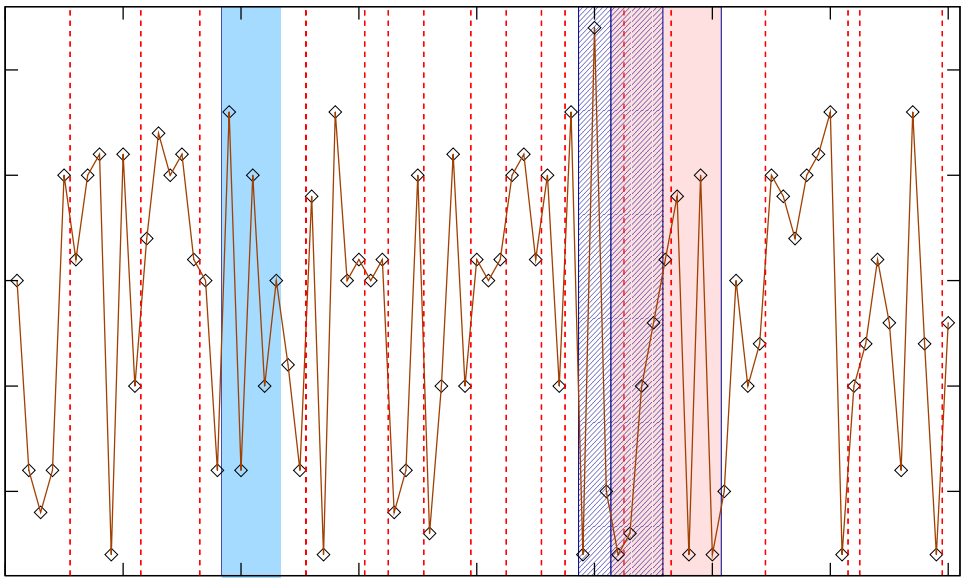


Figure 3. Intermediate Results

Empirical Results

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

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

‡A modified moving-blocks bootstrap.

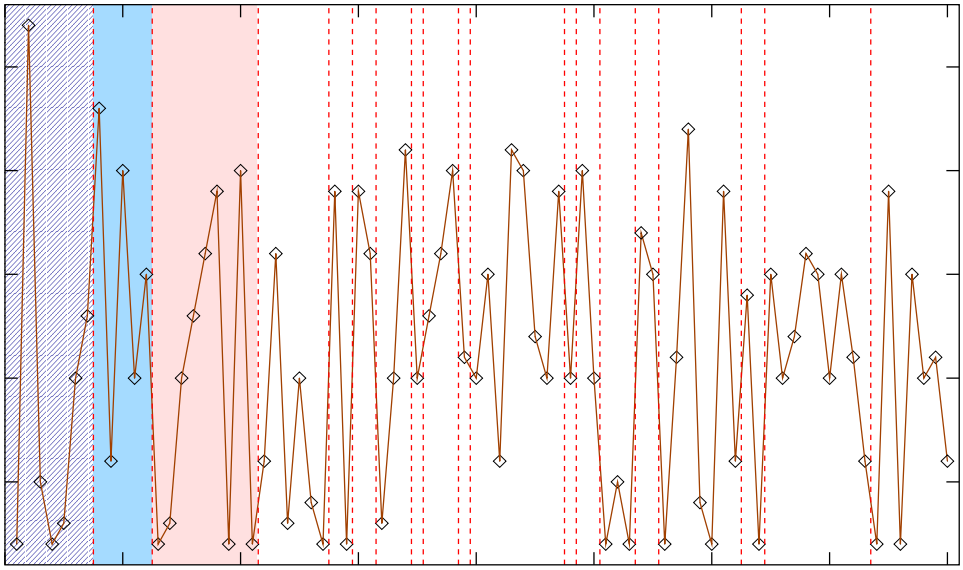


Figure 4. Bootstrapped Time Series

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

Figure 5. Bootstrapped Text

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

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

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

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

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

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

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

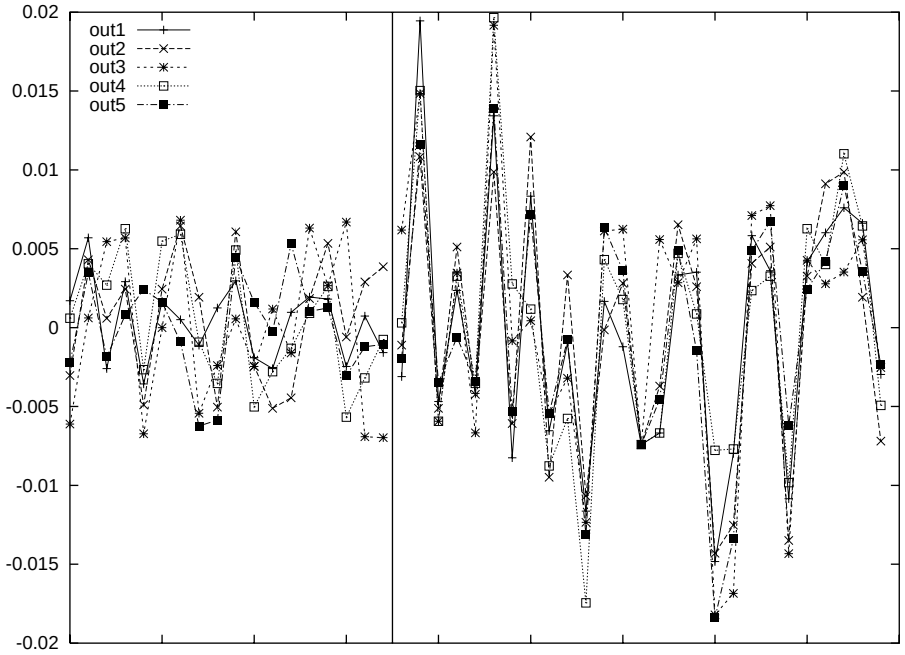


Figure 6. Frequency Differences

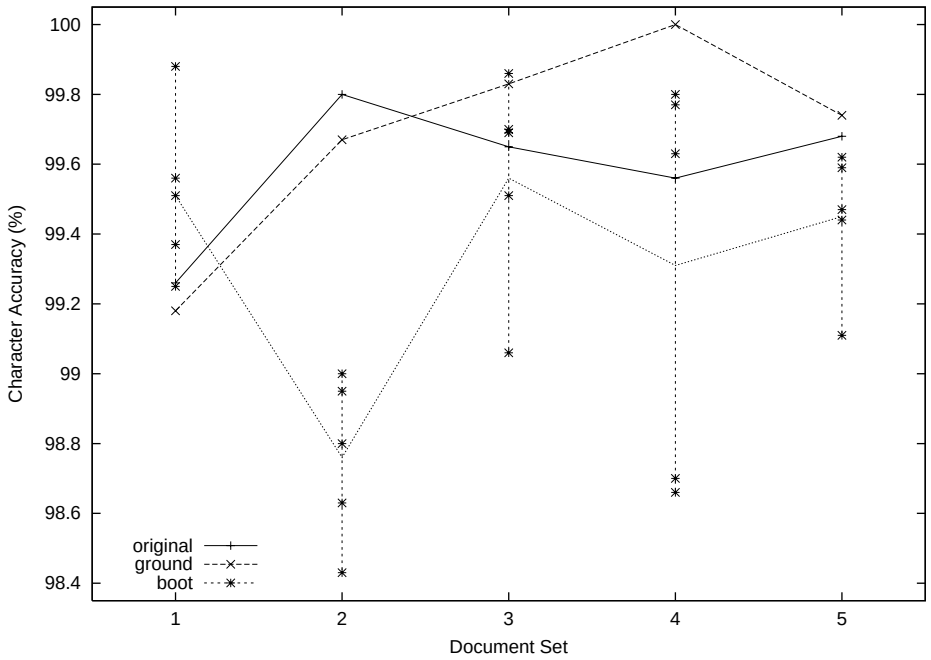


Figure 7. Character Accuracy

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

Model Validation

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

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

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

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

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

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

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

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

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

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

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

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

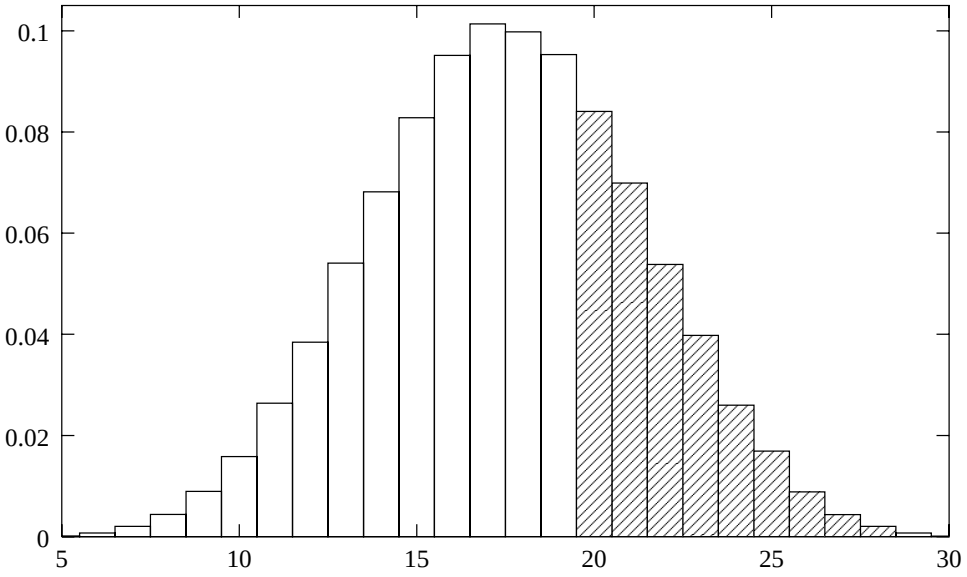


Figure 8. Reference distribution for T.

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

Summary

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

References

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

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

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

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

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

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

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

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

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

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

ANOTHER "NEW" APPROACH FOR "VALIDATING" SIMULATION MODELS

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

ABSTRACT

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

INTRODUCTION

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

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

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

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

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

OUR SETTING

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

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

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

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

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

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

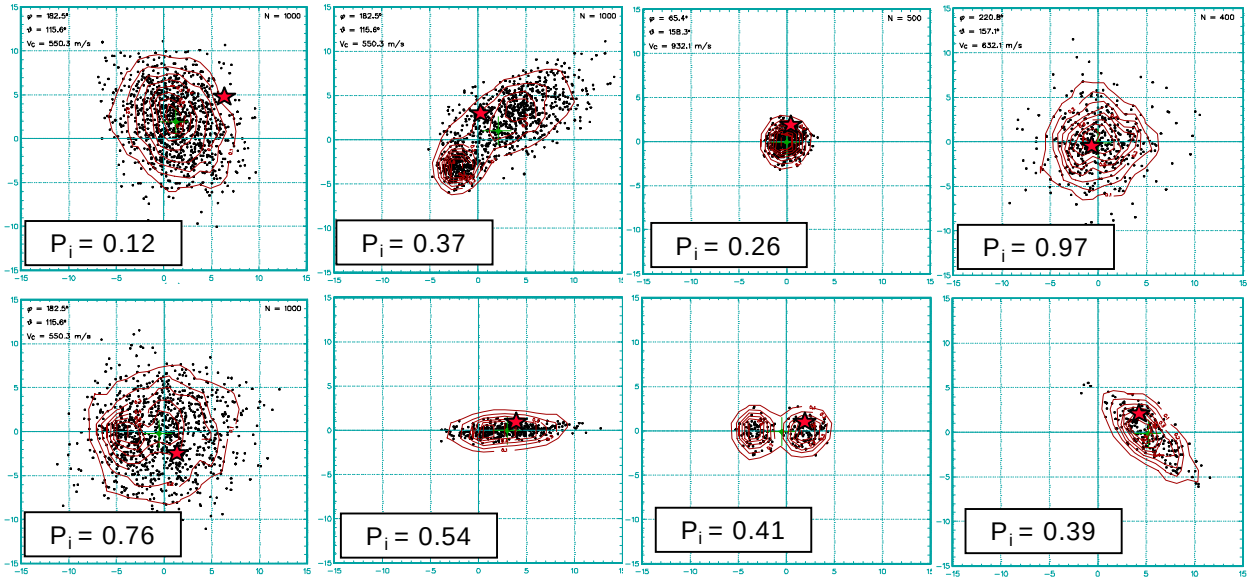


Figure 1. Ensemble Distribution of Tail Probabilities

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

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

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

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

Other shortcomings of t-test approaches are that they treat distinct combinations of factor settings as being statistically equivalent (e.g., all true differences between M&S means and observation means are assumed to be normally distributed with constant moments), and they make no use of any information related to the particular factor setting values themselves. Similar comments apply to nonparametric approaches (excluding, of course, the standard normal distribution assumption). Regression-based approaches could be pursued to attempt to incorporate

factor value information, but they still must contend with the other obstacles reported above. In addition, it may be difficult to find linear regression models or other simple regression formulations that accurately depict the influence of the factors on the M&S results.

FISHER'S COMBINED PROBABILITY TEST

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

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

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

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

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

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

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

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

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

One well-known property of Fisher's methodology is that a solitary small *p*-value can by itself lead to rejection of the null hypothesis — even when all N-1 of the remaining *p*-values are nominally large. Thus one "outlier" value

can dominate completely. This has motivated the construction of a number of alternative more tempered meta-analysis procedures for *p*-value amalgamation: Folks (1984), Rice (1990) and Olkin (1995). Nonetheless, we continue to endorse application of the traditional Fisher procedure as it forces us to confront directly the issue of whether a single test data outcome should "invalidate" the M&S under scrutiny. In addition, it should provoke the discussion of how and why "outlier" *p*-values resulted, whether and with what fidelity the underlying physical causes are represented within the M&S, whether other tested factor combinations logically could generate similar "outliers", etc.

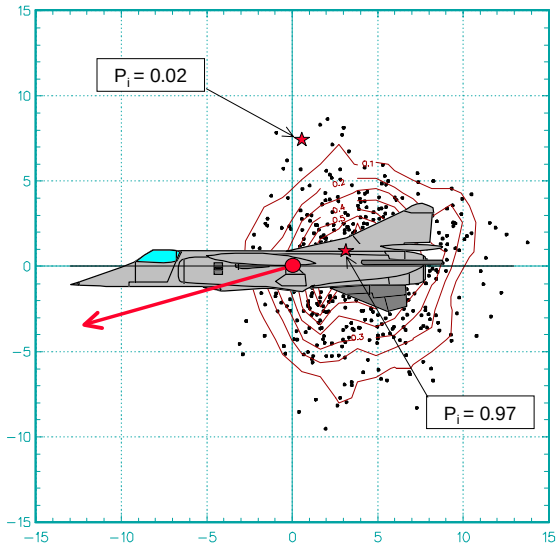


Figure 2. One-Sided "Tail" Probability

GOODNESS-OF-FIT PROCEDURES

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

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

STATISTICAL PROPERTIES

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

Here we report results from a simulation study aimed at characterizing the statistical properties, i.e., Type I errors and powers, of Fisher's combined probability test, the various GOF tests listed in Section 4, and the one-sample paired *t*-test described in Section 2. The simulation study focused on a univariate setting, with $N = 25$ test observations each associated with a known M&S distribution of predictions. For simplicity, and to facilitate comparisons of the Fisher test to the *t*-test when the latter is known to be optimally powerful, we took each M&S distribution to be a standardized n(0,1) normal distribution (with mean 0 and standard deviation 1). Note that we could begin with the more realistic situation in which these distributions initially are dissimilar (recall Figure 1). But, since in any simulation study we have the advantage of knowing what the true distributions are, we can always invoke appropriate transformations (e.g., inverse cdf functions) to force the reference distributions to the prescribed n(0,1) form. Within this construct, we represented various alternative hypotheses, i.e., "test data" distributions, by generalized normal distributions (with varying means and variances, and even raised to different arithmetic powers).

To obtain the analog of a single box in Figure 1, no random draws were necessitated to establish the reference M&S distribution and only one random draw was taken from the "test" distribution. In essence we assumed that the modelers could run the M&S infinitely many times and exactly duplicate the theoretical standardized normal distribution. This process was then repeated $N = 25$ times to generate a complete real-world situation, i.e., all of the paired M&S and "test" data available to support an analytical investigation of M&S validity. Each such "situation" was then replicated a large number of times (typically 100,000) to produce estimates of Type I error and power.

Type I error turns out to be not much of an issue. All of the statistical procedures recovered nominally prescribed significance levels, nearly exactly or, for some of the GOF tests, within acceptable errors due to small sample size effects. Such an outcome would by expected for the Fisher and GOF procedures regardless of how the M&S reference distributions are defined. For the *t*-test, we basically contrived this outcome by prescribing normal distributions.

First we consider two-sided comparisons between the Fisher test and the *t*-test, all for a nominal significancle level of $\alpha = 0.05$. (Similar results hold for other choices of α and for one-sided tests.) Let $S_i \sim$ n(0,1) and $T_i \sim [\text{n}(\mu_i,\sigma_i^2)]^p i$ respectively denote the M&S (H_o) and "test" (H_1) distributions, $i = 1, 2, \ldots, 25$. For those circumstances that the *t*-test was <u>not</u> designed to address, it does not do well. For instance, when the variance is allowed to be *any* value other than the null hypothesis specification of 1, its statistical power (i.e., probability of rejecting H_0) is miniscule — equal to $\alpha = 0.05$. This holds regrettably even for relatively large σ, e.g., set to 2 or 5. In each case, however, the Fisher test has a power essentially equal to 1.00, i.e., P(Fisher) = 1.00. Similar results hold when the "test" distribution distorts H_o via a simple power transformation. When $T_i \sim [\text{n}(0,1)]^3$, P(Fisher) = 0.87 >> P(*t*) = 0.03. But how does the Fisher procedure fare relative to the *t*-test when H_1 is just a shift in the mean away from H_o, i.e., for those situations for which the *t*-test is by theory optimal? (The *Z*-test actually is optimal since $\sigma = 1$ is prescribed, but since *N* is as large as 25 there is only an insignificant reduction in power for the *t*-test.) It turns out the Fisher procedure is extremely efficient within this class of alternative hypotheses. For the three examples $\mu_i =$ 0.1, 0.5, [(*i*-1)/24], the corresponding (P(Fisher),P(*t*)) values are (0.07,0.08), (0.63,0.67), and (0.60,0.63), respectively. Taken together, these results suggest the Fisher methodology should always be preferred to the "standard" *t*-test.

How do the GOF procedures compare to the Fisher approach? Results vary with the specific H_1 under examination and the particular GOF test. In some cases one can find a GOF test whose power is similar to that of the Fisher methodology. For instance, when $T_i \sim$ n(0.1,1) we have already observed that P(Fisher) = 0.07. In contrast, P(GOF) ranges from 0.04 to 0.05 for the different choices of the test procedure. If we increase the mean somewhat, say to $\mu_i = 0.8$, both sets of powers increase, but much more so for the Fisher approach: P(Fisher) = 0.94 and 0.18 < P(GOF) < 0.41. To illustrate the effect of changing variances, we considered $T_i \sim \text{n}(0,1.8^2)$ resulting in P(Fisher) = 0.94 and 0.63 < P(GOF) < 0.95. These results again suggest the superiority of the Fisher test.

But we should not forget the hypothetical example presented in Section 4, which clearly establishes that the Fisher test is not uniformly more powerful than GOF tests (across all possible H_1 designations). The fundamental characteristic of the H_1 considered in that example was a "compressed" distribution — with "test" observations tending to occur near the middle of the reference H_o distribution established by the M&S. To explore this notion further, we considered two H_1 distributions: $T_i \sim n(0,0.6^2)$ and $T_i \sim n(0.4,0.6^2)$. GOF procedures were markedly superior for the former:

$$P(\text{Fisher}) = 0 < P(t) = 0.05 << 0.55 < P(\text{GOF}) < 0.83.$$

For the latter, the *t*-test was actually best, although some of the GOF tests were relatively efficient:

$$P(\text{Fisher}) = 0.04 < 0.21 < P(\text{GOF}) < 0.38 < P(t) = 0.41.$$

SUMMARY & DISCUSSION

Taken in total, our simulation results suggest that some combination of Fisher and/or various GOF procedures generally outperforms the "standard" *t*-test, with relatively little penalty associated with those circumstances for which the *t*-test is slightly more powerful. For the most common alternative hypothesis, concerned with optimistic M&S predictions, the Fisher procedure is recommended.

From an M&S programmatic perspective, it is important to note that the Fisher and GOF tests accommodate formal statistical comparisons, even when data are sparsely distributed across a highly-dimensional factor space. Thus, the M&S VV&A process can be supported rigorously by statistical quantifications without demanding that a large sample of "test" events be dedicated to any individual point in the factor space.

The degree to which such an evaluation provides meaningful VV&A information depends greatly on the representativeness and extent of the factor combinations that yield "test" data. Ideally, the M&S offers the opportunity to study the sensitivity and importance of various factors, individually and in combination. Such insights should play a central role in judiciously allocating the locations within the factor space that are actually tested. But statistical efficiency should not be the sole concern. A practical emphasis should be placed on those regions of the factor space that are most likely to be encountered in tactical conditions, and on those specific circumstances that stress touted performance enhancements and new capabilities.

As always, the statistical methodologies discussed and endorsed in this paper should not be applied in any automatic fashion, and due attention must be given to the important distinction between "statistical significance" and "practical significance". Statistical conclusions of "inconsistencies" between M&S and "test" results should not in and of themselves lead immediately to a declaration of an "invalid" model. Considerable thought should be given to what "valid" and "invalid" mean within the intended sphere of M&S application. For instance, should a single "outlier" data observation by itself imply that the M&S is "invalid"? The focus should be on ways to better understand and improve the M&S, vice on formal pronouncements of "valid" or "invalid" M&S.

ACKNOWLEDGEMENTS & DISCLAIMERS

Portions of Arthur Fries' research were undertaken at the Institute for Defense Analyses (IDA) under IDA Central Research Project C9016 and under various tasks sponsored by the Office of the Director of Operational Test and Evaluation (ODOT&E) in the Office of the Secretary of Defense (OSD) within the U.S. Department of Defense (DoD). The author gratefully acknowledges appreciate helpful comments and inputs provided by A. Rex Rivolo, IDA, and the numerous supporting analyses conducted by LeAnna Guerin, IDA summer intern.

The views expressed in this paper are solely those of the author. No official endorsement by IDA, ODOT&E, OSD or DoD is intended or should be inferred.

REFERENCES

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991), "Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments," *Journal of the American Statistical Association*, 86, 953-963.U.S.

Army (1997), AR 5-11, *Management of Army Models and Simulations*, Washington, DC.

Fisher, R.A. (1932), *Statistical Methods for Research Workers*, Oliver and Boyd, Edingburgh.

Folks, J.L. (1984), "Combination of Independent Tests," in *Handbook of Statistics, 4, Nonparametric Methods*, P.R. Krishnaiah and P.K. Sen (eds.), North-Holland, New York.

Law, A.M. and Kelton, W.D. (1991), *Simulation Modeling and Analysis*, Mc-Graw Hill, New York.

Morris, M.D. (1991), "Factorial Sampling Plans for Preliminary Computational Experiments," *Technometrics*, 33, 161-174.

Olkin, I. (1995), "Meta-Analysis: Reconciling the Results of Independent Studies," *Statistics in Medicine*, 14, 457-472.

Rice, W.R. (1990), "A Consensus Combined *P*-Value Test and the Family-wide Significance of Component Tests, *Biometrics*, 46, 303-308.

Sacks, J., Schiller, S.B., and Welch, W.J. (1989a), "Designs for Computer Experiments," *Technometrics*, 31, 41-47.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. (1989b), "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409-435.

Saltelli, A., Andres, T.H., and Homma, T. (1993), "Sensitivity Analysis of Model Output: An Investigation of New Techniques," *Computational Statistics and Data Analysis*, 15, 211-238.

Saltelli, A., Tarantola, S., and Chan, K.P.-S. (1999), "A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output," *Technometrics*, 41, 39-56.

GRAPHICAL ANALYSIS OF COMMUNICATIONS LATENCY IN A LARGE DISTRIBUTED SIMULATION

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

ABSTRACT

The Theater Missile Defense System Exerciser (TMDSE) is a large geographically-distributed simulation developed by the Ballistic Missile Defense Organization (BMDO). TMDSE is being used to investigate Joint Data Network (JDN) interoperability between the members of the Theater Missile Defense (TMD) Family of Systems (FoS) as they develop. One area of interest is the time delays (latencies) inherent to the simulation and how well they represent the latencies expected in tactically deployed systems. This paper shows how simple statistical graphics implemented on a modern laser printer can produce comprehensive, easily understood characterizations and analyses of such communications latencies for 100,000 or more data points.

INTRODUCTION

The Theater Missile Defense (TMD) Family of Systems (FoS) is comprised of many systems at varying levels. The complete FoS has systems at the early warning sensor level, the weapon and sensor level, and the command and control level, see Figure 1. For this full FoS to operate as efficiently as possible, the various systems within each level and across levels need to be interoperable. To achieve this interoperability, each system has to implement the appropriate information exchange protocols, and within the specific protocol, the appropriate messages.

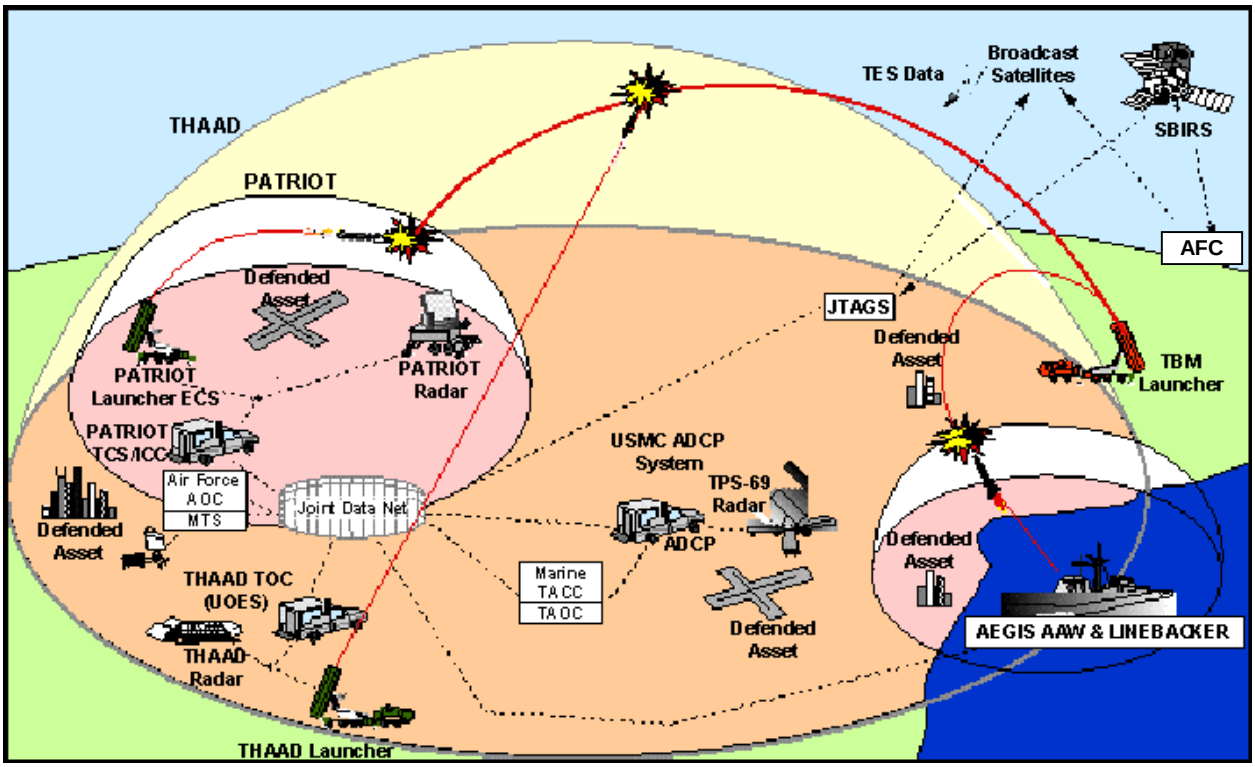


Figure 1. Theater Missile Defense Family of Systems Architecture.

The Theater Missile Defense System Exerciser (TMDSE) is a large geographically-distributed hardware-in-the-loop (HWIL) simulation developed by the Ballistic Missile Defense Organization (BMDO). The tactical segments participating in a recent TMDSE HWIL test were as follows:

- Early warning sensor level
  - U.S. Army Joint Tactical Ground Station (JTAGS)
  - U.S. Air Force Aerospace Fusion Center (AFC)

- Weapon and sensor level
  - U.S. Navy AEGIS LINEBACKER
  - U.S. Army Theater High Altitude Area Defense (THAAD)
  - U.S. Army Phased Array Tracking to Intercept of Target (PATRIOT), both Post-Deployment Build (PDB) -4 and PDB-5
- Command and control level
  - U.S. Air Force Theater Air Command and Control Simulation Facility (TACCSF), Control and Reporting Center (CRC) (Missile Tracking System [MTS] only)
  - U.S. Marine Corps (USMC) Tactical Air Operations Center (TAOC), Air Defense Communication Platform (ADCP) and TPS-59 only
  - U.S. Navy AEGIS Anti-Air Warfare (AAW)

Geographical locations of these tactical segments are shown in Figure 2. TMDSE stimulates current or future versions of tactical segments by providing a simulated threat environment and communications connectivity.

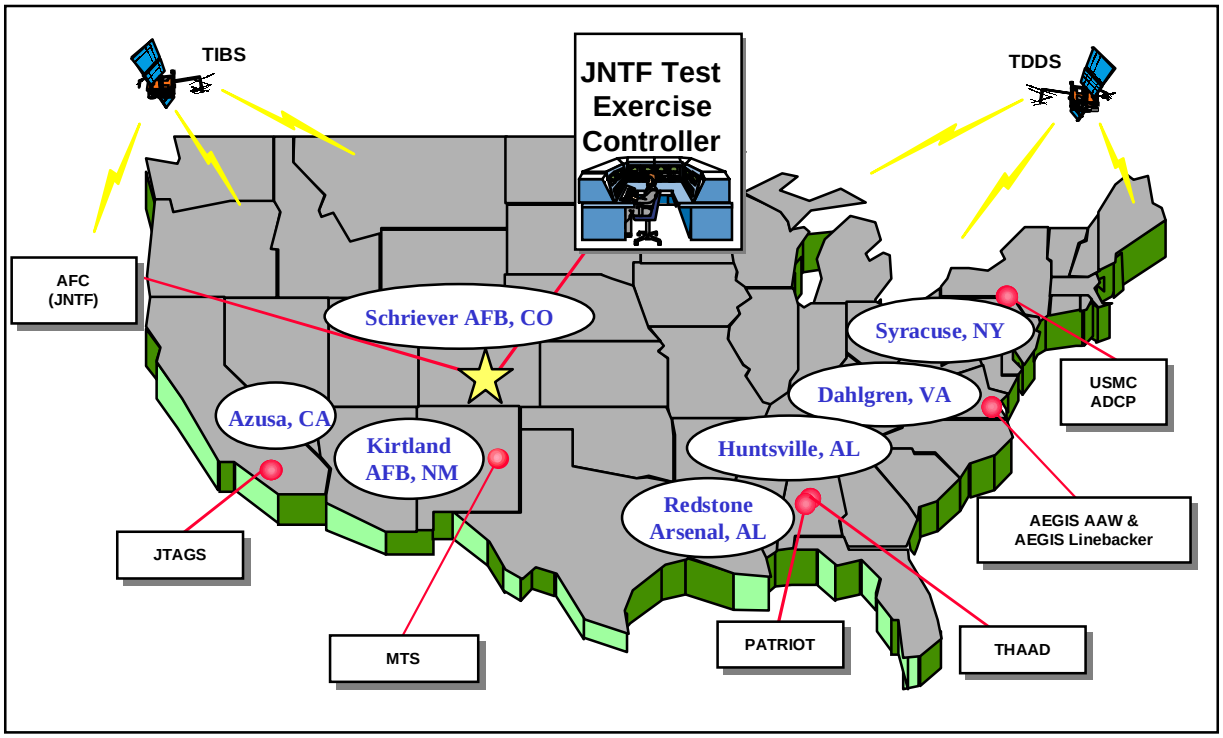


Figure 2. Geographical Location of TMDSE Tactical Segments.

