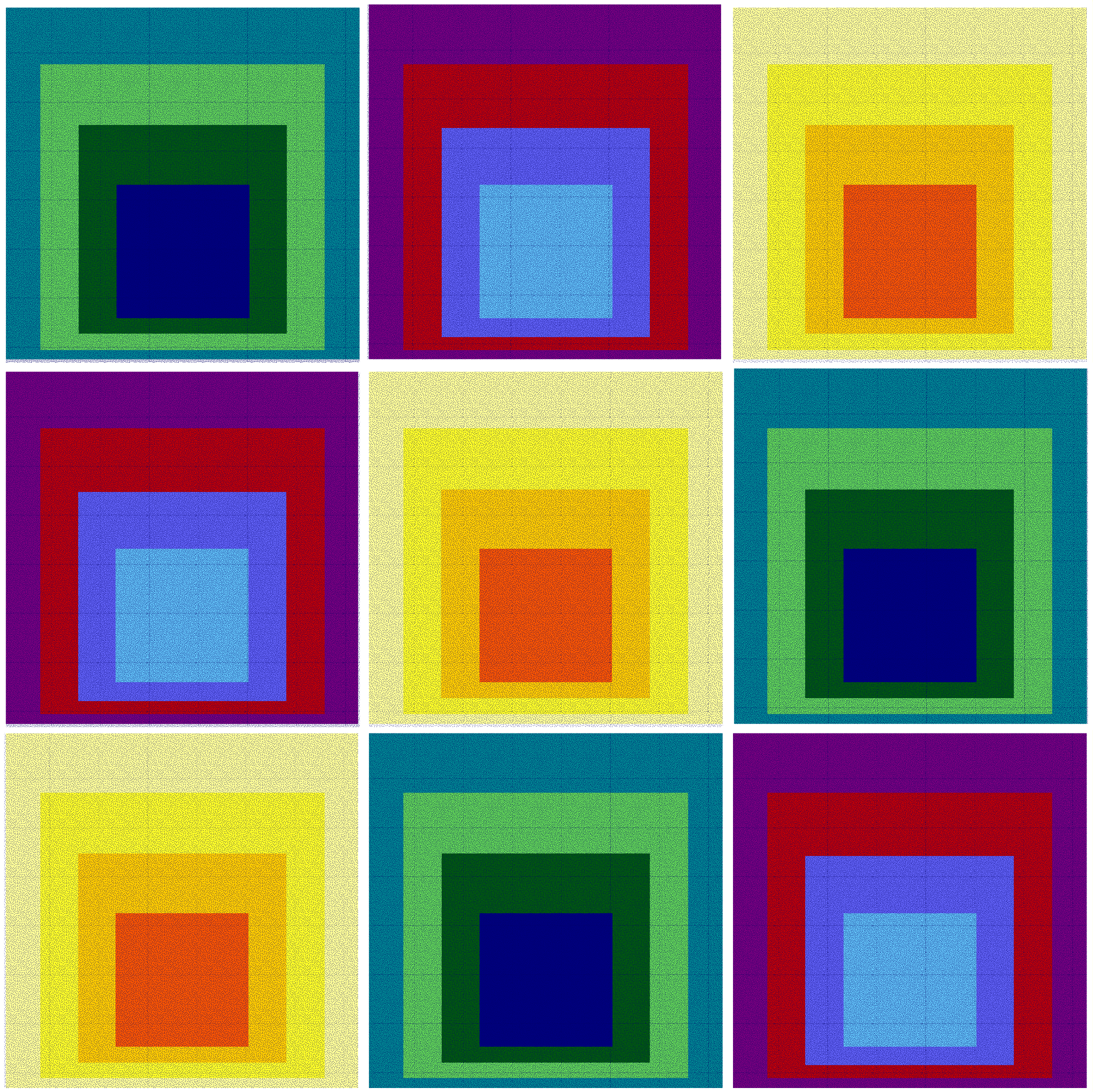
Report of Activities: 2004

Statistical Engineering Division



U.S. Department of Commerce
National Institute of standards and technology
Information technology laboratory
Gaithersburg, md 20899 USA

U.S. Department of Commerce
Technology Administration
National Institute of Standards and Technology
Information Technology Laboratory

REPORT OF
ACTIVITIES OF THE
STATISTICAL ENGINEERING DIVISION

January 2005

Covering Period: January 2004 – December 2004

Certain commercial equipment, instruments, or materials are identified in this document to specify the scientific procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Covers:
With homage to Josef Albers

Contents

1 Division Overview

Nell Sedransk, Chief
Statistical Engineering Division, ITL

The Statistical Engineering Division (SED) of the Information Technology Laboratory (ITL) of the National Institute of Standards and Technology (NIST) conducts fundamental and applied statistical research on problems in metrology and collaborates on research in other Divisions of ITL, in other Laboratories of NIST and with NIST's industrial partners.

The research and collaboration programs of the Statistical Engineering Division are driven by the needs of high priority NIST research projects, by the requirements for a statistically sound foundation for metrology and measurement services that NIST provides to its customers, and by scientists' and engineers' needs for rigorously developed statistical tools for metrology.

This report provides summaries of significant projects selected to indicate the spectrum of SED contributions to NIST and to statistical metrology in 2004. These presentations of multidisciplinary collaborations here are just that: brief précis of intensive interactions with scientists and engineers. References to research publications are provided as sources for expanded descriptions and technical details of projects that are often larger and more complex than can be captured here. Also, descriptions of other activities of SED that cannot be included here can be found on the SED home page at: http://www.itl.nist.gov/div898/. The *e-Handbook of Statistical Methods* can be found directly at: http://www.nist.gov/stat.handbook/.

The role of SED is integrated into research throughout NIST; SED staff actively collaborate with more than 90% of the scientific Divisions at NIST, on the Gaithersburg and Boulder campuses. NIST prizes excellence in research; consequently high priority, high impact scientific research almost invariably calls for sophisticated expertise in multiple disciplines, including statistics. The most common role for SED staff is as part of an interdisciplinary research team, contributing to the definition of research objectives, the formulation of statistical strategies and the development of new statistical methodology and statistical computation for process characterization and the analysis of experimental data. The combined statistical expertise in SED is necessarily extremely broad, encompassing many sub disciplines of statistics (including experimental design, generalized linear models, stochastic models, Bayesian inference, time series analysis, reliability analysis, statistical signal processing, image analysis, spatial statistics, quality control, exploratory data analysis, data mining, statistical computation and graphics, etc.). However, the SED objective is always to integrate statistics fully into the research plan to strengthen the fundamental research design and to implement the most powerful statistical tools for drawing inferences and for estimating uncertainties. Success in these collaborations is largely due to the deep involvement of SED staff with the science itself via their scientist colleagues.

For the world of internationally metrology, SED focuses on the development of statistical theory, methodology and practical tools for metrologists to use worldwide. The prominence of statistical metrology continues to grow with increased attention to international intercomparisons and international acceptance of standards among national metrology laboratories. Whether the requisite technical expertise lies in probabilistic modeling, in design of experiments, in theory and methodology of inference, in computationally intensive statistical tools or in Bayesian inference and modeling, part of the mission of SED is to expand the statistical methodology available to NIST scientists, to US industry and to metrologists worldwide. As rapidly advancing technology leads to new kinds of measurement instrumentation and measurement processes, implementation of these depends upon advances in statistical metrology. This research also contributes in a fundamental way to the discipline of statistics.

SED supports the NIST mission through collaboration on projects that reflect the highest NIST priorities. The vignettes included here are organized according to the principal NIST focus areas and initiatives and the special focus areas for SED. In addition to primary research roles, CORE activities of SED also support the certification for NIST Standard Reference Materials and calibration services and provide statistical metrology education to NIST scientists and engineers. Each year, undergraduate and graduate students join SED as part of a highly successful program of statistical research opportunities in metrology and statistics.

The professional staff comprises three Groups of mathematical statisticians with graduate degrees, as listed in Section 2. Two of the Groups are located in Gaithersburg, Maryland, and the third is in Boulder, Colorado. Also integral to SED activities are Visiting Faculty appointees from several universities, Guest Researchers and Students who join SED to participate in research experiences.

Thank you for reading. Your comments are most welcome.

Nell Sedransk, Ph.D.
Chief, Statistical Engineering Division
820 W. Diamond Avenue, 820/353
Gaithersburg, Maryland 20899-8980

Email: nell.sedransk@nist.gov
Phone: (301) 975-2839

2 Staff

Nell Sedransk, Ph.D. Division Chief

Metrology Statistics and Computation Group
Nien-Fan Zhang, Ph.D. Group Manager
A. Ivelisse Avilés, Ph.D.
Will Guthrie, M.S.
Charles Hagwood, Ph.D.
Alan Heckert, M.S.
Walter Liggett, Ph.D.
John Lu, Ph.D.
Juan Soto, M.S.

Statistical Modeling and Analysis Group
James Filliben, Ph.D. Group Manager
Dennis Leber, M.S.
Stefan Leigh, M.S.
Hung-kung Liu, Ph.D.
Andrew Rukhin, Ph.D.
Blaza Toman, Ph.D.
James Yen, Ph.D.

Boulder Statistics Group
Jack Wang, Ph.D. Group Manager
Kevin Coakley, Ph.D.
Jolene Splett, M.S.
Sarah Streett, Ph.D.
Dom Vecchia, Ph.D.

Visiting Faculty and Guest Researchers
Duane Boes, Ph.D. Colorado State University
Abderahman Cheniour, ISTIL Engineer, NIST Guest Researcher
M. Carroll Croarkin, M.S. NIST Guest Researcher
Dipak Dey, Ph.D. University of Connecticut
Jan Hannig, Ph.D. Colorado State University
Adriana Horníková, Ph. D. NIST Guest Researcher
Hari Iyer, Ph.D. Colorado State University
Don Malec, Ph.D. Bureau of Census
Joan Rosenblatt, Ph.D. NIST Guest Researcher
Tom Ryan, Ph.D. University of Michigan
Bill Strawderman, Ph.D. Rutgers University
Grace Yang, Ph.D. University of Maryland

Administrative Staff
Stephany Bailey
Lorna Buhse

3 Fundamental Statistical Metrology

Rapid advances in measurement science, technology and metrology continually present new challenges to develop appropriate statistical methodology. Taking a place side by side with scientific experimentation in science and metrology laboratories is computational experimentation or virtual measurement. Whether large scale simulations or multi-dimensional interpolation/extrapolation from extensive data bases, these synthetic or predictive measurements require extended formulations of uncertainty and new statistical tools.

Simultaneously, the external driver of international commerce requirements through the multi-nation Mutual Recognition Arrangement gives urgency to rigorous statistical methods for metrology, especially uncertainty analysis, for international intercomparisons on which to base mutual international recognition of National Metrology Laboratories' measurement capabilities.

Research in *Fundamental Statistical Metrology* focuses on providing both a unified rigorous framework and the statistical tools for implementation. Currently these efforts are concentrated on Bayesian metrology and on analysis of intercomparisons and methods for linking several intercomparisons. A major initiative in the development of a comprehensive statistical approach to computational experimentation for measurement science and metrology is just beginning. As always, the context for development of statistical methodology is the collaborative research at NIST: the specific science and technology developments already underway and the vision for metrology in the future.

3.1 Estimating Common Vector Parameters in Interlaboratory Studies

Andrew L. Rukhin
Statistical Engineering Division, ITL

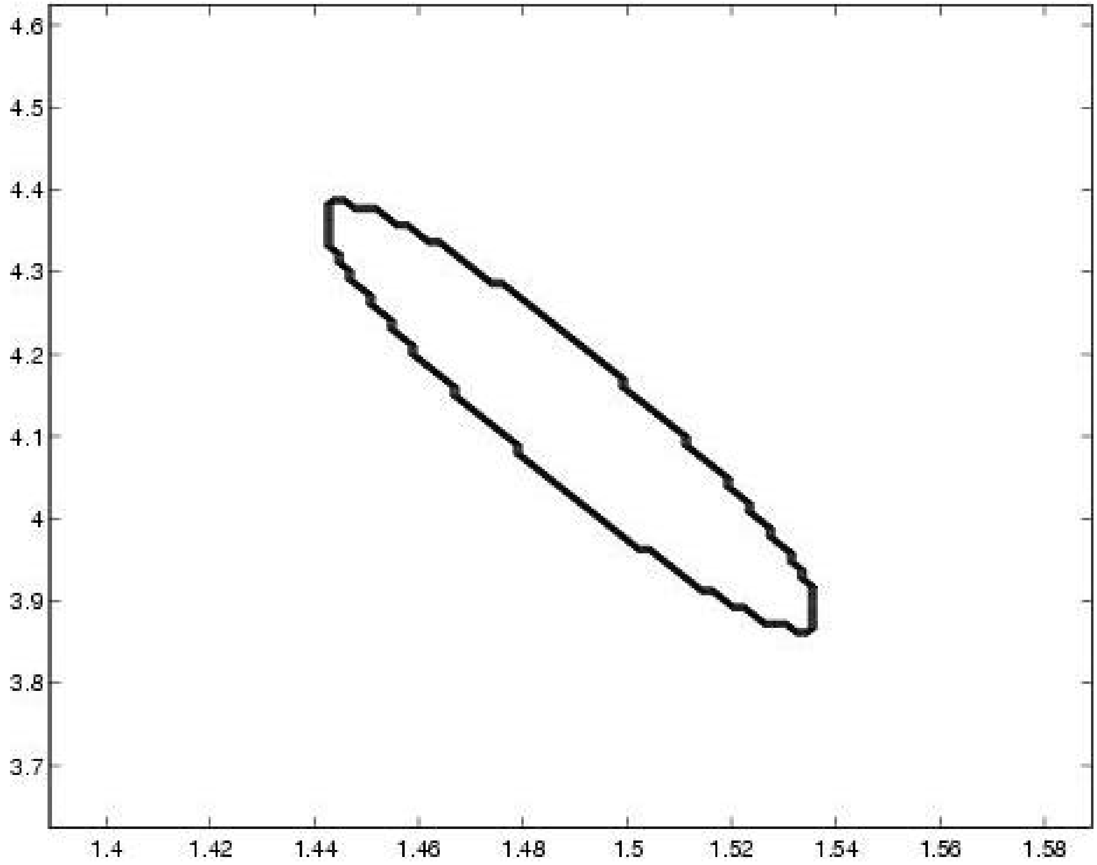


Figure 1. Confidence ellipsoid for the diffusivity and thermal heat in the thermal diffusivity example, $q = 2, p = 28,$ and $Y_{DL} = [0.084, -0277; -0277, 0.914]$.

Statistical modeling and analysis of interlaboratory comparisons pose several fundamental questions about determination of the consensus value and its associated uncertainty. An appropriate choice of statistical model can be difficult, especially when measurements are made across a range of values of a physical characteristic, i.e., the reference value is a curve or a multivariate vector.

Data sets consisting of samples of curves occur increasingly often in applications, especially in interlaboratory studies known as Key Comparisons in which the key comparison reference value (KCRV) is to be determined. The sample curves, being independent realizations of an underlying stochastic process, present common features that a researcher wants to investigate. In many applications, these features have irregular shapes that cannot be adequately captured by traditional statistical models. We derive statistical procedures for vector KCRV evaluation (a discretized version of a possibly irregular underlying curve) along with estimates of the uncertainty of this value by using a model taken from meta-analysis methodology under the assumption of Gaussian distributions and equivalent qualification of all participating laboratories. A class of matrix-weighted vector means for the KCRV and a method of assessing the uncertainty of the resulting KCRV estimates are obtained. In particular, we analyzed the estimation problem of the covariance matrix of the KCRV, approximate confidence ellipsoids are constructed for these estimators.

One of the motivating examples for this study was the Key Comparisons of accelerometers (CCAUV.V-K1, von Martens et al., 2002) that was organized to compare measurements of sinusoidal linear accelerometers over the range of frequencies from 40Hz to 5kHz. (Each accelerometer measured charge sensitivity at the specified frequencies and at different acceleration amplitudes.) Two types of accelerometers (single-ended design and back-to-back) were employed at each of twelve NMIs (including NIST), with the Physikalish-Technische Bundesanstalt, Germany serving as the coordinating laboratory. Each participating NMI reported its own laboratory means and the within lab sample covariance matrices (Type A uncertainties). The KCRV for charge sensitivity as a function of frequency is determined by our method. In the original study the KCRV was found separately for each type of accelerometer and for each specified frequency.

Another example is related to the study of Pyroceram 9606, a glass ceramic material especially suited for high temperature applications. This material is being used for performance evaluation of instruments measuring thermal properties such as thermal conductivity, thermal diffusivity, and specific heat (heat capacity). All these characteristics are temperature dependent, so the reference value must be a function of temperature. Twenty-eight thermal conductivity experiments in different countries have been performed on this material, and a consensus value was needed. Indeed, the data from different laboratories were of widely differing quality. We have used unpublished data on diffusivity and thermal heat provided by R. Zarr (NIST).

In accordance with the goals discussed, we formulated the following mathematical model in the situation where multiple (correlated) q-dimensional measurements are made by each of p laboratories. In our model, the i-th laboratory repeats its vector measurements n_i times, and the vector data $\{X_{ij}\}$ for $i = 1,\dots, p$ and $j = 1,\dots, n_i$ follow a one-way random-effects MANOVA model, which may be both unbalanced and heteroscedastic, i.e.,

$$X_{ij} = q + \ell_i + e_{ij},$$

with mutually independent $\ell_i \sim N_q(0, A)$ and $e_{ij} \sim N_q(0, B_i)$, $j = 1, \ldots, n_i$. The vector q plays the role of the common mean or the KCRV, ℓ_i is the between-laboratories effect, and the e's are the measurement errors. The unknown $q \times q$ matrix A may have rank smaller than q and q represents an unknown q-dimensional structural parameter common to all laboratories. The covariance matrices B_i and A are the nuisance parameters. The goal is to estimate the structural parametric vector q, and to provide a standard error for this estimate.

The usual estimators of the laboratory means and of their covariance matrices are $X_i = \overline{x}_i = \sum X_{ij}/n_i$, and $S_i = \sum_{j=1}^{n_i} (X_{ij} - X_i)(X_{ij} - X_i)^T/[(n_i - 1)n_i]$ with $X_i, S_i,\ i = 1, \ldots, p$, being sufficient statistics. Reduction by sufficiency to the sample means X_i and sample covariance matrices S_i makes this problem more specific.

The maximum likelihood estimator of q has the form,

$$\hat{q} = \sum_{i=1}^{p} \hat{w}_i X_i,$$

where the matrix weights $\hat{w}_i$ are found as maximizers of the likelihood function. Unfortunately, the likelihood can have many local extrems, and iterative algorithms may converge to one of them. Also, for moderate p the estimator of the variance of $\hat{q}$ obtained from the inverse of the Fisher information as well as statistics of the form, $\left[\sum_j (S_i + A)^{-1}\right]^{-1}$, underestimate the true variance of $\hat{q}$ even when $q = 1$.

For this reason, alternative, simpler procedures are desired. It makes sense to employ the available statistics S_i to approximate the within-trials covariance matrices. Thus, we restrict the class of estimators to those with matrix weights of the form

$$W_i = W_i(Y) = (S_i + Y)^{-1},$$

for some non-negative definite matrix Y so that an estimator $\tilde{X}$ of q from this class has the following representation

$$\tilde{X}_Y = \tilde{X} = W^{-} \sum_{i=1}^{p} W_i X_i = \sum_{i=1}^{p} w_i X_i,$$

with $W^{-} = \left(\sum_{i=1}^{p} W_i(Y)\right)^{-}$, being a generalized inverse of $\sum_{i=1}^{p} W_i$. Estimators of this form include the analogue of one of the traditional estimators of the common vector mean suggested by Graybill and Deal in the case $q = 1$,

$$\tilde{X}_0 = \left[\sum_{i=1}^{p} S_i^{-1} \right]^{-} \sum_{i=1}^{p} S_i^{-1} X_i ,$$

and also the sample mean

$$\tilde{X}_\infty = \frac{1}{p} \sum_{i=1}^{p} X_i .$$

We suggest two different methods of choosing the matrix Y by extending the procedure suggested by DerSimonian and Laird (1986) when $q = 1$ and the algorithm suggested by Mandel and Paule (1970). We demonstrate the relationship of these estimators to the restricted maximum likelihood estimator and to the maximum likelihood estimator. An estimator of the covariance matrix of these statistics similar to the one suggested in a more general setting of linear models by Horn, Horn and Duncan (1975) is also put forward.