The Joint Data Network (JDN) is a subset of the total FoS communications structure, and provides the supporting communications architecture. Figure 3 depicts the TMDSE test architecture used. TMDSE is comprised of the TMDSE Control Segment (TCS), the Tactical Communications Environment Segment (TCES), and the Remote Environments (REs). The TCS is comprised of the TEC located at the JNTF and REs located with the Tactical Driver Segments (TDSs). The TCS controls and stimulates the environment within which the TMD FoS operates and provides overall control and coordination of the TMDSE system for testing. The TCS provides the physical interface and functional capabilities to interface with each of the Tactical Drivers (TDs) and the TCES Link-16 Gateways, which emulate the JDN communications networks.

TMDSE and the participating tactical nodes exchange environment information (ground truth information such as time space position information (TSPI) on Theater Ballistic Missiles (TBMs) and TMD interceptors) using Distributed Interactive Simulation (DIS) Protocol Data Units (PDUs). The tactical nodes exchange JDN messages via TCES. The TCES emulates a Joint Tactical Information Distribution System (JTIDS) network for Tactical Digital Information Link (TADIL)–J messages over dedicated T-1 landlines and provides a backup capability for the Tactical Information Broadcast Service (TIBS) and TRAP (Tactical Receiver and Related Applications) Data Dissemination System (TDDS) messages. The primary route for TIBS and TDDS messages is through a live

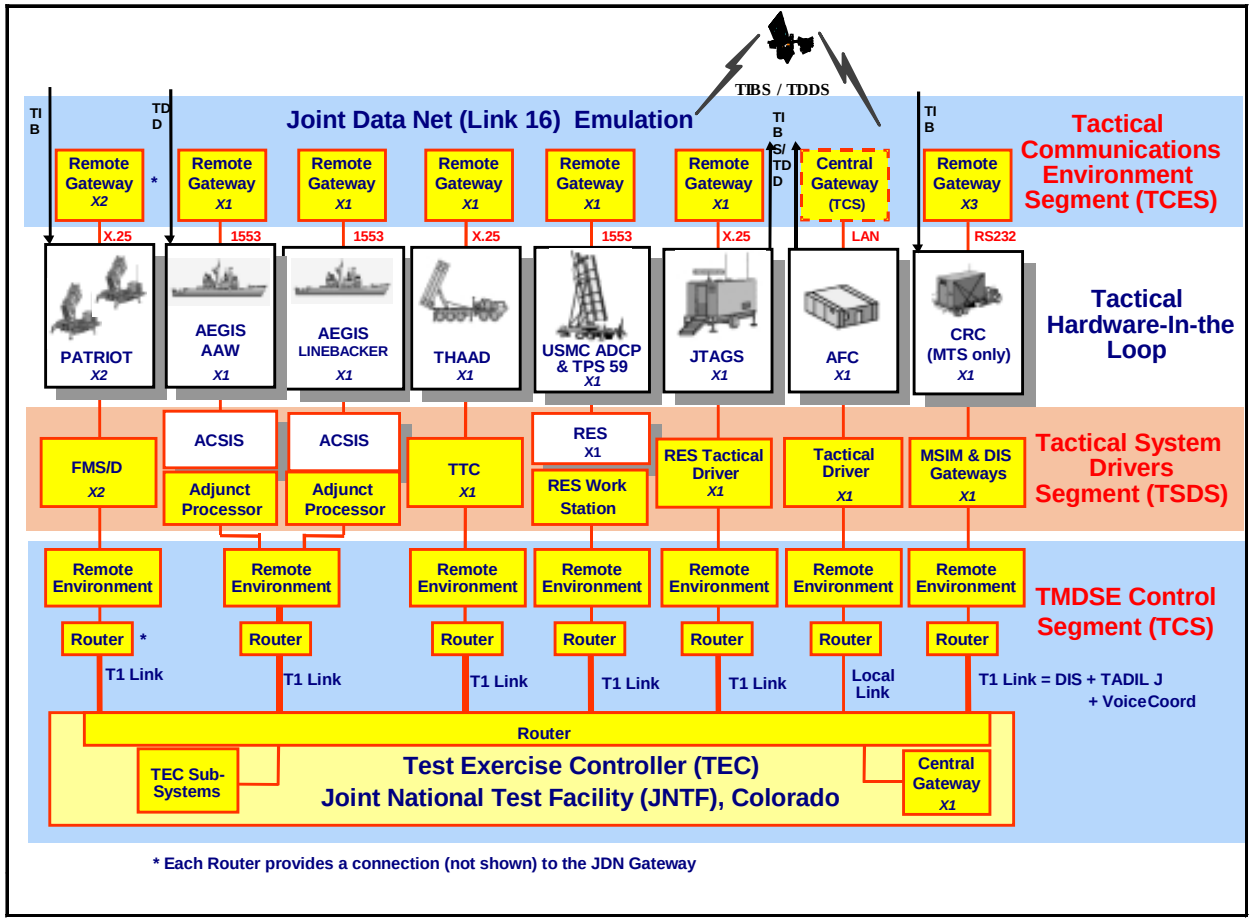


Figure 3. TMDSE Test Architecture.

Satellite Communications (SATCOM) feed; the secondary route is via TCES over a dedicated T-1 landline. Positioning of the TMDSE segments (simulated locations on the virtual battlefield) and launches of threat TBMs are scripted according to an approved scenario, but responses to those threat launches is via the distributed segment simulators and TADIL-J messages over the surrogate JDN.

COMMUNICATIONS LATENCY IN TMDSE

Among the issues addressed during testing was communications latency within TMDSE. Messages were recorded and time-stamped at the transmitting node and its corresponding gateway as well as at receiving gateways and nodes. After matching each sent message with its receipts, latencies could be calculated throughout the system for each receipt. Of course some messages could not be matched end-to-end, and some messages were probably mismatched. Table 1 gives for five key types of TADIL-J message receipts the number of messages for which latencies could be calculated (by location of receipt for each test run of interest). Possible dependence of latencies on transmitting and receiving segments are also of interest, but since not all message types were recorded at all segments, the effects of transmitting and receiving segments are nested within message type. Considering all 1143 factor combinations (including transmitting and receiving node), the range of sample sizes per factor combination is extreme, as shown by the summary in Figure 4. The overall sample sizes are so large that using simple linear modeling to understand trends breaks down because everything is statistically significant. This is shown by the ANOVA in Table 2 for a simple linear model of latencies from transmitting node to receiving node. Moreover, detailed tables of latency summaries are unwieldy, stretching to several hundred pages. Fortunately, it is not difficult to display the latency data graphically in a way that makes meaningful analysis clear.

GRAPHICAL APPROACH TO OVERALL TRANSMITTING-NODE-TO-RECEIVING-NODE LATENCIES

Boxplots provide an effective way to display latencies by transmitting segment, message type, and receiving segment as in Figure 5 (80 percent boxplots are used: whiskers stretch from the 10^{th} to the 25^{th} percentile and 75^{th}

Table 1. Key TADIL-J Messages for Latency Calculation, by Location, Message Type and Test Run.

Location of Received Message	Message Type	Number of Messages for Latency Calculation		
		Test Run I	Test Run II	Test Run III
Transmitting Gateway	**a**	2126	2033	1774
	b	1072	973	976
	c	478	528	507
	d	124	95	166
	e	1210	1324	1831
Receiving Gateway	**a**	17668	16640	15884
	b	8584	7664	8014
	c	3802	4220	4174
	d	981	745	1300
	e	9619	10524	14562
Receiving Node	**a**	12755	12229	12352
	b	6474	5817	6341
	c	3802	3062	3181
	d	981	529	974
	e	5964	6141	8803

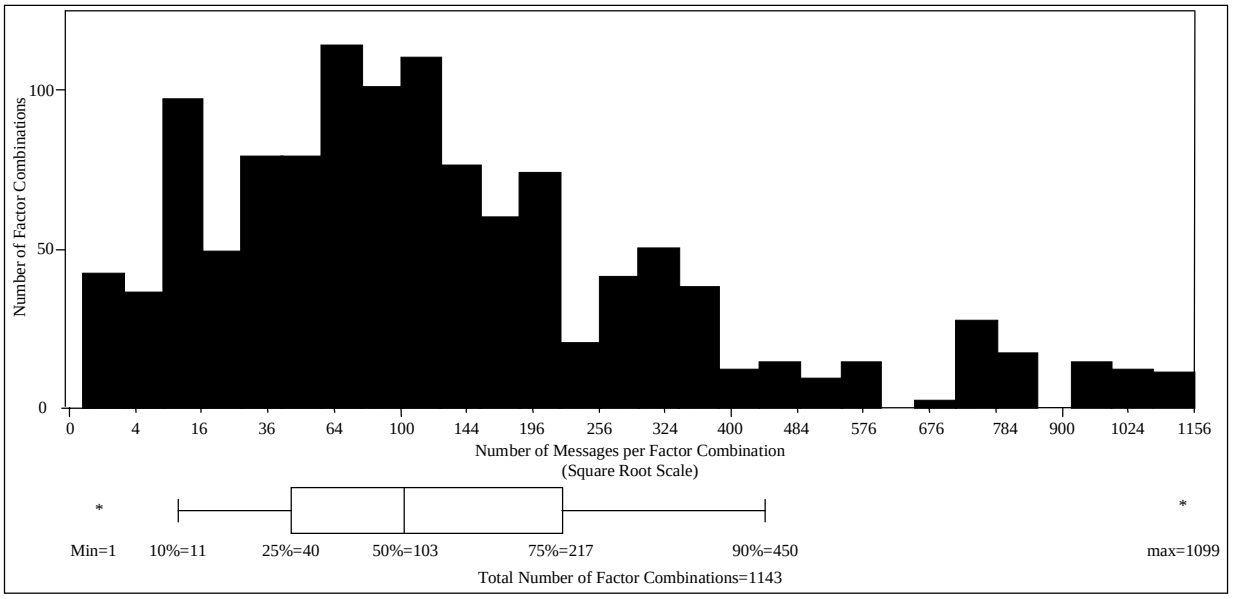


Figure 4. Distribution of Number of Messages per Factor Combination.

Table 2. Analysis of Variance for a Simple Model (Everything Is Statistically Significant)

Source	DF	F Ratio	Prob>F
Test Run	2	529.2	<0.0001
Message Type	4	63.8	<0.0001
Test Run x MsgType	8	8.5	<0.0001
Transmitter	7	1897.4	0.0000
Receiver	6	19.8	<0.0001
Test Run x Transmitter	14	196.9	0.0000
Test Run x Receiver	12	5.2	<0.0001

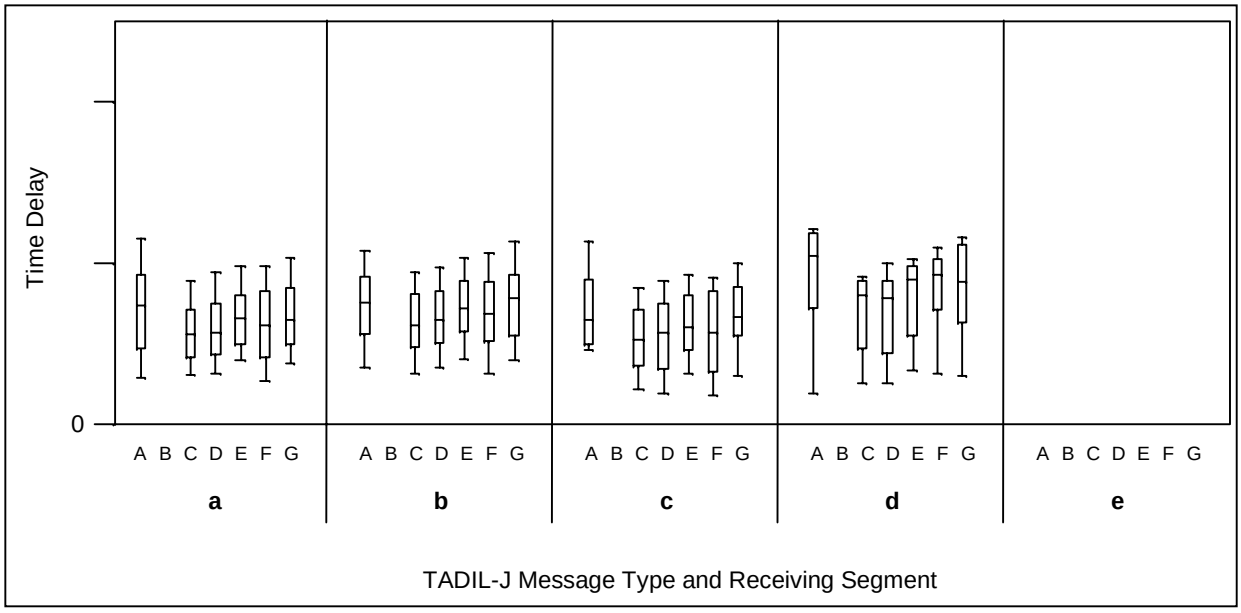


Figure 5. Time Delays for Key Messages Transmitted by Segment "B" on Test Run I, by Message Type and Receiving Segment.

percentile to 90th percentile of the latency distribution for each receiving segment; the 50th percentile is marked with a horizontal line). Missing data for segment "B" with all message types is due to the fact the segment "B" is transmitting the messages and therefore has no latency distribution. Missing data for all receiving segments with message type "e" is due to the fact that segment "B" does not transmit message type "e." Shrinking Figure 5, repeating it for each transmitting and receiving segment on each trial, and arranging the resulting displays in a tabular fashion gives Figure 6. Figure 6 is even more effective when printed on tabloid-sized paper, but it is still usable on letter-sized paper. Patterns of messages not present at certain segments are clear as is the overall stability of latencies. Close examination of Figure 6 reveals many apparent small trends but few appear to be substantial. Most of these small trends, including the suspicious long whiskers, are due to the small sample sizes identified in Figure 4, and these can be easily tabulated. Three larger trends stand out but will not be discussed in detail here.

GRAPHICAL DECOMPOSITION OF TRANSMITTING-NODE-TO-RECEIVING-NODE LATENCIES

For each Test Run in Figure 6 (first three rows) the transmitting-node-to-receiving-node latencies can be decomposed into delays from transmitting node to transmitting gateway, time delays from transmitting gateway to receiving gateway (this is where the geographical distribution is, but T1 lines are being used), and time delays from receiving gateway to receiving node. This is done in Figures 7-9 which show that almost all the transmitting node to receiving node latencies is due to time delays from the transmitting nodes to the transmitting node gateways and that time delays from transmitting gateways to receiving gateways are negligible. Counting the summary transmitting-node-to-receiving-node rows in Figures 7-9 separately, Figures 7-9 concisely summarize more than 300,000 data points in a much more useful fashion than any table or analysis of variance could possibly do.

GRAPHICAL ADJUSTMENT FOR APPARENT NODE CLOCK ERRORS

Data in Figures 5-9 have actually been adjusted for apparent errors in clock synchronization at nodes "E"-"H" on some test runs. Figure 10, analogous to Figure 6 but without any adjustment, shows obvious systematic anomalies in the unadjusted data. There are clearly no serious synchronization problems between gateways since unadjusted time delays between gateways in Figures 7-9 are so small. The unadjusted gateway-to-node time delays in Figure 11 show consistent differences between time delays at "A"-"D" and those at "E"-"G." These permit approximate graphical estimates of what clock adjustments should be. The unadjusted node-to-gateway time delays in Figure 12 provide a check on the Figure 11 estimates and give additional estimates for segment "H" adjustment. (Several computational methods to determine appropriate adjustments were also attempted, but the graphical method worked best.)

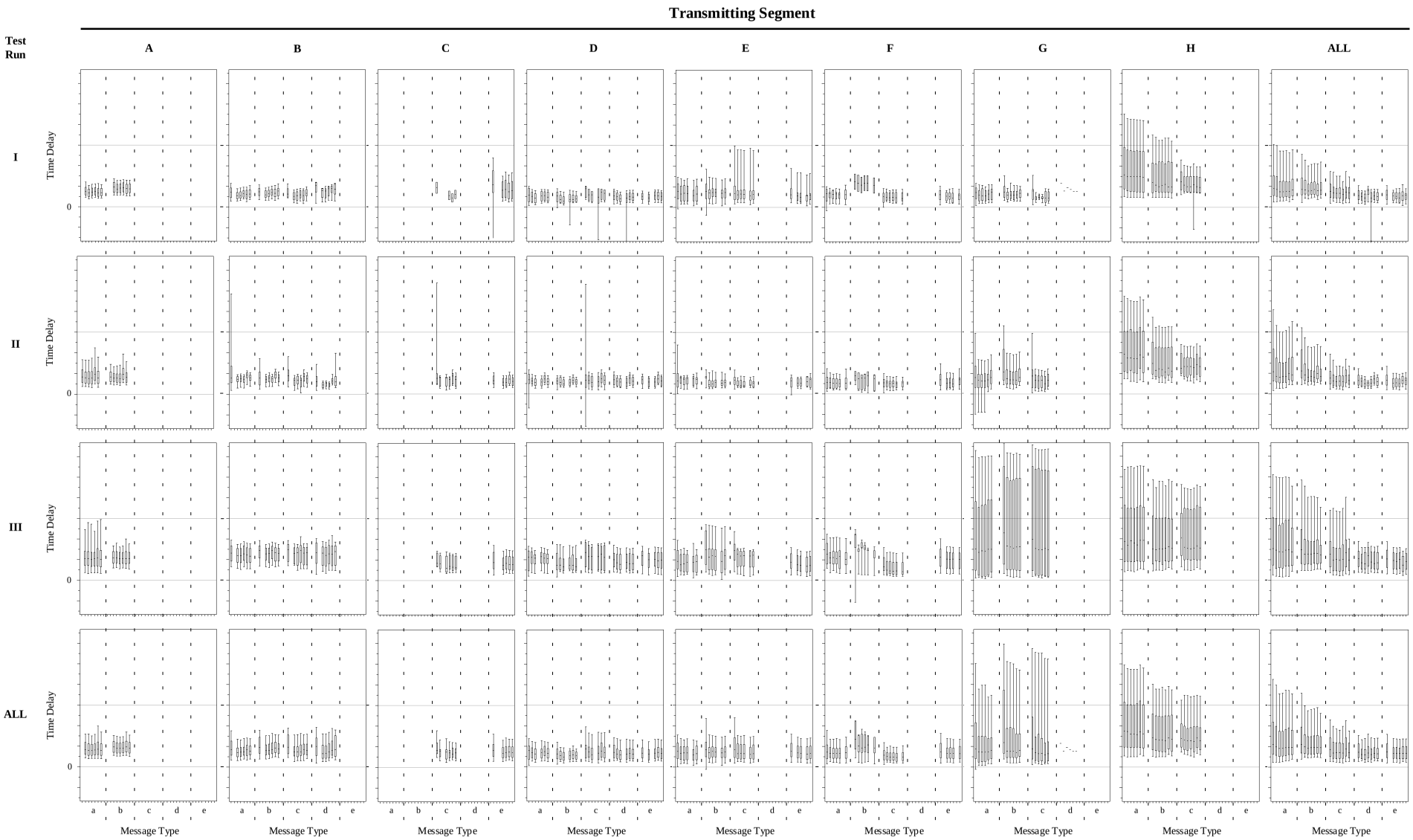
SUMMARY

This paper has shown how simple statistical graphics—miniaturized and implemented on a modern printer (1200 dots per inch with tabloid paper capability)—can produce rich, easily understood analyses of large data sets. Such graphical techniques have not yet been much used in DoD analyses, but similar techniques have been frequently exploited by in other fields and as the following brief discussion indicates.

Tukey's "Orange Book"[1] popularized the boxplot and ignited a continuing emphasis on the graphical side of statistics, and his 1993[2] paper provided new insights on boxplots and other aspects of statistical graphics. Tufte's three extraordinary books[3,4,5] broadened the audience for clear quantitative display. Among many other insights, Tufte tauted the ability of the human eye to make "a remarkable number of distinctions within a small area" and promoted high "data density" ("number of entries in data matrix" ÷ "area of data graphic") and the use of small multiples ("graphics can be shrunk way down") as in this paper. Three bullets of three words each per PowerPoint slide is a more standard goal for DoD briefings, and high data densities typically face stiff resistance. This is true even though the densest common quantitative graphic—a map—is readily understood. More efforts like the one in this paper may gain wider acceptance of dense graphics in DoD. Display of graphics panels in tabular form has been common for some time in the form of scatterplot matrices (now available in most statistics packages) and more recently in trellis displays,[6] but the tabular displays used here have more in common with recent graphical data mining techniques[7,8,9,10,11] than with trellis displays.

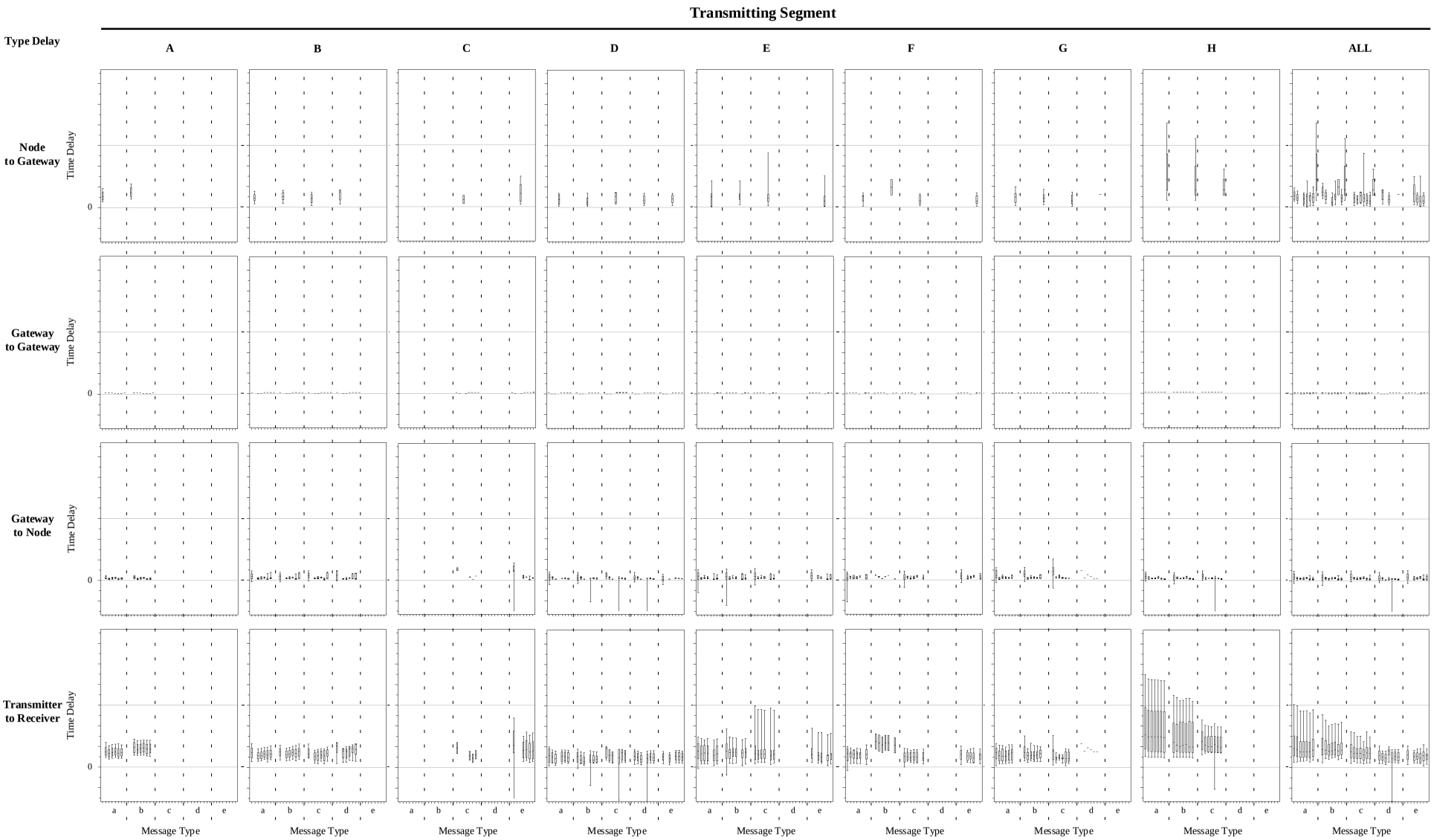
REFERENCES

1. Tukey, J.W. Exploratory Data Analysis. Reading: 1977.
2. Tukey, J.W. "Graphical Comparisons of Several Linked Aspects: Alternatives and Suggested Principles." Journal of Computational and Graphical Statistics, vol. 2 nbr. 1, pp. 1-49, 1993.
3. Tufte, E.R. The Visual Display of Quantitative Information. Cheshire, Connecticut: 1983, pp161-175.
4. Tufte, E.R. Envisioning Information. Cheshire, Connecticut: 1990, pp. 67-79
5. Tufte, E.R. Visual Explanations. Cheshire, Connecticut: 1997.
6. Becker, R.A., Cleveland, W.S. and Shyu, M-J. "The Visual Design and Control of Trellis Display." Journal of Computational and Graphical Statistics, vol. 5 nbr. 12 pp. 123-155, 1996.
7. Church, K.W. and Helfman, J.I. "Dotplot: A Program for Exploring Self-Similarity in Millions of Lines of Text and Code." Journal of Computational and Graphical Statistics, vol. 2 nbr. 2, pp. 153-174, 1993.
8. Eick, S.G. "Graphically Displaying Text." Journal of Computational and Graphical Statistics, vol. 3 nbr. 2, pp. 127-142, 1994.
9. Becker, R.A., Clark, L.A. and Lambert, D. "Cave Plots: A Graphical Technique for Comparing Time Series." Journal of Computational and Graphical Statistics, vol. 3 nbr. 3, pp. 277-283, 1994.
10. Keim, D.A. "Pixel-Oriented Visualization Techniques for Exploring Very Large Data Bases." Journal of Computational and Graphical Statistics, vol. 5 nbr. 1, pp. 58-77, 1996.
11. Fayyad, U.M. and Smyth P. "Cataloging and Mining Massive Datasets for Science Data Analysis." Journal of Computational and Graphical Statistics, vol. 8 nbr. 3, pp. 589-610, 1999.



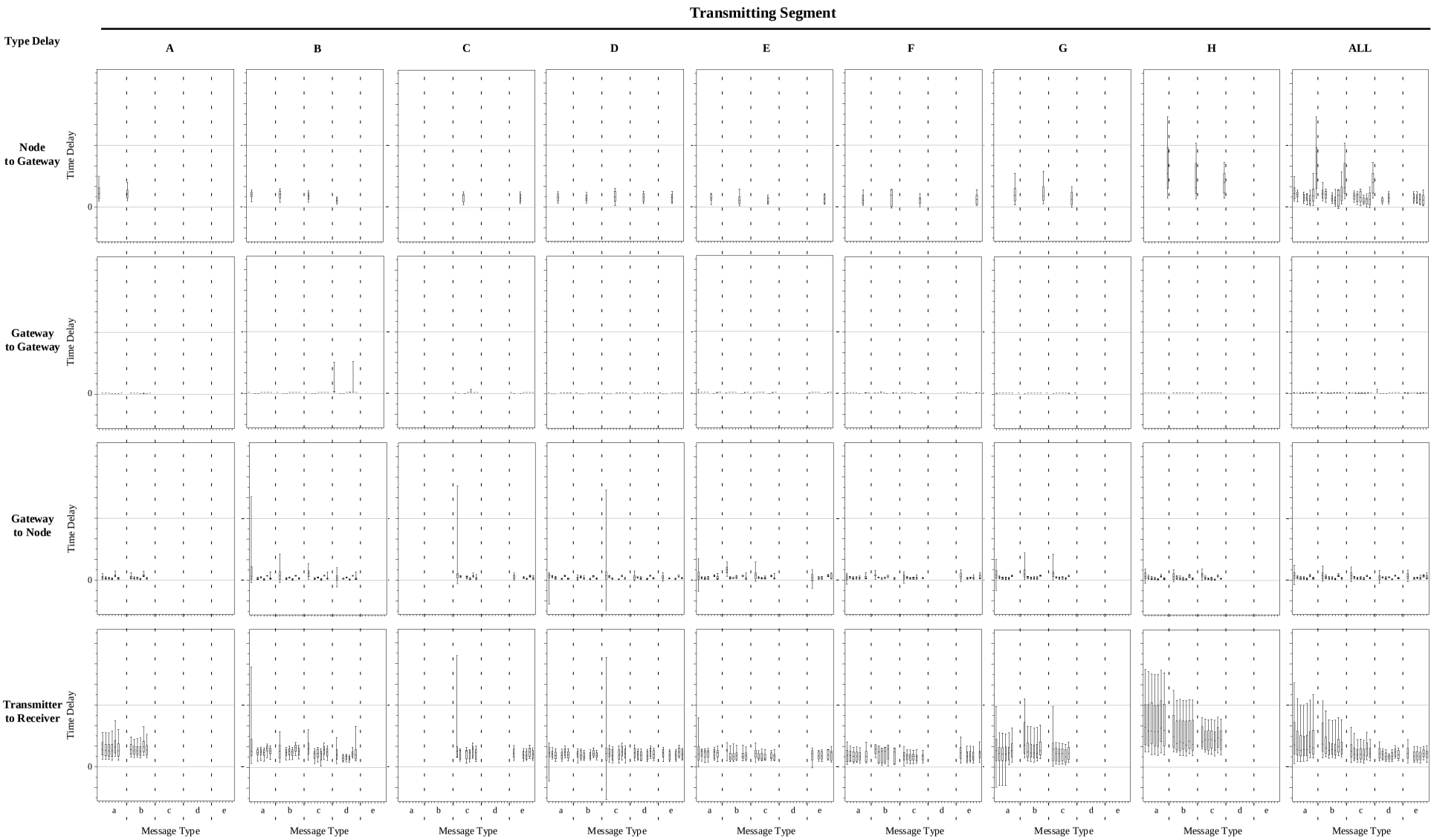
Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 6. Transmit to Receive Time Delays by Test Run, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Adjusted for Apparent Node Clock Errors.



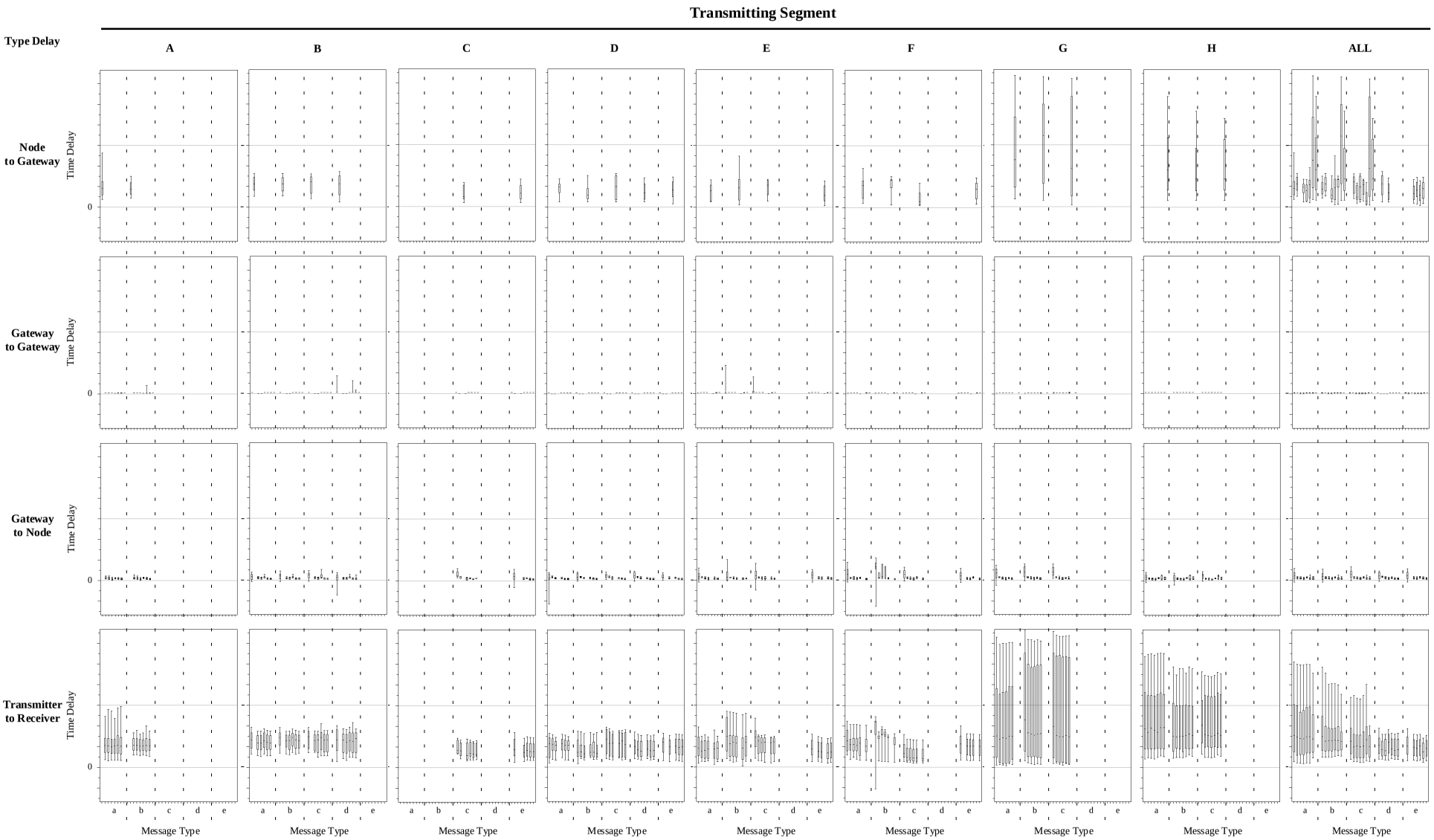
Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 7. Time Delays on Run I by Type Delay, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Adjusted for Apparent Node Clock Errors.



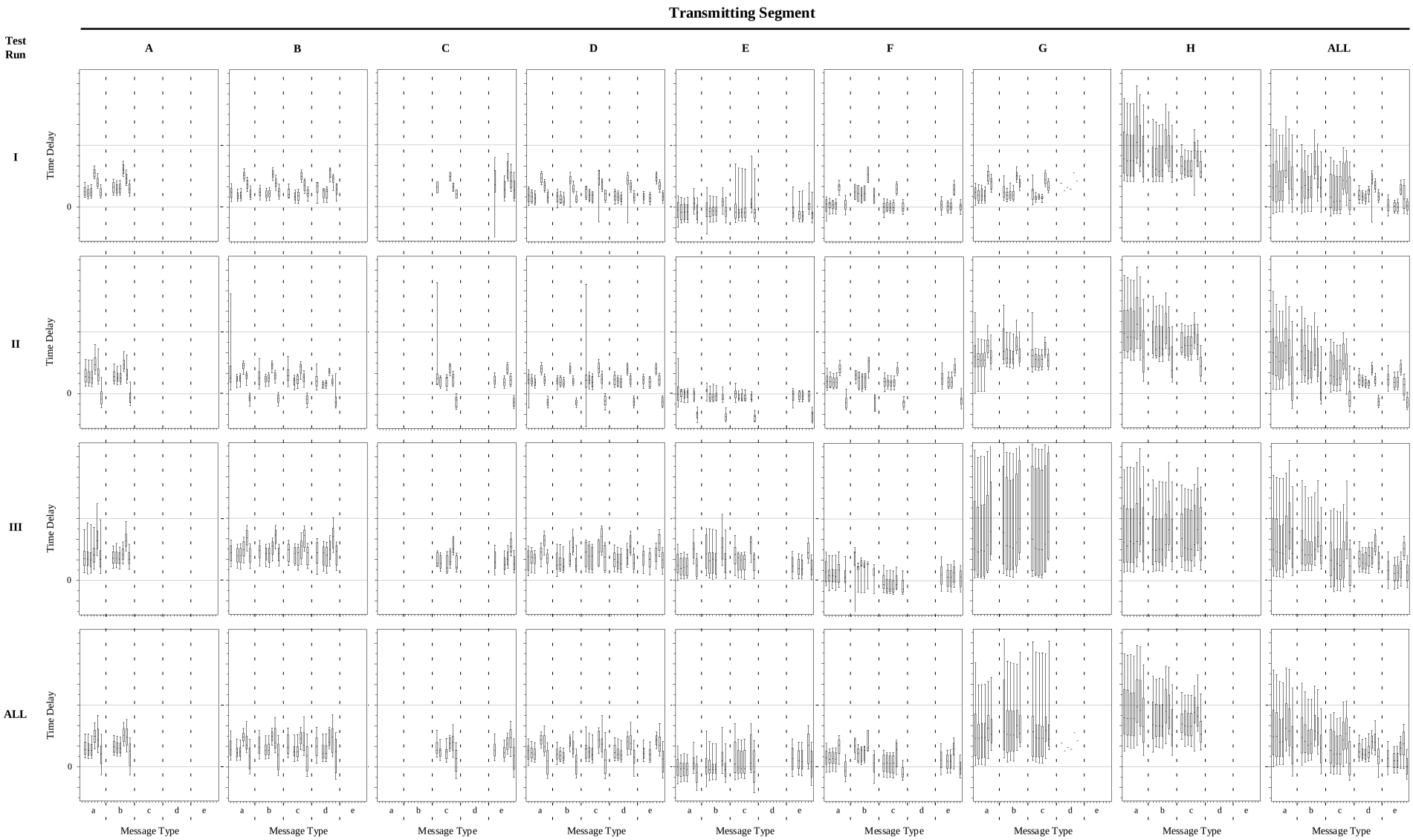
Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 8. Time Delays on Run II by Type Delay, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Adjusted for Apparent Node Clock Errors.



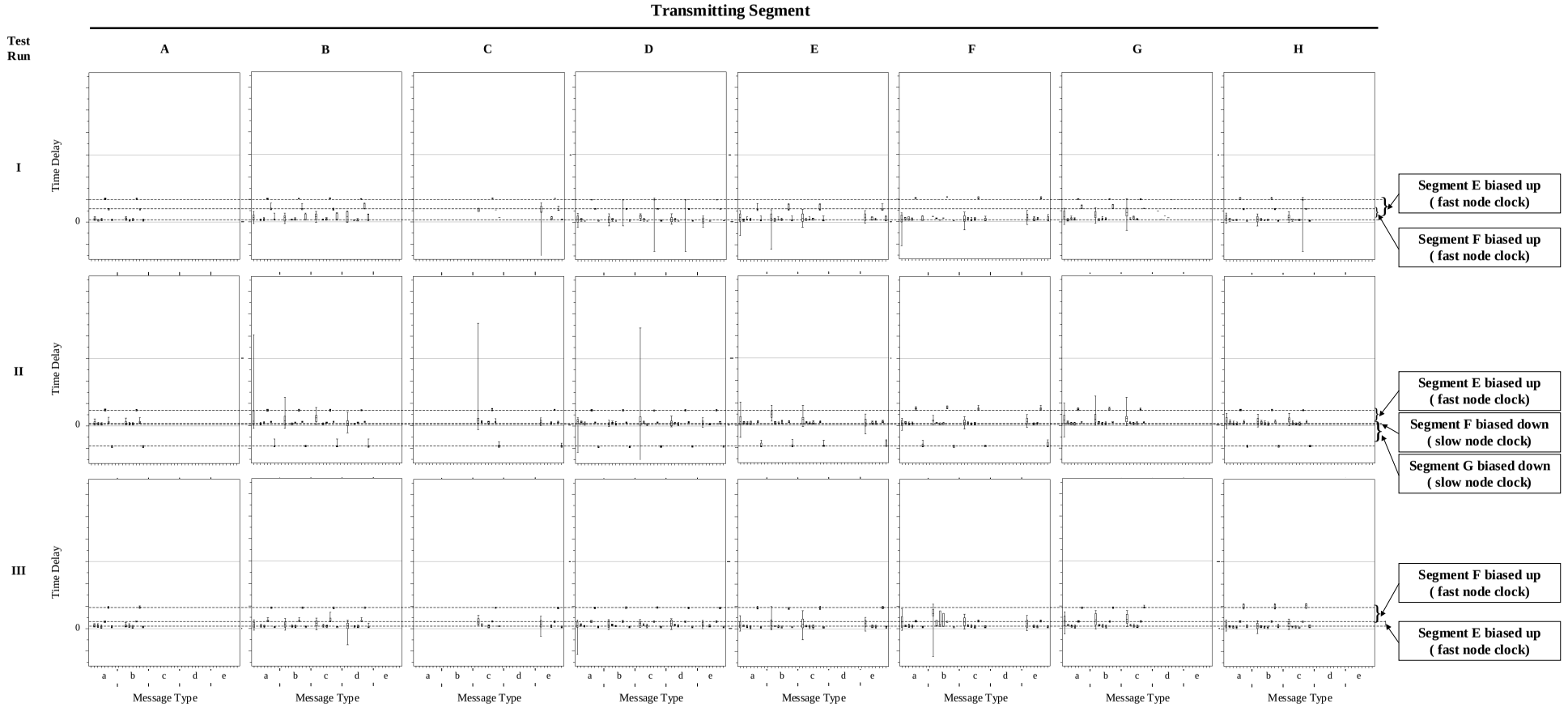
Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 9. Time Delays on Run III by Type Delay, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Adjusted for Apparent Node Clock Errors.



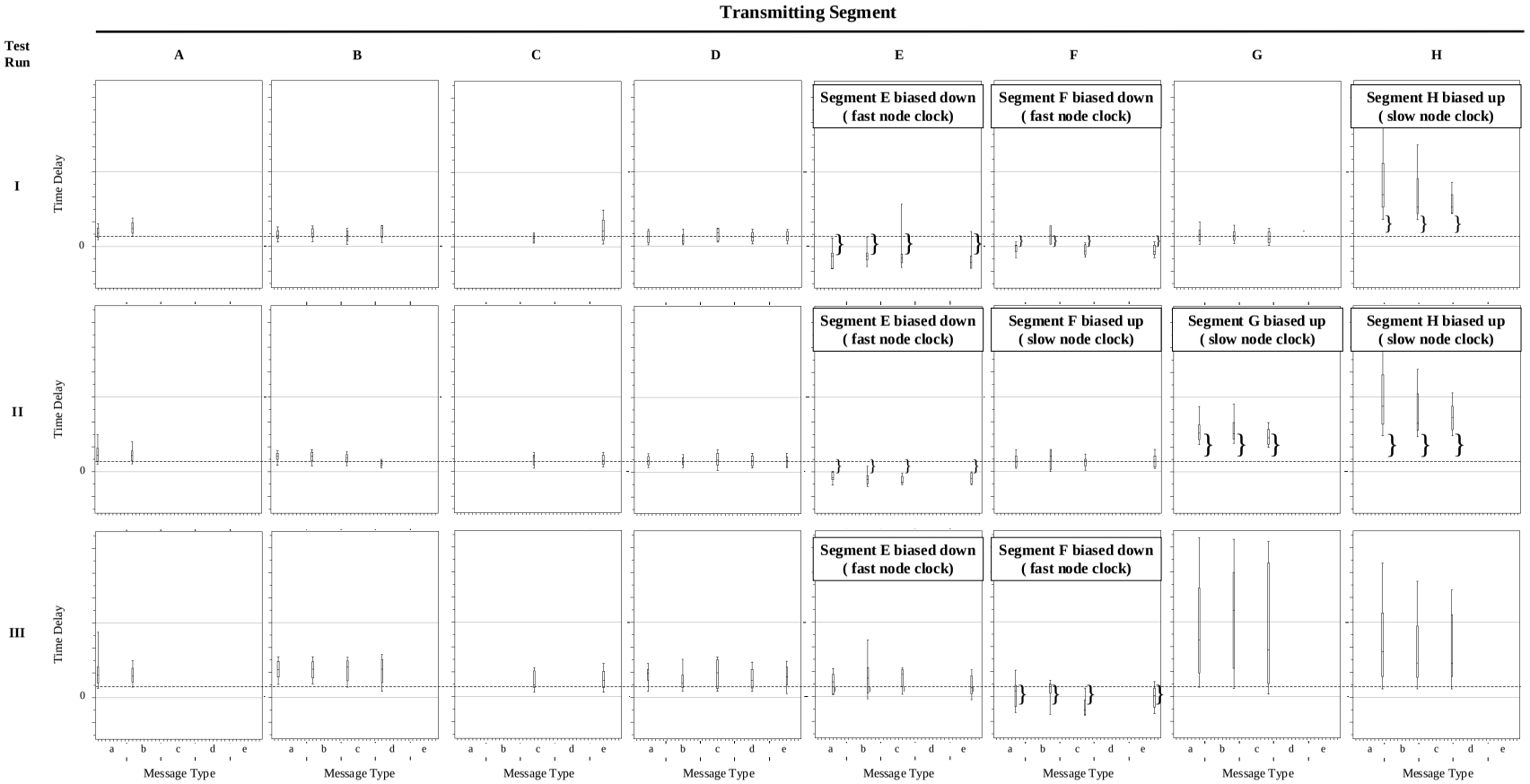
Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 10. Transmit to Receive Time Delays by Test Run, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Not Adjusted for Apparent Node Clock Errors.



Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 11. Gateway to Node Time Delays by Test Run, Type, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Not Adjusted for Apparent Node Clock Errors.



Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.

Figure 12. Node to Gateway Time Delays by Test Run, Type, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Not Adjusted for Apparent Node Clock Errors.

WEDNESDAY, OCTOBER 18

WILKS BANQUET ADDRESS (2000 - 2045)

Lessons From the History of Wargaming
Matthew Caffrey (Air Command and Staff College)

It's been said, experience is learning from your mistakes, wisdom is learning from the mistakes of others. For almost 200 years modern wargaming has been providing decisive insights and tragic mirages. Do we have the wisdom to use wargaming more effectively by learning from our predecessor's experiences? Professor Caffrey, Air Command and Staff College, Air University, looks back at the history of wargaming and suggests lessons for the future.

THURSDAY, OCTOBER 19

GENERAL SESSION II (0830 - 1000)

Challenges for Categorical Data Analysis in the 21st Century
Alan Agresti, University of Florida

As we begin the 21st century, the state-of-the-art in categorical data analysis, as in all branches of statistics, is vastly different than at the start of the 20th century. This article focuses on some of the primary developments of the past ten to twenty years and discusses areas likely to see significant progress in coming years. Topics on which I focus are methods for ordered categorical responses, repeated measurement and clustered data, and small samples and sparse data. As the literature evolves, the variety of options for data analysis continues to increase dramatically. As a consequence, perhaps the greatest challenge now is maintaining adequate communication among statisticians and between statisticians and other scientists.

A Spatial-Temporal Statistical Approach to Problems In Command and Control

Noel Cressie, David A. Wendt, and Gardar Johannesson
Department of Statistics
The Ohio State University, Columbus, OH
Andrew S. Mugglin
Medtronic, Inc., Minneapolis, MN
and
Birgir Hrafnkelsson
Marine Research Institute, Iceland

Abstract

The integration, visualization, and overall management of battle-space information for the purpose of Command and Control (C2) is a challenging problem. For example, how can one assimilate incoming data rapidly in a highly dynamic environment, to allow the battle commander to make timely and informed decisions? In this paper, we present a spatial-temporal statistical approach to estimating the battlefield, based on noisy data from multiple sources. Specifically, we examine the danger-potential field generated by an enemy's weapons in the spatial domain and extend it to incorporate the temporal dimension. We propose that maps of fields of this sort are very effective decision tools for the battle commander; methods for rapid updating of the maps is an area of current research. This includes visualization of the predictions and the uncertainty associated with them. It is the quantification of uncertainty in C2 predictions that distinguishes our statistical approach from deterministic approaches. In this paper we describe an object-oriented combat-simulation program from which we generate noisy battlefield data; in the absence of real data, we apply our methodology to these simulated data.

1. A Brief Introduction to C2

Command and Control (C2) and related variants (C3, C4I, etc.) are umbrella terms used to describe a body of research and applications dedicated to preparing all branches of the armed forces for the "War After Next" [2]. Applications incorporated into the greater C2 framework include, but are not limited to, command applications, operations applications, intelligence applications, fire-support applications, logistics applications, and communications applications. C2 needs are shared by all branches of the armed services. Consequently, appli-

cations should be sufficiently flexible to work across services and across allied forces, as needed. In order to be applicable to future wars, applications need to be able to convert a flood of data from a wide variety of sources into information and knowledge in a timely manner. These concepts are now taught in a focused way by the U.S. Military, during officer training [3].

In preparing for the Command Post of the Future, tools should be developed with the following three goals in mind: to rapidly visualize the battlespace, to rapidly analyze the battlespace, and to rapidly understand the battlespace. When visualizing the battlespace, the prototypical commander wants a variety of information, and urgent information needs to be highlighted so that time-critical decisions can be made. In addition, the location and status of friendly and enemy forces need to be available to the commander. In order to facilitate the rapid visualization and understanding of the battlespace, to provide the ability to receive and send information while mobile, and to begin converting that information into knowledge, is critical to the battle commander. Facility for intelligent alerting and reporting, as well as customizable tactical display elements, need to be integrated into any C2 application.

Systems developed for C2 applications need to keep several general capabilities in mind. The military currently suffers not from a lack of data but from a flood of data. Data arrive from multiple sources and possibly multiple nationalities. Data arrive in visual, verbal, and other sensor formats, and also from historical data bases. At the same time, there is a distinct deficiency of information and knowledge. C2 applications should aim at developing decision aids that can turn massive amounts of data into highly useful information and that reduce the number of viable options available to the commander at key decision-making junctures. Further, even though there are large amounts of data available, developed systems must be able to deal with missing and corrupted data. In addition to traditional causes of missing or corrupted data, military applications face the potential problem of hostile corruption. In a war, it should be anticipated that data could be actively intercepted, altered, or destroyed by unfriendly forces.

Given the wide variety of battle scenarios possible in future wars, systems developed for C2 need to be scaleable between large-scale operations and unit tactics. At the same time, these systems need to provide all war fighters with a common picture of any particular battlespace. Finally, and perhaps most crucially, information processing needs to be completed quickly. To the modern war fighter, time is of the essence and it is projected that speed of processing will become even more vital in the future.

Statisticians have a unique perspective on the challenges presented by C2 and thus have an opportunity to contribute to research and applications. Statistical techniques can help make the transition from the flood of data to meaningful information and knowledge for decision making. For example, consider the area of situational awareness. The statistician might consider probabilistic frameworks, scaleable algorithms, and sequential decision-making. In particular, one

might examine methods of estimating and predicting the threat or danger to a region, posed by enemy constituents. Additional areas of interest are change and anomaly detection, measures of information and understanding, and decision theory. Statisticians should examine source data, developing methods to address confidence, accuracy, and completeness of such data. They might also consider the validity of results after information processing and the quality of database information. Other research areas should focus on the representation of uncertainty or confidence in statistical results and might involve algorithms that allow statistical inference to be decentralized. Note that, in addition to providing processing stability in a hostile environment, such decentralization should lead to increased inference speed through parallel, distributed processing.

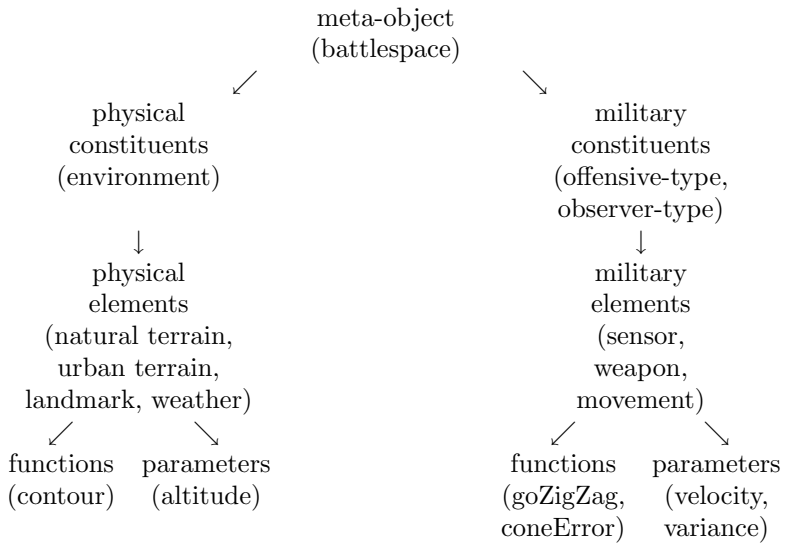
In conclusion, Command and Control (C2) is a broad field, providing a wide variety of options for statistical research and applications. In this paper, we give particular attention to one aspect, namely optimal mapping of the regional danger posed by enemy constituents in the battlespace. In Section 2, we develop an abstraction of C2 for statistical modeling and analysis and define the spatial-temporal danger-potential field. In Section 3, we discuss the object-oriented combat-simulation program we have developed in some detail. In Section 4, we examine the data for C2 decisions and the possible degradations such data might suffer. We consider analysis of the data with regard to estimating danger-potential fields in Section 5. Finally, in Section 6, we summarize our efforts and discuss future directions.

2. Abstraction of C2 for Statistical Modeling and Analysis

Before considering C2 from a statistical point-of-view, it is important to examine the nature of the battlespace, D, and the data that might emerge from it, as well as the state process of interest. By the state process, or Y-process, we mean the underlying spatial-temporal process describing the battle, assuming perfect, noiseless, complete knowledge. It was decided that, regardless of the choice of Y, a flexible simulation tool would be developed to produce imperfect, noisy, incomplete data Z. To define the tool properly, a set of definitions and rules were established. In order to keep the structure as flexible as possible, we settled on the following hierarchical design for any battlespace object.

At the largest size, the battlespace was defined to be a meta-object. For a particular simulation experiment, the battlespace contains all the other objects in the hierarchy. The next smallest class of objects is made up of *constituents*. In general, a constituent is an object that cannot be further combined (in the scope of the experiment) with another object of that class in a meaningful way. Specific examples of constituents include the environment in which the battle occurs, tanks and other offensive objects, and radar towers and other passive sensing objects. It should be noted that for battles of a larger scale, a constituent might actually be a group of 'smaller' objects. For example, a group of three tanks on patrol together might be considered a constituent in a battle scenario of broad enough extent. In the scale considered in this paper, single tanks are considered

Figure 1: Hierarchical Design of Battlespace Object



to be constituents. Constituents are made up of *elements*. Elements are a smaller class of objects that have a specific sort of activity assigned to them. A tank constituent, for example, might have one or more human elements, mobility elements, sensor elements, and weapon elements. Elements of the environment constituent might include terrain and weather. The smallest classes of objects in our hierarchy are *functions* and *parameters*. Functions and parameters work together to define specifically how elements perform their activities. Functions may have deterministic or random aspects associated with them, and parameters provide the necessary numeric arguments for functions. Again, for a movement element of a tank constituent, functions might include a deterministic 'head straight for your target' function ('goStraight' object) or a 'random walk toward your target' function ('goZigZag' object). Either of these functions might accept parameters such as maximum speed, mean angle, and the standard deviations on speed and angle. Other functions might include those describing targeting error ('coneError') or other scenario-specific activities. Functions need not necessarily be limited to acting on parameters. For instance, in more complicated models, elements of a given constituent might interact according to some element-level function. It is conceivable that even higher-level interactions or functions might also occur, but our analysis will focus only on the parameter-level functions for the time being. In the design of this hierarchy, additional object classes could be inserted for more complex simulations. For example, a *component* class could be inserted between *constituents* and *elements*.

It should be noted that one or more commanders may be associated with

each level of this hierarchy. An overall battle commander would be associated with the battlespace at the meta-object level; at the military-constituent and military-element levels, a commander would receive orders from the overall commander but would also direct actions associated with that constituent or element. Even at the lowest level of the hierarchy, one can imagine a function commander who is responsible for performing the tasks described by the function and who follows orders from higher levels of the hierarchy.

At this point, a brief notational comment should be made. For the purposes of formulating the battlespace model and developing the analysis, constituents are notated as vectors. Offensive-type constituents, such as tanks, are notated as $\mathbf{w}_k$, for $k = 1, 2, \ldots$. Observer-type constituents, such as radar towers, are notated as $\mathbf{v}_i$, for $i = 1, 2, \ldots$. In initial work, this vector is interpreted strictly as the Cartesian coordinates of the constituent in question, but it can be thought of more generally as the state of the constituent in question. Elements of a particular constituent are assumed to be located at the same location as their 'parent' constituent, although this assumption could be generalized if the constituents are distributed. When time is incorporated into the analysis, these vectors become functions of t. Specifically, in spatial-temporal-analysis settings, offensive-type constituents are notated $\mathbf{w}_k(t)$ and observer-type constituents are notated $\mathbf{v}_i(t)$. This notation refers to the location (or state) of the constituent in question at time t.

Consequently, all our analyses are done on continuous fields rather than on grids or lattices, although the grid is used for some numerical and visualization algorithms. For convenience, we have started with a region that is a rectangle of fixed size. In this paper, we focus on simple, flat terrain elements, but later work will include the effects of more complex terrain elements.

To carry out a battlespace simulation, we need to define the number, position, and type of constituents that would exist within a battle. This could be done in advance or assigned randomly in an obvious way (e.g., Poisson distribution [4] for numbers of objects, constrained and scaled uniform distribution or Beta distribution [4] for location coordinates). For initial experiments, two general classes of constituents were settled on: weapons and observers. We shall discuss observers in greater detail later (see Section 4, where data collection and degradation is considered), but we note here that observers are defined by functions and parameters associated with their viewing area and their error types.

Weapons, in general, have more complicated functions and parameters than do observers. For the simplest scenarios, only functions and related parameters associated with movement and offensive elements need to be considered. For more complicated scenarios, defensive functions and parameters need to be defined as well. Movement parameters include such things as autoregressive parameters [1] and error types, assigned waypoints, and commands. Offensive functions and parameters include targeting functions (such as 'coneError') and the explosion parameters used below. Defensive parameters include the amount

of damage that an element can suffer and still be operational and the amount of resistance an element carries against damage. We recognize that many additional parameters might be of interest to the experimenter, and so provision for their inclusion later was incorporated into the design of the simulation.

All functions were stipulated with an eye to keeping them and the simulation tool as flexible as possible. Some parameters, for example explosion parameters, were assumed to be specified and deterministic, while other parameters, for example instantaneous velocity, were given probability distributions. For a single time unit, the velocity parameter might be distributed as a scaled Beta random variable. Targeting is assigned a radius-angle error distribution [6]. This idea is discussed further in Section 4 but, in general, it was decided to apply error to the radius and angle from the weapon to the target, rather than to the Cartesian coordinates of the weapon-to-target displacement. For initial simulation experiments, waypoint parameters were chosen to be purely deterministic, with the idea that later they may be a random function based on commander orders and affected by environmental parameters. Waypoints are pre-specified locations in space that a given mobile object is ordered to reach at pre-specified times. One way to model randomness in waypoints is to consider the location-based error arising from landmark-based orders. That is, a constituent might be instructed to move to a specific landmark element within the natural- or urban-environment constituent, rather than to a specific set of coordinates. If the location of the landmark element is known with error, a level of uncertainty emerges.

Recognizing that maps are an intuitive way to present knowledge, we now focus on mapping some sort of summary of the Y-process based on imperfect, noisy, incomplete data, Z. As an example, consider the damage potential posed by an enemy unit or set of enemy units. This should be particularly interesting to a commander, and it exhibits a lot of space-time variability. From the damage potential, we wish to estimate a danger-potential field generated by a set of enemy units and to examine the uncertainties associated with it. We consider our study illustrative of the general problem of producing statistically optimal maps that evolve with the changing battlespace.

We see the damage potential of an enemy weapon as analogous to gravitational potential; that is, the damage that a weapon could do to a target, can be thought of as being equal to the damage potential of a weapon element times an armor parameter that depends on the target's ability to protect itself. Thus, the damage potential of a given weapon element is equal to the damage it could do to any target constituent with a unit armor parameter.

Clearly, the assumption that a missle applies damage at a precise confined location is unrealistic. Consequently, all damage was considered to be of an explosive type, that is, affecting a continuous region and being a non-increasing function of distance from the impact point. The following formula describes one possible form of the damage potential at a distance r from the impact point:

$$\delta(r) = \begin{cases} \alpha(1-(r/R)^{p_1})^{p_2}, & \text{for} \quad 0 < r < R \\ 0, & \text{else,} \end{cases} \tag{1}$$

where α, R, p_1, and p_2 are all explosion parameters defined for the weapon element in question. Under this definition, a single location in the battlespace can be affected by damage resulting from nearby impacts in the space, and the damage potential will vary with the distance from the impact.

Before continuing, consider the following notational conventions. Specifically, let $\mathbf{w}_{kl}$ denote the lth location impacted by weapon $\mathbf{w}_k$, allowing a single weapon to have possible multiple 'hits' in a short, specific time interval. For the purpose of this paper, attention will be focused on single-impact weapons with the impact location denoted $\mathbf{w}_{k1}$. Further, denote $f(\mathbf{w}_{k1}|\mathbf{s}, \mathbf{w}_k)$ as the probability density function of an impact at location $\mathbf{w}_{k1}$ given the weapon is located at $\mathbf{w}_k$ and is aiming at location $\mathbf{s}$. Similarly, let $\mathbf{v}_{ij}$ denote the jth location observed by observer i in a specific time interval. Notice that enemy weapon constituents may not always be distinguishable by an observer and, thus, we do not know in general to which weapon each $\mathbf{v}_{ij}$ refers. However, for the purpose of this paper, we shall assume that the observe can identify the weapon without ambiguity and we thus let $\mathbf{v}_{ik}$ denote the location observed by observer i for weapon k.

We define the *danger potential*, generated by a single weapon element at location $\mathbf{w}_k$, as the expected damage at any location $\mathbf{s}$:

$$g(\mathbf{s}; \mathbf{w}_k) = \int_{\mathbf{w}_{k1}} \delta(r_{\mathbf{s},\mathbf{w}_{k1}}) f(\mathbf{w}_{k1}|\mathbf{s}, \mathbf{w}_k) d\mathbf{w}_{k1}; \qquad \mathbf{s} \in D. \tag{2}$$

The distribution, $f(\mathbf{w}_{k1}|\mathbf{s}, \mathbf{w}_k)$, is based on the same radius-angle probability distributions given in Section 4. It should be noted that, given the parameters of the k-th weapon, $g(\mathbf{s}; \mathbf{w}_k)$ can be computed ahead of time, thus reducing the amount of time it takes to generate a danger-potential map over D.

We further assume that danger is summable. That is, we define the danger potential to an object at a specific location, from a set of enemy weapon elements, as the sum of the individual danger potentials of each weapon element. Such a definition makes sense if the individual damage potentials are summable, as we now illustrate. The danger potential at $\mathbf{s}$ is,

$$\begin{aligned} g(\mathbf{s}; \{\mathbf{w}_k\}) &= \int_{\mathbf{w}_{k1}} \sum_k \delta(r_{s,\mathbf{w}_{k1}}) f(\mathbf{w}_{k1}|\mathbf{s}, \mathbf{w}_k) d\mathbf{w}_{k1} & (3) \\ &= \sum_k \int_{\mathbf{w}_{k1}} \delta(r_{s,\mathbf{w}_{k1}}) f(\mathbf{w}_{k1}|\mathbf{s}, \mathbf{w}_k) d\mathbf{w}_{k1} & (4) \\ &= \sum_k g(\mathbf{s}; \mathbf{w}_k), & (5) \end{aligned}$$

where $\sum_k \delta(r_{s,\mathbf{w}_{k1}})$ is the damage potential of the multiple weapons $\{\mathbf{w}_{k1}\}$.