The results of a Monte Carlo simulation study confirmed a good approximation of the pivotal ratio by an F-distribution. To illustrate the techniques, we implemented them in the mentioned accelerometers key comparisons study (CCAUV.V-K1) and in interlaboratory study of thermal diffusivity and conductivity. As in the simulation study, the Mandel-Paule procedure and the DerSimonian-Laird method both gave the same answer, which practically coincides with the PTB solution given by von Martens et al. (2002). The figure shows the confidence ellipsoid for the diffusivity and thermal heat in the thermal diffusivity study.

The suggested matrix-weighted vector means are useful for the vector KCRV estimation. The method of assessing the uncertainty of these estimates provides joint confidence ellipsoids for the parameters involved.

3.2 Linear Statistical Models with Type B Uncertainty: A Bayesian View of Annex H.5 of the *Guide to the Expression of Uncertainty in Measurement*

Blaza Toman
Statistical Engineering Division, ITL

Annex H.5 of the *Guide to the Expression of Uncertainty in Measurement* presents a class of statistical models and analysis techniques that are commonly called the Analysis of Variance (ANOVA). These models are useful for accounting for the effects of factors that cause the measurand in an experiment to change over time or over experimental conditions. The proposed procedures assume that the observations are not subject to type B uncertainties. A natural question then is: Can these models be used in the presence of type B uncertainties? This article answers the question in the affirmative and provides a natural interpretation of the results. The example data from the Annex is used for an illustration.

Annex H.5 of the *Guide* uses the following example to illustrate the use of ANOVA methods.

Example :
A 10V Zener voltage standard is calibrated against a stable voltage reference over a two-week period. On each of J days during the period, K independent repeated observations of the potential difference of the standard are made. Denote the observations by v_{jk}. The following table, reproduced from the Annex, contains the summarized data.

Day	Daily mean $\bar{v}_{j.}$ in V	Daily std s_j in μV
1	10000.172	60
2	10000.116	77
3	10000.013	111
4	10000.144	101
5	10000.106	67
6	10000.031	93
7	10000.060	80
8	10000.125	73
9	10000.163	88
10	10000.041	86

Table 1. Summary of the voltage standard calibration data from Table H.9 of Annex H.5 of the *Guide*.

The classical ANOVA model used in the Annex makes the assumption that the observations v_{jk} are realizations of random variables V_{jk} which can be represented by the equation

$$V_{jk} = m + a_j + e_{jk} \tag{1}$$

where m is an unknown constant (measurand), the effects of day, a_j , j=1,...,J , are random variables, and the e_{jk} are random variables representing the within-day variability. The a_j are assumed to be independently normally distributed with mean 0 and standard deviation s_a and the e_{jk} are independently normally distributed with mean 0 and standard deviation s. Model (1) can also be written in a hierarchical representation as

$$\begin{aligned} &V_{jk} | q_j, s^2 \sim N\left(q_j, s^2\right) \\ &q_j \,|\, m, s_a^2 \sim N\left(m, s_a^2\right). \end{aligned} \qquad (2)$$

The notation, $q_j \,|\, m, s_a^2$ represents conditioning. That is, the probability distribution of q_j, given m and s_a^2, is normal with mean m and variance s_a^2. In this representation, the first (or top) level of the hierarchy accounts for the within-day variability and the second level accounts for the between days variability. The two representations differ only in notation, leading to an identical analysis. The test of significance for the day effect is an F-test of the hypothesis that $s_a = 0$. The test in essence compares the size of an estimate of s_a with the size of an estimate of s . For the above data, the hypothesis test concludes that the day effect is statistically significant at the 5% level.

The *value* of the experiment, that is an estimate of m, is the grand mean $\bar{V}_{..}$. The *standard uncertainty* (as defined in the *Guide*) based on this model is an estimate of

$$\sqrt{\frac{1}{J}\left(\frac{s^2}{K} + s_a^2\right)}.$$

For the data given in the above table, the Annex produces the quantities value = 10000.097 V and uncertainty = 18 μV, giving 9 degrees of freedom for the uncertainty estimate. Note that when no day effect is assumed, the value is still $\bar{V}_{..}$ but the uncertainty is now an estimate of $s/\sqrt{JK}$ and for the above data this gives uncertainty = 13 μV with 49 degrees of freedom.

The analysis based on this model does not take into account any sources of type B uncertainty and the Annex specifically states, that such uncertainties must be negligible or "must be of a type that can be taken into account at the end of the analysis". To be of such a type, the size of the uncertainties should be the same for various days. This is a strong restriction, which is likely to make the usual classical ANOVA analysis unsuitable for use in many applications. A Bayesian approach for such cases is suitable, as it is fully capable of modeling the type B uncertainty.

Under the Bayesian paradigm, parameters of the model such as m are given state-of-belief prior probability distributions which are combined with the data in the form of the likelihood function via Bayes theorem to obtain a posterior distribution of the parameter. This distribution summarizes all knowledge about the parameter available after the data are collected, and if the parameter is a measurand, is used to obtain quantities such as the *value* (posterior mean), the *standard uncertainty* (the posterior standard deviation) and a probability interval.

The experiment given in the Example can be represented by the following Bayesian model, written in the hierarchical form, which allows for day effects as well as for type B uncertainty sources. Specifically, let for *j*=1,…,10, *k*=1,…,5

$$\begin{gathered} V_{jk} \,|\, q_j, s_j^{\;2} \sim N(q_j, s_j^{\;2}) \\ q_j \,|\, d_j, t_j^{\;2} \sim N(d_j, t_j^{\;2}) \\ d_j \,|\, m, s_a^2 \sim N(m, s_a^2) \\ s_a \sim U(0, c) \\ m \sim N\left(m, w^2\right). \end{gathered} \tag{3}$$

The first level in this model accounts for within day variability as in the classical ANOVA, and is the likelihood function. In the Bayesian model we do not require that $s_j = s \;\; \forall j$ as in the classical model, that is, different days can have different within day variability. This is a useful relaxation of assumptions permitted by the flexibility of the Bayesian paradigm. The second level of the hierarchy accounts for the type B uncertainty. That is, there is uncertainty in the day means q_j that is represented by a normal distribution with standard deviation equal to t_j, the size of the type B uncertainty. The form of this distribution is usually determined subjectively by the experimenter. The third level of the hierarchy accounts for the variability between days just as in model (2). The last two distributions represent the experimenter's prior knowledge about m and s_a. Usually, such knowledge is very limited, a fact that in this model would be denoted by letting the value of *c* be large and by letting $w^2 \to \infty$. Such assumptions on the prior distributions allow the likelihood function to dominate, resulting in a posterior distribution that depends mostly on the data.

Applying Bayes theorem results in posterior distributions of the various parameters of the model, namely for the μ, the s_a and the d_j. Unfortunately, the integrals needed to apply Bayes theorem to this model are not of a form that is readily evaluated in a closed form. Numerical methods must be employed. The simplest procedure is to apply Markov Chain Monte Carlo (MCMC) methods using the free software WinBUGS.

To illustrate the method, consider the data in Table 1, and assume that the values in the third column are the <u>combined</u> <u>standard uncertainties</u> instead of the simple standard deviations and further assume that type B uncertainty accounts for 40% of the combined standard uncertainty. Performing the Bayesian analysis yields an estimate for the *value* (the posterior mean of m) as 10000.097V with the *standard uncertainty* (posterior standard deviation of m) of 22μV.

This result can be compared to the classical ANOVA result; that is, the *value* of 10000.097V with *uncertainty* of 18μV. A different comparison can be done by analyzing model (2) using Bayesian methods. That is, using model (3) with type B uncertainty set at 0. This produces a *value* of 10000.097V and *uncertainty* of 19μV. This shows that Bayesian ANOVA can give results very similar to those obtained with classical ANOVA when a noninformative prior distribution is used.

As in classical ANOVA, it is possible here to test whether there is a difference in the daily mean voltages or whether the simpler model

$$
\begin{aligned}
V_{jk} \mid q_j, s_j^{\ 2} &\sim N(q_j, s_j^{\ 2}) \\
q_j \mid m, t_j^{\ 2} &\sim N(m, t_j^{\ 2})
\end{aligned}
\qquad (4)
$$

fits well. Under the Bayesian paradigm, such a test can be performed using so-called posterior predictive p-values. Namely, it is possible to obtain a predictive distribution of the $\overline{V}_j$, that is, the probability distribution of the sample day average given the observed data, under model (3) and under model (4). Using the predictive distribution, it is now possible to compute $p_j = P\left(\overline{V}_j \textit{ is more extreme than } \overline{v}_j\right)$, the posterior predictive p-value, for each j. A small value of p_j indicates that for that particular *j*, the model is not likely to be true given the observed data. If majority of the p_j are small then the model is not appropriate.
Consider the results for the data given in Table 1.

Day j	Predictive p-value model (3)	Predictive p-value model (4)
1	0.143	0.012
2	0.441	0.412
3	0.219	0.031
4	0.348	0.213
5	0.488	0.491
6	0.233	0.042
7	0.294	0.116
8	0.393	0.302
9	0.261	0.091
10	0.244	0.047

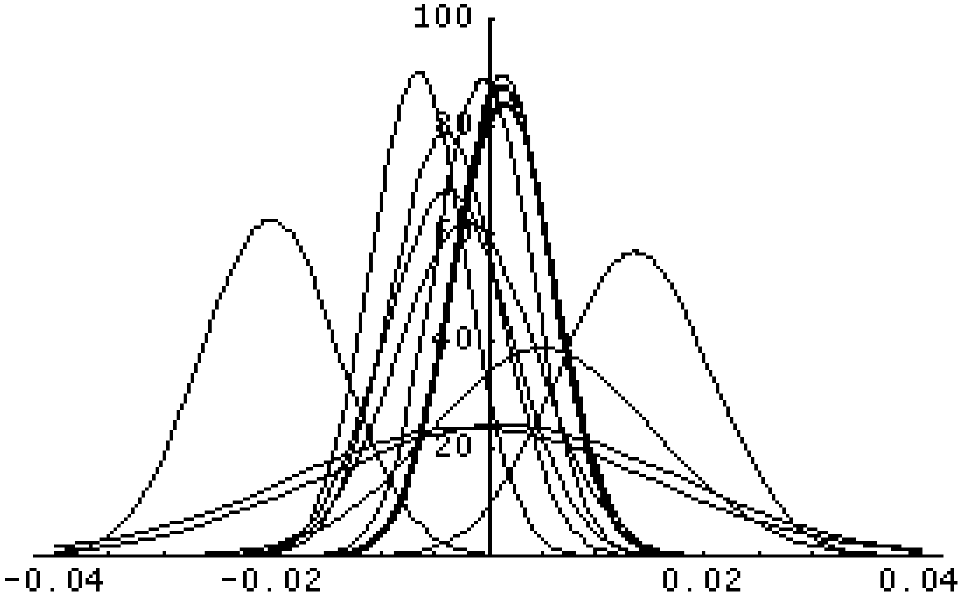
Table 2. Predictive p-values

It appears that if we take “small” to mean less than 0.05, then four out of the ten days appear to have significantly different means. This is a large enough number that model (4) should be judged not appropriate. Note that this is the same decision as is that made using the classical ANOVA F-test. The full model, i.e., model (3), appears to fit quite well.

The One-way ANOVA model discussed in Annex H.5 of the Guide *can be used in the presence of type B uncertainty. This requires state-of-knowledge probability distributions and is best done in a Bayesian framework. It is shown here that there are Bayesian procedures which are similar to the usual hypothesis testing in ANOVA models and that very similar numerical answers can be obtained. It is important to note, however, that the assumptions underlying the two methods and the interpretation of the results are quite different. The Bayesian methodology presented here can be easily extended to other linear models such as regression.*

3.3 Bayesian Models for Key Comparison Analysis

Blaza Toman
Statistical Engineering Division, ITL



Probability distributions of the individual laboratories' measurands with the distribution of the KCRV.

Key Comparison experiments pose numerous challenges to the statistical analyst. Most importantly, Type B uncertainty must be included in the statistical model in a meaningful way that satisfies both the scientist and the statistician. Furthermore, the scientific objectives of the experiment must be reflected in the statistical summaries and these are generally required to conform to the definitions of the Guide to the Expression of Uncertainty in Measurement. This article summarizes a Bayesian approach to this analysis.

Key Comparison experiments can be roughly divided into two categories, depending on whether they are designed to have a common measurand for all laboratories, or multiple measurands, with each laboratory having its own measurand. The objectives of the statistical analysis are different for the two categories. In single measurand experiments, a Key Comparison Reference Value (KCRV) and laboratory-to-laboratory comparisons generally have a straightforward physical interpretation. It is possible, however, that the measurand is unstable and changing between measurements. Thus the statistical model must be able to account for such changes. In multiple measurand experiments, the definition of a KCRV is not straightforward. The statistical model must provide a meaningful interpretation of such a summary.

The following two statistical models form the basis of a Bayesian Key Comparison analysis.

Multiple Means Model
A Key Comparison experiment is a multi-laboratory study. If we treat all laboratories' data totally independently from each other, that is, if we assume that there are no relationships between the measurands or the uncertainties of the various laboratories, we can proceed as follows. For $i = 1,\ldots,k$, where k is the number of laboratories, we have

$$\begin{aligned} Y_i \mid q_i, s_i^2 &\sim N(q_i, s_i^2) \\ q_i \mid m_i, t_i^2 &\sim N(m_i, t_i^2) \\ m_i \mid m_i, w^2 &\sim N(m_i, w^2) \end{aligned} \qquad (1)$$

Stage one in the hierarchy is used to quantify the usual sampling variability. Stage two represents the Type B uncertainty. Stage three gives a prior distribution on the measurand and possibly on s^2. In a Key Comparison experiment, it is generally not possible to use an informative prior distribution on m, so allowing $w^2 \rightarrow \infty$ leads to an approximate posterior distribution

$$m_i \mid y_i, t_i^2 \sim N\left(y_i, t_i^2 + s_i^2\right) \qquad (2)$$

and so the posterior mean and posterior standard deviation of each laboratory are approximately the ISO guide recommended quantities, the *value* and the *standard uncertainty*.

Single Mean Model

Suppose now that there is a common measurand. Model (1) can be modified to reflect this fact:

$$\begin{aligned} Y_i \mid q_i, s_i^{\,2} &\sim N(q_i, s_i^{\,2}) \\ q_i \mid m_i, t_i^{\,2} &\sim N(m_i, t_i^{\,2}) \\ m_i \mid m, g^{\,2} &\sim N(m, g^{\,2}) \end{aligned} \qquad (3)$$

$$\begin{aligned} m &\sim N\left(m, w^2\right) \\ g &\sim U\left(0, 100\right). \end{aligned}$$

The prior distributions on the m_i are now hierarchical, the common mean μ being the measurand of the entire experiment. The quantity $g^{\,2}$ determines the strength of the relationship between the laboratories. If $g^{\,2}$ = 0, then all laboratories are truly measuring the same quantity μ. In such a case, all of the data are pooled to obtain an approximate posterior mean and posterior variance

$$m_p \;= \frac{\sum_1^k y_i \left(t_i^{\,2} + s_i^2\right)^{-1}}{\sum_1^k \left(t_i^{\,2} + s_i^2\right)^{-1}} \qquad (4)$$

$$w_p = \frac{1}{\sum_1^k \left(t_i^{\,2} + s_i^2\right)^{-1}}. \qquad (5)$$

For other values of $g^{\,2}$, its size controls the pooling of the data in the estimation of the m_i. The larger the size, the less pooling there is. Another interesting property of this model is that $g^{\,2}$ is essentially an additional Type B uncertainty term, one common to the laboratories. When the value of γ is not specified, and a prior distribution on it is introduced as above, it is in fact estimable from the data. This is useful, as many sources of Type B uncertainty are truly random laboratory effects.

Next, consider the analysis of a single measurand experiment. Because of the dual purpose of the Key Comparison experiment, that is estimation of the common measurand and estimation of lab-to-lab differences, the choice of model for the analysis is not completely straightforward. The problem arises from the fact that there is usually not a complete agreement among the scientists as to the accuracy of everyone's Type B uncertainty estimates and that the form of analysis has to be agreed upon by all participants.

First, consider using model (3). In this case, the meaning of the KCRV is clear and so is its estimation. It is plainly an estimate of the common measurand μ and can be provided by the mean of the posterior distribution. The uncertainty of the KCRV is the posterior standard deviation of μ. The fact that this model accounts for the additional Type B uncertainty provided by the γ term is an advantage for KCRV estimation as many participants do not fully accept their partners' estimates of Type B uncertainty. The best method of comparison between the laboratories is not so clear however. The problem is that the posterior means of the m_i are likely to be quite different from the y_i. Similarly, the posterior standard deviations of the m_i are also likely to be different from the ISO Guide recommended values of $\sqrt{t_i^2 + s_i^2}$. This would generally not be acceptable to the scientists in a Key Comparison experiment as each laboratory wants to rely solely on their own data (and their own estimates of Type B uncertainty) in the estimation of their m_i. Thus a compromise solution would be to use model (3) for KCRV estimation and model (1) for lab-to-lab comparisons.

Next, consider an experiment where each laboratory has its own measurand. In such a case, model (1) would be used. Laboratory-to-laboratory differences could be measured by the differences in the m_i. In some situations a KCRV will be requested. The question then is how to estimate it since it has no natural interpretation. One possible solution is the following method, based on the so-called *linear opinion pool*, which dates back to Laplace. In this method, *k* probability distributions $p_i(\ )$ are combined as

$$p(\ ) = \sum_{i=1}^{k} w_i p_i(\ ) \qquad (6)$$

where the weights w_i add up to one. In the present application, the *k* laboratories posterior distributions for m_i could be combined into the mixture distribution of a new random variable *m*,

$$f(m) = \sum_{i=1}^{k} w_i f\left(y_i, t_i^2 + s_i^2\right) \qquad (7)$$

where $f(\ )$ is the normal density. In most cases, the weights would be taken to be $1\!/\!_k$, representing a view that the laboratories' data are of equal quality. The mean of this distribution, that is, the average of the *k* laboratories' measurements, would be taken as the KCRV. The standard deviation of this distribution

$$u(\bar{y}) = \sqrt{\frac{\sum_i u^2(y_i)}{k} + \frac{\sum_i \left(y_i - \bar{y}\right)^2}{k}} \qquad (8)$$

is the standard uncertainty of the KCRV. The linear opinion pool is an easily understood and easily performed method. The $u(\bar{y})$ can be thought to represent the total variability in the population of measurands of the Key Comparison. This can be viewed as the true measure of uncertainty in such a Key Comparison because of the assumed equality of the laboratories in terms of their competence.

Bayesian models form a flexible basis for Key Comparison experiments analysis. They allow sensible representation of Type B uncertainty and result in analyses that can conform to the requirements of the ISO Guide.