Though danger potential $g(\mathbf{s};\mathbf{w}_k)$ was defined above in a purely spatial setting, there is a natural extension to the spatial-temporal setting. Consider an offensive constituent at location $\mathbf{w}_k(\tau)$ and time τ. Then the expected damage potential at location $\mathbf{s}$ and time $t > \tau$, not only depends on the probability of applying damage, but also on the probability of the weapon's location at unobserved times. That is, in addition to taking an expectation over the targeting distribution, as in (2), the expectation of the damage at a given spatial-temporal location is also based on knowledge of the weapon's location at some previous time (τ). The resulting danger-potential field at $\mathbf{s}$ and t is:

$$
\begin{aligned}
g(\mathbf{s},t;\mathbf{w}_k(\tau),\tau) &= \int_{\mathbf{w}_{k1}(t)} \delta(r_{\mathbf{s},\mathbf{w}_{k1}(t),t}) f_1(\mathbf{w}_{k1}(t)|\mathbf{s},t,\mathbf{w}_k(\tau),\tau) d\mathbf{w}_{k1}(t) \quad (6)\\
&= \int_{\mathbf{w}_{k1}(t)} \delta(r_{\mathbf{s},\mathbf{w}_{k1}(t),t}) \quad (7)\\
&\quad \times \int_{\mathbf{w}_k(t)} f_2(\mathbf{w}_{k1}(t),\mathbf{w}_k(t)|\mathbf{s},t,\mathbf{w}_k(\tau),\tau) d\mathbf{w}_k(t) d\mathbf{w}_{k1}(t)\\
&= \int_{\mathbf{w}_{k1}(t)} \int_{\mathbf{w}_k(t)} \delta(r_{\mathbf{s},\mathbf{w}_{k1}(t),t}) f_3(\mathbf{w}_{k1}(t)|\mathbf{s},t,\mathbf{w}_k(t),t)\\
&\quad \times h_1(\mathbf{w}_k(t)|t,\mathbf{w}_k(\tau),\tau) d\mathbf{w}_k(t) d\mathbf{w}_{k1}(t) \quad (8)\\
&= \int_{\mathbf{w}_k(t)} g(\mathbf{s},t;\mathbf{w}_k(t),t) h_1(\mathbf{w}_k(t)|t,\mathbf{w}_k(\tau),\tau) d\mathbf{w}_k(t)\,. \quad (9)
\end{aligned}
$$

The probability density function, $h_1(\mathbf{w}_k(t)|t,\mathbf{w}_k(\tau),\tau)$, represents the conditional probability that the weapon is at $\mathbf{w}_k(t)$ at time t given that it was at spatial-temporal location $(\mathbf{w}_k(\tau),\tau)$, and it introduces a dynamic aspect to the analysis. These equalities hold assuming that danger is instantaneous, that the old location has no effect on targeting if the current location is known, and that the movement of the weapon does not depend on the generic spatial index $\mathbf{s}$. The probability depends not only on the movement function and the parameters associated with the weapon element, but also on the evolution of the spatial-temporal battlespace. For instance, if the weapon element is damaged, its mobility may be affected. In the simple example discussed in this paper, the weapon cannot be damaged, so that part of the dynamic aspect may be ignored. Expanding this conceptualization, one might re-write (8) and (9) as:

$$
\begin{aligned}
g(\mathbf{s},t;\mathbf{W}) &= \sum_k \int_{\mathbf{w}_{k1}(t)} \int_{\mathbf{w}_k(t)} \delta(r_{\mathbf{s},\mathbf{w}_{k1}(t),t}) f_3(\mathbf{w}_{k1}(t)|\mathbf{s},t,\mathbf{w}_k(t),t)\\
&\quad \times h_2(\mathbf{w}_k(t)|t,\mathbf{W}) d\mathbf{w}_k(t) d\mathbf{w}_{k1}(t) \quad (10)\\
&= \sum_k \int_{\mathbf{w}_k(t)} g(\mathbf{s},t;\mathbf{w}_k(t),t) h_2(\mathbf{w}_k(t)|t,\mathbf{W}) d\mathbf{w}_k(t)\,, \quad (11)
\end{aligned}
$$

where $\mathbf{W} \equiv \{(\mathbf{w}_k(\tau), \tau) : \tau < t\,;\ k = 1, 2, \ldots\}$ is the set of all past information on all weapons and $h_2(\mathbf{w}_k(t)|t, \mathbf{W})$ is the conditional probability density function describing the probability that a weapon is at a specific location in space-time, given knowledge of all weapons at various locations in space-time.

Figure 2: Notation Conventions

$\mathbf{v}_i, \mathbf{v}_i(t)$	The location (state) of observer i (at time t).
$\mathbf{v}_{ik}, \mathbf{v}_{ik}(t)$	The location of weapon k observed by observer i (at time t).
$\mathbf{w}_k, \mathbf{w}_k(t)$	The true location (state) of weapon k (at time t).
$\mathbf{w}_{kl}, \mathbf{w}_{kl}(t)$	The l^{th} location impacted by weapon k (at time t).
$\mathbf{s}, \mathbf{s}(t)$	A hypothetical location (at time t).
$r_{\mathbf{sw}_k}$	The Euclidean distance, $\|\|\mathbf{w}_k - \mathbf{s}\|\|$, between weapon k at $\mathbf{w}_k$ and the target location $\mathbf{s}$.
$r_{\mathbf{w}_k\mathbf{w}_{kl}}$	The Euclidean distance, $\|\|\mathbf{w}_{kl} - \mathbf{w}_k\|\|$, between weapon k at $\mathbf{w}_k$ and the impact location $\mathbf{w}_{kl}$.
$r_{\mathbf{sw}_{kl}}$	The Euclidean distance, $\|\|\mathbf{w}_{kl} - \mathbf{s}\|\|$, between the impact location $\mathbf{w}_{kl}$ and the target location $\mathbf{s}$.
$\theta_{\mathbf{sw}_k}$	The angle, $\arctan[(w_{ky} - s_y)/(w_{kx} - s_x)]$, between weapon k at $\mathbf{w}_k$ and the target location $\mathbf{s}$.
$\theta_{\mathbf{w}_k\mathbf{w}_{kl}}$	The angle, $\arctan[(w_{kly} - w_{ky})/(w_{klx} - w_{kx})]$, between weapon k at $\mathbf{w}_k$ and the impact location $\mathbf{w}_{kl}$.
$\theta_{\mathbf{sw}_{kl}}$	The angle, $\arctan[(w_{kly} - s_y)/(w_{klx} - s_x)]$,between the impact location $\mathbf{w}_{kl}$ and the target location $\mathbf{s}$.
etc.	

In the proposed conceptualization, data are collected primarily as observations of military-constituents locations. That is, at a given time point, each observer constituent provides a list of Cartesian coordinates representing its observed locations of other constituents. Note that these observed locations are almost certainly noisy and perhaps compromised through censoring or enemy interference. The nature of the error associated with these data is discussed further in Section 4.

Given the data, our goal in this article is to estimate and map the true danger potential everywhere in the spatial-temporal region of interest. Information about the parameters of the weapon constituents and observer constituents is combined to generate estimates of the danger potential. An advantage of the spatial-temporal statistical approaches discussed in this paper is the flexibility of the questions that might be answered using the data. Simple questions about the location of a weapon or a set of weapons are not the only ones that are answerable. In addition, one might pose questions such as, "How often are enemy weapons within 10 miles of the border?", or "Does the danger posed to friendly regions appear to increase significantly over time?", or "If friendly mobile objects need to go from A to B, what is the path of least danger?", or "Where are the enemy's mobile weapons headed?" All of these questions can be answered through some *nonlinear* functional of $\{\mathbf{w}_k(\tau) : \tau \leq t\,;\ k = 1, 2, \ldots\}$, and thus

techniques such as Kalman filtering [8] will typically yield biased estimates. An area of research for us is to develop smoothers/filters/forecasts for which any nonlinear functional (e.g., danger-potential field) is approximately unbiased.

3. Object-Oriented Combat-Simulation Program

A battle simulation program has been writen in the interpreted, object-oriented language R (`http://www.stat.cmu.edu/R/CRAN/`). The R program implements a dialect of the programing language S, developed at Bell Laboratories.

We can think about the battlespace as consisting of military *constituents* (e.g., tanks, radar stations, etc.) that interact and change in time in surrounding physical constituents (the environment). The constituents are constructed in a hierarchical fashion, using *elements* with specific functionality (e.g., a gun is one element of a tank, a sensor is another).

The simulation program consists of defining different *classes* of constituents and elements, and writing functions that act on both constituents and elements (e.g., a tank has a gun element with functionality defined by its parameters, which include range, accuracy, and explosive power).

The simulation consists of an inner and an outer loop. The outer one loops through time and the inner one loops through the military constituents (we have assumed for the moment that the physical constituents are constant during the period of simulation).

The inner loop consists of the following major steps:

Scanning the Battlespace: The scanBS Function. Most constituents have some way of observing the battlespace in order to detect other constituents. Currently, a constituent can have any number of elements of the class sensor.

Making a Decision: The askCommander Function. Every constituent has a commander that makes decisions concerning the current destination (if the constituent is mobile) and applying weapons (if available). The commander of a mobile constituent has, in most cases, a set of waypoints to follow to reach its final destination. At the same time, the commander is receiving information from the sensors about the status of the surrounding battlespace at every time point. Depending on the 'character' (e.g., aggressive, defensive, etc.) of the commander, a decision is made.

Change Location: The move Function. A constituent moves according to its current speed and destination (both determined by the constituent commander). There are a number of ways a constituent can move at each time-step, including a method that assigns a little random error to both speed and direction at each time-step.

Attack: The attack Function. A constituent can have one or more elements of the class weapon. Applying a weapon consists of (1) shooting at a

given target, but with error, and (2) evaluating the damage caused by the weapon, to all constituents (including the target) in the battlespace within damage range.

Update Attributes: The update Function. Keeps track of the history by collecting, at each time-step, parameters of interest and storing them (e.g., the value of the life and armor parameters; see Section 3.4)

Some of these major steps do not necessarily apply to all constituents.

Currently, two classes of millitary constituents have been defined. The basic constituent is a stationary object of *class* defObj with a commander. An extension of the defObj class is the mobileObj. The mobileObj *inherits* all the attributes and functions from the defObj. The only addition to the defObj is a function and parameters to make the object mobile.

Two classes of elements have currently been defined, namely, sensor and weapon. An object of class sensor has a function to observe what is in its surroundings; that is, it observes the position of natural and military constituents within the sensor's range, but with error. An element of class weapon has a function to "shoot", with error, at a given target and functions to compute the damage caused to other constituents using their life and armor values. All these error distributions are of the radius-angle type, where the radius distribution is assumed independent of the angle distribution. Furthermore, the *radius* distribution is always *truncated* so that no random radius is bigger than $r_{\max}$, and the *angle* distribution is always *wrapped* so that no random angle is outsie $(-\pi, \pi]$.

Other constituent classes can be created by extending the two basic classes; for example, the constituent class, tank, has been specified by extending the mobileObj class and adding a radar (a sensor with specific sensor function) and a gun (weapon object with attributes specified to model the workings of a gun on a tank).

3.1 Sensor

Any military constituent can have one or more elements of class sensor. Each sensor element checks for constituents within the range of the sensor. Each sensor is provided with a sensor function and a list of parameters that define the procedure used to generate noisy observations of the constitutents' location within the sensor's range. The sensor function can be easily adapted to diverse types of sensors.

coneErrorSensor Let $r_{\mathbf{v}_i\mathbf{w}_k}$ and $\theta_{\mathbf{v}_i\mathbf{w}_k}$ be the true distance and angle, respectively, between the sensor and the observed constituent, with respect to the sensor. If $r_{\mathbf{v}_i\mathbf{w}_k}$ is within the sensor's range, then suppose we observe $(r_{\mathbf{v}_i\mathbf{v}_{ik}}, \theta_{\mathbf{v}_i\mathbf{v}_{ik}})$ and express the error of that observation through:

$$r_{\mathbf{v}_i\mathbf{v}_{ik}} = r_{\mathbf{v}_i\mathbf{w}_k} e^{\varepsilon_r} \quad \text{and} \quad \theta_{\mathbf{v}_i\mathbf{v}_{ik}} = \theta_{\mathbf{v}_i\mathbf{w}_k} + \varepsilon_\theta, \tag{12}$$

where

$$\varepsilon_r \quad \sim Gau(mean = 0\,,\ variance = \sigma_r^2), \tag{13}$$

$$\varepsilon_\theta \quad \sim Gau(mean = 0\,,\ variance = \sigma_\theta^2)\,, \tag{14}$$

and when necessary truncation and wrapping occurs in (12). Then a realization of the constituent's location, $\mathbf{v}_{ik} = (v_{ikx}, v_{iky})$, is given by $v_{ikx} = v_{ix} + r_{\mathbf{v}_i\mathbf{v}_{ik}} \cos(\theta_{\mathbf{v}_i\mathbf{v}_{ik}})$ and $v_{iky} = v_{iy} + r_{\mathbf{v}_i\mathbf{v}_{ik}} \sin(\theta_{\mathbf{v}_i\mathbf{v}_{ik}})$, where $\mathbf{v}_i = (v_{ix}, v_{iy})$ is the location of the sensor.

bivarNormSensor If $r_{\mathbf{v}_i\mathbf{w}_k}$ is within the sensor's range and $\mathbf{w}_k = (w_{kx}, w_{ky})$ is the true location of the weapon constituent, then suppose we observe $\mathbf{v}_{ik} = (v_{ikx}, v_{iky})$ and express the error of the observation through:

$$v_{ikx} = w_{kx} + \varepsilon_x \quad and \quad v_{iky} = w_{ky} + \varepsilon_y, \tag{15}$$

where independently,

$$\varepsilon_x, \varepsilon_y \sim Gau(mean = 0, variance = \sigma^2)\,. \tag{16}$$

3.2 The Commander

Every class of constituent has a commander, who we shall call the constituent commander. For a constituent of the class, defObj, the simplest commander does nothing. For a mobile constituent of the class, mobileObj, the simplest commander just follows the waypoints given to reach its final destination. In general, a commander element is a function that alters the movement of the constituent and gives weapons a target to shoot at.

Other types of commanders that have been constructed include an aggressiveCommander and a defensiveCommander. The aggressiveCommander chases and then attacks every enemy seen on any sensor, but the defensiveCommander sticks to its given waypoints and shoots at any enemy within range (but does not chase them).

3.3 Mobility

For a constituent of the class, mobileObj, and all other classes that inherit the mobileObj class, there is a function that changes the position of the constituent at each time-step.

Different functions can be specifed as to how the constituent can accomplish this. In general, if s_t is the speed of the constituent at time t and ϕ_t its angle of trajectory, then the speed and angle at time $t+1$ are given by

$$s_{t+1} \quad = \quad \rho_s s_t + (1-\rho_s) S_{t+1}, \tag{17}$$

$$\phi_{t+1} \quad = \quad \rho_a \phi_t + (1-\rho_a) \Phi_{t+1}, \tag{18}$$

where the $\rho_s \in [0,1]$ and $\rho_a \in [0,1]$ are given, and S_{t+1} and Φ_{t+1} are given by a move function. A number of move functions have been constructed. The two most frequently used are:

goStraightLNorm There is no error involved in the angle, only in the speed:

$$\Phi_{t+1} \quad = \quad \phi, \tag{19}$$

$$log(S_{t+1}) \quad \sim \quad Gau(mean = \mu_s,\ variance = \sigma_s^2). \tag{20}$$

goZigZag In this case, there is error in both the angle and the speed:

$$\Phi_{t+1} \sim Gau(mean = \phi, variance = \sigma_\phi^2), \tag{21}$$

$$log(S_{t+1}) \sim Gau(mean = \mu_s,\ variance = \sigma_s^2)\,, \tag{22}$$

and when necessary wrapping occurs in (21).

3.4 Weapons

A military constituent can have any number of weapon elements. Other constituents provide targets for the weapons.

Using a weapon consists of four major parts (functions):

Shooting: The weapon has a function, and a list of parameters, that specify how the actual impact point of the weapon is computed given the target position and the position of the weapon.

Damage: At the impact point (provided by the shooting function), a damage factor (potential damage) to the constituents within the maximum damage range are computed by the damage function.

Life Reduction: The damage factors are used by the life function to reduce the life value of the constituents within the maximum damage range.

Armor Reduction: Similarly, the damage factors are used by the armor function to reduce the armor value of the constituents within maximum damage range.

We now give more details on the four major steps described above. Let $\mathbf{w}_k = (w_{kx}, w_{ky})$ and $\mathbf{s} = (s_x, s_y)$ denote the position of the weapon and the target, respectively, and let $r_{\mathbf{w}_k\mathbf{s}}$ and $\theta_{\mathbf{w}_k\mathbf{s}}$ denote the distance and angle of the target with respect to the weapon. The two most frequently used shooting functions are:

coneShooting: The impact point is given by

$$w_{k1x} = w_{kx} + (r_{\mathbf{w}_k\mathbf{s}} e^{\varepsilon_r}) \cos(\theta_{\mathbf{w}_k\mathbf{s}} + \varepsilon_\theta), \tag{23}$$

$$w_{k1y} = w_{ky} + (r_{\mathbf{w}_k\mathbf{s}} e^{\varepsilon_r}) \sin(\theta_{\mathbf{w}_k\mathbf{s}} + \varepsilon_\theta), \tag{24}$$

where independently,

$$\varepsilon_r \sim Gau(mean = 0,\ variance = \sigma_r^2), \tag{25}$$

$$\varepsilon_\theta \sim Gau(mean = 0,\ variance = \sigma_\theta^2), \tag{26}$$

and when necessary truncation occurs with $(r_{\mathbf{w}_k\mathbf{s}} e^{\varepsilon_r})$ and wrapping occurs with $(\theta_{\mathbf{w}_k\mathbf{s}} + \varepsilon_\theta)$.

bivarNormShooting: The impact point is given by

$$w_{k1x} = w_{kx} + \Delta x_{\mathbf{w}_k\mathbf{s}}, \tag{27}$$
$$w_{k1y} = w_{ky} + \Delta y_{\mathbf{w}_k\mathbf{s}}, \tag{28}$$

where independently,

$$\Delta x_{\mathbf{w}_k\mathbf{s}} \sim Gau(mean = s_x - w_{kx}, SD = \alpha + \beta r_{\mathbf{w}_k\mathbf{s}}), \tag{29}$$
$$\Delta y_{\mathbf{w}_k\mathbf{s}} \sim Gau(mean = s_y - w_{ky}, SD = \alpha + \beta r_{\mathbf{w}_k\mathbf{s}}), \tag{30}$$

with α and β given. Note that impact is written in terms of the location of the firing weapon plus a deviation, where the deviation is based on the target location $\mathbf{s}$. The function assumes that the further the target is from the weapon, the larger the variance of the impact location.

Recall that after firing a weapon aimed at $\mathbf{s}$, the explosion actually occurs at $\mathbf{w}_{k1}$. Let $r_{\mathbf{s}\mathbf{w}_{k1}}$ denote the distance between the location of the potential target and the explosion location caused by the weapon, let α denote the maximum damage that the weapon can cause, and let R denote the maximum damage range from the point-of-impact. The function that is currently used to define the damage is given below.

powerDamageScaling: The damage-factor function, $\delta(r_{\mathbf{s}\mathbf{w}_{k1}})$, is given by,

$$\delta(r_{\mathbf{s}\mathbf{w}_{k1}}) = \begin{cases} \alpha(1 - (r_{\mathbf{s}\mathbf{w}_{k1}}/R)^{p_1})^{p_2} & \text{for } 0 < r_{\mathbf{s}\mathbf{w}_{k1}} < R \\ 0 & \text{else}, \end{cases} \tag{31}$$

Let ℓ_c denote the current value of the life parameter of the constituent in question, ℓ_n the new life value after weapon impact, a_c the value of the current armor strength of the constituent, and δ the damage factor from the explosion at the location of the constituent (as given by the damage-factor function). The current life-reduction function is defined as follows.

scalingLifeReduction: The new life value is given by,

$$\ell_n = \ell_c - \delta/a_c; \tag{32}$$

if $\ell_n \leq 0$, the constituent is considered 'dead' (destroyed).

The current armor-reduction function is defined as follows.

negativeArmorReduction: The new armor value, a_n, is given by

$$a_n = a_c - \delta; \tag{33}$$

if $a_n < 0$, then a_n is put equal to 0.

The program allows for user-defined functions and is designed in such a way that it is easy to add new functions to elements.

3.5 The Environment

In the current version, a very simple physical environment is assumed; a flat terrain with no obstacles. This can be generalized; the object-oriented simulation program is very flexible.

3.6 Running the Program

Running the simulation consists of constructing all military constituents, which can include both offensive-type and observer-type constituents. After the simulation, each constituent stores its history, including the exact position of the constituent at all time points. Similarly, the sensor elements (if any are installed) store everything that they see. It is also possible to construct observer-type constituents after the simulation has run, and let them observe the battlespace at some specific time points, or at all time points. An example of five mobile tanks and three sensors (two radar and one satellite) in the battlespace is given in [9].

4. Data for C2 Decisions

Once a simulation of the Y-process has been completed, it is time to consider the generation of the data Z. Running the simulation gives the experimenter access to the true battle information but, in real-world situations, the information is degraded in some manner. To reflect this, we degraded the simulated Y-process in several ways. Censoring might occur in both the time- and space-domains for various reasons. Terrain features might limit the observation region of an observer, and technical difficulties might prevent or delay observation of data. Location-based error was also deemed to be a fairly likely situation. Less likely, but worth considering, was false data provided by the enemy.

Spatial and temporal censoring can occur under a variety of circumstances. Some examples might include adverse sensing conditions (e.g., sunspot activities, jamming) that could prevent or delay any data transmission during the affected time period. Such censoring might be generated by using exponentially distributed start, stop, and transmission delay times. It is also likely that certain sensors may not transmit continuously. These sensors may transmit at systematic intervals (for instance, one data report every 30 seconds) or at event-invoked times. One might consider a reliability probability for such sensors. That is, for a predetermined set of time points, there is a probability that at each of these times the sensor will transmit its data. Additionally, any given sensor might fail during the course of the battle. Spatial censoring might occur as the result of terrain (e.g., a hill) or meterological (e.g., clouds) effects. Finally, there may be systematic spatial censoring resulting from spatial regions outside of the sensing range of all available sensor constituents or elements. Any or all of these types of censoring may occur in a single battle scenario.

Location-based error is worth particular mention. It seems to us to be ubiquitous in all battle scenarios. Moreover the radius-angle error distributions turn up repeatedly. While standard bivariate normal error might be applied to any location (and is indeed applied to satellite-type sensors), radar-type sensors and weapon-target applications were given radius-angle error distributions.

Let (x, y) be the true Cartesian coordinates of the observer location (assume known) and (x_0, y_0) the true Cartesian coordinates of the location of interest (unknown). Notate the true angle and radius from the observer location to the location of interest as θ_0 and r_0, respectively. Now suppose that $\widehat{\theta}$ and $\widehat{r}$ are independent, random, noisy observations of θ_0 and r_0. If $\widehat{\theta}$ and $\widehat{r}$ are unbiased for θ_0 and r_0, the nonlinear relationship between (θ_0, r_0) and (x_0, y_0) leads to a bias in $(\widehat{x}, \widehat{y})$, the estimates of (x_0, y_0) based on $(\widehat{\theta}, \widehat{r})$. (To see this, note that $E(\widehat{x}) = E(x + \widehat{r}\cos\widehat{\theta}) = x + E(\widehat{r})E(\cos\widehat{\theta}) = x + r_0 E(\cos\widehat{\theta})$. But $E(\cos\widehat{\theta}) \neq \cos\theta_0$, so $E(\widehat{x}) \neq x_0 = x + r_0\cos\theta_0$. This also holds for y_0, by symmetry.) However, a Taylor approximation such as the one described in Section 5 below, shows that by adjusting the mean of $\widehat{r}$, an approximately unbiased estimate for (x_0, y_0) can be achieved. That is, assume that the reported angle $\widehat{\theta}$ is:

$$\widehat{\theta} = \theta_0 + \varepsilon_\theta \,, \tag{34}$$

where ε_θ is distributed with mean 0 and variance σ_θ^2; and assume that the reported distance $\widehat{r}$ is randomly distributed with (adjusted) mean $r_0/(1 - \frac{1}{2}\sigma_\theta^2)$ and variance σ_r^2. As is shown in (**??**)-(**??**) of Section 5, these assumptions lead to an approximately unbiased estimate of (x_0, y_0). The angle $\widehat{\theta}$ and radius $\widehat{r}$ are then converted back into the standard Cartesian coordinates for the purpose of mapping the results.

In our simulations, a small set of distributions was considered for the radius and the angle; we always assume the radius distribution and the angle distribution to be independent. For the radius, we considered the truncated lognormal, truncated gamma, and scaled beta distributions; these were chosen because it was felt that in each case a maximum range seemed applicable. For the angle, only two distributions were considered, namely, the wrapped normal and the Von Mises distribution [6]. The default distributions for the radius-angle error distributions were the truncated lognormal for the radius, independent of the wrapped normal for the angle.

5. Analysis

For the purpose of this paper, the goal of the analysis is to estimate the true danger-potential field of a battlespace over time. Note that the deterministic danger-potential field is well defined if the impact-location distribution and the location of the firing weapon is known. Since the former is likely to be known from intelligence sources, our focus is on the latter. Consequently, we wish to estimate the true danger potential in the face of unknown weapon locations. We discuss two possible approaches, whose differences we shall investigate in the future. One approach is to apply Bayesian methods [5]. In this case, one might generate the distribution of danger-potential fields implied by the distribution of weapon locations. An alternate, much faster approach is to apply some sort of 'plug-in' estimation for the locations of the offensive constituent(s).

Regardless of the approach used, an estimate of the posterior probability distribution of the locations of the weapons, given the observations, would be

useful. When dealing with the radius-angle errors described earlier, the distributions of the observations can be complicated to represent. However, second-degree Taylor approximations can be used to approximate the mean and the (2×2) covariance matrix of the observations, as we shall now demonstrate.

Before examining this approximation, the following facts are worth noting. Second-degree Taylor approximations show that, for ε_θ small, $\cos(\varepsilon_\theta)$ is approximately equal to $1 - \frac{1}{2}\varepsilon_\theta^2$, and $\sin(\varepsilon_\theta)$ is approximately equal to ε_θ. Similarly, $\cos^2(\varepsilon_\theta)$, $\sin^2(\varepsilon_\theta)$, and $\cos(\varepsilon_\theta)\sin(\varepsilon_\theta)$ are each approximately equal to $1 - \varepsilon_\theta^2$, ε_θ^2, and ε_θ, respectively. Assuming the same distribution on the distance and angle as discussed above, consider the observation, $\mathbf{v}_{ik} \equiv (v_{ikx}, v_{iky})$. Write $\mathbf{v}_{ik}$ as $\mathbf{v}_i + \mathbf{\Delta}_{\mathbf{v}_{ik}}$, where $\mathbf{\Delta}_{\mathbf{v}_{ik}} = (\Delta_x, \Delta_y)$, $\Delta_x = r\cos(\theta)$, $\Delta_y = r\sin(\theta)$, and $\theta = \theta_{\mathbf{v}_i\mathbf{w}_k} + \epsilon_\theta$. Combining these equalities and assuming independence of r and θ, we obtain

$$\begin{aligned} E[v_{ikx}] &= E[v_{ix} + r\cos(\theta_{\mathbf{v}_i\mathbf{w}_k} + \epsilon_\theta)] \\ &= v_{ix} + E[r]E[\cos(\theta_{\mathbf{v}_i\mathbf{w}_k} + \epsilon_\theta)] \qquad (35) \\ &= v_{ix} + \frac{r_{\mathbf{v}_i\mathbf{w}_k}}{1 - \frac{1}{2}\sigma_\theta^2}\left\{\cos(\theta_{\mathbf{v}_i\mathbf{w}_k})E[\cos(\epsilon_\theta)] - \sin(\theta_{\mathbf{v}_i\mathbf{w}_k})E[\sin(\epsilon_\theta)]\right\}, \end{aligned}$$

where the last equality is a consequence of the assumptions following equation (**??**). Applying the Taylor approximations, we obtain

$$\begin{aligned} E[v_{ikx}] &\simeq v_{ix} + \frac{r_{\mathbf{v}_i\mathbf{w}_k}}{1 - \frac{1}{2}\sigma_\theta^2}[\cos(\theta_{\mathbf{v}_i\mathbf{w}_k})E[1 - \frac{1}{2}\epsilon_\theta^2] \qquad (36) \\ &\quad -r_{\mathbf{v}_i\mathbf{w}_k}\sin(\theta_{\mathbf{v}_i\mathbf{w}_k})E[\epsilon_\theta]] \\ &= v_{ix} + r_{\mathbf{v}_i\mathbf{w}_k}\cos(\theta_{\mathbf{v}_i\mathbf{w}_k}) \qquad (37) \\ &= w_{kx}\,. \end{aligned}$$

A similar approach can be used to compute the expected values of v_{iky}, v_{ikx}^2, v_{iky}^2 and $v_{ikx}v_{iky}$. With these expected values, the approximate expected value and covariance matrix of $\mathbf{v}_{ik}$ can be computed, yielding

$$E[\mathbf{v}_{ik}] \simeq \mathbf{w}_k\,, \qquad (38)$$

and

$$\mathrm{var}[\mathbf{v}_{ik}] \simeq [\sigma_r^2 - r_{\mathbf{v}_i\mathbf{w}_k}^2(1-\gamma) - \sigma_\theta^2(\sigma_r^2 + r_{\mathbf{v}_i\mathbf{w}_k}^2\gamma)]\mathbf{p}\mathbf{p}' + [\sigma_r^2 + r_{\mathbf{v}_i\mathbf{w}_k}^2\gamma]\sigma_\theta^2\mathbf{q}\mathbf{q}', \qquad (39)$$

where $\mathbf{p} \equiv (\cos\theta_{\mathbf{v}_i\mathbf{w}_k}, \sin\theta_{\mathbf{v}_i\mathbf{w}_k})'$, $\mathbf{q} \equiv (\sin\theta_{\mathbf{v}_i\mathbf{w}_k}, -\cos\theta_{\mathbf{v}_i\mathbf{w}_k})'$, and $\gamma = (1 - \frac{1}{2}\sigma_\theta^2)^{-2}$. Note that the covariance matrix above involves the unknown angle $\theta_{\mathbf{v}_i\mathbf{w}_k}$; it is suggested that the observed angle $\theta_{\mathbf{v}_i\mathbf{v}_{ik}}$ be used in its place.

More generally, consider the following model for data obtained by observing the offensive constituent $\mathbf{w}_k$ from an observation constituent, $\mathbf{v}_i$. The resulting observation, $\mathbf{v}_{ik}$, can be expressed as:

$$\mathbf{v}_{ik} = \mathbf{w}_k + \boldsymbol{\varepsilon}_{ik}, \tag{40}$$

where $\boldsymbol{\varepsilon}_{ik}$ is a random vector with mean zero and a (2×2) covariance matrix depending on the type of error distribution associated with the observer, as well as the relative position of the observation constituent with respect to the offensive constituent. In the case of the observation constituents being radar sites, we assume a radius-angle error distribution and the covariance matrix of $\boldsymbol{\varepsilon}_{ik}$ is given by (**??**).