3.4 A Note on the Fallacy of Certain 2-Sigma and 3-Sigma Outlier Rejection Rules in Key and Interlaboratory Comparisons

James J. Filliben
Statistical Engineering Division, ITL

Charles Guttman
Polymers Division, MSEL

Robert Kelly
Analytical Chemistry Division, CSTL

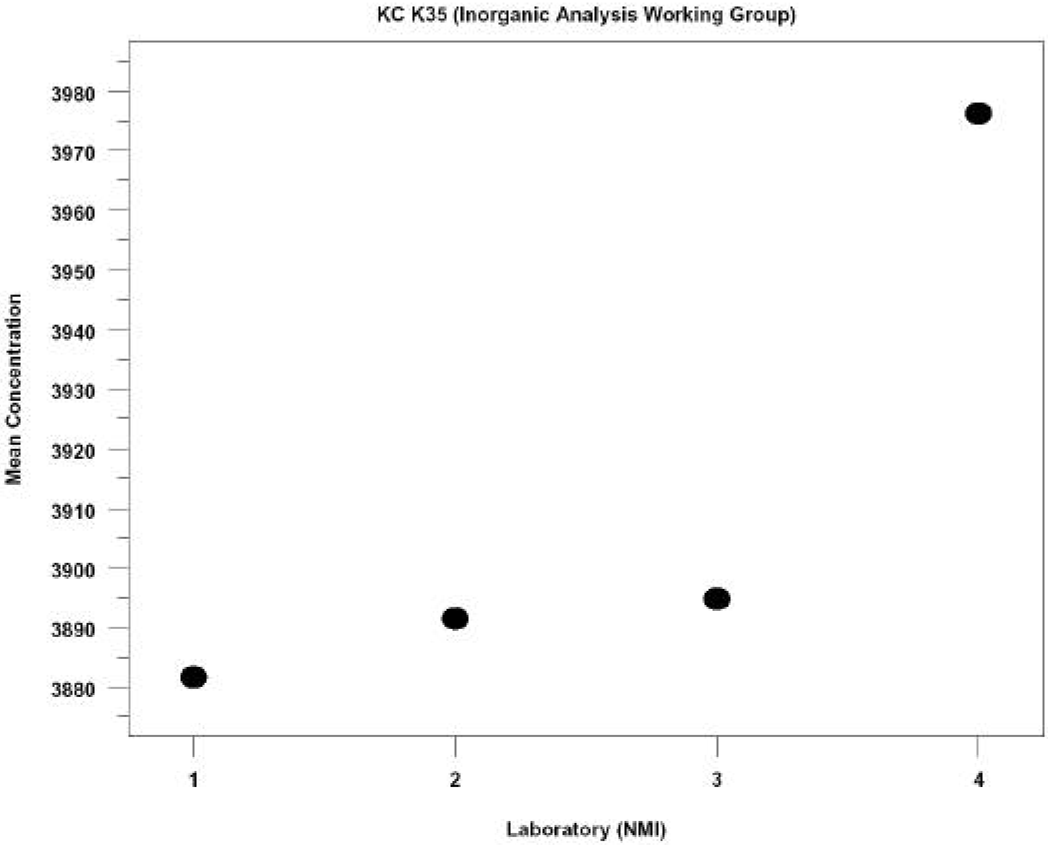


Figure 1. Representative output from a 4-laboratory Key Comparison

In accord with its worldwide reputation for providing state-of-the-art chemical measurements of broad application, the NIST Analytical Chemistry Division (ACD) participates in a wide variety of interlab and key comparison (KC) experiments. One such KC was in connection with CCQM-K35: Consultative Committee for Amount of Substance (Metrology in Chemistry) and K35: Low Sulfur in Fuel. NIST is currently the pilot laboratory in this ongoing KC. Bob Kelly is an active member of this CCQM Inorganic Working Group. This note flowed out of data analysis carried out by Bob Kelly (and confirmed by SED) on such inorganic KC data.

The central question that this note deals with is "When may a participating laboratory in a key comparison be declared an 'outlier'?" Frequently in KCs, the set of participating labs is quite heterogeneous in nature. Given that, at what point does a lab become so "different" that it is deemed unworthy of inclusion in the final analysis--especially in connection with the computation of KCRVs (Key Comparison Reference Values)?

Declaring a laboratory an outlier is a special case of the larger KC problem of how to separate participating labs into 2 categories: those that are "acceptable" and those that are "outliers" (relative to the acceptable group). There is a myriad of statistical outlier tests that may be brought to bear to assist in this partitioning. This note deals with one class of such tests, and points out anomalous properties of the tests that would result in inappropriate KC conclusions.

NIST is serving as the pilot laboratory for the CCQM-K35 (Low-level Sulfur in Fuel) Key Comparison. This KC had 4 participating laboratories. Looking at simulated--but representative--mean sulfur concentrations per lab as presented in Figure 1, the question naturally arises as to whether laboratory 4 should be identified as an "outlier". The corollary sub-question in this regard is whether laboratory 4 should be excluded from consideration when computing the KCRV (Key Comparison Reference Value)? This was the question that Bob Kelly of the Analytical Chemistry Division was faced with while doing the data analysis in connection with NIST's pilot laboratory role for the Low Sulfur in Fuel KC. While dealing with this question, Bob uncovered some interesting characteristics of a commonly employed outlier-identification test, and SED’s involvement was to verify Bob's conclusions and offer some additional considerations. The developed conclusions apply to KCs and Interlab Comparisons in general.

A common outlier statistic that could be applied to the data in Figure 1 would be the following standardized deviation:

$$(y^* - \text{ybar})/s$$

where y* is the outlying observation in question, ybar is the sample mean (of all of the data), and s is the sample standard deviation (of all of the data). In the statistics community, this statistic is known as the Grubbs statistic. The value of this Grubbs statistic is, of course, the number of sample standard deviations that the observation in question is away from the sample mean.

Let us consider a few simple outlier detection "tests/rules". One such non-rigorous, but commonly employed (in some circles), outlier-detection rule is the "2s rule" (with rough connotations to a 5% error rate):

> "Conclude that y* is an outlier if y* is beyond 2
> sample standard deviations from the sample mean."

What would this rule yield for the pseudo-data (3882, 3892, 3895, 3976) at hand?

The data yield the following summary statistics:

y* = 3976
ybar = 3911.25
s = 43.523
Grubbs = 1.48772

that is, the extreme-appearing observation (3976) is 1.48772 sample standard deviations from the sample mean. Applying the above decision rule of rejecting a lab only if it is beyond 2 standard deviations would thus lead to the conclusion that y* is not an outlier.

This conclusion is suspect at best. The subsidiary question arises in the use of this statistic as to how big the statistic can get under any circumstances, or equivalently, what is the maximum number of sample standard deviations that an observation--any observation--can be offset from the sample mean? Surprisingly, for small numbers of observations (laboratories), this maximum value is very small. In fact, it can be shown that this maximum value can be computed for n = 4 observations (labs) by considering the following simple "worst-case" prototype data set: 0, 0, 0, 1. This data set has the following summary statistics:

y* = 1
ybar = .25
s = .50
Grubbs = 1.5

and so for 4 observations (labs), no observation (lab) will ever be more than 1.5 sample standard deviations from the sample mean. Thus for a 4-lab key comparison, if the 2s rule is utilized, it will be impossible to ever reject a lab as an outlier. Thus, what appears graphically obvious in Figure 1, namely, that lab 4 is "different" from the other 3 labs, would be quantitatively impossible to substantiate by applying the simplistic 2s rule.

What happens to the 2s rule for sample sizes other than 4? At what sample size (if any) does the 2s rule make sense--that is, have some probabilistic justification? To answer this question, see Figure 2, which presents the maximal value of the Grubbs statistic (or equivalently, the maximum number of sample standard deviations any observation can ever be away from the sample mean) as a function of the number of labs n.

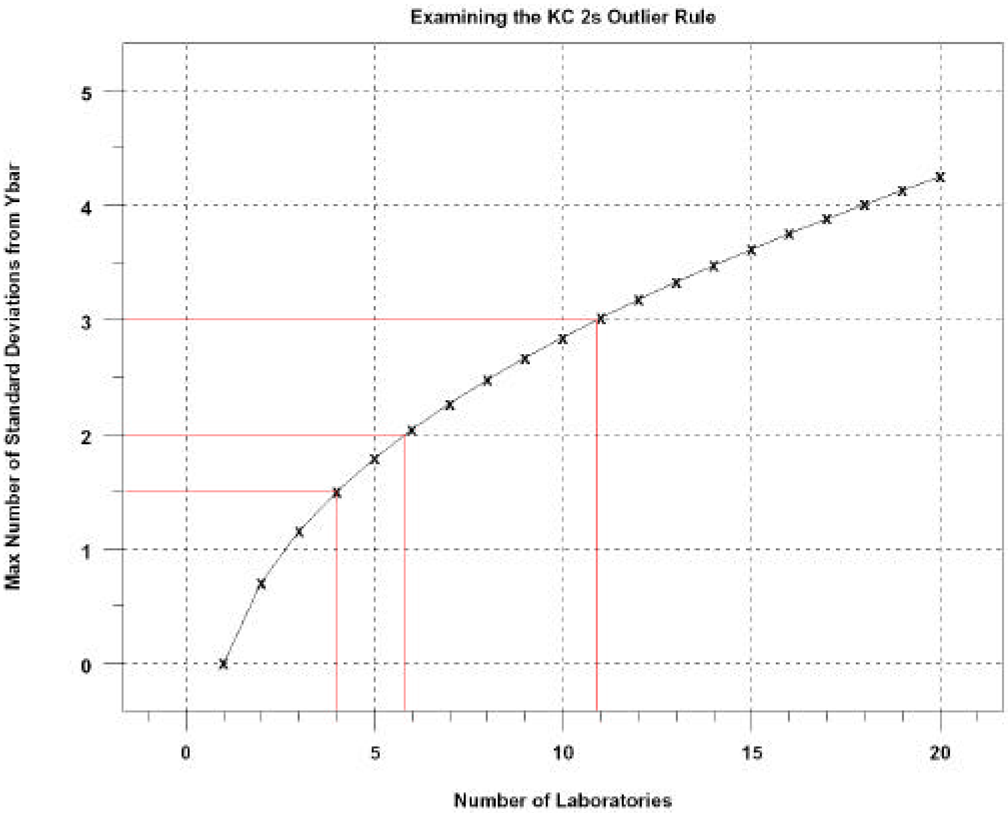


Figure 2. Theoretical maximum of the standardized deviation.

How was the above maximum curve determined? It can be shown--with some elementary calculations similar to the previous n = 4 case--that the general form for the maximum (in units of sample standard deviations) deviation from the sample mean is given by

max (y*-ybar)/s) = ((n-1)/n)*sqrt(n)

As a tribute to the inquisitiveness of the NIST scientist, the above result was derived by Bob Kelly in his consideration about his KC outlier problem. After-the-fact, it was jointly thought that this result must certainly be existent somewhere in the vast statistics literature. Sure enough it was, with the most distant reference implicit in a Pearson paper in 1939, and the most recent reference a Ron Shiffler Technometrics note in 1988. Be that as it may, the application and implication of such a result in the modern-day context of key and interlaboratory comparisons is worth the intellectual revisit.

The maximum curve as presented in Figure 2 is very revealing. It shows that for key and interlab comparisons, the value of 2 is not even possible unless the number of labs is at least 6, and hence, based on the 2s rule, no key or interlab study will ever reject any lab if the number of participating labs in the study is 5 or less.

Similarly, if we formulate a second outlier detection rule: a 3s rule (which is even more "conservative" with rough connotations to a 1% error rate):

> "Conclude that y* is an outlier if y* is beyond 3
> sample standard deviations from the sample mean."

then from Figure 2, no key comparison or interlab experiment will ever reject any laboratory--no matter how "bad" looking--if the number of participating labs is between 1 and 10.

These mathematical limits, though elementary in nature, are nonetheless not well-known, nor are the implications of these limits well-appreciated in the context of outlier detection in key and interlab comparisons. In most KCs, the number of participating labs is very frequently less than 10, and is on occasion as small as 3 or 4 (as with Bob Kelly's Low Sulfur in Fuel problem), and so the somewhat common "rule of thumb" about using 2 (or 3) standard deviations as an outlier cutoff is totally misleading and entirely inappropriate.

Clearly other outlier rules should be utilized. There are a variety of such rules, but the simplest extension to the standardized deviation statistic at hand is to set aside the 2s and 3s considerations and use the more rigorously derived Grubbs distribution for this same standardized deviation statistic. The distribution of this statistic was formally derived by Frank Grubbs in a series of classic papers (1950, 1969, and 1972). Given that (and noting the assumption of normality), one could formulate a third outlier detection rule--which has statistical rigor over rules 1 and 2:

> Conclude that y* is an outlier if y* is beyond t
> sample standard deviations from the sample mean

where (for y* > ybar), t is the 95 percent point of the modified t distribution as derived by Grubbs. For n = 4 labs, the critical values of the Grubbs statistic are as follows:

95% Point	1.48125
99% Point	1.49625

and so for the original problem at hand, recalling that the value of the standardized deviation for the original data points (3882, 3892, 3895, and 3976) was 1.48772, we would compare 1.48772 with the Grubbs critical values and conclude that lab 4 with the value 3976 is an outlier at the 5% level, but not at the 1% level. The tabulated Grubbs tables are of course fine, but they would be a bit more practically illuminating if they also presented the 100% Point (the theoretical max) which is easily computable as shown in Figure 2. Thus the above critical value for the 4-lab Grubbs statistic might be pr ofitably augmented as:

95% Point	1.48125
99% Point	1.49625
100% Point	1.50000

and this would thus "calibrate" the analyst to the fact that a finite upper bound does exist (a useful fact all by itself), and so expecting the larger and more traditional values such as 2 and 3 are not just a low-probability event, but rather are a zero-probability event. Table 1 below presents such an augmented version of the Grubbs percent points for the number of labs n = 3 to 20.

Number of Labs	0%	50%	75%	90%	95%	99%	100%
3	0	1.115355	1.144822	1.153118	1.154305	1.154685	1.154701
4	0	1.3125	1.40625	1.4625	1.48125	1.49625	1.5
5	0	1.440714	1.571221	1.671386	1.715037	1.763679	1.788854
6	0	1.539087	1.690863	1.82212	1.887145	1.972817	2.041241
7	0	1.619517	1.784803	1.938135	2.019968	2.139105	2.267787
8	0	1.687583	1.862097	2.031651	2.126645	2.274364	2.474874
9	0	1.746501	1.92763	2.109448	2.214809	2.386316	2.666667
10	0	1.798399	1.984508	2.175992	2.289819	2.481723	2.84605
11	0	1.84469	2.034635	2.233855	2.354635	2.56386	3.015114
12	0	1.886416	2.079384	2.284915	2.411488	2.63553	3.175426
13	0	1.924355	2.119751	2.330511	2.461981	2.698821	3.328201
14	0	1.959105	2.156475	2.371631	2.507284	2.755256	3.474396
15	0	1.991132	2.19013	2.409022	2.548278	2.806027	3.614784
16	0	2.020809	2.221162	2.443258	2.585656	2.852007	3.75
17	0	2.048441	2.249928	2.474799	2.619948	2.893973	3.88057
18	0	2.074275	2.276721	2.504011	2.651582	2.932449	4.006938
19	0	2.098518	2.301776	2.531184	2.680922	2.967917	4.129484
20	0	2.121345	2.325292	2.556576	2.708231	3.000786	4.248529

Table 1. Percent Points of the standardized deviation statistic
(the Grubbs statistic): (y* - ybar) / s

Note how little the difference is between the 95 % point of the statistic and the absolute maximum (= the 100% point) when the number of labs, n, is small (3 to 5, say).

In summary, the 2s and 3s rules have severe bounds and limitations, which can easily lead to incorrect KC/interlab conclusions when the number of participating labs is small. The Grubbs critical values are a far superior way to proceed in such cases.

The problem of outlier identification in key and interlab comparisons is a serious one that requires thoughtful consideration. There are many possible statistical tests and approaches that could (and should) be used in the determination of whether a lab is an outlier or not. The most important non-consideration is the sobering fact that just because a lab is "different" does not make a lab wrong. Statistically speaking, it is relatively easy to test for "different"; it is much harder to test for "wrong". The fact of the matter is that the "outlying" lab could be correct and all the "conforming" labs incorrect, or possibly all of the labs are incorrect and the absolute truth lies elsewhere. Given these caveats, key and interlab comparisons still serve a unique and irreplaceable "cross-calibration" role in the larger scientific community.

The crux of the above note deals with the less-than-obvious upper bounds on the commonly-used outlier statistic: the standardized deviation. This note points out the sheer impossibility of rejecting labs--as bad as they may be--when one uses the 2s and 3s rules and the number of participating labs is small. Since a small number of participating labs is the rule rather than the exception in practice, this precautionary note will serve to force general key and interlab analyses to set aside the 2s and 3s rules, and make heavier use of more rigorous and more appropriate tests for outlying labs. Such insights into the small-sample limitations and inappropriateness of simplistic rules have served to maintain NIST/ACD's leadership role in serving as pilot labs for chemistry metrology KCs in general, and for the Low Sulfur in Fuel KCs in particular.

3.5 Comments on GUM Supplement 1

Nien Fan Zhang, Will Guthrie, Hung-kung Liu, John Lu, Andrew Rukhin,
Blaza Toman, and Jack Wang
Statistical Engineering Division, ITL

NIST has been requested to review and comment on the draft of Supplement 1: Numerical Methods for the Propagation of Distributions (GUM Supplement 1) to the Guide to the Expression of Uncertainty in Measurement (GUM), prepared by the Joint Committee for Guides in Metrology (JCGM) of the International Bureau for Weights and Measures (BIPM).

The supplemental guide's intention was to generalize the law of propagation of uncertainty as given in the GUM to the propagation of probability distributions through a model of measurement with multiple input quantities and a single output quantity. The proposed methods of distribution propagation are based on Monte Carlo simulation.

The Statistical Engineering Division has made the following general comments. A separate summary of more specific comments is also available.

1. Overview

The draft of Supplement 1 of the GUM [1] is a laudable effort to develop and describe a method for uncertainty propagation using probability distributions of the input variables of a measurement equation. The method can be described as follows:

The Supplement's Method

1. The quantity of interest is function f( ) of input quantities $X_1,\ldots,X_N$, that is $Y = f(X_1,\ldots,X_N)$.

2. It is necessary to obtain value, uncertainty and coverage interval for Y.

3. Any relevant information, including any available measurements $x_{i1},\ldots,x_{in_i}$, for i=1,…,N, which are realizations of random variables with means $X_1,\ldots,X_N$, is to be used to construct a distribution of the $X_1,\ldots,X_N$.

4. Monte Carlo samples, $x_{ib_1},\ldots x_{ib_{n_i}}$, for i=1,…,N are generated.
5. The $y_j = f(x_{1b_j},\ldots,x_{Nb_j})$, j = 1,…,M, are computed.
6. The quantity $\bar{y}$ is used to estimate the value.
7. The standard deviation of the y_i's is used to estimate the uncertainty.
8. The quantiles of the y_i's are used to estimate the coverage interval.

The Supplement's method differs from the GUM method in two major ways. First of all, the value is an estimate of the E(Y) rather than an estimate of $f(E(X_1),\ldots,E(X_N))$. For example if

$Y=X_1/X_2$ then the Supplement uses $\frac{1}{M}\sum_{j=1}^{M}\frac{x_{1j}}{x_{2j}}$ for the value and the GUM uses $\frac{\overline{x}_1}{\overline{x}_2}$. Secondly, the uncertainty of Y is computed using the entire distributions of the input variables in the Supplement's method rather than just the uncertainties as in the GUM.

Clearly, when the probability distributions of the input variables are correctly assigned, the Supplement's method is preferable to the GUM as it uses more available information and is thus more exact.

The problem is that it may be extremely difficult to assign the distributions of the Xs (or that the distributions may be totally subjective or arbitrary) and that if the distributions used in the Monte Carlo uncertainty analysis are not properly chosen, the statistical properties of this analysis are likely to be far worse than the corresponding analysis by propagation of uncertainty. Indeed, the latter method enjoys a certain level of robustness based on lack of distributional assumptions.

Thus Section 4 of the Supplement is key to the entire methodology. In our view, it needs to be rewritten to be much more educational. In particular, it has to provide clear instructions as to how a probability distribution is to be constructed.

When related measurements for a particular input variable are available, two different statistical paradigms, the frequentist and Bayesian approaches, can be used to obtain a distribution. Under the Bayesian approach, the distribution will be for the input variable itself, while under the frequentist approach, the distribution will be for an estimator of the input quantity. Sections 3 and 4 of this document address potential problems with the Supplement method separately from these two points of view.

Since using the methodology of Supplement 1 in practice may prove to be quite challenging, it is our opinion that the set of examples should be expanded, possibly to contain the examples from the GUM.

2. Motivation for GUM Supplement 1

The motivation for the GUM Supplement 1 as posed in the Introduction is cogent, and the problems addressed are important: to provide solutions when the tacit assumptions underlying computations presented in the GUM cannot be met. Specific cases in point are when the measurement model is complicated, when the input quantities themselves have large uncertainties, or when a symmetric (often, Gaussian) probability distribution cannot be presupposed for the measurement models.

The GUM Supplement 1 proposes a particular alternative; the first question is when this alternative is a suitable one; the second question is what improvements might be required and/or whether there are better practical alternatives already available.

The Monte Carlo simulation that is proposed for the computation of uncertainty in all these cases is one approach, with a better chance for success in some cases than in others. In each of these cases alternatives are available, and in every case there are caveats for the proper use of the Monte Carlo methods. For Supplement 1 to be used wisely, the circumstances for appropriate use of Monte Carlo methods need to be articulated clearly. A common issue for all cases is the critical decision about the Monte Carlo distribution – whether a "nonparametric bootstrap" (i.e., empirical data distribution) or a parametric bootstrap (i.e., specified parameterized probability distribution) or even the prior distribution functions used in a Bayesian model will be selected. In every case, the Monte Carlo methods are quite sensitive to mis-specification, and the consequent distortions in the stated uncertainty and in the expanded uncertainty intervals can be surprisingly large.

In the case of complicated non-linear measurement models, the need to use simulation depends upon the functional complexity and the relative uncertainty in the input quantities, and also depends on the degree of their interdependence (i.e., requiring simulation of their joint distribution rather than the separate marginal distributions). When the relative uncertainty in inputs is fairly small (say <10%), and there is little interdependence, then even strongly nonlinear functions may gain little from simulation methods. The computational intensity issue may also be approached via approximation software – even quite sophisticated approximation methods can be used easily.

However, when the relative uncertainty of inputs is significantly larger or when the inputs are themselves interrelated, the gain from simulating their joint behavior can be quite large. The caveat here is that a sufficiently accurate specification of the necessary joint distributions is usually difficult, especially when empirical data (past or present) is limited, and without an appropriate joint distribution the simulation method may not give reliable results. A fully Bayesian approach applies naturally in this case.

Perhaps the strongest motivation for using (parametric) Monte Carlo methods is when the inputs to the measurement model have asymmetric (non-Gaussian) distributions. This case has the potential of substantial improvement of the uncertainty intervals (with coverage at the stated level of confidence or credibility). The caveat here is once again the requirement of the correct specification of the (asymmetric) distribution(s). A much broader class of distributional models needs to be entertained than are currently presented in Section 4. Also, more rigorous procedures (e.g., distributional fitting techniques, etc.) for selection and/or testing of distributional assumptions are crucial. (The need to specify joint rather than marginal distributions applies here as well) In fact, if the distributions used in a Monte Carlo uncertainty analysis are not properly chosen, the statistical properties will likely be worse than the corresponding analysis by (traditional) propagation of uncertainty, which depends upon first and second moments only.

The alternative nonparametric Monte Carlo methods (e.g., “nonparametric bootstrap”) can be much more attractive than parametric Monte Carlo methods, provided there is sufficient data to justify a resampling process. In general, the caveat for nonparametric Monte Carlo methods is the requirement of an adequate (from large to very large) sample size.

The Monte Carlo methods proposed here have been studied extensively and there is a rich literature with respect to their computation, their properties and their suitability (or limitation) in various circumstances; this relevant work is not reflected in the Supplement 1. A collection of references, both applied and theoretical is attached.

3. Frequency-based approach to deriving probability distributions.

Summary

When related data for an input variable are available, that is when there are measurements $x_{i1},\ldots,x_{in_i}$, for i=1,…,N that are realizations of random variables with means $X_1,\ldots,X_N$, the GUM appears to favor frequentist approaches to obtaining the uncertainty of the $X_1,\ldots,X_N$. Using this statistical paradigm, an estimator (some function of the random variables whose realizations are the $x_{i1},\ldots,x_{in_i}$) and its frequency-based distribution are used to obtain information about the unknown parameters $X_1,\ldots,X_N$. Under this paradigm, the $X_1,\ldots,X_N$ do not actually have a distribution. This fact has serious consequences; for example, statistical confidence intervals cannot be interpreted as coverage intervals. Nevertheless, this approach can have the advantage of computational simplicity and may be preferable when none of the input variables involves Type B uncertainty.

The method, as presented in the Supplement, can be viewed as a simplifies version of the parametric bootstrap approach for distribution estimation. The statistical literature in this area is vast (see [2] and [3]). It is known that this simple approach does not behave satisfactorily in small sample situations and it is our view that the user of the Supplement should be warned about this. A further point is that this behavior is even worse for multivariate cases and the requisite sample size is much larger. The poor performance is evidenced by unsatisfactory coverage probability of the resulting confidence intervals of the $X_1,\ldots,X_N$. (Note that these are not coverage intervals of the $X_1,\ldots,X_N$.) The examples that follow illustrate these points.

There are many other, more modern statistical methods for deriving distributions of estimators. Kernel density estimation, for example, and many other nonparametric estimation methods provide very flexible models when there is enough data. The Supplement should consider these points and provide some references to help the user.

The Supplement approach with low sample size.

In a univariate normal case, that is when the sample observations $x_1,\ldots,x_n$ come from a normal distribution with mean X, a 95 % confidence interval for X , obtained based on the samples suggested by the Supplement from $N(\overline{x},s^2)$, is

$$\overline{x} \pm 1.96\frac{s}{\sqrt{n}}$$

where $\overline{x}$ and s are the sample mean and standard deviation of the observations. The width of the interval suggested by the Supplement is $2\times 1.96 s/\sqrt{n}$. The width of the exact 95 % interval for X is $2\times t_{0.975,n-1}\, s/\sqrt{n}$, where $t_{b,n}$ is the b quantile of the Student's t distribution with n degrees of freedom. The width of the interval based on the Supplement is only 46 % of the width of the exact interval when n=3. The percentages are 71 % and 87 % when n=5 and n=10, respectively. Thus, the Supplement suggested intervals based on small n have insufficient coverage.

The multivariate case.

The GUM example Annex H.2 provides a good illustration of poor performance in the multivariate case. The measurands of interest are the three impedance components, which are functions of the input variables V, I, and f . For illustrative purposes, only one of the measurands, Z = f(V,I) = V/I, is considered. The measurements are assumed to be distributed as a bivariate normal with mean (V,I) and covariance matrix $\begin{bmatrix} s_V^2 & r s_V s_I \\ r s_V s_I & s_I^2 \end{bmatrix}$. The coverage probability of the bootstrap interval for Z can be examined using statistical simulation. This probability depends on the ratios s_V/V, s_I/I, r, and n. For illustration, V= 4.999, I = 19.661, s_V = 0.0071764, s_I = 0.0211778, r = -0.3553, and n = 5 are used as in the ISO GUM Annex H.2. Thus, the value of Z in the simulation study is 4.999/19.661. Independent realizations (V_j, I_j), j = 1,...,5 were generated according to their joint normal distribution specified above. The sample mean vector $(\bar{V}, \bar{I})$ and the covariance matrix s were calculated based on these 5 observations. The 95 % bootstrap interval for Z, using 10000 bootstrap samples from $N((\bar{V}, \bar{I}), s)$, was obtained. The percentage of time that the intervals contained 4.999/19.661 was recorded. The simulated coverage probability was found to be 0.8778. Using the confidence interval based on the first-order Taylor expansion (with correct coverage factor), the coverage probability for the same problem was found to be 0.9495. Even for a linear measurand such as V + I, the coverage probability for the nominal 95 % interval based on Supplement was only 0.8758.

This illustration shows that the proposed method can produce confidence intervals that do not have the expected frequentist properties even for simple cases.

4. Belief-based approach to obtaining probability distributions.

It is well accepted by users of the GUM that when no related data are available, the probability distributions are to be interpreted as summaries of all available knowledge about the input variables. It may not be as well known that it is possible to obtain probability distributions of the $X_1,\ldots,X_N$ directly (not through an estimator) by Bayesian statistical methods. This is true regardless of the types of uncertainty, Type A or Type B, that are present in the problem. This important point should be made clear in Section 4 of the Supplement.

The Bayesian process of obtaining a distribution for the $X_1,\ldots,X_N$, when measurements $x_{i1},\ldots,x_{in_i}$ are available, is as follows.

- First, a prior distribution of the $X_1,\ldots,X_N$, is constructed. This distribution summarizes all information available before the data are collected. The Maximum Entropy method, alluded to in the Supplement, is only one possible way of obtaining a prior distribution. Other possible choices are Jeffreys noninformative priors, ML-II priors, hierarchical priors and others. See [4] for guidance and further references.
- Second, a likelihood function for the data must be specified.
- Third, the posterior distribution of the $X_1,\ldots,X_N$ can be obtained using Bayes theorem.

It may be argued that using the Bayesian approach for all of the input variables is the only way to keep a clear interpretation of the resulting coverage interval of Y. However, even when applying the Supplement's method using only belief-based distributions, problems still arise from the fact that the results for Y can be quite sensitive to the choice of the distribution of the Xs.

Thus in our view, the method in Supplement 1 should not be used without study of sensitivity to the choice of the input variable distributions. This means that whenever possible one should identify a class of possible distributions of each X_i and then perform the Monte Carlo with various members of this class. Agreement of the results for Y across such varied Monte Carlo samples would mean that the sensitivity to the specific distributions was not too great.

The following example shows how the method in Supplement 1 could be applied using the Bayesian paradigm, as well as some possible consequences.

Suppose there are measurements $x_1,\ldots,x_n$ of observations of a random variable that is assumed to have a normal (Gaussian) distribution with expected value X and variance s^{2}. Note that this assumption specifies the likelihood function. The measurements are: 102.22190, 99.29446, 101.59621, 100.81106, and 98.67992. The expected value X is thought to be related to the unknown value Y of the measurand by the following linear equation

$$Y = \frac{X - B_0}{B_1}.$$

The intercept *B_0* and slope *B_1* are input variables representing sources of uncertainty of Type B.

First, all probability distributions are specified. For the purpose of this example, two different sets of distributions for the *Bs* with the same means and variances are used.

1. B_0 ~ N(0, t^2), B_1 ~ N(0, ?^{2}).
2. B_0 ~ U(-t√3, τ√3), B_1 ~ U(1 – ? √3, 1 + ? √3).

with t = 0.25 or 4.00 and ? = 0.05 or 0.2.

This is to illustrate that common values of the first two moments can be represented by very different probability distributions. These distributions are based purely on expert opinion. For the distribution of X, a common Bayesian approach is adopted. That is, we specify a noninformative prior distribution on X and on σ

$$\left(X, s^{2}\right) \sim 1/s^{2}.$$

Next, using Bayes theorem the posterior distribution of X , that is Student's t

$$X \mid x_1,\ldots,x_n \sim t_{n-1}(\bar{x}, s^{-1}, (n-1)),$$

is obtained. This distribution can now be used in the Monte Carlo simulation. The posterior mean of X is $\bar{x}$ and the posterior variance is $\left((n-1)/(n-3)\right)s^2$, where $\bar{x}$ is the sample mean and s^2 is the sample variance.

It is useful to point out that the Supplement method could be very easily carried out using the free software BUGS [5] as follows:

```
{b0 ~ dnorm(0,0.06)
b1~dnorm(1.0,400)
mean~dnorm(0,1.0E-4)
sig ~ dgamma(1.0E-4,1.0E-4)
Y <- (mean - b0)/b1
for(i in 1:5){
x[i]~dnorm(mean, sig)}
}
```

The results of the calculations are given in the following two tables:

(t,w)	Mean (median)	Uncertainty (95%HPD)
(0.25, 0.05)	100.760 (100.527)	5.160 (91.342, 111.559)
(0.25, 0.20)	105.100 (100.599)	24.494 (72.130, 164.720)
(4.00, 0.05)	100.757 (100.514)	6.581 (88.602, 114.273)
(4.00, 0.20)	105.098 (100.595)	24.932 (71.401, 165.682)

Table 1. Results using the Gaussian distribution for the Bs

(t,w)	Mean (median)	Uncertainty (95%HPD)
(0.25, 0.05)	100.741 (100.469)	5.348 (92.277, 110.145)
(0.25, 0.20)	104.914 (100.378)	22.378 (75.371, 150.654)
(4.00, 0.05)	100.702 (100.449)	6.690 (88.678, 113.963)
(4.00, 0.20)	104.914 (100.398)	22.762 (73.643, 152.171)

Table 2. Results using the Uniform distribution for the Bs.

For comparison, the value and uncertainty were also computed using the GUM approach. The GUM estimates are the same for the two sets of distributions and are given in the following table.

(τ, ω)	**Value**	**Uncertainty**
(0.25, 0.05)	100.521	5.121
(0.25, 0.20)	100.521	20.128
(4.00, 0.05)	100.521	6.493
(4.00, 0.20)	100.521	20.520

Table 3. The GUM results

The tables show that using the distributions rather than just the moments as in the GUM can have quite a large effect on both the value and the uncertainty. This is especially the case for the slope parameter. The value of t = 0.2 represents a rather extreme case but illustrates what can happen.

Tables 1 and 2 further show that for this example, the method in Supplement 1 is not particularly sensitive to the choice of distributions for the offset B_0. It further shows that the method is quite sensitive to the choice of distribution of the slope parameter. There are differences in the value of Y as well as the uncertainties, the difference in uncertainty being especially large: 9% larger for the Gaussian distribution than for the rectangular.

It is further clear that the distributions of Y are quite skewed and thus the arithmetic mean may not be a good estimator of the value. The median may be a better estimator.

These comments on the draft of Supplement 1: Numerical Methods for the Propagation of Distributions to the Guide to the Expression of Uncertainty in Measurement, along with the comments submitted by other reviewers, will help ensure that the new methods that are proposed in this document are metrologically sound and as clearly explained as possible.

Bibliography

[1] Guide to the Expression of Uncertainty in Measurement, 2nd ed (Geneva: International Organization for Standardization) ISBN 92-67-10188-9.

[2] Chernick, M. (1999) Bootstrap Methods, A Practitioner Guide, Wiley, New York.

[3] Efron, B. and Tibshirani, R. J. (1993) An Introduction to the Bootstrap, Chapman & Hall, New York.

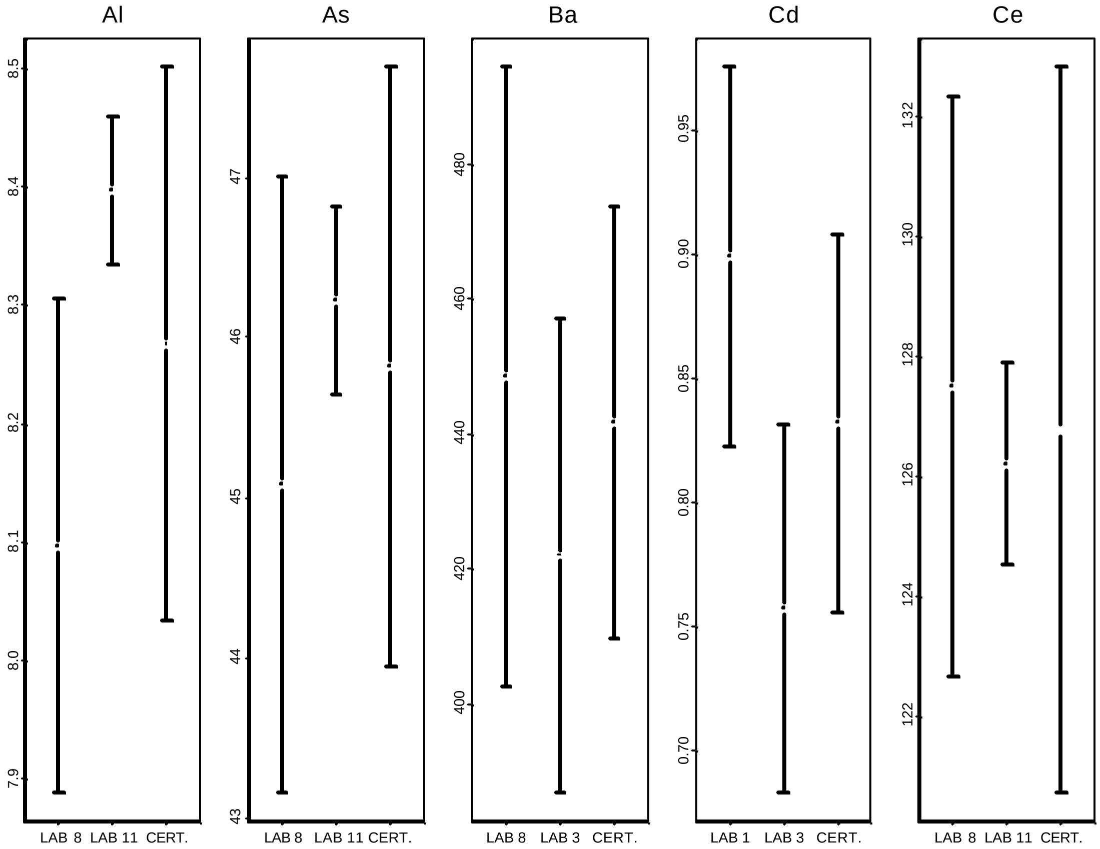
[4] Berger, J. O. (1985) Statistical Decision Theory and Bayesian Analysis, 2nd ed, Springer-Verlag, New York.

[5] http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

3.6 Accounting for Suspected Sources of Bias

Walter Liggett
Statistical Engineering Division, ITL

Rolf Zeisler
Analytical Chemistry Division, CSTL



Concentration determinations for five elements in the material to be issued as SRM 2703 along with expanded uncertainties (multiplier 2). Determinations from individual laboratories and the combined determination for the certificate are shown. Certification is based on determinations shown and SRM 2702 determinations not shown. The uncertainty reported by some laboratories is augmented with an imputed value for the Type B uncertainty.

In value-assignment of reference materials for chemical measurement, a NIST certified value is one for which NIST has accounted for all known or suspected sources of bias. Certification is based on one or more chemical methods with sources of uncertainty that have been investigated. Types A and B standard uncertainties account for the known sources of uncertainty. Sometimes existing scientific understanding indicates the need for an additional standard uncertainty that accounts for suspected sources. Typically, this additional uncertainty is assessed by complimenting the NIST determination with a second determination by an independent, critically-evaluated method, or by a different method realized in an outside collaborating laboratory. This article describes how one might proceed when a determination by a different chemical method is not available. When the material to be certified differs little from another material already certified, this additional uncertainty can be assessed by borrowing from the previous certification. Even better certified values and uncertainties can be obtained when the determination of the new material includes a simultaneous fresh determination of the already certified material. This article discusses certification based on the use of a hierarchical Bayesian approach to weaving together information on the suspected sources of bias.

Consider the determinations for Al (aluminum), especially the determination from LAB 11. This laboratory measured the 2703 material using neutron activation analysis and reported the Type A uncertainty, which accounts mostly for the Poisson variation in the counting, and the Type B uncertainty, which accounts for non-uniformity in the neutron flux, among other things. The error bars show the two types of uncertainty combined as the square root of the sum of squares. The length of the error bar is twice that for a standard uncertainty. The two types of uncertainty reported are not adequate for certification of the 2703 material because other sources of bias are suspected. To account for the uncertainty contributed by these suspected sources, one typically uses a determination by a second, independent method. For Al, a second determination by LAB 8 is shown. However, this determination does not meet the requirement because LAB 8 also used neutron activation analysis. The problem arises from the fact that there may be sources of bias associated with neutron activation analysis that are shared by LAB 8 and LAB 11. For this reason, a combination of the determinations for Al by LAB 8 and LAB 11 is not a sufficient basis for certification.

To comply with the NIST value-assignment criteria, we combined the determinations shown with determinations on SRM 2702. This is reasonable because production of the 2703 material involved little but sieving the 2702 material. Neutron activation analysis and three other independent methods were used to measure the 2702 material as part of the SRM 2702 certification. These determinations made as part of the SRM 2702 certification help us account for suspected sources of bias in the 2703 certification. Moreover, both LAB 8 and LAB 11 measured SRM 2702 at the same time that they measured the 2703 material. These measurements of the 2702 material at the same time as the 2703 measurements provide a link between the two materials with smaller uncertainty than might otherwise be the case.

In terms of the 2703 Al determinations by LAB 8 and LAB 11, we need to adjust the central value using the link established by the simultaneous measurement of the 2702 and 2703 materials and adjust the uncertainty using all the 2702 determinations. We do not actually estimate these adjustments. Rather, we use a Bayesian approach to summarize all the determinations available. The result for Al is labeled CERT.

There are 22 elements to be summarized. For each element, different laboratories, perhaps using different methods, provide determinations of the concentration. In this article, we consider an approach to summarization that is general enough to apply to all elements for which there are determinations. Because the approach is general, the summaries produced can be compared in that they reflect what is actually known on the basis of the available determinations but do not reflect ad hoc choices in the summarization.

There are three different types of chemical determination that enter our Bayesian approach to certification. These are determinations for the 2703 material alone, determinations for the 2702 material alone, and determinations of the two materials jointly. For the purpose of presenting our approach, we model these determinations under the assumption that the Type A and Type B standard uncertainties are known. In actuality, we take into account the fact that the Type A uncertainty is given as an estimate with stated degrees of freedom. We also imputed Type B uncertainties when these were not provided as part of the determination. We model a determination on the 2703 material as

$$\mathrm{N}(x_3|m, s_{\mathrm{A}}^2 + s_{\mathrm{B}}^2 + s^2),$$

where x_3 is the determination, m is the measurand, s_{A} is the Type A standard uncertainty, s_{B} is the Type B standard uncertainty, and s is the standard uncertainty for the suspected sources of bias. In this expression, the quantities m and s are unknown. We model a determination on the 2702 material as

$$\mathrm{N}(x_2|m + d, s_{\mathrm{A}}^2 + s_{\mathrm{B}}^2 + s^2),$$

where x_2 is the determination and d is the unknown difference in measurand between the 2702 material and the 2703 material. We model a joint determination as

$$\mathrm{N}_2\left(\begin{pmatrix} x_3 \\ x_2 \end{pmatrix}\middle|\begin{pmatrix} m \\ m+d \end{pmatrix}, \begin{pmatrix} s_{\mathrm{A}}^2 + s_{\mathrm{B}}^2 + s^2 & s_{\mathrm{B}}^2 + s^2/2 \\ s_{\mathrm{B}}^2 + s^2/2 & s_{\mathrm{A}}^2 + s_{\mathrm{B}}^2 + s^2 \end{pmatrix}\right).$$

Note that when interpreting the Type A and Type B uncertainties, we have reverted to the classical definitions of random and systematic. Note also that we have taken the uncertainty due to suspected sources to be neither perfectly correlated nor uncorrelated, but to have correlation of ½. This particular approach to including joint determinations in a hierarchical Bayesian approach raises the question of whether there is a better approach.