Thus, for observations $\{\mathbf{v}_{ik} : i = 1, \ldots, M\}$ on $\mathbf{w}_k$ from multiple observers, $\mathbf{v}_1, \ldots, \mathbf{v}_M$, the following model applies:

$$\begin{pmatrix} \mathbf{v}_{1k} \\ \vdots \\ \mathbf{v}_{Mk} \end{pmatrix} = G\mathbf{w}_k + \begin{pmatrix} \boldsymbol{\varepsilon}_{1k} \\ \vdots \\ \boldsymbol{\varepsilon}_{Mk} \end{pmatrix}, \tag{41}$$

where

$$G = \begin{bmatrix} I_2 \\ I_2 \\ \vdots \\ I_2 \end{bmatrix}, \tag{42}$$

and I_2 is the (2×2) identity matrix. The $\boldsymbol{\varepsilon}_{ik}$ in (41) are independent (2×1) zero-mean error vectors with covariance matrix, Σ_i, as given by (**??**). Notice that G is a $(2M \times 2)$ matrix, $\mathbf{w}_k$ is a (2×1) vector, and the other vectors represented above are $(2M \times 1)$ vectors. The generalized least-squares estimator for $\mathbf{w}_k$ is then

$$\hat{\mathbf{w}}_k = (G'\Sigma^{-1}G)^{-1}G'\Sigma^{-1} \begin{bmatrix} \mathbf{v}_{1k} \\ \vdots \\ \mathbf{v}_{Mk} \end{bmatrix}, \tag{43}$$

$$\tag{44}$$

where

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_M \end{bmatrix}. \tag{45}$$

This is a rather crude estimate, and we are currently exploring filtering/forecasting methods (e.g., variants of the Kalman filter and Bayesian sequential imputation) that incorporate past information on the weapon's locations.

Recall from (11) how weapons' past locations $\mathbf{W} = \{(\mathbf{w}_k(\tau), \tau) : \tau < t\}$ determine the danger-potential field $g(\mathbf{s}, t; \mathbf{W})$ in space and time. We now consider a forecast of this field at the *present* time t, based on noisy *past* observations, $\mathbf{V} \equiv \{(\mathbf{v}_{ik}(\tau), \tau) : \tau < t;\ k = 1, 2, \ldots\}$ of the weapons' movements. Define

$$\begin{aligned} \widehat{g}(\mathbf{s}, t; \mathbf{V}) &\equiv E[g(\mathbf{s}, t; \mathbf{W})|\mathbf{V}] \\ &= \int_{\mathbf{W}} g(\mathbf{s}, t; \mathbf{W}) f_{\mathbf{W}|\mathbf{V}}(\mathbf{W}|\mathbf{V}) d\mathbf{W}\,, \end{aligned} \tag{46}$$

where $f_{\mathbf{W}|\mathbf{V}}(\cdot)$ is the posterior probability density of the true spatial-temporal locations $\mathbf{W}$, given the data $\mathbf{V}$. Holding all the assumptions from the development of (2)-(**??**) to be true and replacing $g(\mathbf{s}, t; \mathbf{W})$ with (**??**), (**??**) can be written as

$$\begin{aligned} \widehat{g}(\mathbf{s}, t; \mathbf{V}) &= \int_{\mathbf{W}} \sum_k \int_{\mathbf{w}_k(t)} g(\mathbf{s}, t; \mathbf{w}_k(t), t) h_2(\mathbf{w}_k(t)|t, \mathbf{W}) d\mathbf{w}_k(t) \\ &\qquad \times f_{\mathbf{W}|\mathbf{V}}(\mathbf{W}|\mathbf{V}) d\mathbf{W} \qquad (47) \\ &= \sum_k \int_{\mathbf{w}_k(t)} g(\mathbf{s}, t; \mathbf{w}_k(t), t) \\ &\qquad \times \int_{\mathbf{W}} h_2(\mathbf{w}_k(t)|t, \mathbf{W}) f_{\mathbf{W}|\mathbf{V}}(\mathbf{W}|\mathbf{V}) d\mathbf{W} d\mathbf{w}_k(t)\,. \qquad (48) \end{aligned}$$

Because $\mathbf{w}_k(t)$ at t and $\mathbf{V}$ are conditionally independent given $\mathbf{W}$, we can rewrite $h_2(\mathbf{w}_k(t)|t, \mathbf{W}) f_{\mathbf{W}|\mathbf{V}}(\mathbf{W}|\mathbf{V})$ as $h_3(\mathbf{w}_k(t), \mathbf{W}|t, \mathbf{V})$. Hence,

$$\begin{aligned} \widehat{g}(\mathbf{s}, t; \mathbf{V}) &= \sum_k \int_{\mathbf{w}_k(t)} g(\mathbf{s}, t; \mathbf{w}_k(t), t) \qquad (49) \\ &\qquad \times \int_{\mathbf{W}} h_3(\mathbf{w}_k(t), \mathbf{W}|t, \mathbf{V}) d\mathbf{W} d\mathbf{w}_k(t) \\ &= \sum_k \int_{\mathbf{w}_k(t)} g(\mathbf{s}, t; \mathbf{w}_k(t), t) h_4(\mathbf{w}_k(t)|t, \mathbf{V}) d\mathbf{w}_k(t)\,, \qquad (50) \end{aligned}$$

where $h_4(\mathbf{w}_k(t)|t, \mathbf{V})$ is the conditional probability density function describing the probability that a weapon is at a specific location at time t, given the past observations of all weapons at various locations. The calculations leading to (50) require modification if $\mathbf{V}$ also contains current observations $\{(\mathbf{v}_{ik}(t), t) : k = 1, 2, \ldots\}$; this results in a filtered (rather than a forecasted) space-time danger potential. These calculations will be reported on elsewhere. It should also be noted that at times where observations are not taken or are incomplete, the danger potential will tend to spread out as the uncertainty about the locations of the weapons increase. For an example of a space-time danger field generated by the computer program described in Section 3, see [9].

6. Discussion and Future Directions

In examining the ways in which Statistics can contribute meaningfully to problems in Command and Control, our research efforts at The Ohio State University [7] are based on our belief that the proper quantification of uncertainty is crucial to providing the commander with as accurate and precise information as possible. We also believe that maps are an effective way of representing the resulting knowledge and uncertainty. We have developed a hierarchical design and notation for discussing the hypothetical battlespace, which has led to the postulation of danger potential as an information tool of interest to the battle commander.

The danger-potential field has a number of desirable properties. First, the fields are summable, so that the effect of additional weapons can be easily incorporated. In addition, danger potential extends naturally to spatial-temporal fields. Since the form of the danger potential of a specific weapon can be computed in advance and can be represented concisely as a function of the weapon location, a spatial-temporal picture of the danger-potential field can be developed quickly, assuming the locations of the weapons are known.

The weapons' locations are often unknown and we are exploring two approaches to incorporating the location uncertainty into the danger-potential field: plug-in estimation and full Bayesian inference. Covariance-matching kriging methods are being considered for plug-in estimation and sequential-imputation methods are being considered for Bayesian inference.

Acknowledgement

This research was supported by the Probability and Statistics Program of the Office of Naval Research under grant number N00014-99-1-0001.

Bibliography

[1] Chatfield, C., *The Analysis of Time Series: An Introduction*. Chapman & Hall, London, 1989

[2] Command and Control Applications Compendium 2000 (http://www.mcu.usmc.mil/ccss/ccsc/C2%20Compendium/default.htm)

[3] Command and Control Systems Course (http://www.mcu.usmc.mil/ccss/CCSC/)

[4] DeGroot, M. H., *Probability and Statistics*. Addison-Wesley, Reading, 1989

[5] Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., *Bayesian Data Analysis*. Chapman & Hall, London, 1995

[6] Mardia, K. V., Jupp, P. R., *Directional Statistics*. John Wiley & Sons, Chichester, 2000

[7] Probability and Statistics in Command and Control (http://www.stat.ohio-state.edu/∼C2/)

[8] Ramachandra, K.V., *Kalman Filtering Techniques for Radar Tracking.* Marcel Dekker, New York, 2000

[9] Wendt, D., Cressie, N., and Johannesson, G. A spatial-temporal statistical approach to command and control problems in battle-space digitization, in *Battlespace Digitization and Network Centric Warfare*, ed. R. Suresh. Society of Photo-Optical Instrumentation Engineers (SPIE) Proceedings, forthcoming in 2001.

SPECIAL SESSION ON THE DIGITAL GOVERNMENT (1030 - 1200)

Statistics and a Digital Government for the 21st Century
Cathy Dippo, Bureau of Labor Statistics

Fostering information technology research for and technology transfer to non-R&D government agencies is the focus of a new Digital Government program at the National Science Foundation. The major Federal statistical agencies have been involved in the development of the program and recruiting academic research partners to apply for NSF grants for four years. Our efforts have been so successful, that $4.2 million of the first year's program funds were awarded in 4 grants for research related to statistics. Several more grants have just been awarded as part of the second round of awards. The partnerships with information technology academic researchers have also led to their applying for grants under NSF's Information Technology Research (ITR) program. Some of the partnerships were developed through participation in what is now called the Digital Government Consortium, which is about to change its membership rules to encourage broader participation by non-R&D agencies.

Web Dissemination of Disclosure-Limited Analyses of Confidential Data
Alan Karr, National Institute of Statistical Sciences and Sallie Keller-McNulty, Los Alamos National Laboratory

This talk describes Web-based systems, under development at NISS, for dissemination of disclosure-limited statistical analyses of confidential data (for example, data collected by Federal agencies such as the Bureau of Labor Statistics). The essential features of these systems are that: (1) The results of statistical analyses are disseminated (as opposed to restricted access via data centers or dissemination of public use microdata); (2) Disclosure risk for each query is computed dynamically, in light of previous queries; (3) Strategies may be applied to reduce risk; (4) Analyses are possible that integrate user data with the confidential data.

Statistics in Intrusion Detection
Jeffrey Solka, Naval Surface Warfare Center

This talk will focus on several recent efforts in the application of statistics to computer security. These efforts have focused on the application of modern statistical visualization procedures to network traffic data, the inference of machine functionality via cluster analysis, and the application of importance sampling to n-gram based intrusion detection. This work is currently being funded by the Office of Naval Research.

LUNCHEON PRESENTATION (1200 - 1330)

John Tukey (1915-2000): Deconstructing Statistics[1]

James R. Thompson
Department of Statistics
Rice University, Houston, Texas 77251-1892
thomp@rice.edu

Abstract

John Tukey ushered in a new age of statistics. He taught us to question the fundamental assumptions of Fisher. He was the first significant postmodern statistician. John gave us a new paradigm of data analysis. He brought statistics into the age of computer visualization as an alternative to model based inference. He created a revolution whose ultimate consequences are still unclear.

Keywords: Exploratory Data Analysis, Graphics, Robustness.

1 Introduction

G.K Chesterton observed that there was a fundamental difference between the Anglo-Saxon and European (i.e., French) modalities of model-based action. The Europeans would build up a logical structure, brick on brick, to a point where conclusions, possibly dreadful ones, became clear and then act upon them. The Anglo-Saxons would go through the same process but then, before implementing the conclusions, would subject them to "common sense." "Yes, yes," they would say. "These seem to be the conclusions, but they simply aren't consistent with what we know to be true. Better have a reexamination of the assumptions."

Such an attitude has a long tradition in British letters and science. We recall Samuel Johnson when looking at a carefully crafted argument for determinism, including the absence of free will, dismissing the argument with, "Why, Sir, we know the will is free." And again, when considering Bishop Berkeley's neoPlatonist rejection of objective reality, Johnson kicked a stone down the high street as a conclusive counter-example.

Sometimes, even the British fall captive to the consequences of flawed models. For example, the Reverend Thomas Malthus's model opined that human populations would grow exponentially until they hit the linear (in time) growth of food supplies, and then crash, and start the process all over again. The ruling powers in

[1]This research was supported in part by a grant (DAAD19-99-1-0150) from the Army Research Office (Durham.

England quickly started acting on the implications of this model. The Irish Potato Famine was one of the results. Arguing that it would do no good to feed the starving Irish peasants from the plentiful, Irish produced, grain supplies, as this would simply postpone the inevitable, the English calmly watched while over a million people starved. The Malthusian model led to the (inappropriately named) theory called Social Darwinism, in reaction to which Marxism, with its own devastating consequences, was a response.

John Tukey was perhaps the most important questioner of models in Twentieth Century science. His logical roots extend all the way back to the Franciscan *Prince of Nominalists* William of Ockham (1280-1349).

There is a danger in disdaining the consquences of reasonable models. The AIDS epidemic is a case in point. Strong model based arguments were made early on for closing the gay bathhouses. They remain open in the United States to this day. Now there seems to be evidence that the huge pool of HIV infectives in the United States, via cheap air travel, drives a non stand-alone epidemic in Europe. But these are all model based conclusions, and model based conclusions are not the fashion today. An anti-modeling philosophy fits in well with the temper of the times. An American president has answered the ancient question, "What is truth?" with "It depends on what the meaning of 'is' is." We live in a time of grayness: "Some things nearly so, others nearly not."

Modern statistics was started in 1662 by John Graunt (Graunt, 1662), a London haberdasher and late Royalist captain of horse, without academic credentials, who came up with the incredible notion of aggregating data by quantitative attribute as a means of seeing the forest instead of the trees. Instead of discarding his ideas as downmarket because of the lack of credentials of the discoverer, the "respectable" establishment of his time, in the person of William Petty, simply tried to co-opt them as their own discovery.

The "Victorian Enlightenment," in which the statistical community played an important part, received its initial impetus from the Darwinian revolution in biological science. Statisticians were iconoclasts, questioning assumptions, using data based modeling to attack assumptions based on long held views. Francis Galton and Karl Pearson were key players, with Fisher coming along later as the premier consolidator and innovator of his statistical age.

A characteristic of these bold Victorians was a belief in their models. The old models of Lemarque, say, were wrong. The new models of the Victorians, however, were, if not precisely correct, a giant step toward reality. We recall Galton's confidence (Galton, 1879) in the universality of the "normal distribution" (so-called because everything followed it):

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified and deified by the Greeks, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason.

Students who have had their grades on tests subjected to a "normal curve" are victims of blind faith in Galton's 110 year old "maxim." (There is no obvious pooling of attributes which causes the Central Limit Theorem to operate in this case.)

2 Enter John Tukey and EDA

An obvious scientific progression would be for Galton's maxim to serve as an hypothesis which would compete with some new maxim which would then evolve into an improvement on the notion of universal normality. Tukey questioned the normal assumptions. He gave counter-example after counter-example where the normality assumption was false. But he never presumed to replace it with any model based alternative.

The same can be said for just about everything in the Fisherian pantheon of statistical models. Tukey deconstructed them all. But he did not replace them with a new set of explicit models. He simply dispensed with the Aristotelian paradigm of a chain of ever better models and changed the terms of discourse to the construction of a set of techniques for data analysis which were apparently "model free." Those who attended one of John's shortcourses in Exploratory Data Analysis (generally team taught with a junior colleague, to whom John gave most of the good punch lines) have heard the mantra:

> Exploratory Data Analysis lets the data speak to you without the interference of models.

3 The Box Plot

Let us consider one of the main tools of EDA: the *schematic* or *box* plot. The rules for its construction are rather simple:

1. Find the median (M) of the data.

2. Find the lower quartile and lower quartile of the data (lower hinge (LH) and upper hinge (UH)).

3. Compute the Step length: S = 1.5 (UH - LH).

4. Determine the Lower and Upper Inner Fences according to LIF = LH - S and UIF = LH + S.

5. Determine the Lower and Upper Outer Fences according to LOF = LH - 2S and UOF = UH + 2S.

The use of the *schematic* plot essentially replaces the normal theory based one-dimensional hypothesis tests of Fisher. Let us consider a sampling of 30 income levels from a small town: 5600, 8700, 9200, 9900, 10100, 11200,13100,16100,19300, 23900, 25100, 25800, 28100, 31000, 31300,32400, 35800, 37600,40100,42800, 47600, 49300, 53600, 55600, 58700, 63900, 72500, 81600, 86400, 156400.

$$M = \frac{1}{2}[31300 + 32400] = 31850. \tag{1}$$

$$LH = \frac{1}{2}[13100 + 16100] = 14600. \tag{2}$$

Similarly, for the *upper hinge*, we have

$$UH = \frac{1}{2}[53600 + 55600] = 54600. \tag{3}$$

$$S = 1.5 \times (54600 - 14600) = 60000. \tag{4}$$

$$UIF = 54600 + 60000 = 1146000. \tag{5}$$

$$UOF = 54600 + 2 \times 60000 = 174600. \tag{6}$$

In the Tukey paradigm, an income value outside an inner fence would be suspicious, i.e., we would not be very confident that it really was from the mass of incomes considered. And a value that is outside an outer fence would almost certainly not be from the mass of incomes considered. We see, here, that only one income is outside an inner fence, and it is inside an upper fence, namely the income of 156,400. So, by the use of the schematic plot, we would say that it appeared that the individual with an income of 156,400 is not a part of the overall population considered. Our judgement is ambiguous, since the observation falls inside the upper outer fence. Such a rule, as given in *Exploratory Data Analysis*, is didactic and, unlike the Fisherian significance tests, apparently model free.

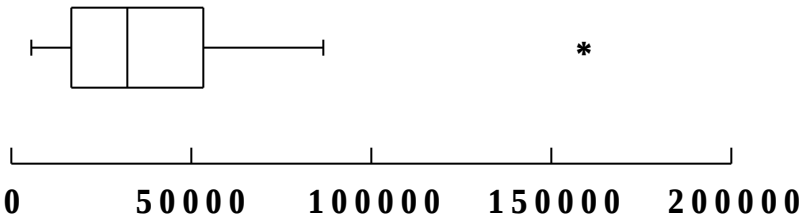


Figure 1. Box Plot of Income Data.

Let us note, however, some obvious assumptions. Firstly, the technique assumes that the data comes from a symmetrical distribution. (Of course, such symmetry might be approximately obtained from the raw data by an appropriate transformation.) Then, we need to notice some tacit assumptions underlying the placing of the

fences. If the underlying distribution is normal, the probability of an observation being outside an inner fence is $.007 \approx 1\%$. This is, interestingly, about the same one would obtain with a 1% significance test of parametric form. If the data is normal, the probability of an observation lying outside the outer fences is 2 in a million (a nice round number). It is hard to imagine that Tukey was not thinking in terms of the normal distribution as his standard. The inner fence and outer fence boundaries are not appropriate for distributions much tailier than the normal.[2] The technique based approach of the box plot, then, is very similar to what one would get if one used the normal assumption and specifically derived a likelihood ratio test, or the like.

If the schematic plot reduces essentially to a model based parametric procedure, what have we gained by its use? The answer would appear to be "graphical perception" rather than robustness. Moreover, the schematic plot does not correct for sample size. With a sample size of 30 from the same normal distribution, the probability that none of the observations will fall outside the inner fences is 81% rather than 99%. For 30 samples from the same normal distribution, the probability that all the observations will fall inside the outer fences is 99.994% rather than 99.9998%. Suppose the number of incomes we considered was 300 instead of 30. Then the probability of all the normal variates falling inside the inner fences is only 12.16%. The probability of all falling inside the outer fences is 99.94 %. The box plot is a wonderful desk accessory (and is installed as such in most statistical software packages), particularly when one is plotting data from two sources side by side, but it needs to be used with some care.

4 The 3RH Smooth

The work in nonparametric regression has been voluminous. Almost all has actually been within the context of Fisherian orthodoxy. Tukey attacked it *de novo*.

Consider a small company that makes van modifications transforming vans into campers. In Table 1 we show production figures on 30 consecutive days.

[2]For the Cauchy distribution, the probability of an observation falling outside an inner fence is 15.6%. The probability of an observation falling outside an outer fence is 9%.

Table 1. Production.	
1	133
2	155
3	199
4	193
5	201
6	235
7	185
8	160
9	161
10	182
11	149
12	110
13	72
14	101
15	60
16	42
17	15
18	44
19	60
20	86
21	50
22	40
23	34
24	40
25	33
26	67
27	73
28	57
29	85
30	90

We graph this information in Figure 2. The points we see are real production figures. There are no errors in them. Nevertheless, most people would naturally smooth them. Perhaps the figures are real, but the human observer wants them to be smooth, not rough. One could look upon such a tendency to want smoothness as being some sort of natural Platonic notion hardwired into the human brain. We can rationalize this tendency by saying that the real world has smooth productions contaminated by shocks such as sudden cancellations of orders, sickness of workers, etc. But the fact is that most of us would hate to make plans based on such a jumpy plot.

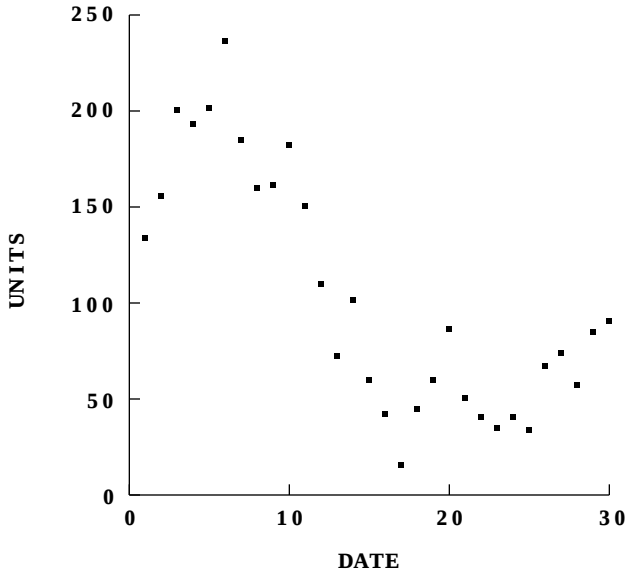


Figure 2. Daily Production Data.

The fact is that the world in which we live does tend to move rather smoothly in time. If there were so much turbulence that this was not the case, it would be hard to imagine how any sort of civilized society could ever have developed. So we ought not despise our apparently instinctive desire to see smooth curves rather than jumpy ones. Tukey achieved this by a running median (3R) smooth. Unlike the earlier hanning smooth (H) which replaced an observed value by a weighted (.25, .50, .25) average of it and its fellows, the 3R smooth will not drive the smooth to oversmoothed unreality when not checked. The 3R smooth is naturally self-limiting and requires no human manager. However, it is somewhat inclined to exhibit patches where one transitions from one locally flat function to another. So Tukey "polished" the 3R smooth with one or two hanning smooths. The result was an easy solution to the nonparametric regression problem, one that owed naught to normal theory or Fisherian orthodoxy. We show a 3RHH smooth of the production numbers in in Table 2 graphed in Figure 3.

Table 2. 3RHH Smooth.					
Day	Production	3	33=3R	3RH	3RHH
1	133	133	133	133	133
2	155	155	155	159	159
3	199	193	193	185	182
4	193	199	199	198	196
5	201	201	201	201	199
6	235	201	201	197	195
7	185	185	185	183	183
8	160	161	161	167	170
9	161	161	161	161	162
10	182	161	161	158	155
11	149	149	149	142	140
12	110	110	110	118	118
13	72	101	101	96	97
14	101	72	72	76	77
15	60	60	60	59	60
16	42	42	42	47	49
17	15	42	42	43	45
18	44	44	44	48	49
19	60	60	60	56	55
20	86	60	60	58	55
21	50	50	50	50	50
22	40	40	40	43	44
23	34	40	40	39	39
24	40	34	34	37	40
25	33	33	40	45	47
26	67	67	67	60	59
27	73	67	67	69	69
28	57	73	73	79	79
29	85	85	85	88	86
30	90	90	90	90	90

In the case of the box (schematic) plot, we have seen a technique of John Tukey which appears somewhat derivative of a Fisherian test of significance. The 3RHH smooth has nothing to do with Fisherian thinking. It is brand new with Tukey and is built on Gestaltic notions rather than anything dealing with normal theory.

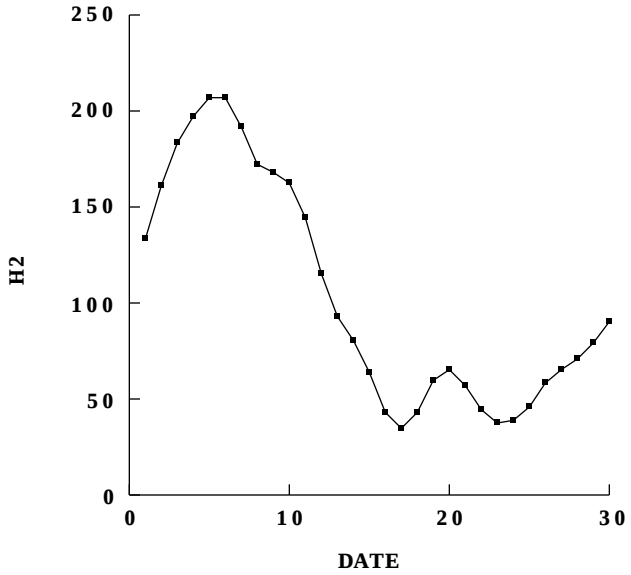


Figure 3. The 3RHH Smooth.

5 Conclusions

Most of the postmodern philosophers proclaim a new evangel and then wind up spouting old stuff from Nietzsche, Marx, Heidegger, Freud, etc. Tukey was a true revolutionary. Some of his material was, quite naturally, derivative from the work of Fisher, Pearson, and other statisticians from an earlier age. But most of the Tukey paradigm was new. In the case of smoothing, jackknifing, prewhitening, multiple comparisons, analysis of spectra, analysis of seismic data, citation indexing, projection-pursuit, Fast Fourier Transforming, data spinning, election forecasting, probably in most of the areas in which he worked, John Tukey was doing new things, things that nobody had done before. He deconstructed Fisher and Pearson and Gauss, but he then built his own structures.

What John did not do, unfortunately, was leave a roadmap of the models behind his paradigm. It can be said that W. Edwards Deming's legacy suffers from the same problem. Deming, like Tukey, was rather didactic, telling his disciples to do this or that because he, Ed Deming, had said to do it. But Deming was an Aristotelian, and, though he did not himself publish his modeling taxonomy, it is possible to discern it if one looks carefully. (Thompson and Koronacki (1992) have attempted to infer Deming's modeling structure.) Tukey is different. Nearly a pure nominalist, he spurns modeling and attacks problems almost sui generis. Yet there is a pattern. We can see some of John's implied axioms. Here are a few:

1. The eye expects adjacent pixels to be likely parts of a common whole.

2. The eye expects continuity (not quite the same as (1)).

3. As points move far apart, the human processor needs training to decide when points are no longer to be considered part of the common whole. Because of the ubiquity of situations where the Central Limit Theorem, in one form or another, applies, a natural benchmark is the normal distribution.

4. Symmetry reduces the complexity of data.

5. Symmetry essentially demands unimodality.

6. The only function which can be identified by the human eye is the straight line.

7. A point remains a point in any dimension.

Much more than this is required if one is to attempt to infer John's scientific philosophy. It might be an appropriate life's work for somebody, or for several somebodys. Though I have had a go at deciphering Deming, I despair of doing this for Tukey. And I am reasonably sure that if John were asked about the wisdom of somebody trying to come up with a Rosetta Stone for understanding the philosophy of John Tukey, he would laugh at the idea. And yet, if statistics is to survive, we have to come to grips with the fact that the statistical paradigm was changed by John Tukey in revolutionary ways, ways always brilliant, frequently destabilizing. He wanted to leave us a toolbox, and he did. But until we understand better the implications of the Tukey Revolution, we cannot possibly answer questions about the place of statistics in Twenty-First Century Science.

References

Brillinger, D.R., Fernholz, L.T. and Morgenthaler, S., eds. (1997), *The Practice of Data Analysis: Essays in Honor of John W. Tukey* Princeton: Princeton University Press, 40-45.

Cleveland, W.S.(editor in chief), *The Collected Works of John W. Tukey* (Eight volumes published from 1984 to 1994). Monterey, CA: Wadsworth Advanced Books and Software.

Galton, Francis (1889). *Natural Inheritance*. London: Macmillan & Company, 66.

Graunt, John (1662). *Natural and Political Observations on the Bills of Mortality*. London.

Thompson, J.R. (2000). "Is the United States Country Zero for the First-World AIDS Epidemic?." *The Journal of Theoretical Biology*, pp. 621-628.

Thompson, J.R. (1992), *Statistical Process Control for Quality Assurance*. London: Chapman % Hall.

Tukey, J.W. (1962), "The future of data analysis," *Annals of Mathematical Statistics*, 33, 1-67.

Tukey, J.W. (1973). *Citation Index,Index to Statistics and Probability, v.2* Los Altos, CA: R&D Press.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison Wesley Publishing Company.

Tukey, J.W. (1979), "Discussion of a Paper by Emanuel Parzen," *Journal of the American Statistical Association*, 74, 121-122.

CLINICAL SESSION I (1330 - 1530)

A Statistical Analysis of Course of Action Generation ?
Barry Bodt, Joan Forester, Charles Hansen, Eric Heilman, Richard Kaste, and Janet O'May, Army Research Laboratory

A recent approach to course of action (COA) generation employs a genetic algorithm (Fox-GA) to generate a pool of diverse, high quality courses of action. This pool is intended as an initial set which planners can and adjust before nominating a subset to the commander. Evaluation of the recommended COAs by Fox-GA has to date involved either expert opinion as to suitability or results from the abstract wargame internal to Fox-GA. (A single scenario is simulated using different COAs over terrain at the National Training Center, Fort Irwin.) Further, since the concept prototype was a component of a larger research program in military displays, much of the emphasis on Fox-GA development has been in perfecting usability.

In a limited scenario, we seek to examine Fox-GA with respect to two basic questions: (1) How do the user controls on Fox-GA affect the quality of the COA reflected in the battle outcome? (2) How well do Fox-GA recommended COAs perform in a more widely accepted simulation, Modular Semi-Automated Forces (ModSAF)? Some information regarding the variability of ModSAF results also factor in the discussion.

The author's principal concern is how to choose a reasonable response measure to support evaluation within FoxGA and comparison with ModSAF. A few measures are proposed. Within the context of the questions we seek to answer about Fox-GA, what advice can the panel offer?

A Clinical Paper On Efficient Search Strategies in High-Dimensional Complex Models

Major Thomas M. Cioppa, United States Army, Doctoral Candidate, United States Naval Postgraduate School, Department of Operations Research, Monterey, CA 93940

Dr. Thomas W. Lucas, Associate Professor, United States Naval Postgraduate School, Department of Operations Research, Monterey, CA 93940

Simulation models have become increasingly complex, often with a large number of factors (variables) requiring examination. A comprehensive exploration of these factors may not be feasible even as processing speed increases or with parallel computing. For example, a model with 26 factors, each at two levels, with only one replication would require over 67 million observations in a traditional full factorial design (2^{26}). This could be considered a "small" problem since many computer simulations have thousands of factors. If one is willing to assume negligible interactions, then a fractional-factorial design is satisfactory, but still the required number of observations may be quite large. These "traditional" experimental designs and associated theory have been well researched (Box, Hunter, and Hunter (1978) and Hicks (1993)). Experimental designs relating specifically to computer simulations and models with a large number of factors were studied by Jacoby and Harrison (1962), Hunter and Naylor (1970), Kleijnen (1975), Biles (1979), Welch et al (1992), and Bates et al (1996).

As the number of factors available for experimentation has increased, methodologies were developed to reduce the number of observations required to identify significant main effects. Earlier work by Dorfman (1943) with group screening was extended by Watson (1961) and Li (1962). Subsequent screening techniques to identify the most significant factors, denoted as factor screening, were developed by Ott and Wehrfritz (1972), Montgomery (1979), Smith and Mauro (1982), Schruben (1986), and Mauro (1986). More recent factor screening methodologies include sequential bifurcation by Bettonvil and Kleijnen (1996), Latin supercube sampling by Owen (1998), and supersaturated designs by Yamada and Lin (1999). The primary focus of these methodologies is efficiently (defined as reducing the number of observations) identifying the critical factors (defined as contributing the most to the outcome measure). The critical assumption in these methodologies is that interactions are negligible. The objective is to efficiently identify the main effects, and then, if necessary, perform additional experimentation for refinement.