Let the vector of determinations for a particular element be $\mathbf{Y}$. In the case of Al for example, there are four determinations of the 2702 material and two joint determinations of the 2703 and 2702 materials. Multiplying the seven corresponding densities together gives the likelihood $f(\mathbf{Y}|m,d,s)$, where we have suppressed the known values of the Type A and Type B uncertainties in our notation. This likelihood suggests maximum likelihood and Bayesian approaches to certification.

Maximum likelihood suggests adopting an uninformative prior on m and maximizing

$$\int f(\mathbf{Y}|m,d,s)\,dm$$

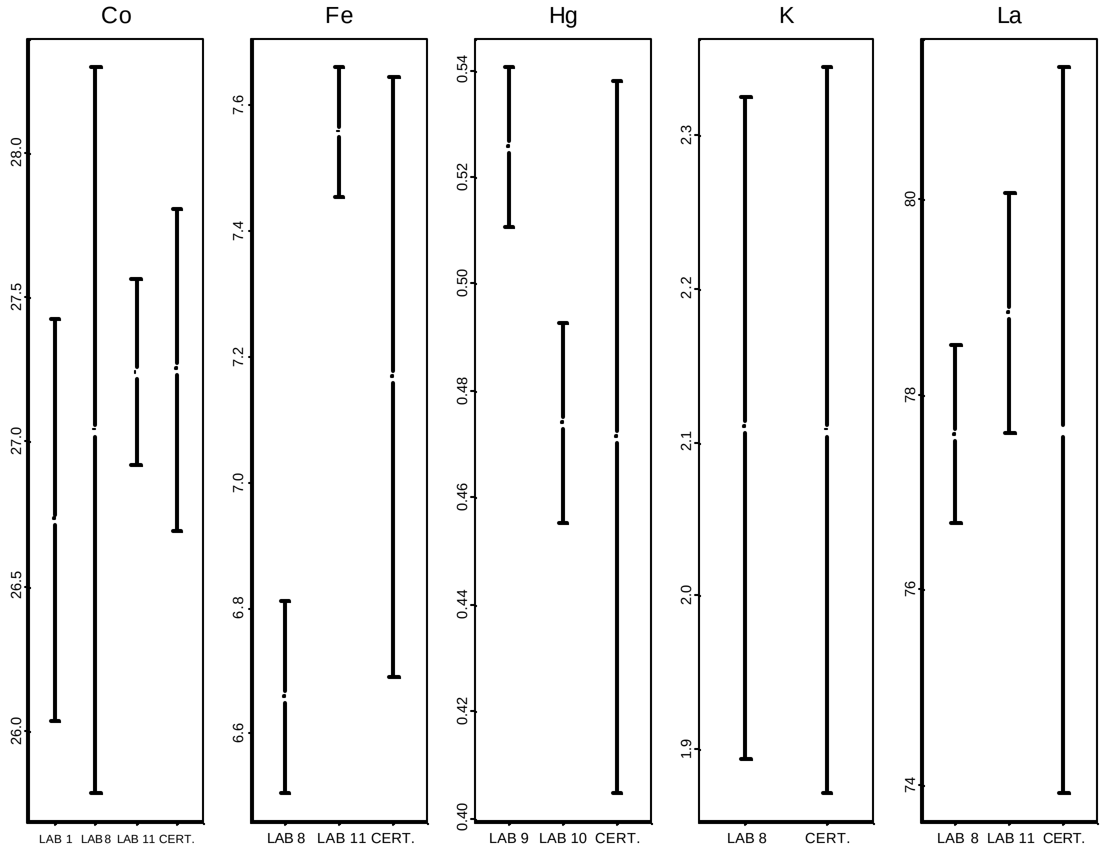
as a way of estimating d and s . Treating these estimates as known, we could use $f(\mathbf{Y}|m,\hat{d},\hat{s})$ to obtain the interval for certification.

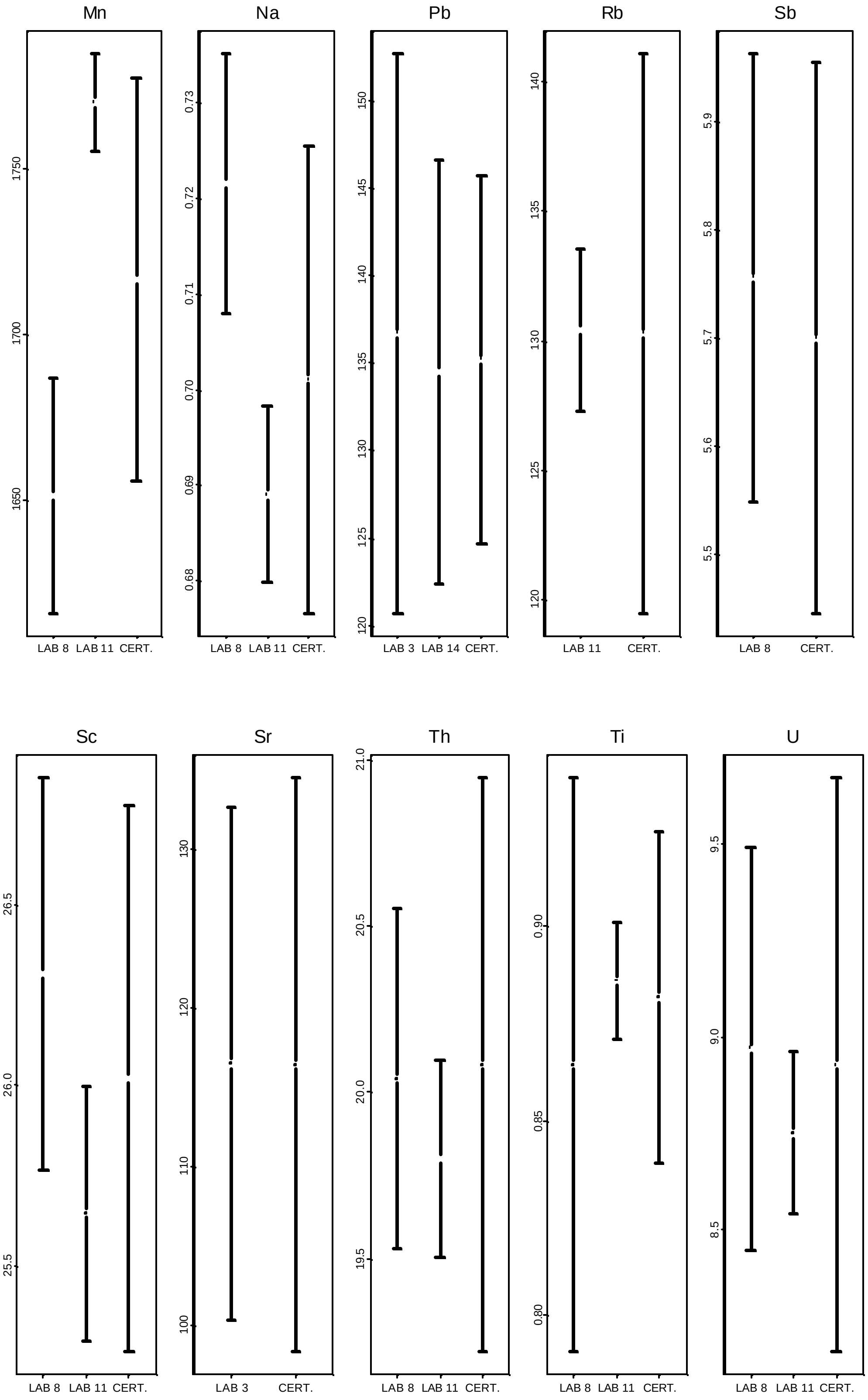
In our Bayesian approach, we adopt an uninformative prior on d and a uniform prior on s that is non-zero only up to a multiple of the uncertainty NIST claimed in its determination for the 2702 material. This prior on s is justified by the fact that in choosing collaborating laboratories, NIST chose only those whose performance is respectable in terms of the state of the art. The posterior density on m is then proportional to

$$\iint f(\mathbf{Y}|m,d,s)\,dd\,ds \ .$$

We compute the mean and variance of the posterior distribution as the basis for the certification. We could have used an informative prior on d to take into account the prior knowledge that the 2703 material is similar to the 2702 material, but we did not.

The following three figures give further examples of how this works out.

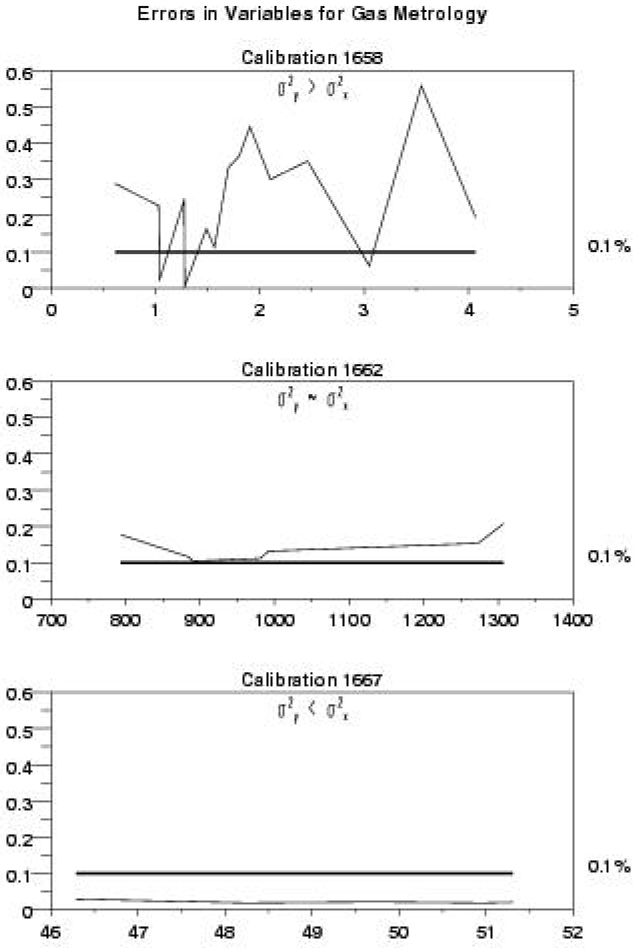



Mn
1650
1700
1750
LAB 8
LAB 11
CERT.
Na
0.68
0.69
0.70
0.71
0.72
0.73
LAB 8
LAB 11
CERT.
Pb
120
125
130
135
140
145
150
LAB 3
LAB 14
CERT.
Rb
120
125
130
135
140
LAB 11
CERT.
Sb
5.5
5.6
5.7
5.8
5.9
LAB 8
CERT.
Sc
25.5
26.0
26.5
LAB 8
LAB 11
CERT.
Sr
100
110
120
130
LAB 3
CERT.
Th
19.5
20.0
20.5
21.0
LAB 8
LAB 11
CERT.
Ti
0.80
0.85
0.90
LAB 8
LAB 11
CERT.
U
8.5
9.0
9.5
LAB 8
LAB 11
CERT.

3.7 Errors in Variables for Gas Standard Calibration

Stefan Leigh, Andrew Rukhin, Alan Heckert
Statistical Engineering Division, ITL

Frank Guenther
Analytical Chemistry Division, CSTL



Relative standard deviation plots derived from actual calibration lines showing the occurrence of each of the three modes for regression analysis: $s_y^2 >> s_x^2$, $s_y^2 \approx s_x^2$ and $s_y^2 << s_x^2$. The solid horizontal 0.1 % line represents the assumed relative standard deviation in X, the standard gas concentration.

In gas metrology at the international standards organizations level, frequent calibration of automated concentration measurements against carefully prepared and certified gas concentration standards is routine. Typically, such data have been analyzed using classical linear methodology: Ordinary Least Squares (OLS) fitting, with associated Working-Hotelling and Fieller (propagation-of-error based) calibration bounds. But it has been observed that often the basic assumptions of linear least squares are not met: most crucially, the assumption of Y variance dominance may be violated. This fact led the ISO Gas Analysis Working Group TC 158 to develop and adopt a new Gas Analysis Standard (ISO 6143) based on the use of Errors in Variables (EiV) methodology for the analysis of gas mixture composition calibrations.

SED/NIST noted multiple troubling features of the new standard: incorrect language is employed; the use of Errors in Variables (EiV) methodology, to supplant OLS, is recommended irrespective of the variance ordering in the data; EiV in the standard is implemented as an unclearly documented `black box' executable code with source code not made available. Finally, and most telling, no reference is made to the large body of hard statistical research in this area, and no reasons given for ignoring classic solutions such as Maximum Likelihood Estimators (MLE) for specific variance ordering scenarios.

Major progress was achieved last year with a clearer understanding of what the ISO code computes for an EiV calibration error, which shows why it is wrong as a universal prescription. An important detail of the ISO algorithm is that, early on, X (abscissa, nonreplicated gas standard values) and Y (ordinate, replicated analytic values) are exchanged, and all subsequent analysis - (weighted) Orthogonal Distance Regression (ODR) and generation of calibration uncertainty - is performed on the role-exchanged variables. A fuzzy justification for this was provided: namely, that since the ODR solution must lie between the OLS solution for the calibration of Y on X and the OLS solution for the calibration of X on Y (a correct assertion), in performing ODR one can freely exchange X and Y, and solve the (tricky) calibration problem (inversion: Y backcasting to X) as a straightforward prediction problem (prediction: Y-as-X to X-as-Y). To be able to solve the EiV calibration problem in this way – interchange of variables and `translation' of inversion of uncertainty interval to prediction of uncertainty interval - would appear to demand an exact symmetry between X and Y, the roles and interpretation of X and Y, and the error structures of X and Y, symmetry that in real gas calibration settings is lacking.

Pointing out errors in others' solutions does not constitute a solution to the problem. We are sure that: (1) careful consideration, on a case-by-case basis, of the underlying error structure for any given facility's calibrations is crucial. We categorically reject the notion that one numerical algorithm ``fits all'' cases, as BAM/ISO maintain. We are proving this through simulation. (2) We continue to maintain that appropriate literature-referenced MLE solutions for each case remains the best current approach to the problem. (3) But we must adapt such solutions to a heteroscedastic setting, through weighting. And (4), for each case we must indicate how to solve the correct calibration uncertainty problem.
The goal is to extend a version of Fieller to each for the weighted EiV.

SIMULATION

Representing 6143, improved algorithms and code have been issued by the National Physical Laboratory (NPL) of the UK. It is the NPL version that is used in our ongoing simulations. SED's focus is outlined above, and all code developed reflects those specific goals. In particular, the EiV model has been adapted to the situation where there are replicated measurements on the response variable, and where one of the error variances (X) is assumed known - conditions closely approximating real-world gas metrology. For these conditions, the maximum likelihood estimator and approximate confidence regions for the unknown parameters have been derived. Corrected/improved formulas for the calibration (backcast) estimation problem were developed based on those presented in Fuller (1989).

Simulation of multiple homoscedastic cases has given us our first concrete look at the relative performances of the competing portmanteau 6143 versus case-based MLE estimators. We organize the homoscedastic case simulations, as is customary in EiV, according to ? – the ratio of Y to X variance. There are three relative dominance scenarios: For ? = 1, the case that should most highly favor 6143, MLE and 6143 estimates of slope are indeed comparable both in terms of bias and variance, which is also the case for calibration value and error. In the case of intercept estimates, however, MLE performance is slightly superior at higher noise levels. For ? < 1, again for slope and backcast value and associated error, 6143 and MLE are broadly comparable. Although for intercept, MLE is incrementally better than 6143 both in terms of bias and standard error at all noise levels. For ? > 1, however, slope MLE's are bias/variance superior at higher noise levels, while intercept and calibration MLE's are superior across the board. These and other forthcoming results still beg the practical question: will the incremental improvements demonstrated by use of appropriate MLE or surrogate be considered substantial enough to be taken seriously by the gas metrology community, which is already several years into the deployment of 6143? Heteroscedastic simulations are currently being undertaken.

Errors in Variables is a complex subject matter, even for statisticians. Identifiability and estimability problems are acute. Often the existing literature is not clearly exposited, tending to linger over arcane counterexamples rather than presenting practitioners with clear guidelines and procedures. We seek to lay down a clear logic for the linear calibration problem with specific reference to NIST Gas Metrology calibration experience. We expect those guidelines to be documented, with explicit reference to broadly accepted inferential principles such as maximum likelihood, method of moments, or Bayesian estimation. We seek to make clear the foundation, the implementation, the built-in assumptions, and the potential limitations of any methods suggested.

ISO Standards play multiple direct roles in national and international commerce. Initiation and/or efficient adaptation of such standards by national industries translate to direct commercial leverage and advantage in OECD and third-world markets. NIST and the other world standards organizations bear direct responsibility for ensuring the technical integrity of such standards and for the implementation of correct and relevant statistical methodologies.

3.8 Use of Median and Uncertainty Evaluation for Nanoparticles Sizing

Z.Q. John Lu, Charles Hagwood
Statistical Engineering Division, ITL

Stephen Hsu, Yanan Liang
Nanotribology Group, Ceramics Division, MSEL

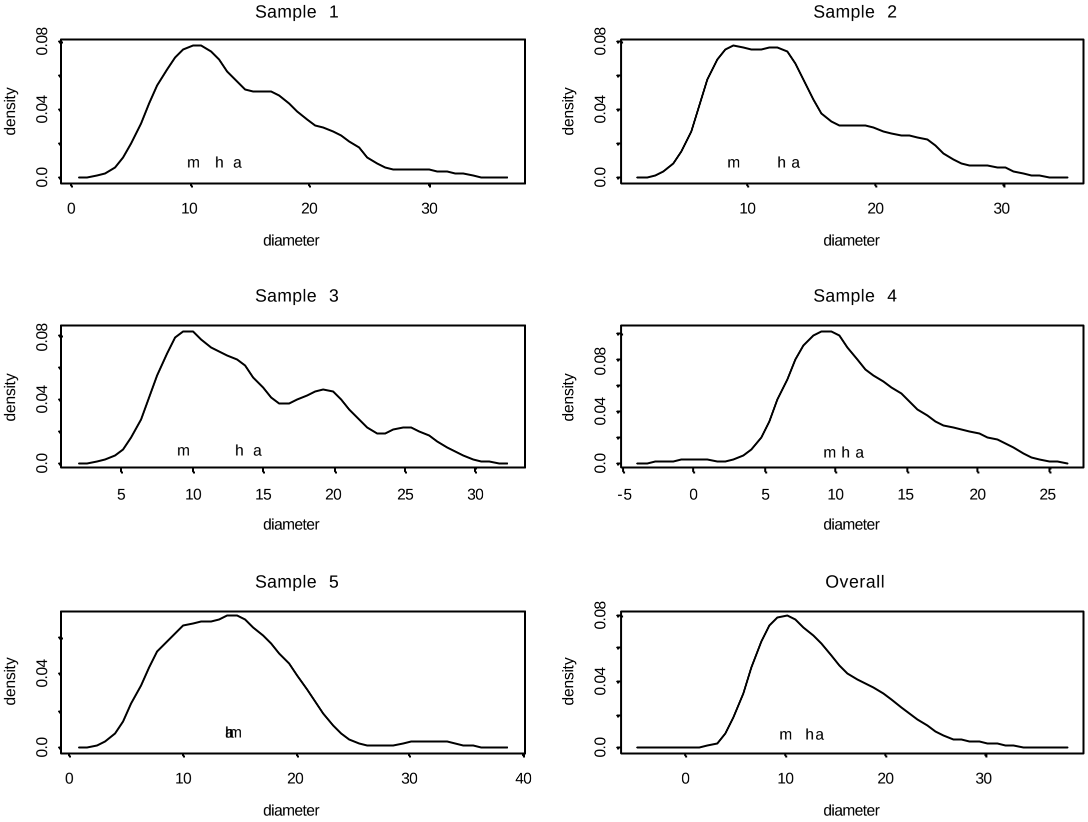


Figure 1. Shown here are plots of kernel density estimation of particle diameters (data R3) for each of the five bottles examined by Dr. Yanan Liang. The last figure is for all data. I have denoted the positions of estimated mean (a), median (h), and mode (m) in each graph. It is seen that median falls always between mean and mode.

Various statistical measures, such as some kind of averages, median, or mode, may be used to provide an overall summary of the “average” size of nanoscale particles, which are used in some standard reference materials that are used in some medical testing developed by the nanotribology group at NIST. Because the particle size distribution is often non-Gaussian with very long right tails, we recommend the use of the median since means will be strongly affected by the large tail values, while the mode is too strongly affected by the statistical algorithms used to locate the peak of the estimated density. Statistical uncertainty evaluation for the median will be described in this entry with the goal of attracting more uses of median measures in metrology studies.

In nanotribology, nanoscale particles are injected into animals to test for certain reactions from friction. Both the dosage and the size of particles are of interest in standard material measurements. For particle size, we notice that there are at least three possible choices: mean, median, and mode. It is well known that the sample mean (average) is strongly influenced by a few large numbers in the right tail and so is less stable. The mode is defined by the peak of a density and describes the size of the most common (abundant) types of particles. The density function is unknown, however, and must to be estimated by some smooth operation of a histogram, such as kernel smoothing (resulting in kernel density estimation, e.g., Silverman 1985). However, estimation of the mode is likely influenced by the degree of smoothness, such as the bandwidth. The median is defined as the particle size that half of the population of particles is larger than and half of the population is smaller. The sample median is a more robust statistical measure than the mean or mode for a non-Gaussian distribution. Thus it is more reliable and stable to use in evaluating the centers of different shapes of particle size distributions.

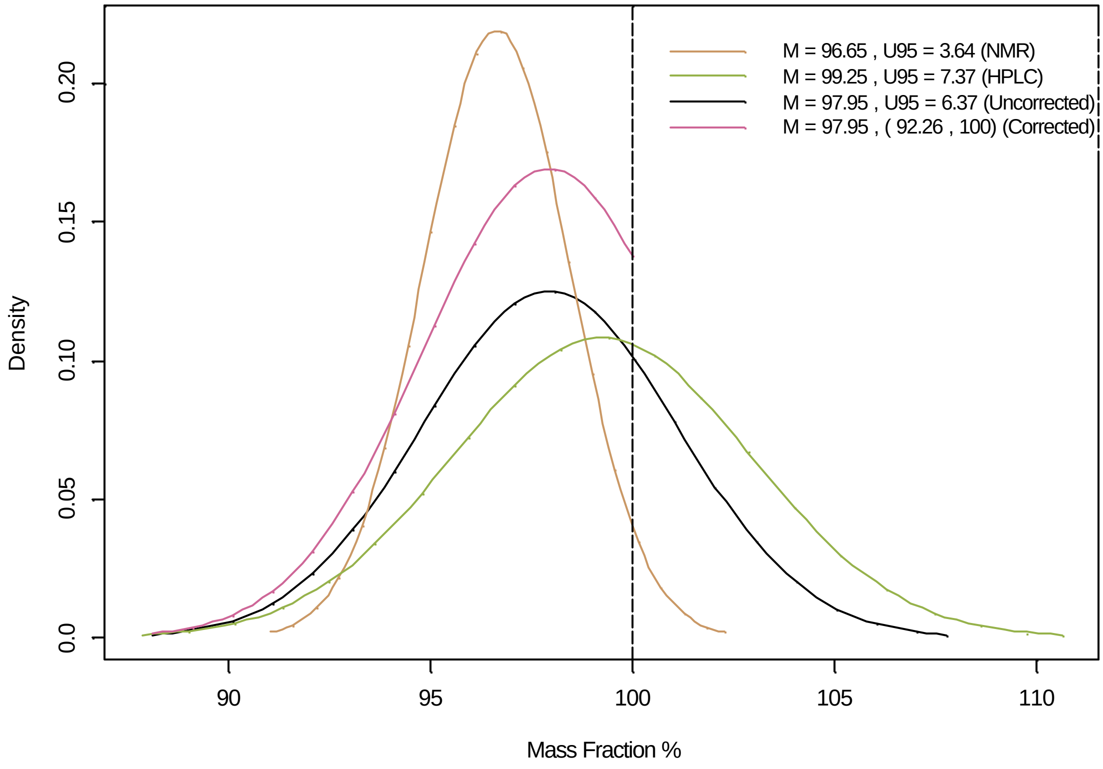
The uncertainty for the sample median can be approximated by its asymptotic variance, which is given by $1/(4\cdot f^2(h)n)$, where n is the sample size and f is the density function evaluated at the median. One can use a kernel density estimate in place of the true density to obtain a fairly robust uncertainty estimate for sample median. Alternatively, one can also apply the bootstrap method for the sample median. Another advantage of using the median instead of the mode is the existence of a rich collection of statistical tests based on ranks for testing homogeneity or heterogeneity (differences) in medians for particle sizes from different bottles or sources. For example, the Mann-Whitney-Wilcoxon test can used for comparing medians from two populations, and Kruskal-Wallis rank tests can be used for testing equality of medians from multiple populations in one-way ANOVA-type fashion.

The offshoot of this metrology study is our determination that the median should be a more widely used measure in metrology for characterizing skewed and long-tailed particle size distributions. The worry over uncertainty evaluation for median is alleviated by recent research on the exact variance approximation and the availability of bootstrap methods and software.

3.9 Estimating Parameters that are Near the Detection Limit

Hung-kung Liu and Grace Yang
Statistical Engineering Division, ITL

Gary W. Cramer and Paul C. DeRose
Analytical Chemistry Division, CSTL



Probability Density Functions for Purity Measurements of Solvent Red 24 Dye.

The verification of purity of chemicals is a prerequisite for valid calibrations as well as for the production of valid reference materials for calibration. The calculation of the composition of the main component is straightforward, but the evaluation of the associated measurement uncertainty is not. Problems arise due to mathematical constraints (the sum of all mass fractions cannot exceed unity) and due to the detection limit of the procedure used. The detection limit of an analytical procedure is regarded as being the lowest concentration of the analyte that can be distinguished with reasonable confidence from a blank. SRM 2037 Solvent Red 24 dye is intended primarily for use in the calibration of spectrophotometric analyses for the determination of the red dyes used to mark off-road diesel fuel for taxation purposes. The NIST measurement result gives the purity of the Red 24 Dye around 98 % with uncertainty about 5 %. We use a Bayesian approach to derive an unsymmetrical uncertainty interval that lies within the 100 % constraint.

A primary method of measurement is a method having the highest metrological quality, whose operation can be completely described and understood, and for which a complete uncertainty statement can be written down in terms of the SI units. When no single primary measurement method proves fitness for purpose, chemical purity must be estimated by assembling information from different sources. Two methods, 500 MHz Proton Nuclear Magnetic Resonance (NMR) and High Performance Liquid Chromatography (HPLC), were used in the purity determination of the Solvent Red 24 Dye for SRM 2037. The tan and purple curves in the above figure are the density functions of measurement results for the NMR and HPLC methods for Red 24 Dye, respectively. For each density, the mean is the certified purity for that method and the standard deviation is equivalent to the estimated standard deviation for a normal distribution. These were assigned for measurement repeatabilities and were combined with balance accuracy uncertainties and estimated instrument method uncertainties using the root-sum-of-squares method. The green curve is the density function of the combined uncertainty distribution, assembling information from both NMR and HPLC. Normally, an expansion factor of k = 2 is applied such that the expanded uncertainties express a symmetrical interval within which the true value is expected to fall with a level of confidence of approximately 95 %. However, the upper limit for the purity of any material is bounded by 100 %. To correct the density function, a Bayesian approach with a prior that distributes the purity uniformly across the constrained mass fraction range was used to give an unsymmetrical interval bounded by 100 % within which the true value is expected to fall with a credibility of 95 %. The black curve is the posterior density function of the combined uncertainty distribution, corrected for the mass fraction upper limit of 100 %. The resulting combined mass fraction percentage (purity) for the Solvent Red 24 dye is 97.95 % with a 95 % uncertainty interval between 92.26 % and 100%.

Different procedures can be used to assign uncertainties in estimating parameters that are near the detection limit. Our method is attractive for two reasons. One, it requires no special choice of error distributions or estimation procedures in order to avoid the problem of having uncertainty intervals lying beyond the constraint. And, two, in our procedure, a straightforward correction is made only at the last-minute of the analysis following a Bayesian argument, using a non-informative prior.

3.10 Estimation of Poisson Intensity in the Presence of Dead Time

Grace Yang
Statistical Engineering Division, ITL

John F. Widmann
Formerly at Fire Research Division, BFRL

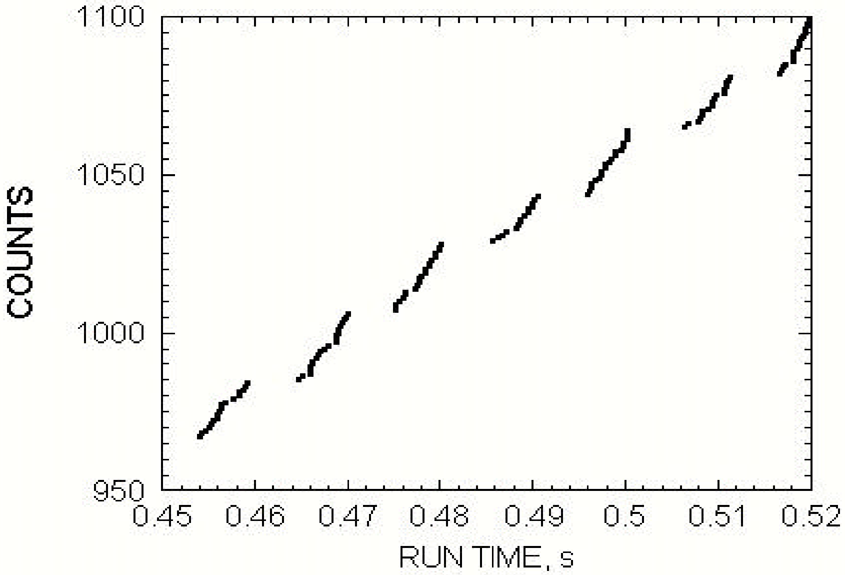


Figure 1: Experimental data collected with the PDI at 5mm above the nozzle and a spray angle of 33.4 degrees.

Understanding the characteristics and dynamics of a spray has numerous implications. Such knowledge helps to design cleaner-burning, more efficient fuel injections in gasoline engines. Other applications include control of hazardous waste incineration, spray coatings, agricultural pesticide sprays, and fire suppression.

At NIST, a non-intrusive technique called phase Doppler interferometry (PDI) is used to investigate the characteristics of a methanol spray flame. The PDI, installed at a fixed location, measures the size and the arrival time of each droplet in the flame that reaches its probe volume. The purpose of our study is to use the arrival time data to estimate the intensity of the spray. However, the data recorded by the PDI are incomplete due to the presence of dead time. Dead time in PDI refers to the periods of inactivity of the PDI processor during which data are not recorded. This is illustrated by the gaps in the experimental data presented in Fig. 1. These gaps occur at random throughout the recording period. Moreover, the duration of each gap is random and unobservable. The arrival times are recorded only during the observation windows and missed during dead times.

Following the spray literature, we model the arrival process of droplets at the PDI by a homogeneous Poisson Process, $\{N(t);\ t = 0\}$ with intensity λ, where $N(t)$ represents the number of droplets that passed through the probe volume (but not necessarily recorded) in the time interval (0,t]. The presence of dead time renders the observed arrival times of the droplets an incompletely observed Poisson process. The observed process consists of the data $\{(N_k, O_k),\ k = 1, 2,\}$, where N_k denotes the number of droplets in the kth observation interval and O_k denotes the total time of the kth observation interval. Suppose m is the number of observation windows. We estimate the intensity λ by

$$\frac{\sum_k N_k - m}{\sum_k O_k}$$

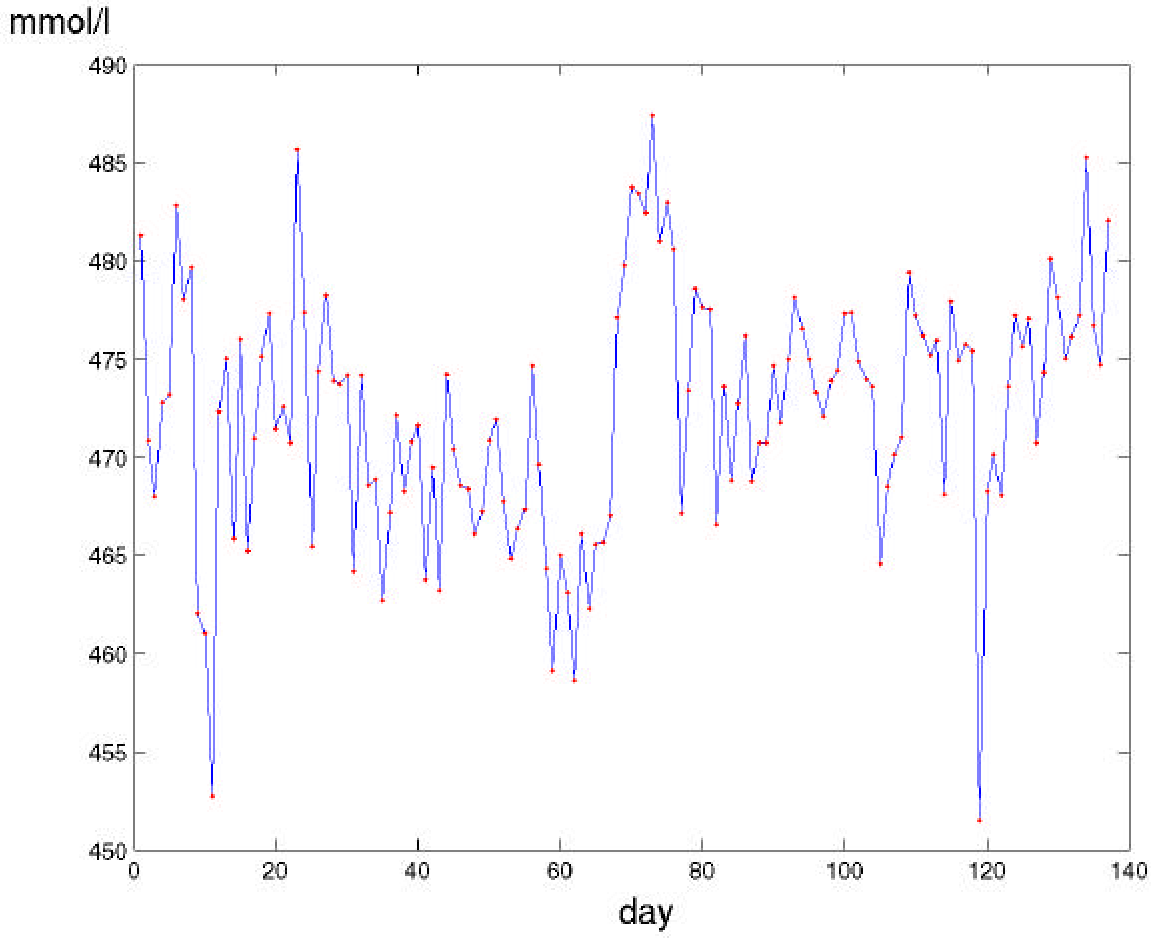
We show that $\{(N_k, O_k),\ k = 1, 2,\}$ forms a strictly stationary sequence, and the estimate is consistent and asymptotically normal as m tends to infinity.

An analytic procedure is provided for estimating the flow rate of droplets in a spray. Large sample properties of the estimator are obtained under various conditions of missing data. The results are useful for bias correction and computation of the uncertainty of the estimate. This is a joint work with J. F. Widmann, S. He and K. T. Fang. The results will be published in the forthcoming issue of the Journal of American Statistical Association. Our present work only considered one-dimensional data. In the pursuit of a better understanding of sprays, stochastic modeling of higher dimensional data that consist of both the arrival time and the size of each droplet is an important problem for future investigation.

3.11 The Effect of Recalibration and Reagent Lot Changes on the Performance of QC Control Charts

Per Winkel
Copenhagen Center for Controlled Trials, HS University Hospital of Copenhagen

Nien Fan Zhang
Statistical Engineering Division, ITL



The figure shows the time series of H-ALB.

Daily quality control (QC) measurements of biochemical quantities were recorded during 4 to 5 months while methods and analyser showed no signs of malfunctioning. The time series of QC values was divided into subseries according to reagents or electrolyte diluent lot and (within diluent subseries) disposable electrode used. ANOVA was used to examine if the mean level changed significantly between subseries. The X chart and the EWMAST or EWMA chart were applied to each time series and each reagents and diluent subseries. We found that the mean levels changed significantly due to diluent lot changes, replacement of disposable electrodes and recalibrations following reagents lot changes. These changes caused spurious autocorrelation as evidenced by the autocorrelation function (ACF) plot. We conclude that the mean level may change due to recalibrations and change of electrode diluent lot that causes an excessive number of false alarms unless new control charts are calculated subsequent to these events.

In a previous joint paper (Winkel and Zhang (2004)) we found that clinical biochemical QC data may be autocorrelated. As a consequence too many false alarms were generated when the exponentially weighted moving average (EWMA) control chart was used. When we used the expon entially weighted moving average chart for stationary processes (EWMAST) chart (Zhang (1998)), so that the magnitude of the autocorrelation was taken into account, the number of false alarms was reduced considerably. However, when the EWMAST chart was applied over an extended period, which includes some recalibrations of the analyser (or using different reagents lot number), it was sometimes found that the mean level changed significantly even though the analyzer was in operational control. It was hypothesized that these shifts in the mean level may be caused by recalibrations of the analyzer.

We used a well-established modern analyzer, the Dimension, as a prototype. When taking a possible effect of recalibrations into account, we wanted to note if the QC data were still autocorrelated. The dilution of each sample prior to the measurement of the concentrations of sodium (NA) and potassium ions (K) is calculated using an aqueous solution of electrolytes of known concentration (the diluent). It was recommended that the corresponding assay be recalibrated whenever a reagents lot is changed. Most of the colorometric assays were recalibrated at least once during the study due to reagents lot changes and the lot number of the diluent was changed in the middle of the study period. Thus, the effects of recalibrating the colorometric assays (the reagents lots of the immunochemical assays were not changed during the study period), changing electrode and changing the electrolyte diluent lot could be studied.

The data material includes the values of daily measurements of 11 components in each of two types of QC material (low and high). Thus, there are 22 quantities. The time series including values from more than one reagents lot was divided into subseries according to the lot used to obtain reagents subseries. The time series of electrolyte values obtained using more than one diluent lot was divided into subseries according to diluent lot used to obtain diluent subseries and each diluent subseries was divided according to disposable electrode used to obtain electrode subseries.

A change in the mean level due to recalibration following a change of reagents lot and a change of diluent lot or disposable electrode (in the electrolyte assays) was suspected. Therefore, by using a one-way ANOVA it was tested if the mean level differed significantly

(p <0.01) between reagents subseries. In the same way, it was tested if the mean level changed between the diluent subseries of each electrolyte time series and if the mean level changed between the electrode subseries within each diluent subseries.

Each time series and each individual subseries (except the electrode subseries which was too short) was tested to note if significant (p <0.05) autocorrelation was present using the ACF plot. We found that the ACF plot, however, may falsely show significant autocorrelations due to a change in the mean level in an otherwise random time series. Therefore the ACF plot was supplemented by a non-parametric run test of randomness since the run test of randomness does not react as much as the ACF plot to mean level changes.

Eleven of the 12 time series of colorometric quantities that comprised more than one reagents subseries were (as whole series) significantly autocorrelated as evidenced by the ACF plot. Among these 11 time series, there are two series that are significantly autocorrelated, as evidenced by both the ACF plot and the non-parametric run test.

The non-parametric run test shows that among 12 time series only two time series are autocorrelated, which contradicts the results based on the ACF plot. In each of the 12 cases, treating the reagents lot as a factor, the one-way ANOVA showed that the mean level changed significantly between reagents subseries. This was interpreted to indicate that the observed autocorrelation based on the ACF for the whole series was spurious and was largely a reflection of the change in the mean level caused by the shift of the reagents lot. We conclude that in this case the ACF plot is not appropriate to test the significance of the autocorrelation when possible mean changes occurred.

Of the 33 reagents subseries/series (excluding the time series that only included measurements using the same reagents lot) 11 were autocorrelated, seven as evidenced by the ACF plot alone and four as evidenced by the ACF plot as well as the non-parametric run test.