In certain areas, for example command and control in military conflict, interactions are prevalent and cannot be assumed negligible, even from the initial onset of the experimentation. The presence of non-monotonicity and chaos in combat models, defined as battles on the edge, is an obscuring factor in the search (Dewar, Gillogly, and Juncosa (1991)). Finding general patterns of behavior is complicated by the difficulties associated with exploring high-dimensional non-linear model surfaces. Existing search methodologies often restrict analysts to comprehensively exploring a tiny hyperplane of model space, finding extreme points, or varying many factors simultaneously by confounding main effects with interactions. The result of the aforementioned approaches

may fail by either identifying a local extrema as a global extrema or failing to identify critical interactions.

Hencke (1998) developed an agent-based combat simulation to analyze information and coordination. His model succeeded in "providing a simulation where the agent's actions are reasonably well behaved, that is reminiscent of well-trained forces." His work is an excellent representation of a high dimensional complex model where interactions are significant and will provide a basis for the proposed experimentation. Assume an imminent conflict will occur between two opposing forces, denoted as Blue and Red. Table 1 shows the factors, each a variable capable of multiple levels, which can be adjusted and will serve as an excellent source for experimentation. Archimedes, an agent-based simulation being developed under the Marine Corps Combat Development Command's Project Albert, contains a much larger factor set than Hencke's model and may serve as an excellent model to further explore a proposed search strategy.

Table 1: Factors for Experimentation

Red Forces	**Blue Forces**
Red agents in cell	Blue agents in cell
box center x	box center x
box center y	box center y
box size x	box size x
box size y	box size y
goal x	goal x
goal y	goal y
probability hit	probability hit
speed	Speed
sensor/shoot range	sensor/shoot range
charge ratio	charge ratio
runaway ratio	runaway ratio
maximum hits	maximum hits

The goal of this research is to determine a methodology and associated theory that has the ability to efficiently search across the breadth of factors to identify not only main effects, but also critical second-order (and perhaps third-order) interactions. A sequential design, where the number of observations may not be known in advance, which combines traditional full and fractional factorial designs, Latin supercube, and supersaturated designs together with group screening, prior information, and expert judgment may provide an effective approach to efficiently identify not only main effects, but also appreciable interactions. Furthermore, once an appropriate set of factors and their interactions have been identified, a perturbation approach to judge the stability of the proposed set will be integrated into the search methodology.

REFERENCES

Bates, R.A., Buck, R.J., Riccomagno, E., and Wynn, H.P., "Experimental Design and Observation for Large Systems," *Journal Royal Statistical Society* 58, No.1, pp. 77-94, 1996.

Bettonvil, B. and Kleijnen, J.P.C., "Searching for important factors in simulation models with many factors: sequential bifurcation," *European Journal of Operations Research* 96, pp. 180-194, 1996.

Biles, W.E., "Experimental Design in Computer Simulation," *1979 Winter Simulation Conference*, pp. 3-9, 1979.

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

Dewar, J.A., Gillogly, J.J., and Juncosa, M.L., "Non-Monotonicity, Chaos, and Combat Models", Final Report, RAND Corporation, Santa Monica, CA, 1991.

Dorfman, R., "The detection of defective numbers of large populations," *The Annals of Mathematical Statistics* 14, pp. 436-440, 1943.

Hunter, J.S. and Naylor, T.H., "Experimental Designs for Computer Simulation Experiments," *Management Science*, Vol. 16, No. 7, pp. 422-434, 1970.

Jacoby, J.E. and Harrison, S., "Multi-variable experimentation and simulation models," *Naval Research Logistics Quarterly* 9, pp. 121-136, 1962.

Hencke, R.B., "An Agent-Based Approach to Analyzing Information and Coordination in Combat," Master's Thesis, United States Naval Postgraduate School, Monterey, CA, September 1998.

Hicks, C.R., *Fundamental Concepts in the Design of Experiments*, Saunders College Publishing, New York, 1993.

Kleijnen, J.P.C., "Screening designs for poly-factor experimentation," *Technometrics* 17, pp. 487-493, 1975.

Li, C., "A sequential method for screening experimental variables," *Journal of the American Statistical Association* 57, pp. 455-477, 1962.

Mauro, C.A., "Efficient Identification of Important Factors in Large Scale Simulations," *Proceedings 1986 Winter Simulation Conference*, pp. 296-305, 1986.

Montgomery, D.C., "Methods for Factor Screening in Computer Simulation Experiments," Final Report, Office of Naval Research, Contract N0014-78-C-0312, Georgia Institute of Technology, Atlanta, GA, 1979.

Ott, E.R. and Wehrfritz, F.W., "A special screening program for many treatments," *Statistica Neerlandica*, Vol. 26, pp. 165-170, 1972.

Owen, A.B., “Latin Supercube Sampling for Very High-Dimensional Simulations,” *ACM Transactions on Modeling and Computer Simulation*, Vol. 8, No. 1, pp. 71-102, 1998.

Schruben, L.W., “Simulation optimization using frequency domain methods,” *Proceedings 1986 Winter Simulation Conference*, pp. 396-369, 1986.

Smith, D.E. and Mauro, C.A., “Factor Screening in Computer Simulation,” *Simulation* 38, pp. 49-54, 1982.

Watson, G.S., “A Study of the Group Screening Method,” *Technometrics* 3, pp. 371-388, 1961.

Welch, W.J, Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J., and Morris, M.D., “Screening, Predicting, and Computer Experiments,” *Technometrics*, Vol. 34, No. 1, pp. 15-25, 1992.

Yamada, S. and Lin, D.K.J., “Three-level supersaturated designs,” *Statistics and Probability Letters* 45, pp. 31-39, 1999.

Statistical Analysis of Atmospheric Properties for Estimation of Infrared Radiance of Ballistic Missiles
CPT Scott Nestler, United States Military Academy

Atmospheric properties, like temperature and density, can greatly affect the amount of IR energy that is reflected off an incoming ballistic missile. While many models to predict mean atmospheric conditions exist, there are no global models that account for the variability in these properties. This shortcoming makes it difficult to assess uncertainty due to atmospheric conditions. For this reason, a model that is adjusted for known extreme values is needed for use in describing the global behavior of atmospheric parameters. This study is in support of the Missile and Space Intelligence Center's development of a Bounded Earth Atmospheric Model (BEAM). This study will attempt to create such a model through statistical analyses on an existing atmospheric model. It is expected that BEAM will primarily be used by designers of IR sensors used in missile defense systems.

CONTRIBUTED SESSION V (1330 - 1530)

Stochastic Properties for Uniformly Optimally Reliable Networks
Yontha Ath, California State University, Dominguez Hills and Milton Sobel, University of California, Santa Barbara

The field of Optimal Network Reliability seems to be a pretty hot topic among a small group of engineering people. Although it is highly statistical in nature, few statisticians have become interested in it. It has many applications, e.g. to military environments as well to local area computer networks that operate in parallel. Such a network is modeled by a probabilistic graph G in which the points or nodes (i.e., communication centers, satellites, telephones, missile sites or computer stations) are perfectly reliable but the edges (representing telephones lines, communication links such as microwaves or multiply-connected cables) operate independently of one another, each with a given common known probability $p(0<p<1)$. The network G is in an operating state iff the surviving edges induce (contain or consist of) a spanning connected subgraph of the given graph G. We consider only the all-terminal reliability. For an undirected network with n nodes and e edges, what is the most reliable network for the given pair (n,e)?It is quite interesting that a unique optimal graph does exist for many of the pairs (n,e) studied thus far. In this talk, we present several new conjectures for a graph to be uniformly optimally reliable (UOR). We also present some new counterexamples to the known conjectures on UOR graphs in the literature. Finally, we use Dirichlet methodology and apply it to the above network problems. For example, we consider different models for the lifetime of the edges and we are interested in the Expectation and Variance of the resulting (random) time until the (all-terminal communication) system gets severed (i.e., the network gets disconnected).

Reliability Described by Belief Functions, A Second Look
George Hanna, Army Materiel Systems Analysis Activity

Glenn Shafer and A. M. Breipohl published the paper Reliability Described by Belief Functions in the 1979 Proceedings of the Annual Reliability and Maintainability Symposium of the IEEE. The paper uses Dempster-Shafer belief functions to assess the reliability of a circuit based on indirect data. The weakest aspect of the method is the construction of the belief functions. The belief functions are assumed to be consonant and one-sided, i.e., non-zero belief is associated only with statements of the form the reliability exceeds a particular value. Surely there will be occasions when the evidence provides support for two-sided statements about reliability. Furthermore, the assignment of confidence levels to belief values though not objectionable is not particularly compelling.

Since the time of the paper, two developments provide value to a reconsideration of Dempster-Shafer belief functions for assessment of reliability. First, is the cost to the Army of obtaining direct data to support verification of reliability requirements for weapons under development. Some items, such as missiles, are inherently expensive to test. For other items money available for testing may be limited due to general reduction in Army funding. Even when the Army can afford direct verification tests of a developmental system there is a strong desire to reduce the cost of verification. Consequently, there is high level interest in new methods to assess both reliability and other performance characteristics using other than direct system test data. In fact the Army has sponsored development of a methodology putatively based on Dempster-Shafer theory to assess reliability using other than direct data. The second development has been one of the reinterpretations of the Dempster-Shafer theory. Specifically the interpretation described in the book A Mathematical Theory of Hints: An approach to the Dempster-Shafer theory of evidence written by Jurg Kohlas and Paul-Andre Monney and published in 1995. The Kohlas and Monney approach to statistical inference employs a process model to relate an observable outcome to an unobservable random variable and a parameter of interest. Process models lead naturally to Dempster-Shafer belief functions by providing a literal meaning to the concept of a basic probability assignment that was introduced in Shafer's book, a mathematical theory of evidence and apparently named because of its formal properties.

The presentation will introduce the process model approach to developing belief functions. The belief function for binomial data based on a stress-strength process model will be discussed in detail. A simple model relating indirect and direct data will be used to illustrate evidentiary combination. The final part of the presentation will consider some of the obstacles to the use of process models and Dempster-Shafer theory for reliability assessment.

Damage Assessment Using Test Data and Expert Testimonies
Yuling Cui and Nozer D. Singpurwalla, George Washington University

This talk lays out a general framework for damage assessment using test data and expert testimonies. Here test data are binary, while expert testimonies arise either in the form of informed judgments, or science based models, or both. Our proposed approach is "normative", in the sense that it is based entirely on the calculus of probability. This is ongoing work.

System Reliability for Precision Missilery
Mike Danesh, Aviation and Missile Research, Development and Engineering Center

This paper presents a top-level summary for the development of a complex missile system. It is based upon a systematic approach that provides insight into the integration of scientific, engineering and mathematical technologies necessary for development of a robust missile system. It has been developed based upon a decade of experience with the PATRIOT system and has been endorsed by established experts.

As part of this talk, a mathematical tool called ExCAP (Confidence Expanded Assessment Package) will also be presented. ExCAP expands on conventional assessment methods by adding capabilities developed in the field of evidential reasoning.

SPECIAL SESSION ON RELIABILITY (1600 - 1730)

NRC Workshop on Reliability for DoD Systems--An Overview of the Statistical Content
Francisco Samaniego, University of California, Davis

The National Research Council of the National Academy of Sciences hosted a DoD-sponsored workshop on Reliability in June, 2000. Invited speakers addressed issues in the areas of system design and performance monitoring, reliability growth, techniques for combining data from related experiments, approaches to the analysis of field-performance data, fatigue modeling and related inferences, software reliability issues, and reliability economics. From my perspective as Chair of the Organizing Committee and of the workshop itself, I will discuss the highlights among the statistical ideas, approaches, visions and recommendations presented at the workshop. Some of their implications for the development of a new "Reliability, Availability and Maintainability (RAM) Primer" will also be discussed.

NRC Workshop on Reliability for DoD Systems--A DoD Perspective
Ernest Seglie, Office of the Secretary of Defense, Operational Test & Evaluation

In 1998 National Research Council issued a study of Statistics, Testing and Defense Acquisition. It recommended that, "The Department of Defense and the military services should give increased attention to their reliability, availability, and maintainability data collection and analysis procedures because deficiencies continue to be responsible for

many of the current field problems and concerns about military readiness." (Recommendation 7.1,Page 105)

In response to this recommendation and other more pointed observations on the state of current practice in DoD compared to industry best practices, the Department asked the National Academies to host a Workshop on Reliability. A second workshop on software intensive systems will be held next year. The first workshop was held 9 and 10 June 2000 at the National Academy in Washington. Speakers and discussants were from industry, academia, and the Department of Defense. The suggestions from the workshop will be part of an effort to implement the recommendation to produce " new battery of military handbooks containing a modern treatment of all pertinent topics in the fields of reliability and life testing" (Recommendation 7.12,Page126). DoD Handbook 5235.1-H, Test and Evaluation of System Reliability, Availability, and Maintainability: A Primer was last updated in 1982.

Implications of this effort to providing reliable systems for the warfighter will be addressed.

FRIDAY, OCTOBER 20

GENERAL SESSION III (0830 - 1000)

September 2000

QUANTILE/QUARTILE PLOTS,CONDITIONAL QUANTILES,COMPARISON DISTRIBUTIONS

Emanuel Parzen
Texas A&M University, Department of Statistics

1. Histograms and Order Statistics

Histograms are traditionally plotted by statisticians to identify distributions that fit a sample (data set) $Y_1, \ldots, Y_n$.

Order statistics (sample values arranged in increasing order are denoted $Y(1;n) \leq \cdots \leq Y(n;n)$.

2. Sample Quantile Function

For identifying distributions that fit data we recommend the sample quantile function

$$Q^{\sim}(u) = F^{\sim^{-1}}(u) = Y(j;n), \quad (j-1)/n < u \leq j/n,$$

inverse of the sample distribution function

$$F^{\sim}(y) = E^{\sim}[I(Y \leq y)] = (1/n)\Sigma_{t=1}^{n}\, I(Y_t \leq y)$$

3. Population Ensemble Quantile Function

When we regard $Y_1, \ldots, Y_n$ as a random sample of Y, we define population distribution function $F(y)$, $-\infty < y < \infty$, and quantile function $Q(u)$, $0 \leq u \leq 1$,

$$F(y) = P[Y \leq y] = E[I(Y \leq y)]$$
$$Q(u) = F^{-1}(u) = \inf\{y : F(y) \geq u\}$$

4. Density Quantile, Quantile Density

If F is continuous, $F(Q(u)) = u$ for all u and (differentiating)

$$f(Q(u))\, Q'(u) = 1.$$

Quantile density $q(u) = Q'(u)$.
Density quantile $fQ(u) = f(Q(u))$.

Asymptotic distribution of sample quantiles which makes statistical inference possible is given by

$$\sqrt{n}\; fQ(u)(Q^{\sim}(u) - Q(u)) \rightarrow_d \; B(u)$$

where $B(u)$, $0 < u < 1$, is Brownian Bridge, zero mean Gaussian process with covariance $E[B(s)B(t)] = \min(s,t) - st$.

Analogous limit theorems can be proved for estimators of conditional quantile functions $Q_{Y|X=x}(u)$.

5.Tail Classification of Distributions

Tail classification of distribution functions can be described by exponents of regular variation α_0 and α_1:

$$fQ(u) = u^{\alpha_0}\ L_0(u), \qquad u \text{ near } 0,$$
$$fQ(u) = (1-u)^{\alpha_1}\ L_1(u), \quad u \text{ near } 0.$$

In terms of α we define

$$\begin{aligned} \alpha > 1,&\ \text{long tail;} \\ \alpha = 1,&\ \text{medium tail;} \\ 0 \leq \alpha = 1,&\ \text{short tail;} \\ \alpha < 0,&\ \text{infinitely short tail.} \end{aligned}$$

6.Continuous Sample Quantile

In practice we prefer continuous sample quantile

$$Q^{\sim c}((j-.5)/n) = Y(j;n),\ j = 1, \ldots, n,$$
$$Q^{\sim c}(0) = Y(1;n),\ Q^{\sim c}(1) = Y(n;n)$$

and defined by linear interpolation at other u values. The continuous quantile command in Splus is related to our definition by

$$Q^{\sim c}(p) = Q^{\sim c}_{\text{Splus}}(p + ((p-.5)/(n-1)))$$

Note $Q^{\sim c}_{\text{Splus}}((j-1)/(n-1)) = Y(j;n),\ j = 1, \ldots, n.$

7.Quantile Function of Transform

Assume $Y = g(W)$ where $g(w)$ is quantile-like (is non-decreasing and continuous from the left).

$$F_Y(y) = F_W(g^{-1}(y))$$
$$Q_Y(u) = g(Q_W(u))$$
$$Q_{Y|X=x}(u) = g(Q_{W|X=x}(u))$$

8.Distribution Transform

When F is continuous, $F_Y(Y)$ is Uniform $(0,1)$ since

$$Q_{F_Y(Y)}(u) = F_Y(Q_Y(u)) = u$$

We call $F_Y(Y)$ distribution transform or probability integral transform. More important is mid-distribution transform $F^{\text{mid}}_Y(Y)$, defining mid-distribution function

$$F^{\text{mid}}_Y(Y) = F_Y(Y) - .5p_Y(Y).$$

Randomized distribution is

$$F^{\text{rand}}_Y(y) = F_Y(y) - U(y)p_Y(y), \quad U(y) \text{ Uniform } (0,1).$$

9.Conditional Quantile Function

Bivariate data (X, Y) is often modeled by conditional mean $E[Y \mid X = x]$. To model conditional distribution

$$F_{Y|X=x}(y) = P[Y \leq y \mid X = x]$$

we recommend estimating conditional quantile function

$$Q_{Y|X=x}(u) = \inf \{y : F_{Y|X=x}(y) \geq u\}.$$

For jointly normal (X, Y) conditional distribution of Y given $X = x$ is normal,

$$Q_{Y|X=x}(u) = \mu_{Y|X=x} + \sigma_{Y|X=x}\Phi^{-1}(u).$$

10.Bayes Theorem Conditional Quantile Function

It is not true that $Q_Y(F_Y(y)) = y$ for all y, but for almost all values of the random variable Y

$$Q_Y(F_Y(Y)) = Y$$

Therefore by formula for quantile function of a tranform

$$Q_{Y|X=x}(u) = Q_Y(Q_{F_Y(Y)|X=x}(u))$$

We have reduced estimating the conditional quantile of Y to estimating the conditional quantile of the distribution transfrom $F_Y(Y)$ which we next interpret as a comparison distribution and estimate as a comparison density. One can apply this formula to Bayes estimation of a parameter θ from data X.

11.Comparison Distribution, Comparison Density

To compare two distributions F and G, a universal problem of statistical inference, we define concepts of comparison distribution $D(u; F, G)$, $0 \leq u \leq 1$, and comparison density

$$d(u; F, G) = D'(u; F, G).$$

When F and G are continuous, and $G << F$ ($f(y) = 0$ implies $g(y) = 0$),

$$D(u; F, G) = G(F^{-1}(u)),$$
$$d(u; F, G) = g(F^{-1}(u))/f(F^{-1}(u))$$

When F and G are discrete, and $G << F$ (probability mass function $p_F(y) = 0$ implies $p_G(y) = 0$),

$$d(u; F, G) = p_G(F^{-1}(u)/p_F(F^{-1}(u)),$$

$$D(u; F, G) = \int_0^u d(s; F, G)\, ds$$

12. Exact u, PP plots

We call u F exact if u is in the range of F, $u = F(y)$ for some y. Then $Q(u) = \inf\{y : F(y) = u\}$, and $F(Q(u)) = u$. For u F exact

$$D(u; F, G) = G(F^{-1}(u));$$

for other u, $D(u; F, G)$ is defined by linear interpolation between its value at exact u. Change comparison distribution

$$\text{Change } D(u; F, G) = G(F^{-1}(u)) - F(F^{-1}(u)).$$

Graph of D, called PP plot, connects linearly

$$(0, 0),\ (F(y_j), G(y_j)),\ (1, 1)$$

where y_j are jump points of F (assume F discrete).

13. Comparison Approach to Estimating Conditional Quantiles

Bayes' theorem for conditional quantiles is written

$$Q_{Y|X=x}(u) = Q_Y(Q_{F_Y(Y)|X=x}(u)) = Q_Y(s)$$

We propose to find

$$s = Q_{F_Y(Y)|X=x}(u)$$

by computing the function of s, $0 < s < 1$,

$$u = F_{F_Y(Y)|X=x}(s) = P\left[F_Y(Y) \le s \mid X = x\right]$$

When Y is continuous

$$u = F_{Y|X=x}(Q_Y(s)) = D(s; F_Y, F_{Y|X=x})$$

When Y is discrete and s is F_Y exact

$$F_Y(Y) \le s = F_Y(Q_Y(s)) \text{ if } fY \le Q_Y(s),$$

$$u = P\left[F_Y(Y) \le s \mid X = x\right] = P\left[Y \le Q_Y(s) \mid X = x\right] = D(s;\ F_Y,\ F_{Y|X=x})$$

14. Bayes Comparison Theorem for Conditional Quantiles

We can verify the following extension of Bayes theorem for conditional quantile functions:

$$Q_{Y|X=x}(u) = Q_Y(D^{-1}(u; F_Y, F_{Y|X=x}))$$

Proof for Y discrete:

For discrete Y with ordered values $y_j, Q_Y(s) = y_j$ for s in interval $F_Y(y_{j-1}) < s \le F_Y(y_j)$.

For u in interval $F_{Y|X=x}(y_{j-1}) < u \le F_{Y|X=x}(y_j), s = D^{-1}(u; F_Y, F_{Y|X=x})$ varies linearly between $F(y_{j-1})$ and $F(y_j)$, and $Q_Y(s) = y_j$.

For interval $F_{Y|X=x}(y_{j-1}) < u \le F_{Y|X=x}(y_j), Q_{Y|X=x}(u) = y_j = Q_Y(s)$.

15. Formula for Conditional Mean From Conditional Quantile Function

Our computational process for computing conditional quantiles involves estimating in succession as a function of x

$$\begin{aligned} d(s; F_Y, F_{Y|X=x}), &\quad 0 < s < 1 \\ u = D(s; F_Y, F_{Y|X=x}), &\quad 0 < s < 1 \\ s = D(u; F_{Y|X=x}, F_Y), &\quad 0 < u < 1 \\ Q_{Y|X=x}(u) = Q_Y(s), &\quad 0 < u < 1 \end{aligned}$$

The relation between the above functions is illustrated by formulas for the conditional mean.

Quantile formulas for conditional mean. When Y continuous

$$\begin{aligned} E[Y \mid X = x] &= \int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy \\ &= \int_{-\infty}^{\infty} y \frac{f_{Y|X=x}(y)}{f_Y(y)} dF_Y(y), y = Q_Y(s) \\ &= \int_0^1 Q_Y(s) \frac{f_{Y|X=x}(Q_Y(s))}{f_Y(Q_Y(s))} ds \\ &= \int_0^1 Q_Y(s) d(s; F_Y, F_{Y|X=x}) ds \\ &= \int_0^1 Q_Y(s) dD(s; F_Y, F_{Y|X=x}), u = D(s; F_Y, F_{Y|X=x}) \\ &= \int_0^1 Q_Y(D^{-1}(u; F_Y, F_{Y|X=x})) du \\ &= \int_0^1 Q_{Y|X=x}(u) du \end{aligned}$$

When Y discrete, similiar formulas start with

$$E[Y \mid X = x] = \sum_y y p_{Y|X=x}(y).$$

16. Logistic Regression Estimation of Conditional Comparison Density

Let $s_j = j/m, j = 0, 1, \ldots, m$ (often $m = 20$). Approximately

$$\begin{aligned} &d(s_j; F_Y, F_{Y|X=x}) \\ &= (s_j - s_{j-1})^{-1} P[Q_Y(s_{j-1}) < Y \leq Q_Y(s_j) \mid X = x] \end{aligned}$$

which can be estimated for fixed s_j as a function of x by logistic regression (for which there exists many parametric and non-parametric methods).

17. Rejection Sampling Computation of Conditional Quantile of Distribution Transform

Combining these estimates as a function of x for fixed s_j one can form a piecewise constant estimate as a function of s for fixed x.

From the comparison density $d(s; F_Y, F_{Y|X=x}), 0 < s < 1$ one can estimate by rejection sampling the values of the comparison quantile

$$s = D^{-1}(u; F_Y, F_{Y|X=x}), 0 < u < 1.$$

18. Mid-Distribution Transform

A unifying role in non-parametric data analysis is played by the mid-distribution transform (equivalent to tied ranks of a sample)

$$\begin{aligned} F_Y^{\text{mid}}(Y) &= F_Y(Y) - .5p_Y(Y) \\ E\left[F_Y^{\text{mid}}(Y)\right] &= .5, \\ \text{VAR}\left[F_Y^{\text{mid}}(Y)\right] &= (1/12)(1 - E\left[p_Y^2(Y)\right]). \end{aligned}$$

The importance of $F_Y^{\text{mid}}(Y)$ in non-parametric statistical data analysis is illustrated by the formula for a score statistic to test $H_0 : F_{Y|X=x} = F_Y$:

$$T(J) = \int_0^1 J(u)d(u; F_Y, F_{Y|X=x})du.$$

When Y is discrete, with distinct values $y_1, \ldots, y_k$,

$$T(J) = \Sigma_{j=1}^k (p_{Y|X=x}(y_j)/p_Y(y_j)) \int_{F_Y(y_{j-1})}^{F_Y(y_j)} J(u)du.$$

Approximately

$$\begin{aligned} T(J) &= \Sigma_{j=1}^k p_{Y|X=x}(y_j) J(F_Y^{\text{mid}}(y_j)) \\ &= E\left[J(F_Y^{\text{mid}}(Y)) \mid X = x\right]. \end{aligned}$$

Linear rank statistics can be represented

$$T^{\sim}(J) = E^{\sim}\left[J(F_Y^{\sim \text{mid}}(Y)) \mid X = x\right].$$

19. Location Scale Quantile Models

The sample quantile $Q^{\sim}(u)$ is a non-parametric estimator of the population quantile $Q(u)$. A parametric estimator can be formed from a location-scale model

$$Q(u) = \mu + \sigma Q_0(u)$$

where $Q_0(u)$ is known.

Maximum likelihood estimators of μ and σ, denoted $\hat{\mu}$ and $\hat{\sigma}$, yield estimator $\hat{Q}(u) = \hat{\mu} + \hat{\sigma}Q_0(u)$.

Asymptotically efficient estimates, denoted $\mu_{n,L}$ and $\sigma_{n,L}$, can be formed as a linear functional of order statistics (sample quantile function); they can be computed by continuous parameter regression analysis from the asymptotic representation of the sample quantile as a linear regression

$$f_0 Q_0(u) Q^{\sim}(u) = \mu f_0 Q_0(u) + \sigma f_0 Q_0(u) Q_0(u) + \sigma B(u).$$

20. Location Scale Models for Conditional Quantiles

$$Q_{Y|X=x}(u) = \mu_{Y|X=x} + \sigma_{Y|X=x} Q_0(u)$$

Estimators

$$\hat{Q}_{Y|X=x}(u) = \hat{\mu}_{Y|X=x} + \hat{\sigma}_{Y|X=x} Q_0(u)$$

can be formed in the same way as in the unconditional case from linear functions of our non-parametric estimators denoted $Q^{\sim}_{Y|X=x}(u)$, of the conditional quantile $Q_{Y|X=x}(u)$.

To compute these parametric estimators we assume $Q_0(u)$ known. We can compare several choices of $Q_0(u)$ by plotting $Q^{\sim}_{Y|X=x}(u)$ and $Q^{\wedge}_{Y|X=x}(u)$ on scatter diagrams.

Estimation of $\mu_{|X=x}$ is alternative to non-parametrically estimating $E[Y \mid X = x]$.

21. Confidence Intervals for $Q_Y(u)$ and Conditional Quantile $Q_{Y|X=x}(u)$

Confidence intervals for a parameter $Q(u)$ can be formed from the asymptotic distribution of $Q^{\sim}(u)$:

$$\sqrt{n}(Q^{\sim}(u) - Q(u)) \rightarrow_d B(u)/fQ(u)$$

This formula has a severe disadvantage , it requires estimation of $fQ(u)$.

From the values of the sample quantile functions $Q^{\sim}_Y(u)$ (Similarly for $Q^{\sim}_{Y|X=x}(u)$) one can obtain a confidence interval for $Q(u)$ using facts such as

$$\sqrt{n}(F^{\sim}_n(Q(u)) - u) \rightarrow_d B(u)$$

One can find functions $c_1(u)$ and $c_2(u)$ such that with probability greater than α, for all u,

$$u - (c_1(u)/\sqrt{n}) < F^{\sim}_n(Q(u)) < u + (c_2(u)/\sqrt{n}),$$
$$Q^{\sim}_n(u - (c_1(u)/\sqrt{n})) < Q(u) < Q^{\sim}_n(u + (c_2(u)/\sqrt{n}))$$

From parametric estimates of a location-scale model

$$\hat{Q}(u) = \hat{\mu} + \hat{\sigma} Q_0(u)$$

or

$$\hat{Q}(u) = \mu_{n,L} + \sigma_{n,L} Q_0(u)$$

one can derive a simultaneous confidence interval (Rosenkrantz (2000))

$$\hat{Q}(u) - c_1(u,n) \leq Q(u) \leq \hat{Q}(u) + c_2(u,n)$$

This location-scale model confidence interval is shorter than the non-parametric confidence intervals above.

22. Five Number Quantile Summary

Quantile function can "compress data" by a five number summary: values of $Q(u)$ at

$$u = .05, .25, .5, .75, .95$$

Median $Q(.5)$
Quartiles $Q(.25),\ Q(.75)$
Mid Quartile $QM = .5(Q(.25) + Q(.75))$
Quartile deviation $QD = 2(Q(.75) - Q(.25))$
QD approximation to $Q(.5)$
Normal distribution $QD = 2.7\sigma$.

23. Quantile Quartile Function $Q/Q(u), 0 < u < 1$

To use quantile functions to identify distributions fitting data we propose

$$Q/Q(u) = (Q(u) - QM)/QD = Q_{Y^Q}(u),$$

defining transform of data

$$Y^Q = (Y - QM)/QD$$

Claim: plot of $y = Q/Q(u)$ contains all the insights of a box plot. Add to plot dotted lines

$$\begin{array}{ll} \text{horizontal} & y = -1, -.5, 0, .5, 1 \\ \text{vertical} & u = .05, .25, .5, .75, .95 \end{array}$$

Five number summary of distribution becomes

$$\begin{array}{rl} QM, & \text{location} \\ QD, & \text{scale} \\ Q/Q(.5), & \text{skewness} \\ Q/Q(.05), & \text{righttail} \\ Q/Q(.95), & \text{lefttail} \end{array}$$

Elegance of $Q/Q(u)$ is its universal values

$$Q/Q(.25) = -.25$$
$$Q/Q(.75) = .25$$

Tukey outliers correspond to $\mid Q/Q(u) \mid > 1$.
Tukey criteria for outlier value $Q(u)$ outside fences:

$$Q(u) > Q(.75) + 1.5(Q(.75) - Q(.25)), Q(u) - QM > QD,$$
$$Q(u) < Q(.25) - 1.5(Q(.75) - Q(.25)), Q(u) - QM < QD.$$

Diagnostics of left and right tail behavior:

$$\begin{array}{rll} \text{Short}: & -5 < Q/Q(.05) & Q/Q(.95) < .5 \\ \text{Medium}: & -1 < Q/Q(.05) < -.5 & .5 < Q/Q(.95) < 1 \\ \text{Long}: & Q/Q(.05) < -1 & 1 < Q/Q(.95). \end{array}$$

Diagnostics of skewness:

$$\begin{array}{lll} Q/Q(.5) > 0, & \text{mean} < \text{median} < \text{mode}, & \text{left} - \text{skewed} \\ Q/Q(.5) < 0, & \text{mode} < \text{median} < \text{mean}, & \text{right} - \text{skewed}. \end{array}$$

24. Conditional Quantile Five Number Summary

From the conditional quantile $Q_{Y|X=x}(u)$ at $u = .05, .25, .5, .75, .9$, we recommend five number summary:
location of conditional distribution

$$QM_{Y^Q|X=x} = (QM_{Y|X=x} - QM_Y)/QD_Y$$

scale of conditional distribution

$$QD_{Y^Q|X=x} = QD_{Y|X=x}/QD_Y$$

skewness of conditional distribution

$$Q/Q_{Y|X=x}(.5) = (Q_{Y|X=x}(.5) - QM_{Y|X=x})/QD_{Y|X=x}$$

right tail of conditional distribution

$$\begin{aligned} &Q/Q_{Y|X=x}(.95) = \\ &(Q_{Y|X=x}(.95) - QM)_{Y|X=x})/QD_{Y|X=x} \end{aligned}$$

$Q/Q_{Y|X=x}(.05)$ defined similarly.

25. Conditional Quantile Scatter Plots, Diagnostic Plots

We recommend plotting as a function of x for $u = .05, .25, .5, .75, .95$

$$y = Q_{Y^Q|X=x}(u)$$

on a scatterdiagram of (X^Q, Y^Q). Also plot $y = QM_{Y|X=x}$. On separate graphs, plot as function of x, $QD_{Y^Q|X=x}$, $Q/Q_{Y|X=x}(.05), Q/Q_{Y|X=x}(.95)$.

26. Unification of Conventional Statistical Methods

Conventional t test, Kruskal-Wallis test, goodness of fit) methods of two sample and multi-sample data analysis can be extended and unified by representing the data as (X, Y) and computing and graphing

conditional quantiles of Y given $X = x$,

conditional comparison quantiles of $F_Y^{\text{mid}}(Y)$,

conditional comparison distribution of $F_Y^{\text{mid}}(Y)$.

These are functions of u for each of the discrete x values of X. We plot summaries as a function of x plotted at $F_X^{\text{mid}}(x)$.

To summarize data in many samples we represent (X_j, Y_j) where X_j denotes population (denoted $k = 1, \ldots, c$) from which response Y_j was observed. The values of Y for a fixed value X can be summarized by a conditional quantile function $Q_{Y|X}{}^{\sim}(u)$. The pooled sample of Y values is summarized by an unconditional quantile $Q_Y{}^{\sim}(u)$. As an alternative to running box plots to compare and summarize the different samples we propose conditional quantile/quartile plot which connects linearly points $(F_X^{mid}(k),\ Q/Q_{Y|X=k}(u)),\ k = 1, \ldots, c$. One plots this curve for $u = .05, .25, .5, .75, .95$. One also plots constant lines $Q/Q_Y(u),\ u = .05, .25, .5,\ .75, .95$.

References

Gilchrist, Warren (2000). *Statistical Modeling with Quantile Functions*, Chapman and Hall/CRC Press: Boca Raton.

Handcock, Mark and Martina Morris (1999). *Relative Distribution Methods in the Social Sciences*, Springer: New York.

Heckman, Nancy and R.H. Zamar (2000). "Comparing the Shapes of Regression Functions," *Biometrika*, 87, 135-144.

Parzen, Emanuel (1977). Discussion to Stone (1977).

Parzen, Emanuel (1979). "Nonparametric Statistical Data Modeling," *Journal of the American Statistical Association,* (with discussion), 74, 105-131.

Parzen, Emanuel (1989). "Multi-Sample Functional Statistical Data Analysis," *Statistical Data Analysis and Inference Conference in Honor of C.R. Rao,* (ed. Y. Dodge), Amsterdam: Elsevier, 71-84.

Parzen, Emanuel (1991). "Unification of Statistical Methods for Continuous and Discrete Data," *Proceedings Computer Science-Statistics INTERFACE '90*, (ed. C. Page and R. LePage), Springer Verlag: New York: 235-242.

Parzen, Emanuel (1992). "Comparison Change Analysis," *Nonparametric Statistics and Related Topics* (ed. A.K. Saleh), Elsevier: Amsterdam, 3-15.

Parzen, Emanuel (1993). "Change PP Plot and Continuous Sample Quantile Function," *Communications in Statistics,* 22, 3287-3304.

Parzen, Emanuel (1996). "Concrete Statistics," *Statistics in Quality,* S. Ghosh, W. Schucany, W. Smith, Marcel Dekker: New York, 309-332.

Parzen, Emanuel (1999). "Statistical Methods Mining, Two Sample Data Analysis, Comparison Distributions, and Quantile Limit Theorems," *Asymptotic Methods in Probability and Statistics* (ed. B. Szyszkowicz), Elsevier: Amsterdam.

Rosenkrantz, Walter (2000). "Confidence Bands for Quantile Functions: A Parametric and Graphic Alternative for Testing Goodness of Fit," *The American Statistician,* 54, 185-190.

Stone, Charles (1977). "Consistent Nonparametric Regression," *Annals of Statistics,* 5, 595-645.

Innovative Bayesian Designs in Clinical Trials
Donald Berry (University of Texas, MD Anderson Cancer Center)

I will give some background on Bayesian designs for clinical trials and four examples of Bayesian designs used in actual trials. These examples embody the principles of early stopping, allocating treatments to maximize benefit to patients in and out of the trial, and two variations on a theme (a drug for stroke and another for non-small cell lung cancer): seamless phases II and III with sequential sampling and using surrogate endpoints.

CONTRIBUTED SESSION VI (1015 - 1115)

Modeling of Tank Gun Accuracy Under Two Different Zeroing Methods

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

Bruce J. Held
RAND
Santa Monica, CA

Sixth Army Conference on Applied Statistics
Rice University, Houston, TX
20 October 2000

The accuracy of weapon systems firing unguided ammunition, such as most tank and artillery systems, relies on understanding numerous sources of error and correcting for them. Random errors are addressed through weapon and ammunition design, quality production, effective maintenance, and firing processes that minimize the magnitude of these error sources. Bias errors are different because their magnitude and direction can often be estimated and then compensated for while aiming the weapon. In this paper, we examine two common methods used to compensate for the bias errors inherent in aiming and firing tank cannons. There are a number of advantages and disadvantages to each of these processes, but the relative accuracy benefits depend on the magnitude of the error sources that make up the total error budget for the fleet of tanks. Although this budget is comprised of numerous error sources, reasonable estimates of accuracy are attainable with basic models of the predominant factors. This presentation illustrates models for first-shot accuracy under two common zeroing processes. Such a comparison can then be used to determine the policy that makes a particular fleet of tanks most accurate in varying tactical situations and can be a useful tool for identifying and reducing the actual source of bias errors.

In tanks, estimating the bias error is a two-part process. First, boresighting is a process that measures the offset between the tank's fire control system and its cannon, and then corrects for that offset. Typically, boresighting is accomplished by placing an optical device along the centerline of a cannon. The cannon is then moved so that a target point, at a known distance from the tank, is viewed with the aid of the optical device. Assuming that there is no error in the optics of the device and that the device can be placed directly along the cannon centerline, the cannon will be aimed at the target point. Of course there are errors, but modern boresighting devices are relatively accurate.[1] Once the cannon is aimed at the target point, the tank's fire control optics are also aimed at the target point. Since the tank-to-target distance is known and the various angles of the cannon and fire control optics can be measured, the boresighting process provides enough geometrical information to aim the tank's cannon at a target, at any range, with the tank's fire control optics.

While boresighting allows accurate alignment of the cannon with respect to the target prior to firing the cannon, the cannon can bend, displace laterally, and rotate during the actual firing process. The result is that by the time the projectile exits the cannon muzzle, the cannon is no longer aimed so precisely at the target. Additional mechanical interactions and aerodynamics further alter the eventual trajectory of the projectile. The sum of the changes between the predicted trajectory (as based on the statically aimed cannon) and the actual trajectory is often referred to as "jump". Correcting for jump, which normally has a large, nonrandom component, is called zeroing. Zeroing is the second step in the two-part process of estimating and compensating for bias error. It can be accomplished a number of ways; but the two most common in tank gunnery are individual zero and fleet zero. Tanks are individually zeroed when their jump is individually estimated, and a unique correction is calculated and applied to that tank. Fleet zero is a process by which an average jump value is estimated for a number (fleet) of tanks and the correction implied by that average jump is applied to every tank in the fleet.

Currently, when tanks are individually zeroed, it is done manually. In other words, the tanks are taken to a firing range, where they shoot a number of rounds at a target, the impact points are measured relative to the aim point, and a correction is calculated and applied based on the average miss distance and direction. Since each ammunition type reacts differently, the tank must be individually zeroed with each ammunition type it will fire. This method of zeroing was used in the United States into the 1980s.

---

[1] Held, B. J., and D. W. Webb. "A Comparison of Muzzle Boresights for Tank Cannon." BRL-MR-3977, U.S. Army Ballistics Research Laboratory, Aberdeen Proving Ground, MD, 1992. Prior to using dedicated boresight devices, one method of boresighting in the United States was to tape string across the muzzle of cannon to form a crosshair. Binoculars were then used to view down the cannon from the breech end, across the crosshair at the muzzle, to the target point.

However, with the advent of depleted uranium ammunition and the 120mm cannon, new calibration policies had to be considered. The new ammunition was considered environmentally unfriendly, hence there was a desire to limit its use. Additionally, the new, higher technology and often larger caliber ammunition being introduced was expensive, so finding a way to zero without firing a lot of rounds was important. This requirement led to the fleet zero technique of estimating and compensating for bias errors.

The major assumption behind the fleet zero concept is that all the tanks in a fleet shoot in a similar manner; for example, if Tank X shoots a particular ammunition type high and to the right, the assumption requires that Tank Y would shoot the same ammunition type high and to the right. If this basic assumption is reasonably accurate, then a good estimation of the bias error for Tank X will also be a good bias estimation for Tank Y. Current fleet zero polices depend on this assumption. Today, the U.S. Army has established a fleet zero computer correction factor (CCF)[*] for each fielded ammunition type. The fleet zero facilitates the introduction of new ammunition types since part of the fielding process is to establish and distribute a CCF for the new round. This is accomplished by firing the ammunition, under several firing conditions, from a number of tanks representative of the larger fleet and measuring the miss distances to the aim point. The basic data is then analyzed and generalized to a single CCF for the entire fleet for the new ammunition type. In the case of the individual zero, introduction of a new ammunition type requires each tank in the fleet to fire several of the rounds to establish CCFs unique to each tank.

Using the fleet zero has a number of other distinct advantages over an individual zero policy. The limited number of rounds required to establish a fleet CCF substantially reduces the cost compared to individually zeroing tanks. The controlled nature of the tests used to establish the fleet CCF also indicates that safety and environmental concerns associated with the firing of tactical rounds can be greatly diminished. As individually zeroing tanks requires safe firing ranges near the battlefield and the need to supply additional ammunition for zeroing, the fleet zero policy is more tactically sound since it eliminates these requirements. Finally, one of the most significant advantages fleet zero offers is the ease with which it is implemented at the tank and soldier level. Once a CCF number is passed to a tank owning unit all that is required is that the CCF be entered tank fire control computers. In contrast, individually zeroing tanks requires firing the tanks, measuring the impacts, calculating the CCF, and then entering it on the fire control computer. Each of these steps, which can produce significant error, requires training.

---

[*] The CCF is the value of the correction that is input into the fire control computer of the tank. The value is used by the fire control computer as part of the calculation that aims the cannon.

There are disadvantages to the fleet zero policy, however. The greatest disadvantage is that the fleet zero policy is only as good as the basic assumption that underlies it; hence, fleet zero does not account for differences between tanks and firing conditions. When that assumption is found wanting, tanks that shoot significantly different than the fleet average will shoot inaccurately under the fleet zero policy. Additionally, if the conditions under which a tank fires (e.g., the ammunition temperature) has a significant affect on the magnitude and direction of the bias error, a fleet zero policy may not adequately address the affect. Individually zeroing tanks obviously accounts for the firing differences between them, and if done under conditions similar to those encountered during battle or gunnery drills, will also account for ammunition bias sensitivity to environmental conditions.

Choosing the better zeroing policy for a given set of conditions requires an understanding of how those conditions affect accuracy as they vary. Since it is impractical to obtain this experimentally, a mathematical model is necessary.

Mathematical Modeling of Fleet Zero

For our modeling purposes, the three dominant components of variance associated with tank cannon accuracy are round-to-round variance, occasion-to-occasion variance, and tank-to-tank variance. These components are quantified by the parameter values σ_R^2, σ_O^2, and σ_T^2, respectively. Round-to-round errors are random differences in jump between the same type of rounds fired on the same occasion from the same tank. An occasion is defined as a total firing event during which nothing has happened to the tank that could appreciably affect how it shoots (e.g., maintenance or significant movement). In other words, the total occasion timeframe is short enough that neither the tank nor firing conditions significantly changing is assured. Occasion-to-occasion errors are therefore the random differences between firing occasions (in regard to their mean jumps) while firing the same type of ammunition from the same tank. Finally, tank-to-tank errors are the differences in mean jumps between different tanks firing the same kind of ammunition. Tank-to-tank error is only pertinent to the fleet zero method of zeroing. In addition to the three major errors, ammunition temperature is included in the model as the major known variable not otherwise accounted for in the fire control calculations of current U.S. tanks.

The three variance components and ammunition temperature are the only contributors to the total variability in the Stationary-to-Stationary Ammunition Accuracy Test (SSAAT), which is the U.S. Army

Test Center's protocol for initially estimating the CCF. As displayed in Table 1, the SSAAT calls for four randomly chosen tanks to each fire four three-round occasions. Within tanks, each occasion is fired at one of four specified ammunition temperatures. The CCF is then calculated as a weighted average of the mean impacts from the firings at the four temperature conditions. The weightings are based on the U.S. Army Materiel and Systems Analysis Activity's (AMSAA) estimated global temperature distribution for armored vehicle combat.

Tank	Temperature (^{o}F)			
	-10	30	70	110
1	Occasion 1: Rounds 1-3	Occasion 5: Rounds 13-15	Occasion 9: Rounds 25-27	Occasion 13: Rounds 37-39
2	Occasion 2: Rounds 4-6	Occasion 6: Rounds 16-18	Occasion 10: Rounds 28-30	Occasion 14: Rounds 40-42
3	Occasion 3: Rounds 7-9	Occasion 7: Rounds 19-21	Occasion 11: Rounds 31-33	Occasion 15: Rounds 43-45
4	Occasion 4: Rounds 10-12	Occasion 8: Rounds 22-24	Occasion 12: Rounds 34-36	Occasion 16: Rounds 46-48

Table 1. Design layout for the Stationary-to-Stationary Ammunition Accuracy Test (SSAAT).

Under more general conditions, a linear model for the SSAAT is given by

$$y_{ijkl} = \alpha + T_i + \beta\tau_j + O_{k(ij)} + R_{l(ijk)},$$

where

(1) y_{ijkl} is the azimuth (or elevation) jump for round l fired during occasion k from tank i at temperature j, for $i = 1, \ldots, a$; $j = 1, \ldots, b$; $k = 1, \ldots, c$; and $l = 1, \ldots, n$;

(2) α is the intercept from the temperature dependency model;

(3) T_i is the effect of tank i, assumed to be $N\left(0, \sigma_T^2\right)$;

(4) β is the slope from the temperature dependency model;

(5) τ_j is the j^{th} ammunition temperature level;

(6) $O_{k(ij)}$ is the effect of occasion *k*, nested within the combination of tank *i* and temperature *j*, assumed to be $N\left(0,\sigma_O^2\right)$; and

(7) $R_{l(ijk)}$ is the effect of round *l* from occasion *k* fired from tank *i* at temperature *j*, assumed to be $N\left(0,\sigma_R^2\right)$.

A few assumptions concerning this model deserve further discussion. First, although impact data are bivariate, the model is univariate as a result of the assumption of independence between horizontal (azimuth) and vertical (elevation) impacts. This assumption is invalid for some weapon systems, especially rapid-cadence systems; however, for large-caliber weapons in which the cadence is relatively slow, this assumption has proven consistent over many years of testing. Second, the temperature dependency model previously noted is a simple linear regression model relating jump as a function of ammunition temperature. This relationship between ammunition temperature and jump is the result of (1) the relationship between propellant temperature and in-bore velocity (hot propellant burns quicker, hence the projectile travels faster while in-bore) and (2) the dynamic and consistent motion of the gun tube immediately after shot initiation. The coefficients from this model are estimated from previous live-fire tests of this ammunition type. Over the range of ammunition temperatures in the SSAAT, a simple linear regression has shown to be a basic, yet adequate model.[2] Finally, based upon the belief that the effect of ammunition temperature is the same upon each tank, the model excludes an interaction term involving these two variables.

In the SSAAT model, the average jump at each temperature is given as $\bar{y}_{\tau_j} = \alpha + \bar{T}_\bullet + \beta\tau_j + \bar{O}_{\bullet(\bullet j)} + \bar{R}_{\bullet(\bullet j\bullet)}$, and under the random-effects assumptions, is normally distributed with mean $\alpha + \beta\tau_j$ and variance $\frac{1}{a}\sigma_T^2 + \frac{1}{ac}\sigma_O^2 + \frac{1}{acn}\sigma_R^2$. For a specific set of weights, $\{w_1,...,w_b\}$ satisfying $\sum_{j=1}^{b} w_j = 1$, the CCF is given by $CCF = \sum_{j=1}^{b} w_j \bar{y}_{\tau_j} =$ $\alpha + \beta\sum_{j=1}^{b} w_j\tau_j + \sum_{j=1}^{b} w_j\bar{T}_\bullet + \sum_{j=1}^{b} w_j\bar{O}_{\bullet(\bullet j)} + \sum_{j=1}^{b} w_j\bar{R}_{\bullet(\bullet j\bullet)}$. Therefore, the CCF is also a normally distributed random variable having mean $\alpha + \beta\sum_{j=1}^{b} w_j\tau_j$ and variance $\frac{\sigma_T^2}{a} + \frac{\sigma_O^2}{ac}\sum_{j=1}^{b} w_i^2 + \frac{\sigma_R^2}{acn}\sum_{j=1}^{b} w_i^2$.

[2] Held, B. J., D. W. Webb and E. M. Schmidt, "Temperature Dependent Jump of the 120mm M256 Tank Cannon." BRL-MR-3927, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, 1991.

After the SSAAT is completed, each tank in the fleet incorporates the CCF into its fire control system. Now, tanks in the fleet firing this same ammunition at perhaps a different temperature, $\tau_{j'}$, will have a first-shot jump which can be similarly modeled as $y_{i'j'k'l'} = \alpha + T_{i'} + \beta\tau_{j'} + O_{k'(i'j')} + R_{l'(i'j'k')} - CCF$. This jump is also normally distributed with mean $\beta\left(\tau_{j'} - \sum_{j=1}^{b} w_j \tau_j\right)$ and variance $\left(\frac{a+1}{a}\right)\sigma_T^2 + \left(1 + \frac{1}{ac}\sum_{j=1}^{b} w_j^2\right)\sigma_O^2 + \left(1 + \frac{1}{acn}\sum_{j=1}^{b} w_j^2\right)\sigma_R^2$. This is a general model which can be applied to either the horizontal and vertical jumps. However, note that for horizontal data, it is assumed that $\beta = 0$, since the temperature dependency on jump is assumed to exist only in the vertical direction.*

Mathematical Modeling of Individual Zero

The mathematical model for the SSAAT provides a framework for a model of those rounds fired during an individual zeroing exercise. Despite the fact that only one occasion (at one temperature) is fired per tank, all factors from the SSAAT model are retained. Because $b = c = 1$, the subscripts for temperature and occasion are no longer necessary and, consequently, can be dropped. The subscript *i* is retained as an identifier for tanks and to accentuate the fact that the individual zero is different for every tank. Therefore, rounds fired during the exercise are modeled as $y_{il} = \alpha + T_i + \beta\tau + O + R_{l(i)}$. To avoid confusion with the SSAAT model, we let *l* range from 1 to *m*. The individual zero for tank *i* will simply be the average jump observed during its zeroing exercise, $IZ_i = \bar{y}_{i\bullet} = \alpha + T_i + \beta\tau + O + \bar{R}_{\bullet(i)}$, where IZ_i is distributed normally with mean $\alpha + \beta\tau$ and variance $\sigma_T^2 + \sigma_O^2 + \frac{\sigma_R^2}{m}$.

---

* Mass imbalances in the tank gun system cause a torque about its center of mass, resulting in movement of the gun tube while the projectile is in bore. Since the projectile's acceleration and velocity vary with the propellant temperature, the projectile's in-bore time also varies with the temperature. Because of the torque-induced gun tube motion, the muzzle-pointing direction at shot exit also varies with propellant temperature, resulting in temperature-related gun jump. The gun system's mass imbalance is primarily in the vertical plane, and the gun tube is more constrained in the horizontal plane due to the trunnions; thus, most of the described gun tube motion and resulting temperature-related gun jump is in the vertical plane.

That same tank firing at a later time and possibly at a different temperature, τ', will have its first-shot vertical jump modeled as $y_{il'} = \alpha + T_i + \beta\tau' + O' + R_{l'(i)} - IZ_i$, where $y_{il'}$ is distributed normally with mean $\beta(\tau' - \tau)$ and variance $2\sigma_{O_y}^2 + \frac{m+1}{m}\sigma_{R_y}^2$. Already we see that it is important for the ammunition temperature at the time of zeroing to be as close to the ammunition temperature during combat so that the jump bias from the intended aimpoint is minimal. For horizontal jumps, we follow a similar argument, except that there is no temperature dependency (i.e., $\beta = 0$), and conclude that $x_{ij'k'l'} \sim N\left(0, 2\sigma_{O_x}^2 + \frac{m+1}{m}\sigma_{R_x}^2\right)$.

Hit Probability Comparison of the Two Zeroing Methods

Given the distributions of first-shot jumps under the two zeroing methods, determining which method offers the higher hit probability is of interest. To do this, we will assume that tanks are properly boresighted and aimed at the center of a square target of length L (mils).

Under a fleet zero policy, first-shot hit probability is given in general by

$$P\left(\frac{-L}{2} < X_{FZ} < \frac{L}{2}\right) \times P\left(\frac{-L}{2} < Y_{FZ} < \frac{L}{2}\right)$$

$$= \left\{2\Phi\left(\frac{L}{2\sigma_{FZ_x}}\right) - 1\right\} \times \left\{\Phi\left(\frac{\frac{L}{2} - \beta\left(\tau_{i'} - \sum_{i=1}^{a} w_i\tau_i\right)}{\sigma_{FZ_y}}\right) - \Phi\left(\frac{\frac{-L}{2} - \beta\left(\tau_{i'} - \sum_{i=1}^{a} w_i\tau_i\right)}{\sigma_{FZ_y}}\right)\right\},$$

where $\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$ (the cumulative normal distribution function), $\sigma_{FZ_x} = \sqrt{\left(\frac{a+1}{a}\right)\sigma_{T_x}^2 + \left(1 + \frac{1}{b}\sum_{i=1}^{a} w_i^2\right)\sigma_{O_x}^2 + \left(1 + \frac{1}{bn}\sum_{i=1}^{a} w_i^2\right)\sigma_{R_x}^2}$, and σ_{FZ_y} is defined similarly.

Recall from Table 1 that the current SSAAT calls for $a = 4$, $b = 4$, $c = 1$, and $n = 3$; additionally, the ammunition temperatures (measured in degrees Fahrenheit) are $\tau_1 = -10$, $\tau_2 = 30$, $\tau_3 = 70$ and $\tau_4 = 110$. The weights are $w_1 = .01$, $w_2 = .19$, $w_3 = .64$, and $w_4 = .16$, representing AMSAA's estimated global temperature distribution for armored vehicle combat (i.e., AMSAA expects

1% of future tank battles to take place in a –10° F environment, 19% of future tank battles to take place in a 30° F environment, and so on). With these substitutions, we obtain $x_{i'j'k'l'} \sim N\left(0, 1.25\sigma_{T_x}^2 + 1.12\sigma_{O_x}^2 + 1.04\sigma_{R_x}^2\right)$ in the horizontal direction; for vertical jumps, $y_{i'j'k'l'} \sim N\left(\beta\left(\tau_{j'} - 68\right), 1.25\sigma_{T_y}^2 + 1.12\sigma_{O_y}^2 + 1.04\sigma_{R_y}^2\right)$. Therefore, the first-shot hit probability under a fleet zero policy is given by

$$P(Hit_{FZ}) = \left\{2\Phi\left(\frac{L}{2\sigma_{FZ_x}}\right) - 1\right\} \times \left\{\Phi\left(\frac{\frac{L}{2} - \beta\left(\tau_{i'} - 68^o F\right)}{\sigma_{FZ_y}}\right) - \Phi\left(\frac{\frac{-L}{2} - \beta\left(\tau_{i'} - 68^o F\right)}{\sigma_{FZ_y}}\right)\right\},$$

where $\sigma_{FZ_x} = \sqrt{1.25\,\sigma_{T_x}^2 + 1.12\,\sigma_{O_x}^2 + 1.04\,\sigma_{R_x}^2}$ and σ_{FZ_y} is similarly defined. For some types of ammunition, the elevation jump and propellant temperature are independent. In this case, we have $\beta = 0$ so that $P(Hit_{FZ}) = \left\{2\Phi\left(\frac{L}{2\sigma_{FZ_x}}\right) - 1\right\} \times \left\{2\Phi\left(\frac{L}{2\sigma_{FZ_y}}\right) - 1\right\}$.

For individual zero, the hit probability is given by

$$P\left(\frac{-L}{2} < X_{IZ} < \frac{L}{2}\right) P\left(\frac{-L}{2} < Y_{IZ} < \frac{L}{2}\right) = \left\{2\Phi\left(\frac{L}{2\sigma_{IZ_x}}\right) - 1\right\} \left\{\Phi\left(\frac{\frac{L}{2} - \beta(\tau_{i'} - \tau_i)}{\sigma_{IZ_y}}\right) - \Phi\left(\frac{\frac{-L}{2} - \beta(\tau_{i'} - \tau_i)}{\sigma_{IZ_y}}\right)\right\},$$

where $\sigma_{IZ_x} = \sqrt{2\sigma_{O_x}^2 + \frac{m+1}{m}\sigma_{R_x}^2}$ and σ_{IZ_y} is similarly defined. Most proponents of individual zeroing policy argue that as few as three rounds give an adequate estimate of a tank's zero and keep the cost of individually zeroing the whole fleet to a minimum. Therefore, setting $m = 3$, $\sigma_{IZ_x} = \sqrt{2\sigma_{O_x}^2 + 1.33\,\sigma_{R_x}^2}$. Finally, in the specific case of ammunition for which $\beta = 0$, we obtain an expression analogous to that for fleet zero, $P(Hit_{IZ}) = \left\{2\Phi\left(\frac{L}{2\sigma_{IZ_x}}\right) - 1\right\} \times \left\{2\Phi\left(\frac{L}{2\sigma_{IZ_y}}\right) - 1\right\}$.

Under specific conditions defined by the parameter values, one can determine which zeroing method has a higher hit probability. For example, under the current SSAAT procedure and an individual zeroing exercise with $m = 3$, if one assumes (1) no temperature dependency (i.e., $\beta = 0$), (2) equivalence of horizontal and vertical variances (i.e., $\sigma_{T_x}^2 = \sigma_{T_y}^2$, $\sigma_{O_x}^2 = \sigma_{O_y}^2$, and $\sigma_{R_x}^2 = \sigma_{R_y}^2$), and (3) that the

occasion-to-occasion and round-to-round variances are equal, then $P(Hit_{FZ}) > P(Hit_{IZ})$ if and only if $\sigma_T < .97\sigma_R$. In most cases, however, such a simple relationship cannot be established.

Given the dependency between jump and ammunition temperature, seeing which zeroing method has a higher hit probability when the temperature at the time of the individual zeroing exercise differs from the temperature at the time of combat is of interest. To do this, we can plot the difference in hit probability, $\Delta P_{Hit} = P(Hit_{IZ}) - P(Hit_{FZ})$, as a function of both of these temperatures for assumed values of the variance components and the temperature-dependent slope. Estimates of these parameters are available for most ammunition in the U.S. Army arsenal. As an example, consider a hypothetical training round with a temperature-dependent slope of $\beta = 0.004$ mils per degree Fahrenheit.[2*] Assume also that the variance components relative to the length (L) of the square target are $\sigma_{R_x} = \sigma_{R_y} = \sigma_{O_x} = \sigma_{O_y} = L/6$ and $\sigma_{T_x} = \sigma_{T_y} = 0.97 \times L/6$. Figure 1 is a 3-D plot generated by MATLAB® of ΔP_{Hit} as a function of the firing and zeroing temperature for this baseline case. The black lines are contours for $\Delta P_{Hit} = 0$. Yellow and orange regions indicate temperature conditions for which $\Delta P_{Hit} > 0$ (i.e., when individual zero produces the higher hit probability). On the other hand, blue regions indicate when fleet zero gives better hit probability. From the figure, we see that when firing and zeroing temperatures are equal, individual zero is the preferred policy, especially when the temperatures are toward their extremes. Conversely, if the firing and zeroing temperatures differ greatly, the preferred procedure is fleet zero (blue regions). For this baseline case, we also obtain the curious result that tanks which individually zero at 68° F will have the same hit probability as using a fleet zero for *any* firing temperature.

One can next study the effect of increasing (or decreasing) any of the variance components on ΔP_{Hit}, keeping all other parameters constant.[†] First, consider a change in tank-to-tank dispersion. Because σ_T is negatively correlated with $P(Hit_{FZ})$ but not correlated at all with $P(Hit_{IZ})$, ΔP_{Hit} increases with σ_T. Intuitively, this makes sense – greater tank-to-tank variation increases the number of tanks with large bias and hence the need to individually zero; the opposite holds with less variability. This is confirmed graphically in Figure 2, where relative to the baseline case, a decrease in σ_T to a value

---

[*] Three of the four ammunition types studied in this report had temperature-dependent slopes ranging from .0041 to .0047 mils per degree Fahrenheit; one ammunition type had no temperature dependency. Therefore, a value of .004 mils is reasonable for the purpose of this example.

[†] For simplification of the ensuing example, we will henceforth assume that the horizontal and vertical variance components are equal and will drop the "X" and "Y" subscripts used previously.

of $0.97 * L/8$ makes fleet zero the more favorable policy; the opposite is true if σ_T is increased to $0.97 * L/4$. In fact, in this latter case, we see that there is a range of temperatures (24^o F to 95^o F) in which the tank can be individually zeroed and have a better $P(Hit)$ than under fleet zero, no matter what the actual ammunition temperature is at the time of firing. This range of temperatures will hereafter be referred to as the Individual Zero Preferred (IZP) range.

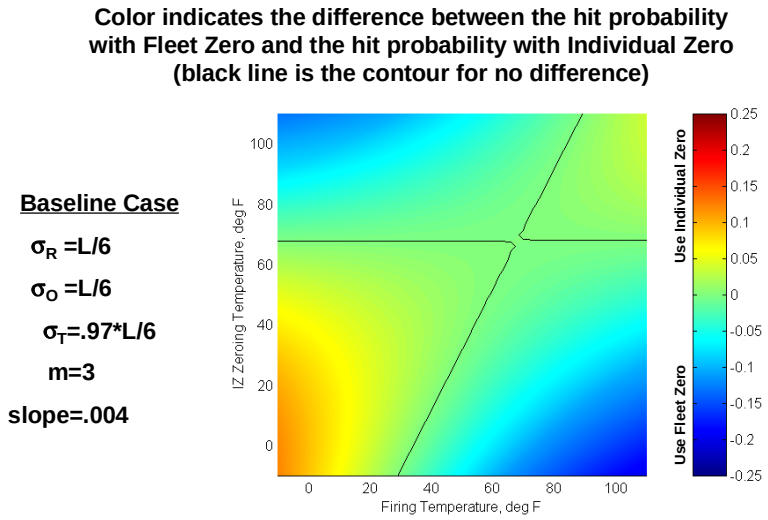


Figure 1. Baseline-case relationship between ΔP_{Hit} and the ammunition temperatures at the time of firing and zeroing.