The ACF plots showed that all four electrolyte time series (L-NA, H-NA, L-K and H-K) and all diluent subseries (except those of L-K) were autocorrelated. Only in case of the L-NA series and the diluent subseries 1 of L-NA did the non-parametric run test show a significant lack of randomness. The ANOVA showed that in all diluent subseries except subseries 1 of L-K did the mean level change significantly between electrode subseries and in all time series except that of L-NA did the mean level change significantly between diluent subseries.

When the X charts were constructed based on the first 50 values and applied to the whole time series in each of 22 quantities, 153 (or 5.64%) values fell outside the control limits while 791 (or 29.1%) fell outside the control limits when the EWMA or EWMAST charts were used.

However, when the control charts based on the first 50 values of each subseries were applied to each subseries, only 19 out of 2714 values or 0.70% fell outside the control limits when the EWMA or EWMAST charts were used. Since 22 quantities are measured per day, this corresponds approximately to an alarm every (2714/22)/19 = 6.49 days. For the subseries showing a significant autocorrelation as based on the ACF, the result was 0.83% (13 out of 1564), while it was 0.52% (6 out of 1149) for those showing no autocorrelation. For the X chart, 12 out of 2714 values or 0.44% fell outside the control limits. This corresponds to an alarm every (2713/22)/12 = 10.3 days. For the subseries showing a significant autocorrelation as evidenced by the ACF plot, the result was 0.58% (9 out of 1149), and for those not showing a significant autocorrelation it was 0.26% (3 out of 1149).

One of the purposes of a control chart is to give a timely warning when something unusual is happening and only then. Thus, when a predictable change in the mean level is taking place, we do not want the control chart to give any warning. For this reason, a new control chart is constructed each time the QC material lot is changed because the mean level of the QC measurements is known to change somewhat when the lot is changed. In this study we used the ANOVA to demonstrate that the mean level of the QC measurements is also changing significantly whenever the analyzer is recalibrated or the electrolyte diluent lot is changed. In the majority of cases the assumption of independence of measurements within groups was met. In a few cases, however, the non-parametric run test of randomness showed that the time series compared in the ANOVA were not random and thus the assumption of independence of the measurements was not met. The conclusions therefore should be tempered accordingly in these cases. Still it is fair to conclude that for the majority of quantities as measured using the analyzer chosen by us, recalibration subsequent to a change of reagents lot does not prevent a significant change in the mean level to take place and a change of diluent lot causes the mean level of the electrolytes to change significantly. Furthermore, the magnitudes of these changes were such that they were counterproductive to the purpose of establishing an early warning system since they created a lot of false alarms.

Our recommendation is that prior to each change of reagents (or diluent) lot for the measurement of the patient samples, QC measurements are made using the new lot and done daily for 30 – 50 days in parallel with the QC measurements made using the old lot to build a historical data base. This data base then could be used to construct a control chart that is applied when the reagents (diluent) lot is changed. Since autocorrelation could change with a change of reagents lot, it is recommended that the autocorrelation structure be updated over time and the EWMAST chart used in place of the EWMA chart in the presence of a significant autocorrelation. Since the magnitude of the autocorrelation may vary over time, it is still an unresolved problem, though, that the test for autocorrelation has to be based on the most recent 50 observations while the correction is applied prospectively. We found that changing the disposable electrode had a significant effect on the mean. It is not practical though to construct a new control chart each time the electrode is changed. Although the electrode is being changed repeatedly and the mean level therefore is changing, the X chart is working well in the sense that not too many false alarms are generated. However, this type of warning system will miss changes in the mean level that are not substantially higher than that caused by changing the electrode.

In summary, it is recommended to have new control charts constructed each time when the reagents lot is changed. The autocorrelation structure should be monitored and whether the EWMA or EWMAST chart is used should depend on whether autocorrelation is present or not. For the electrolytes it is recommended that the X chart be used and that new control charts are constructed each time the diluent lot is changed.

References

Winkel, P. and Zhang, N. F. (2004). Serial Correlation of Quality Control Data – on the Use of Proper Control Charts, Scandinavian Journal of Clinical Laboratory and Investigation, 64, 195-204.

Zhang, N. F. (1998). A Statistical Control Chart for Stationary Process Data, Technometrics, 40, 24-38.

3.12 Fiducial Generalized Confidence Intervals

Jan Hannig
Statistical Engineering Division, ITL
Colorado State University, Fort Collins, CO

H. K. Iyer
Statistical Engineering Division, ITL
Colorado State University, Fort Collins, CO

Paul Patterson
Colorado State University, Fort Collins, CO

Generalized confidence intervals (GCI) have been used in many practical problems where traditional frequentist approaches do not provide useful solutions. In spite of the large number of successful applications of GCIs reported in the literature, it is surprising that there are no published results that discuss either small sample properties of GCIs or their asymptotic behavior. We show that, under reasonable assumptions, GCIs have asymptotically correct frequentist coverage. This result provides a frequentist justification for GCIs.

Tsui and Weerahandi (1989) introduced the concept of generalized P-values and generalized test variables which are useful for developing hypothesis tests in situations where traditional frequentist approaches do not provide useful solutions. Subsequently, Weerahandi (1993) introduced the concept of a generalized pivotal quantity (GPQ) for a scalar parameter θ, using which one can construct an interval estimator for θ in situations where standard pivotal quantity-based approaches may not be applicable. He referred to such intervals as generalized confidence intervals. Since then, several GCIs have been constructed in many practical problems. These intervals do not always have exact frequentist coverage. Nevertheless, results of simulation studies reported in the literature appear to support the claim that coverage probabilities of GCIs are sufficiently close to their stated value that they are in fact useful procedures in practical problems. In spite of the large number of successful applications of GCIs reported in the literature, it is surprising that there are no published results that discuss either small sample properties of GCIs or their asymptotic behavior.

A simple test case for the application of GCIs is the Gaussian two-sample problem with heterogeneous variances where one is interested in a confidence interval for the difference of the two means. This is the well-known Behrens-Fisher problem for which Behrens had proposed a solution in 1929 and later, in 1935, Fisher gave a justification for it based on the fiducial argument. More recently, Weerahandi (1993) derived a GPQ for the difference between the two means and remarked that the resulting interval coincided with the fiducial solution.

In this work we have identified an important subclass of GPQs, which we call Fiducial Generalized Pivotal Quantities (FGPQ) for reasons to be discussed shortly. We also provide some general methods for constructing FGPQs for large classes of problems. Nearly every published GCI can be obtained using FGPQs. More importantly, and perhaps of greater interest to practitioners, we also show that, under reasonable assumptions, GCIs based on FGPQs have asymptotically correct frequentist coverage. This result provides a frequentist justification for GCIs (and also for generalized tests, although our focus here is confidence intervals) when the generalized pivotal quantity is chosen appropriately. Additionally, we provide a number of examples to illustrate these results.

The reason that we chose the name FGPQ is that generalized confidence intervals based on FGPQs are, in fact, obtainable using the fiducial argument of R. A. Fisher within a suitably chosen framework such as the pivotal quantity approach of Barnard, structural inference of Fraser, and functional-model basis for fiducial inference discussed by Dawid and Stone (1982). We establish the connection between FGPQs and fiducial distributions by showing that, given a fiducial distribution for a parameter, there is a systematic procedure for constructing a FGPQ whose distribution is the same as the fiducial distribution. As a byproduct, this connection of FGPQs with fiducial inference leads to a frequentist justification for fiducial inference in many settings. In fact, FGPQs provide a natural framework for associating a distribution with a parameter.

The main result of our work is a theorem to the effect that, under fairly general conditions, GCIs obtained from FGPQs have correct asymptotic coverage. We consider some familiar examples and illustrate the application of the theorem. We then turn our attention to some general methods for constructing FGPQs. First we describe a simple recipe for constructing FGPQs and discuss the scope of application of this recipe. The procedure is illustrated with some examples of GPQs previously not discussed in the literature. We show that these GPQs (actually FGPQs) satisfy the conditions of the main theorem so that the resulting GCIs are guaranteed to have the correct frequentist coverage asymptotically. We also introduce two additional methods for constructing FGPQs that extend the range of problems for which GCIs can be developed. The application of these methods is illustrated with new confidence intervals for some well-known problems. We discuss connections between GPQs and fiducial inference and touch upon non-uniqueness issues associated with GCIs and fiducial intervals. This work has been submitted to JASA.

3.13 An Application of Combining Results from Multiple Methods - Statistical Evaluation of Uncertainty for SRM 1508a

A. Hornikova and N. F. Zhang
Statistical Engineering Division, ITL

M. Welch
Analytical Chemistry Division, ACSL

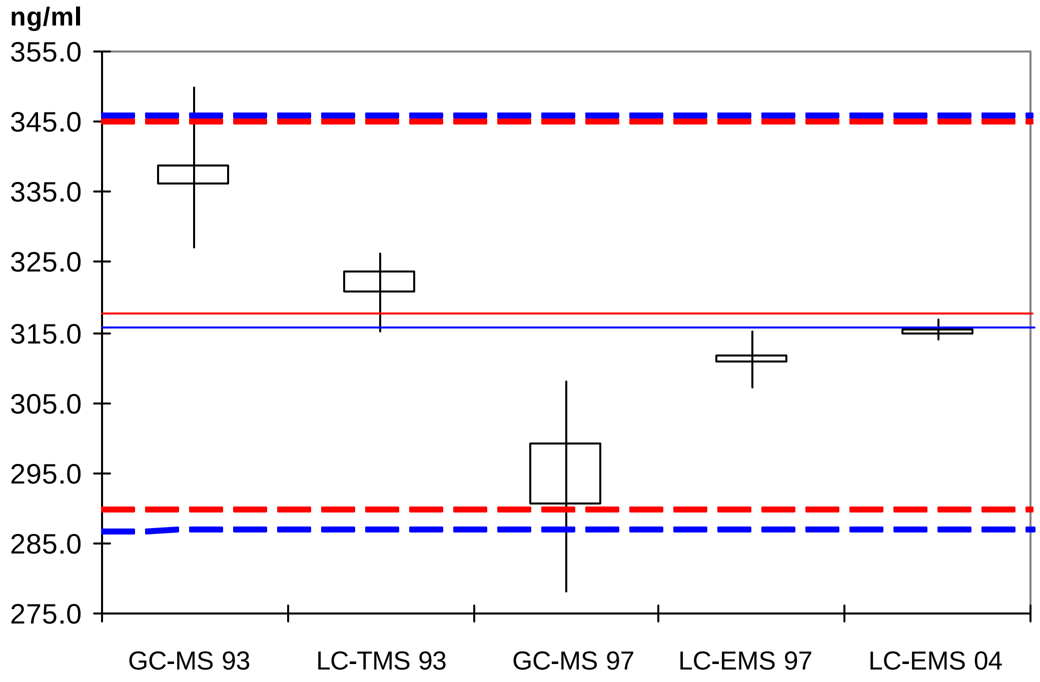


Figure: Box plots for the benzoylecgonine concentrations based on measurement results for Level 3.

The NIST Standard Reference Material (SRM) 1508a is used for validating methods for the determination of benzoylecgonine (cocaine metabolite) in human urine. The SRM was produced in 1993. Currently, the SRM has been re-measured and is going to be recertified. There are three benzoylecgonine nominal concentration values (called levels) for this SRM. For each level, the data consist of the measurement results obtained from three distinct analytical methods at NIST in different years, i.e., methods GC-MS and LC-TMS in 1993, GC-MS and LC-EMS in 1997, and LC-EMS in 2004. The sample sizes of the measurements for these methods are different.

The recertification is based on the combination of the measurement results from these methods for these years. For statistical analysis, we treated each method/year as one individual method. Therefore, for each level we have results from five measurement methods. In this study, we considered several different statistical models and corresponding estimators for the certified value and its uncertainty.
First, for a given level we assume a simple model with equal mean:

$$X_{ij} = m + e_{ij} \qquad (1)$$

where X_{ij} is the jth measurement readout based on the i^{th} method ($i = 1,2,...,p$ and $j = 1,2,...,n_i$) and m is the mean value of the concentration of benzoylecgonine. The e_{ij}'s are the random errors with $\mathrm{var}[e_{ij}] = s_i^2$ for the ith method. In this case, a weighted mean, in particular the Graybill-Deal estimator, can be used to obtain the certified value, with weights based on the reciprocals of the s_i^2's, or their estimators. However, based on the data for some levels, the means corresponding to the measurement methods are significantly different. Thus, the model may not be appropriate and it was our understanding that the uncertainties of the certified values seem too small in those cases.

Another model we considered is the random effects model

$$X_{ij} = m + a_i + e_{ij} \qquad (2)$$

where a_i is the random effect of the ith method with $E[a_i] = 0$ and $\mathrm{var}[a_i] = s_a^2$, and e_{ij} is the random error with $\mathrm{var}[e_{ij}] = s_i^2$. The model in (2) is similar to the model in (1) except that it includes an additional variance component modeling the between subset variability. Note that the means of the X_{ij} are the same for different measurement methods, as is also true for the model in (1). Several estimators were used, including MLE, the Kenward-Rogers estimator, and the Mandel-Paule's estimator, yielding similar results. However, we found that the resultant confidence intervals for the mean are sometimes too wide and therefore are not suitable.

Lastly, we considered a Bayesian approach [1] to estimate the consensus mean from the results of multiple methods. The model assumes that the mean values and variances of the subsets are not necessarily the same. That is, for each level

$$X_{ij} = m_i + e_{ij} \qquad (3)$$

where m_i is the true value of the measurand determined by the ith method. In general, a weighted mean as the estimator of the consensus mean and the corresponding uncertainty are given in Proposition 3 in [1], treating the unweighted mean as a special case. For illustration, in the figure above, the blue and red solid lines represent the consensus means obtained by unweighted and weighted mean estimators, respectively, based on the Bayesian approach. The blue and red dotted lines mark the 2-sigma coverage intervals for the consensus mean, using the unweighted and weighted mean estimators, respectively. Since the Bayesian approach accommodates the differences among the means as well as the variances, it is more flexible and yielded satisfactory results.

We conclude that since the Bayesian method provided convincing results, the estimated consensus means and their uncertainties are used for the certificate of this SRM.

Reference

[1] H. Liu and N. Zhang (2001), Bayesian approach to combining results from multiple methods, Proceedings of the Section on Bayesian Statistical Science, American Statistical Association.

4 Statistics for Systems

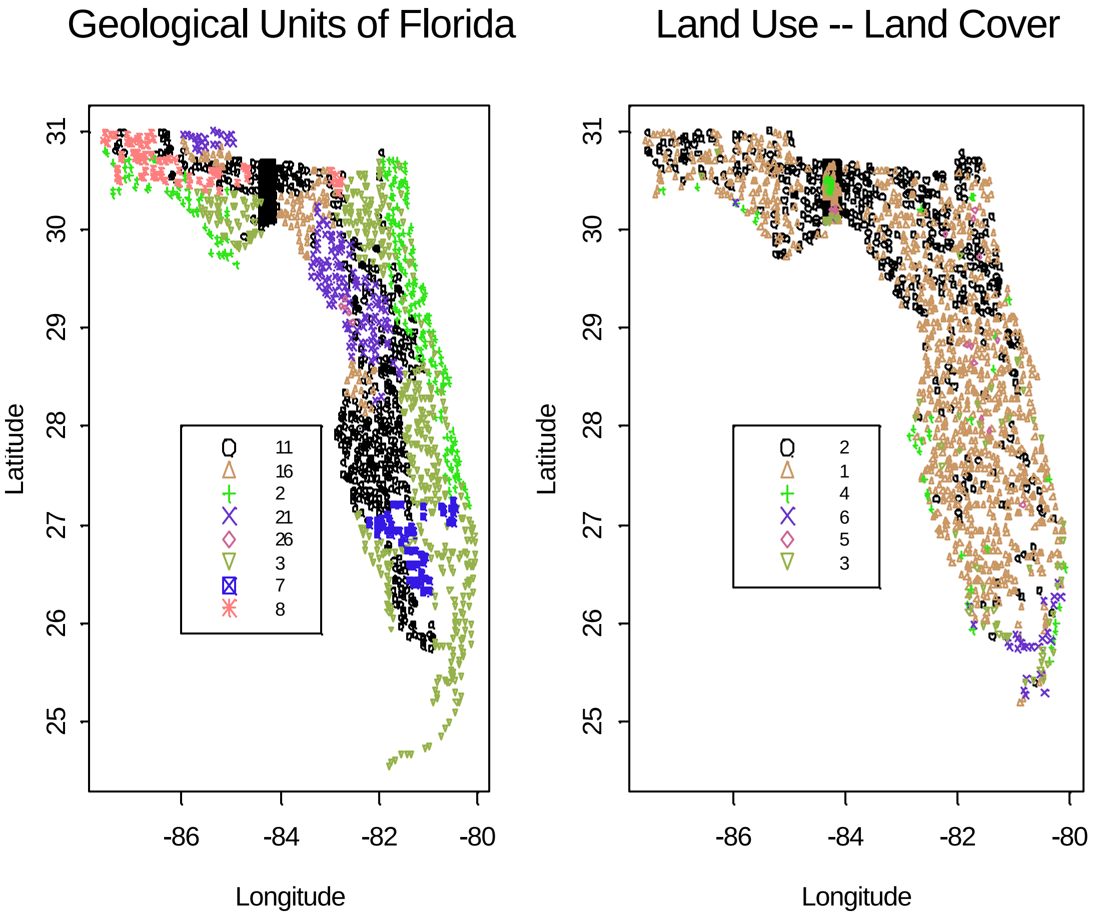
Research at NIST is heavily engaged with complex systems: natural physical systems, engineering systems for instrumentation and measurement, and computational systems for simulation and visualization. The roles of statistics are several: modeling to mimic and understand natural system behavior, data analysis to characterize engineering system properties and software development to govern system performance, and verification and validation of computational systems. Each of these tasks requires immersion in the science or engineering of the system itself as part of a multi-disciplinary team approach with the common goal of bringing to bear deep understanding of the system from scientifically diverse points of view.

Long-term collaboration on the study of ultra-cold subatomic particles has resulted in integration of stochastic elements and random behavior into mathematical models of particle behavior and to the delineation of the critical experiments to test the particle theory. Internal software, embedded in high-precision instruments for self-calibration, and painstaking analysis and decomposition of noise for high-speed electronics have led to a new waveform paradigm now being explored and developed. An overwhelming challenge to combine four massive computational engines in sequence to simulate the complex failure process of the World Trade Center collapse must draw on classical experimental design strategies to organize the simulation experiment to test various theories of the initial failure mode and the causal sequence of structural failure. *Statistics for Systems* is not defined in terms of types of models; rather it is the complete integration of statistics and probability into the science and engineering.

4.1 Statistics for Geochemical Surveys

James Yen
Statistical Engineering Division, ITL

Andrew Grosz
United States Geological Survey



The figure on the left depicts the underlying geological formations that make up Florida's landmass. The figure on the right divides Florida into regions of different land utilization or land cover.

SED and the United States Geological Survey (USGS) have initiated an effort to apply modern statistical methods and expertise towards analyzing outstanding issues in geology and geochemistry that have arisen from USGS studies.

The U.S. Geological Survey is compiling a treasure trove of geochemical data based on thousands of samples taken throughout the United States. These data will include measurements from previous studies, surveys that are ongoing, and re-analyses of existing past samples. The relative quantities of a list of minerals have been measured by several analytical methods such as neutron activation analysis and hydride atomic absorption. In addition, the data may include various sample and site attributes, such as description of the vegetation around the sample site, and whether it was taken from soil sample or from stream sediments. We wish to explore how the different mineral concentrations associate with each other and with various attribute variables.

As an example, excessively high ground concentrations of arsenic can be harmful to human health. It will be useful to estimate baseline arsenic levels in order to determine which areas are suitable for remediation. An area's mineral concentration levels are intimately related to its geologic origins. For example, is the land part of an ancient seabed, or do the predominant rock formations have a volcanic origin? The first picture on the previous page shows the division of the state of Florida into different geologic regions; more exactly, it shows the samples where information about the underlying geologic formations is available. The classifications are nominal. The thick band in the Florida panhandle is a graphical artifact resulting from a stand-alone, high-density study in the Tallahassee area.

There can also be anthropomorphic reasons for elevated arsenic levels. For instance, agricultural areas may contain a residue from repeated applications of pesticides. The picture on the right depicts the sample locations in Florida where land use/land cover information is available. As with the other picture, this figure has nominal classification classes and the thick band in the Tallahassee area. We wish to employ multivariate spatial techniques that will disentangle the different environmental and geologic effects related to the concentration level of metals like arsenic.