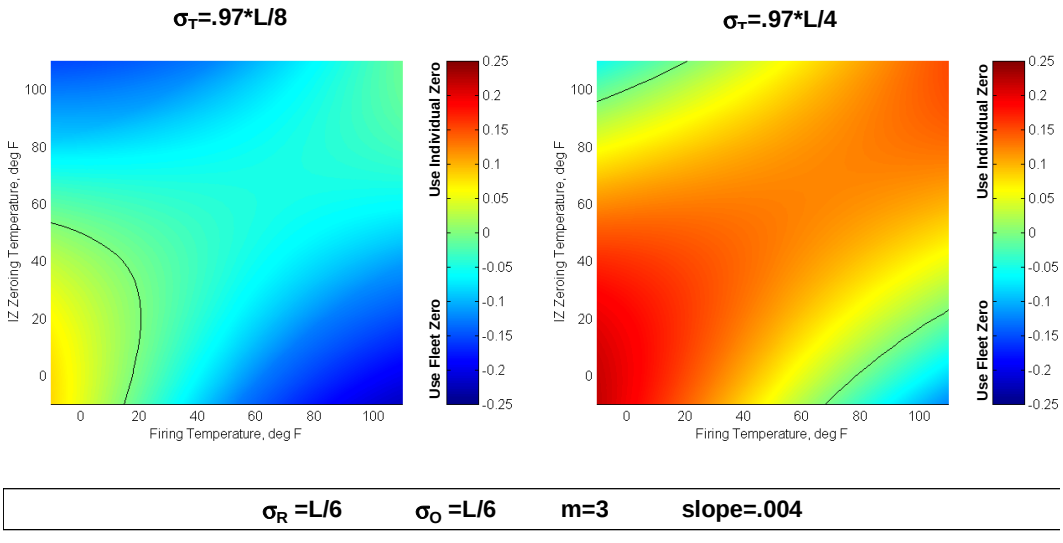


Figure 2 (a) and (b). Effect of changing σ_T, while holding all other parameters fixed at baseline value.

If occasion-to-occasion dispersion is the free parameter, it is not intuitive which zeroing method becomes more preferable since, for example, both $P(Hit_{FZ})$ and $P(Hit_{IZ})$ decrease when σ_O increases. Therefore, one can turn to the 3-D color charts to see which hit probability is influenced the most by a change in σ_O. Figure 3(a) shows that a reduction in σ_O creates an IZP range of approximately 52^o F to 77^o F. On the other hand, as σ_O increases, Figure 3(b) shows fleet zero becoming the more preferable policy. This is easily explained by noting that the expressions for the variance of jump (both in vertical and horizontal) under fleet zero have a σ_O^2 coefficient of approximately 1.12; in the case of individual zeroing, this same coefficient is larger, namely 2.

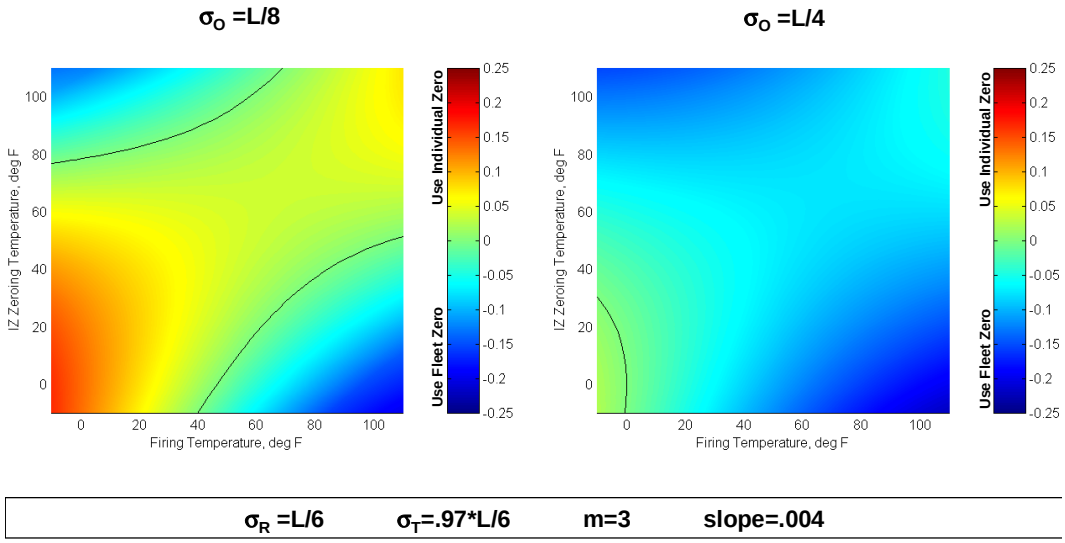


Figure 3 (a) and (b). Effect of changing σ_O, while holding all other parameters fixed at baseline value.

The effect of a different round-to-round dispersion is similar to that of a change in occasion-to-occasion dispersion, although less pronounced. This is seen in Figure 4(a), where lowering σ_R by the same amount as that used for σ_O to generate Figure 3(a) yields a smaller IZP range of 62^o F to 71^o F.

Figure 5 is included to show two extreme cases of operating with a different temperature-dependent slope. Recall the previous discussion (page 9) in which it was shown that under specific conditions (including no temperature dependency), $P(Hit_{FZ}) > P(Hit_{IZ})$ if and only if $\sigma_T < .97\sigma_R$. Figure 5(a) is the trivial extension to this set of conditions, showing that $\sigma_T = .97\sigma_R$ implies $\Delta P_{Hit} = 0$ for all combinations of individual zero and actual firing temperatures. Figure 5(b) is for the case where the temperature dependency is much stronger. Similar to the baseline case, we see that tanks which

individually zero at 68° F will be equally as effective as tanks using fleet zero, regardless of the actual firing temperature. The deep orange regions of this figure indicate that there is great potential for improving hit probability if tank commanders are allowed to individually zero at the same extreme temperature as that expected during battle. However, the blue regions indicate the greater risk involved if a tank is individually zeroed at a temperature much different from that encountered during battle.

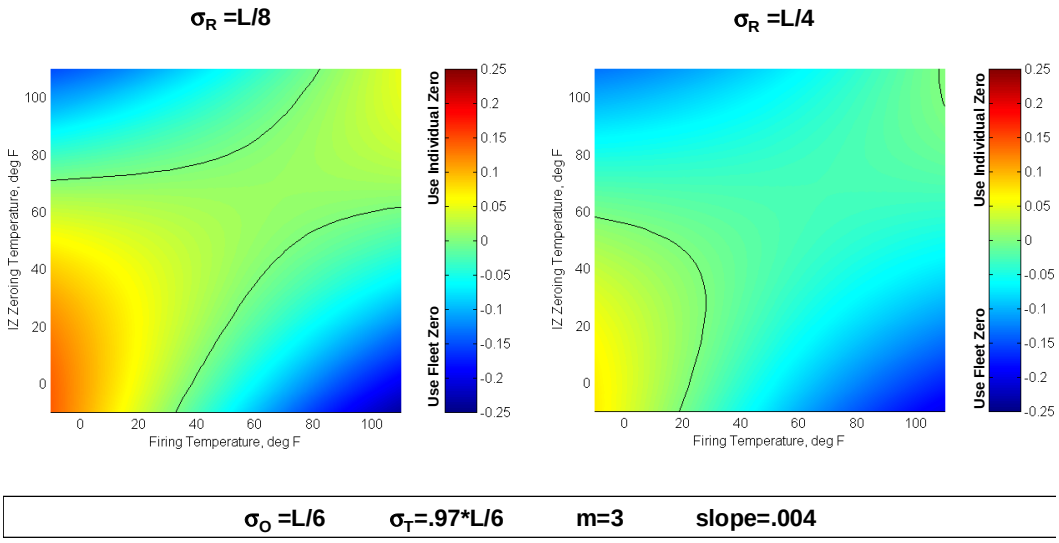


Figure 4 (a) and (b). Effect of changing σ_R, while holding all other parameters fixed at baseline value.

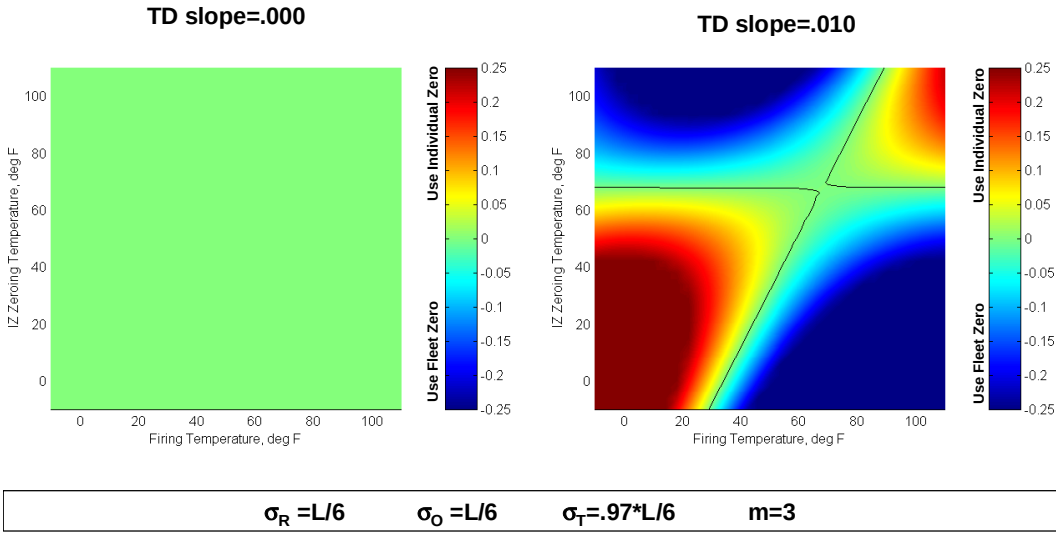


Figure 5 (a) and (b). Effect of changing temperature-dependent slope, while holding all other parameters fixed at baseline value.

Implications

The results of the current modeling effort imply some interesting and important information and policy choices. First, we know from previous testing that the magnitude of the various error sources varies by ammunition type. Ammunition type A may have small tank-to-tank errors, while type B may have large tank-to-tank errors. In fact, the trend appears to be that the more energetic the ammunition, the greater the tank-to-tank errors.[3] This means that an optimal zero policy may vary across ammunition types. For example, ammunition types with small tank-to-tank differences could use the fleet zero, and those types that exhibit large tank-to-tank differences could individually zero those ammunition types when needed.

Conversely, a policy that screens tanks could identify those tanks that do not shoot well with the fleet zero and require that they be individually zeroed. This seems to be the policy that is used today during gunnery exercises. Prior to conducting the record tank tables,[*] each tank fires several rounds of ammunition to ensure that the tank is calibrated using the fleet CCF. When a tank cannot demonstrate a calibration in this manner, the tank is zeroed using an individual zero.

Whether a new zeroing policy would return completely to the individual zero method or to some hybrid where only certain ammunition types are individually zeroed across the fleet or where only certain tanks are individually zeroed, the training impacts will be large. The U.S. Army has trained with the fleet zero for nearly two decades and it was successful in Operation Desert Storm. As a result, there is a significant amount of confidence in the policy and a great deal of institutional support. To change the policy requires convincing most of the Armor Force that a significantly better approach exists. In addition, a substantial training effort would be required to prevent the numerous possible errors associated with individually zeroing tanks from manifesting themselves across the fleet because of improper zeroing.

Significantly, the modeling showed that the greater the jump dependency on ammunition temperature, the more sensitive the accuracy of the tank is to the zeroing decision-making process. This means that attention to temperature-related jump effects should continue when developing ammunition. Fortunately, temperature-dependent jump is measurable and consistent across the fleet of tanks.

---

[3] Pell, Richard F., "Key to Improve Accuracy: Tighter Gun Tube Specs", Letter to the Editor, *ARMOR*, January-February 1996, pg. 3.

[*] Tank tables are training exercises whereby a tank(s) fire at various targets on a range and are scored on number of hits, time, and procedure. On most of the tank tables, the tanks also maneuver

Therefore, corrections for temperature are possible via the fire control system or temperature conditioning while the ammunition is stored in the tank's bustle.

Finally, the importance of having accurate tanks and the difficulties associated with changing zeroing policies suggest that a root cause analysis to determine why some ammunition types or some tanks require different zero policies is warranted. Certainly, a great deal of understanding and knowledge has been developed over the last few decades, but more work seems necessary if the accuracy of U.S. tanks is to keep pace with the demands for greater hit probability at extended range.

Acknowledgments

The authors would like to thank Mr. Paul Durkin of the U.S. Army Test Center and Mr. Ralph Scutti of the U.S. Army Materiel and Systems Analysis Activity for their clarification of some of the technical issues associated with zeroing methods. Also, we are grateful for the support of Dr. Mark Bundy of the U.S. Army Research Laboratory towards this research effort.

ANALYSIS OF FUZZY REGRESSION FOR MODELING SHELF-LIFE OF GUN PROPELLANTS

Iris V. Rivero-Diaz and Kwang-Jae Kim
Penn State University Department of Industrial Engineering
310 Leonhard Bldg., University Park, PA, 16802, USA.
Pohang University of Science and Technology
San 31, Hyoja-Dong, Nam-Gu, Pohang, Kyungbuk 790-784, South Korea.

ABSTRACT

Shelf-life estimation of gun propellants is a situation in which representative data describing the factors influencing its behavior are not clearly established, neither is sufficient nor representative of variations presented from lot to lot. This research applies the concept of fuzzy regression analysis to model the life of the gun propellant and determines the distribution behavior that best fitted the relationship. Fuzzy regression takes into account such aspects as a non-linear relationship between factors, and variations that might exist between lots. It has been found through a simulation study that when representative data quality is inferior fuzzy regression should not be used, and instead classical statistical regression should be performed.

INTRODUCTION

This research is concern with statistically modeling the shelf-life of gun propellants. It is desire to accurately model gun propellant's life in order to predict its stability and as a result assessing how safe the gun ammunition is when it is store during peacetime. Currently, in order to get an estimate of the gun performance, information needed by the U.S. Navy, the Naval Surface Warfare Center performs a stability test of the stored gun propellants. This stability test is called *Master Sample Surveillance Program*. Data collected from it is fitted into a straight line generating a linear regression model of the stability of the gun propellant over time. In assuming that in fact there exists a linear relationship between the age of the gun propellant and the fumes it emits other external factors are ignored such as variations due to differences in characteristics among lots of propellants produced by different manufacturers, and differences that might be encountered for accelerating the aging process of the gun propellants in a heating chamber.

The assumptions made by the program are often of concern when modeling the shelf-life of gun propellants. The methods being utilized currently to fit the model assume that there in fact exists a linear relationship between the age of the propellants and the fumes it expedites. It is also observed that by making a linear regression model, variation on the stability patterns from lot to lot is not accounted.

Along with the discussion of shelf-life estimation of gun propellants it is desire to test a recently developed form of regression called fuzzy regression analysis. Which it is based in situations when boundary conditions are not concretely specified, therefore taking into account variations in the model. In the following section the fuzzy regression analysis will be further described. The next sections explain how the fuzzy regression analysis was applied to the modeling of gun shelf-life estimation and its results. As a final topic, based on the results of the application of fuzzy regression, suggestions and recommendations on the appropriateness of the model to represent gun shelf-life will be made.

FUZZY REGRESSION ANALYSIS

BACKGROUND

Fuzzy regression is a concept that aims to model situations in which estimation is influential towards conclusions made based on the system's model and when forecasting is performed in an uncertain environment[1]. Although it has been tested, compared, and proved that statistical linear regression surpasses the quality of modeling of fuzzy regression it rarely describes the true relation between variables[2]. Therefore, when making predictions on capabilities fuzzy regression offers an option to model linear regression situations in which parameters are not clearly defined or are "too complex or too ill-defined to admit for precise analysis"[3]. Its major drawback is that its quality performance is mainly influenced by such aspects as sample size, and specificity of model which are expected to match correctly with properties such as autocorrelation, and randomness of the studied sample. In these conditions is when fuzzy regression becomes a useful tool in modeling. Fuzzy regression is capable of making model predictions when such variables as reduced sample size, and unclear parameters (for example, boundaries) are considered. Overall, "fuzzy regression represents a phenomenon that is imprecise and vague in nature. In the case where many phenomena are not sharply defined, fuzzy techniques will represent these effects of causal variables in a more realistic way"[1].

A fuzzy set was initially described by Zadeh[4] which stated that samples are categorized as fuzzy when there is not a clear transition point from one sample to other. Therefore, a fuzzy linear function is one in which parameters are given by fuzzy sets. Fuzzy linear functions are intended to be modeled on the premises of Zadeh's extension principle thus when developing and/or formulating a fuzzy linear regression analysis it is intended to account for uncertainties that would not only be due to randomness, but also to fuzziness.

The main objective of fuzzy linear regression is to be a tool in modeling where a relationship among variables is not clearly defined through Zadeh's extension principle[5]. It is important to establish that fuzzy regression analysis is not a statistical method. This is because deviations in this type of modeling are due to the indefiniteness of the variables (or parameters), and not due to measurement errors. Therefore, fuzzy regression brings the opportunity to develop models where strict assumptions such as normality, natural randomness, that are associated with large sample groups can be ignored.

In taking the risk of ignoring marked characteristics of a sample, which attributes would seem not to make any significant relevance to the outcomes of the model, could attempt towards the validity and performance of the model if it is chosen to try to fit the variables in a linear pattern.

Therefore, the setting where fuzzy regression analysis is justified is when decisions involving human criteria are involved, and when a process is not clearly or patterned designed; in a situation where its output is to be predicted, with a small sample size, and with ambiguous data is available. Kim et. al[6] establishes that fuzzy regression improves over statistical regression as the size of the data set diminishes.

FUZZY REGRESSION: OBJECTIVES, DEFINITION, AND FORMULATION

Deviations when modeling fuzzy sets are accounted due to the indefiniteness of the system structure and not to errors recorded in observations influenced by human interaction. These deviations are fuzzy parameters, which set the function to be modeled to be of a vague phenomenon. The vagueness can be modeled by

$$J = c_1 + \ldots + c_n \qquad (1)$$

Fuzzy regression main objective is to minimize the vagueness brought about by the fuzzy parameters. According to studies developed by Hideo Tanaka in 1982[7], fuzzy parameters are defined as:

$$\mu_A(a) = \min j \, [\mu_{Aj} (a_j)]$$

$$\mu_A(a) = \{ 1 - | \alpha_j - a_j | / c_j \quad , \alpha_j - c_j \leq a_j < \alpha_j + c_j$$
$$0 \qquad\qquad , \text{otherwise}$$
$$\text{where } c_j > 0 \tag{2}$$

Then, a fuzzy linear regression can be obtained of the form of:

$$\mathbf{Y} = \mathbf{A}_1 x_1 + \ldots + \mathbf{A}_n = \mathbf{A}x \tag{3}$$

where $\mathbf{A}_i$ have been defined to be the variable describing the fuzzy set.

If we are able to find the value of the fuzzy parameters $\mathbf{A}^*_i=(\alpha_1 , c_i)$, then the basic goal of fuzzy regression, that is to reduce the vagueness introduced into the model by the uncertainty of the parameters, would be simplified to get the ith observed value of the dependent variable as close to its fuzzy estimate, defined as $y^*_i=A^*x_i$, to be at least H. H is a value between 0 and 1 that establishes the degree of fitness to which the model is desired to be represented. To minimize the fuzziness of the dependent variable the problem could be introduced as a simple linear programming problem in the form of:

$$\min \alpha, c\, J = c_1 + \ldots + c_n$$
$$\text{subject to} \qquad c \geq 0,$$
$$\alpha^t x_i + (1 - H) \sum c_j \, | x_{ij} | \geq y_i + (1 - H)e_i,$$
$$\alpha^t x_i + (1 - H) \sum c_j \, | x_{ij} | \geq -y_i + (1 - H)e_i, \qquad i = 1, \ldots, N \tag{4}$$

where H is the desire degree to which we would like to linearly fit the model. Values of x_i and y_i represent the inputs and outputs per observation recorded respectively. α_i is the center value and c_i its corresponding width which introduces the fuzziness and/or vagueness to the problem. It needs to be stated that the values of α are not always positive as a result $\alpha' \geq 0$ is defined to be as.

$$\alpha = \alpha' + d \tag{5}$$

It has also been stated that since the number of constraints 2N is larger than the number of variables in Eq. (4), it would be more convenient to solve it with the dual.

FUZZY REGRESSION: PARAMETERS AND APPROPRIATENESS OF THE REGRESSION

Parameters in fuzzy regression are fuzzy numbers that can be considered as a possibility distribution[9]; the estimated dependent variable is also a fuzzy number, yielding a fuzzy interval of the dependent variable[10].

It has to be noted that since fuzzy regression is a non statistical method there are non specific means to check the appropriateness of the model thus the decision maker will specify a confidence level that will be the equivalent value of H. From this, it can be concluded that the focus of the fuzzy intervals is entirely on the given observation and not on the future sample or predictions[10]. Fuzzy regression is not appropriate for predictions.

The possibility distribution shape of fuzzy model parameters usually utilizes a linear, symmetric, triangular membership function. Fuzzy regression provides for more flexibility in the shape of its distribution than statistical regression.

From several studies performed in sampling and establishing comparison between the general statistical regression processes used in the past, it has been shown that the spread of the fuzzy estimate increases with more data whereas in basic statistical regressions as more data was available the spread of the point estimator decreased. To explain the behavior of the fuzzy estimate we rely in the concept of possibility. Since the fuzzy regression algorithm requires that the fitted fuzzy model explain all possibilities in a given observation set, if the observations

are increased more possibilities are to be explained by the fuzzy model. Possibility is explained but sacrificing precision.

As a summary, there are certain situations in which fuzzy regression is appropriate to be used: when observations and/or data come from fuzzy numbers, when data is obtained from measurements and/or estimated, when there is not sufficient data available to perform a study; Overall, mainly in situations when human knowledge is the source of information used for modeling.

Fuzzy linear regression usefulness could be observed in the estimation of regression parameters when there is a lack of data collected and/or aptness of the regression model is poor.

PROBLEM DEFINITION AND FUZZY REGRESSION APPLICATION

The main concern in this research is to determine an appropriate distribution to model a data set (which observations are considered fuzzy due to their uncertainty and lack of observable patterned behavior). The data to be modeled is the shelf-life of gun propellants. Since the majority of gun propellants are produced before they are completely used, its stability deteriorates during storage. Gun performance is assessed currently by the U. S. Navy through performing stability tests.

The test consists of taking several samples, five pounds each, from individual propellant lots; taking a portion of propellant of about 45 gm, from each five pound sample, which would then be heated in a chamber at 65.5°C. This heating process will accelerate the aging of the propellant causing it to generate red nitrogen oxide fumes. The time it takes to generate fumes is used as a measure of stability of the propellant once it enters the heating chamber. The general observed trend in past experiments has shown that the stability of the propellants tends to decrease in proportion with the fume time. The usefulness and safe life of the gun propellants as a result of these performed stability test is estimated by determining the fume time duration. When fume time decreases below thirty days at 65.5°C temperature the propellant is considered to be unsafe and the lot corresponding to the tested sample is destroyed.

But the method just described to determine the shelf-life of gun propellants has a few weak points. In order to obtain accurate results from this method observation of the samples to age at a 65.5°C temperature takes approximately twenty to eighty years, which is considered too long in order to correctly establish safety guidelines, do inventory management, acquisition decisions, etc. A major flaw from the testing is the assumption of linearity it makes when it is statistically represented, producing a model that it is not adequate in estimating the true shelf-life of the gun propellants. Currently, shelf-life is being modeled with fixed effects of the linear regression model, where fume time is fitted against propellant age. Data was collected from a group of similar gun propellant lots. The basic assumption taken with this regression is that there exists a linear relationship between fume time and age. The shelf-life was defined to be as the age when a "95% one-sided lower prediction limit for the fume time curve intersects the acceptable lower specification level"[11].

Three major concerns from the method just described arise, 1) How certain are we as decision makers that there in fact exists a linear relationship between stability and age of the propellant; 2) In estimating the shelf-life of gun propellants how is the lot to lot variation going to be treated; 3) Are we sure that using an accelerated condition, of submitting the gun propellant to a temperature of 65.5°C, is a reliable procedure to find unsafe propellants.

With the concerns just addressed we will try a new method of modeling situations in which variations could be taken into account, fuzzy regression analysis. As it can be seen from Table 1 the possibility distribution theory better describes a system where uncertainties are to be taken into account in modeling.

With the concerns just addressed we will try a new method of modeling situations in which variations could be

To formulate a possibility regression analysis the linear transformation in Eq. (6) is used:

$$\mathbf{Y} = T\mathbf{x} \tag{6}$$

where T is defined to be a p x n matrix and rank[T]=p. For the purpose of modeling the fuzzy regression analysis **x** is going to be assumed to be governed by a possibility distribution $(\mathbf{a}, D_A)_e$.

$$\mathbf{Y} = T\mathbf{A} \tag{7}$$

Defining the possibility distribution[12]

$$\prod_Y(\mathbf{y}) = \max_{\{x|y=Tx\}} \prod_A(\mathbf{x}) \tag{8}$$

Therefore, the optimization problem would be written as[12]:

$$\begin{aligned} &\min x\ (\mathbf{x}-\mathbf{a})^t D_A (\mathbf{x}-\mathbf{a}) \\ &\text{subject to} \quad \mathbf{y} = T\mathbf{x} \end{aligned} \tag{9}$$

where optimal solution are:

$$\begin{aligned} &\lambda^* = 2(T D_A T^t)^{-1}(\mathbf{y}-T\mathbf{a}), \\ &\mathbf{x}^* = \mathbf{a} + D_A T^t (T D_A T^t)^{-1} (\mathbf{y}-T\mathbf{a}) \end{aligned} \tag{10}$$

Defining the exponential possibility distribution as:

$$\begin{aligned} &\prod_Y(\mathbf{y}) = \exp\{-(\mathbf{y}-T\mathbf{a})^t (T D_A T^t)^{-1} (\mathbf{y}-T\mathbf{a})\}, \\ &Y = (T\mathbf{a}, T D_A T^t)_e \end{aligned} \tag{11}$$

To specially define the possibility distribution $\prod_Y(y)$ of the output could be done by replacing T with $\mathbf{X}^t$:

$$\begin{aligned} &\prod_Y(y) = \exp\{-(y-\mathbf{X}^t\mathbf{a}_c)^2 (\mathbf{X}^t D_A \mathbf{x})^{-1}, \\ &Y = (\mathbf{X}^t\mathbf{a}_c, \mathbf{X}^t D_A \mathbf{x})_e \end{aligned} \tag{12}$$

The optimization problem to find the possibility distribution can be written as:

$$\begin{aligned} &\min_{D_A, a_c} J = \sum_{j=1} \mathbf{x}_j^t D \mathbf{x}_j \\ &\text{subject to} \quad 1)\ \mathbf{x}_j^t D_A \mathbf{x}_j \geq (y - \mathbf{a}_c^t \mathbf{x}_j)^2 / (-\log h_j) \\ &\qquad\qquad 2)\ \mathbf{D}_A > 0 \end{aligned} \tag{13}$$

The constraints presented in the optimization problem by Eq. (13) define the problem to be non-linear, thus in order to be able to use the basic procedure of problem solving of linear programming the following measures should be taken: 1) Find the center vector $\mathbf{a}_c$ through conventional regression; and 2) Substitute the obtained center vector $\mathbf{a}_c^*$ in the first constraint of the optimization problem described by Eq. (13):

$$\begin{aligned} &\min_{D_A, a_c} J = \sum_{j=1} \mathbf{x}_j^t D_A \mathbf{x}_j \\ &\text{subject to} \quad 1)\ \mathbf{x}_j^t D_A \mathbf{x}_j \geq (y - \mathbf{a}_c^{*t} \mathbf{x}_j)^2 / (-\log h_j) \\ &\qquad\qquad 2)\ \mathbf{x}_i^t D_A \mathbf{x}_i = 0 \qquad i \neq j, \qquad i,j \in E \end{aligned} \tag{14}$$

To describe the shelf-life of gun propellants constraint "1" was modified and in the denominator $\ln h_j$ was used instead of $\log h_j$.

In order to apply the fuzzy regression analysis a data set consisting of thirty-six observations was studied. The observations recorded were of the age of the gun propellants and its associated developed fume. Observations recorded were based in a 365 days cycle, therefore for our investigative procedure observations were scaled by 365 (that is, age = age/365; fume = fume/365).

Since it is intended to make use of fuzzy regression analysis as it was explained before, this system is going to be defined as a possibility linear system of the form:

$$Y = A_1x_1 + \ldots + A_nx_n = \mathbf{Ax} \tag{15}$$

Defining **A** as a fuzzy coefficient vector defined by the exponential possibility distribution "with center vector $\mathbf{a}_c$ and a symmetrical positive definite matrix D_A"[11] which takes the following form:

$$\mathbf{A} = (\mathbf{a}_c, D_A)_e \tag{16}$$

The independent variable to be modeled in this experiment was defined to be the age of the gun propellant while the fumes generated represented the dependent variable. Using conventional regression analysis the center vector $\mathbf{a}_c^*$ was obtained to be:

$$\mathbf{a}_c^* = (22.6496, -5.0068) \tag{17}$$

Although it did not seem to reflect a great impact in the ultimate result of our minimization problem, in order to establish a fair comparison with others already applied modeling systems for the gun propellant's shelf-life the center vector values utilized before were also used here:

$$\mathbf{a}_c^* = (3.97535, -0.15136) \tag{18}$$

The linear programming problem defined by Eq. (14) was solved by making use of the LP software "LINDO" by selecting ten random orthogonal condition combinations, since there are ${}_{36}C_2 = (36*35)/2 = 630$ combinations that would result in too many combinations for the experiment. Table 3 shows the (input – output) data given where $\mathbf{x}_j$ is represented as an input vector of the form $(1, x_j)^t$.

In this problem Z is defined to be as the covariance, therefore its value could be assumed as either positive and/or negative. "LINDO" only defines its variables to be nonnegative; Therefore, to account for this the Z variable value will be stated as $Z = Z_0 - Z_1$. An h_j value was set at a value of "0.5" to account for a fair confidence interval.

Then the objective function to be minimized is defined by the following function:

Min $J = 36X + 162.03998Y + 202.87Z$

subject to

$X + 6.12Y + 9.37Z \geq 7.63$
$X + 5.49Y + 7.54Z \geq 0.91$
$X + 4.06Y + 4.12Z \geq 3.93$
$X + 3.77Y + 3.55Z \geq 0.21$
$X + 3.89Y + 3.78Z \geq 4.99$
$X + 3.17Y + 2.51Z \geq 8.01$
$X + 5.90Y + 8.71Z \geq 13.16$
$X + 5.72Y + 8.17Z \geq 0.39$
$X + 4.07Y + 4.13Z \geq 4.78$
$X + 3.74Y + 3.50Z \geq 0.50$
$X + 3.75Y + 3.52Z \geq 2.55$
$X + 3.02Y + 2.29Z \geq 2.61$
$X + 4.42Y + 4.88Z \geq 49.41$
$X + 5.73Y + 8.21Z \geq 19.02$
$X + 5.95Y + 8.85Z \geq 0.07$
$X + 4.04Y + 4.08Z \geq 6.06$
$X + 3.83Y + 3.67Z \geq 0.42$
$X + 3.50Y + 3.05Z \geq 0.40$
$X + 0.10Y + 0.003Z \geq 56.35$
$X + 2.61Y + 1.71Z \geq 9.63$

$X + 0.10Y + 0.003Z \geq 56.35$