There are a host of other questions that bear investigation. For instance, some groups of metals tend to occur in tandem, indicating the need for compositional methods involving log-ratios. Also, many observations of mineral levels are below the threshold of detection; how do we best incorporate those points into our analysis? Finally, there are questions on how best to depict the data involving multiple minerals into maps and how to calculate and incorporate uncertainties into maps.

Future work may include possible joint research efforts with the U.S. Army, DARPA, and the metals industry, with the objective of locating within the United States more viable sources of titanium, a metal for which the critical military and industrial demand greatly exceeds the current supply. On the statistical front, we hope that this work leads to advances in data mining techniques for spatial multivariate data that are applicable to myriad problems other than metals mining. A long-term goal is development of a metrology of mapping that associates appropriate uncertainties with maps.

4.2 Charpy Impact Machine Verification Program

Jolene Splett and Jack Wang
Statistical Engineering Division, ITL

Chris McCowan and Tom Siewert
Materials Reliability Division, MSEL

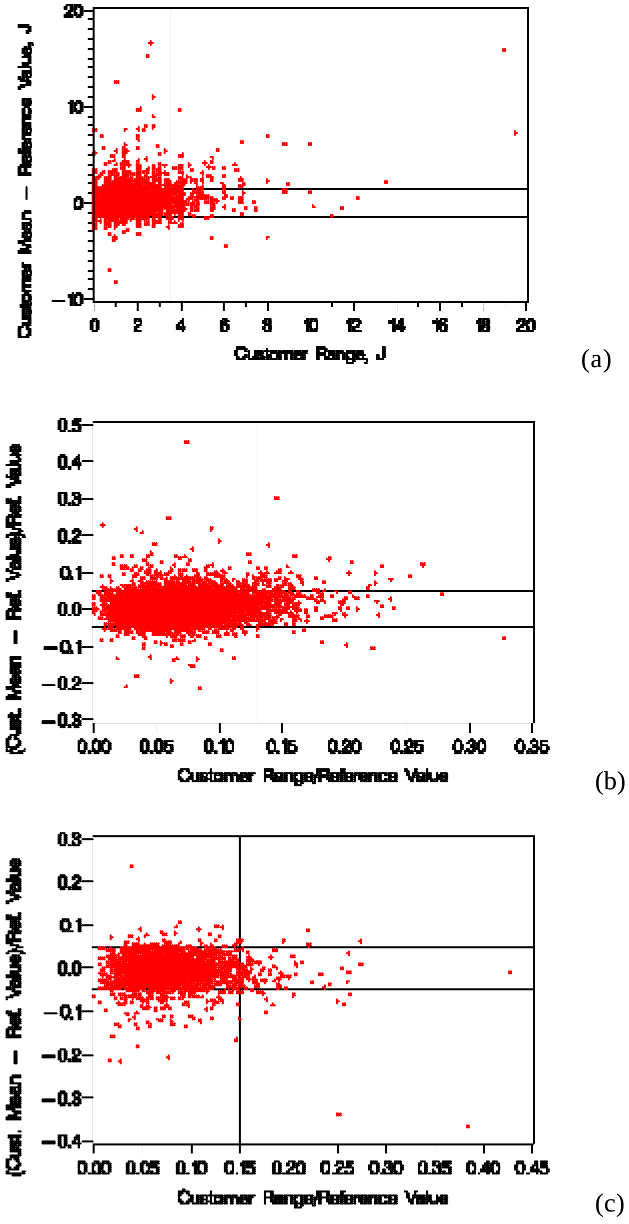


Figure 1. (a) Deviations of customer average from reference value versus customer range for the low energy level. Plots (b) and (c) show normalized deviations of customer average from reference value versus customer range for high and super-high energy levels, respectively. Horizontal reference lines represent current ASTM E-23 verification limits; vertical reference lines represent proposed limits to the customer range.

The Charpy impact test is one of the most common tests used to quantify the breaking strength of materials. The test is implemented by striking a small, rectangular metal specimen with a large pendulum and recording the energy absorbed by the specimen as it breaks.

NIST administers a program to verify the performance of Charpy impact machines by selling specimens with certified breaking strength. The verification program works as follows. NIST obtains a pilot lot of 75 Charpy specimens from a supplier and then measures the breaking strength of the specimens using three master machines. If the measurements meet certain criteria, then the rest of the specimens are machined and sent to NIST. An additional 15 specimens are selected at random from the lot and broken. If the breaking strength of the additional specimens is in agreement with the pilot lot, then the lot is certified as a reference material by NIST. Customers test sets of five specimens and then return the broken specimens and observed values to NIST for analysis. The data are stored in a database for future reference.

The Charpy verification program is conducted in accordance with ASTM Standard E-23. However, the verification limits stated in E-23 are somewhat arbitrary and there is no limit to variation. We analyzed historical data comprised of customer measurements and summarized the results in the paper, “Analysis of Charpy Impact Verification Data: 1993-2003.” The goal of the paper was to provide evidence that the current verification limits were performing well and to propose a limit to variation. Since the true breaking strength of a verification specimen is unknown, it is impossible to evaluate the absolute performance of Charpy machines. We can, however, evaluate machines relative to each other. Thus, we recommended that the range of customer data be less than the 95th percentile of the ranges observed for the historical customer data.

The development of an international Charpy verification program is very desirable given the current global economy. A three-year international study to compare reference machines and verification specimens over time was completed in the spring of 2004. The data were analyzed by SED staff and the results documented in the paper “International Comparison of Impact Reference Materials (2004).” The paper was a joint effort by study participants at four national measurement laboratories. We found that verification specimens appeared to be stable over the three-year test period, and that the Charpy machines under test were significantly different from each other, especially at low energy levels.

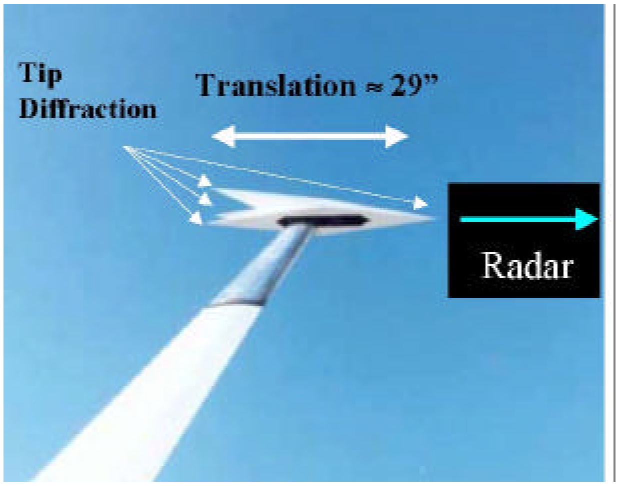
The two papers, co-authored by SED staff, were presented at the Second Symposium on Pendulum Impact Machines: Procedures and Specimens held on November 10, 2004 in Washington D.C. The purpose of the symposium, sponsored by ASTM Committee E28 on Mechanical Testing and its Subcommittee E28.07 on impact testing, was to discuss issues pertaining to Charpy impact machines and address practical problems faced by those who perform impact tests and those who develop the standards.

Accurately determining the breaking strength of metals is critical in the construction of bridges, buildings, and pressure structures. In FY2001, about 1000 customers participated in the Charpy impact machine verification program. The work completed in the past year will hopefully help change verification limits specified in ASTM E-23 and initiate discussion on the development of an international Charpy verification program.

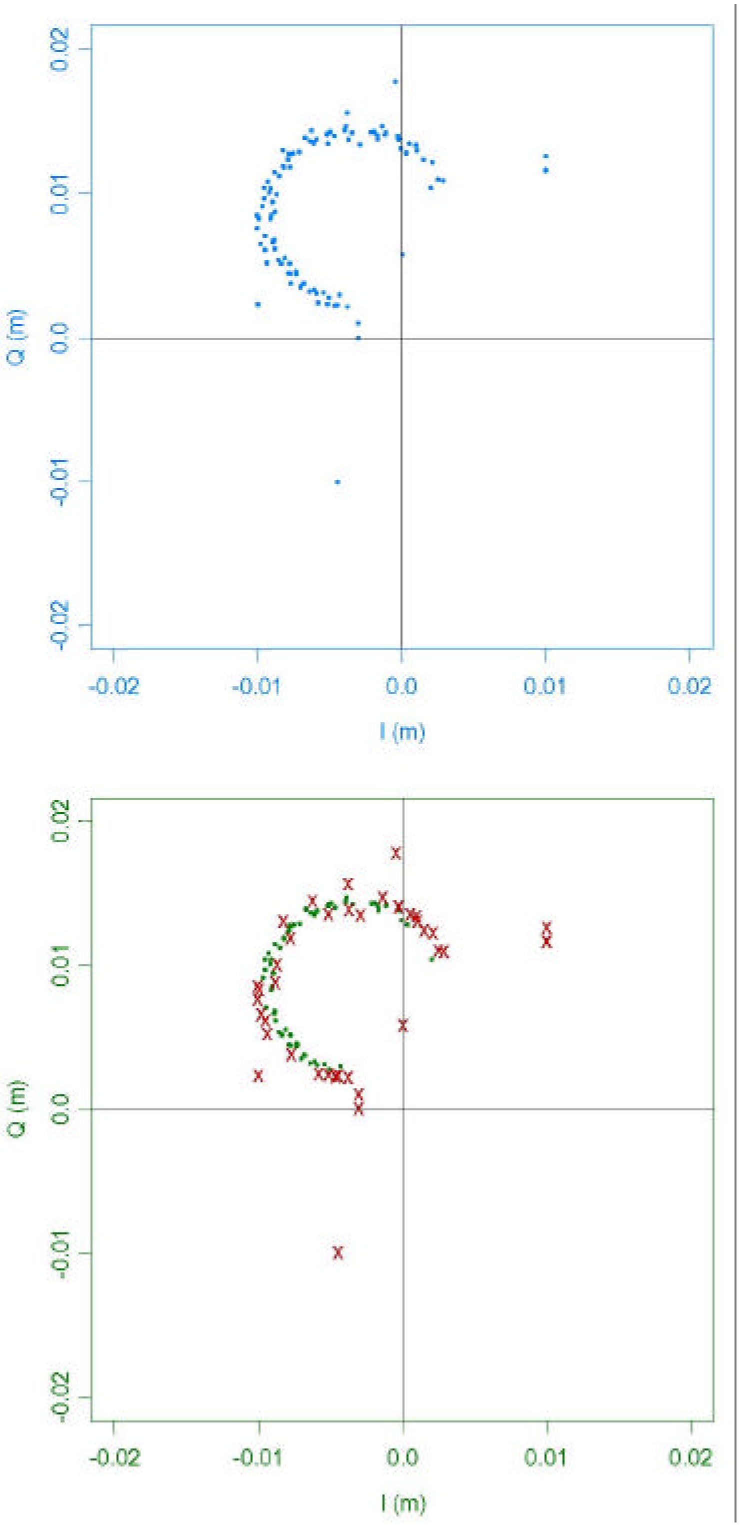
4.3 Robust Separation of Background and Target Signals in Radar Cross Section Measurements

C. M. Wang
Statistical Engineering Division, ITL

L. A. Muth
Electromagnetic Technology Division, EEEL



The top shows the Arrow III artifact target mounted on a pylon (top) to obtain phase-dependent RCS measurements to separate the target and background signals, primarily at VHF and UHF. The bottom shows a 69-inch diameter offset calibration cylinder to be mounted on a rotator on a pylon. When the cylinder is rotated, the phase of the received signal varies. By separating the cylinder and background contributions, calibration accuracy can be improved.



The complex electric field data includes RFI signals (top) that can degrade the determination of the background signal when measuring the Arrow III artifact. The RFI and other outliers (x) are identified (bottom) using the LMS criterion. A more accurate background signal is then determined using the ODR procedure.

Coherent measurements of radar cross section (RCS) on a target moving along the system line-of-sight in free space will trace a circle on a plane. The presence of additional complex background signals (including stationary clutter, target support and averaged target-mount interactions), which do not depend on target position, will translate the origin of the circle to some point on the plane. The presence of outliers (mostly due to rf interference) can introduce significant errors in the determination of the radius and center of the circle. We have implemented a combination of a robust and efficient algorithm to eliminate or reduce the influence of outliers, and then to separate the target and background signals. Concurrently, the influence of noise is also reduced. Thus, we can obtain both a target-independent estimate of the background and a background-free estimate of the RCS of calibration artifacts.

The RCS of a target is, by definition, the squared amplitude of the electric field scattered by a target located at infinity when illuminated by a plane wave. In practice, the reflected electric field is measured monostatically or bistatically at a large distance d between the target and the transmitting and receiving antennae, such that $kd >> 2p$, where k is the transmitted wavenumber. The measured signal S is the resultant of a theoretical electric field scattered by the target, plus fields that originate from the measurement environment, and distortions due to noise, drift, unknown instrumentation problems and rf interference (RFI). The measured complex electric field S can thus be written as

$$S(r,q,b,b) = re^{iq} + be^{ib} + I + n + d + o$$

where r and q are the amplitude and phase, respectively, of the reflected electric-field signal from the target (which could possibly include in-phase error signals), b and b are the amplitude and phase of the electric field originating from the environment (mostly clutter, but could also include calibration and instrumentation effects), I is the target-mount interaction, n is noise, d is drift, and o represents the outliers (e.g., rf interference that inevitably shows up in real data). The first term in the equation describes a circle centered on the origin as q varies from 0 to $2p$, and the second term is a constant that moves the center of the theoretical circle. Obviously, if we can determine and remove the background signal b from the data, the measurement error can be significantly reduced.

The orthogonal distance regression (ODR) can be appropriately used to obtain the parameters of phase-dependent RCS data, since both components have measurement errors. The ODR, however, lacks robustness in the sense that even a single outlier can significantly degrade the accuracy of the estimated parameters. As has been pointed out, our measurements (and all RCS measurements) are usually subject to a large number of outliers. We use a least median of squares (LMS) criterion as an outlier-identification tool in RCS measurements. Once the contamination by outliers has been reduced, we apply ODR to the filtered data to obtain estimates of the parameters r and b.

To develop an LMS-based outlier detection routine for RCS measurements, we use the parametric equation

$$Ax^2 + Ay^2 + Bx + cy + 1 = 0$$

where $A \neq 0$, to represent a circle (assuming that the circle does not pass through the origin). Given a set of circle data $\{(x_i, y_i),\ i = 1, \ldots, n\}$, the LMS estimate of parameters A, B, and C is obtained using

$$\min_{\tilde{A},\tilde{B},\tilde{C}} \text{median} \left\{ (\tilde{A}x_i^2 + \tilde{A}y_i^2 + \tilde{B}x_i + \tilde{C}y_i + 1)^2,\ i = 1,\ \ldots,\ n \right\}$$

Residuals obtained from the LMS fit are reliable indicators of outliers. Specifically, we use the following steps to identify outliers for RCS data. First, obtain parameter estimates $\tilde{A},\ \tilde{B},\ \tilde{C},$ together with a corresponding error scale estimate $\hat{s}$. Secondly, compute the associated absolute standardized pseudo-residual given by

$$R_i = \left| (\tilde{A}x_i^2 + \tilde{A}y_i^2 + \tilde{B}x_i + \tilde{C}y_i + 1) \right| / \hat{s},$$

and remove a point from the data if its residual value R_i is larger than a threshold, such as 2 or 3.

We applied the LMS criterion to data obtained using the Arrow III artifact target to determine the background response. We also used a rotating offset cylinder to improve calibration accuracy. In measurements on low-observable targets, the subtraction of the background signal from the measurement and calibration significantly improves the measurement accuracy. This technique is especially useful for sub-wavelength translations at VHF and UHF frequencies, where spectral techniques are not applicable because the available arc of data is limited.

4.4 Harmonic Phase Standards Drift

Dom Vecchia and Jolene Splett
Statistical Engineering Division, ITL

Jeff Jargon and Don DeGroot
Electromagnetics Division,EEEL

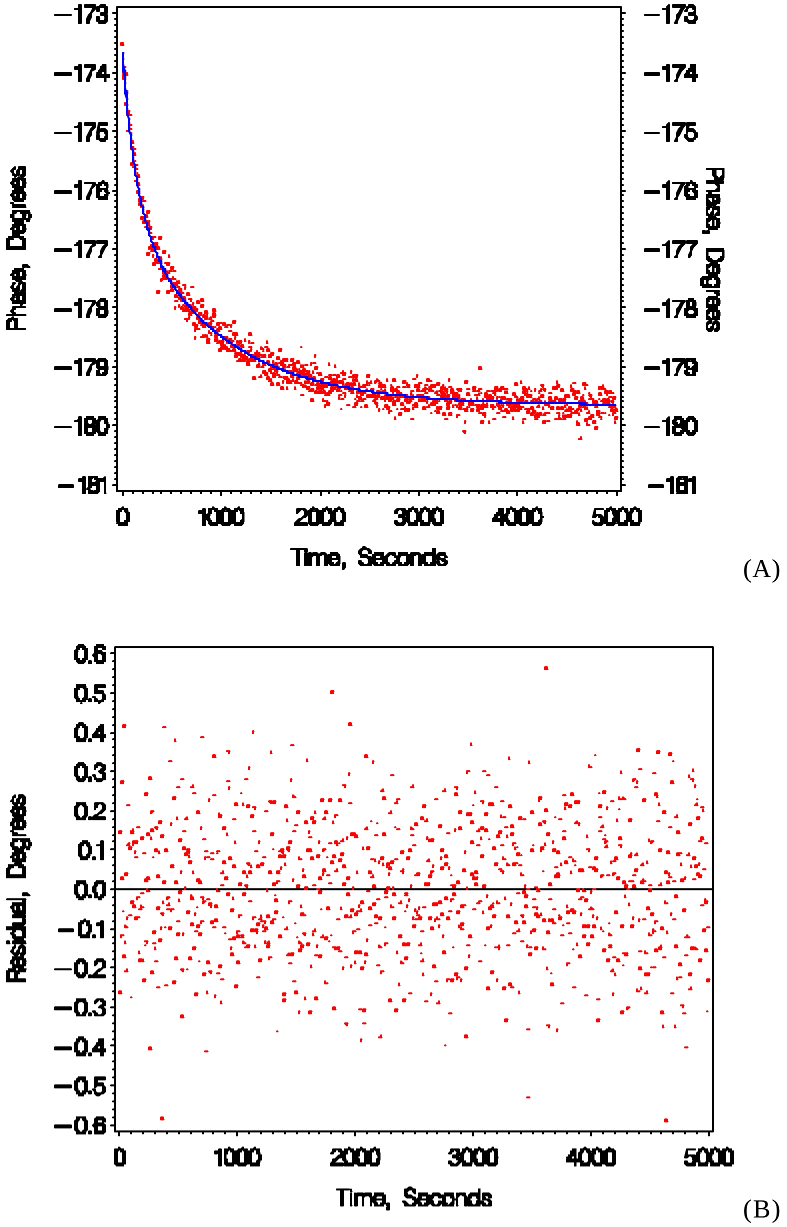


Figure 1. (A) Phase angles of the fifth harmonic of the 20 GHz harmonic phase standard along with the estimated curve using the double-exponential decay model. (B) Residuals from the fitted curve.

Radio-frequency (RF) measurements are applied extensively in the deployment of commercial wireless communication systems. They are crucial to all stages of system development, from device modelling to circuit design and system performance characterization. NIST develops and supports general methods for characterizing nonlinear components, circuits, or systems used in digital wireless communications, and transfers these methods through interactions with industrial research and development laboratories.

Nonlinear vector network analyzers (NVNAs) are instruments capable of characterizing nonlinear devices under large-signal operating conditions, which are typical in wireless communications. To do this, complex travelling waves are measured at the ports of a device both at the stimulus frequency (or frequencies) and at other frequencies that are part of the response, including harmonics and intermodulation products created by the nonlinearity of the device, in conjunction with impedance mismatches between the system and the device. The calibration of a commercial NVNA consists of three steps: a relative calibration, an amplitude calibration that makes use of a power meter standard, and a phase distortion calibration that makes use of a harmonic phase standard. All are performed on a frequency grid related to the source tones and the anticipated nonlinear response of the device.

The commercial harmonic phase standard (HPS), the main components of which are a power amplifier and a step recovery diode, is driven at a fundamental frequency and produces a harmonic series output signal. The HPS, which is used as a transfer standard, is characterized by calibration on an independent measurement system in order to “know” the phase relationship of each harmonic of the HPS output signal to the fundamental frequency. In a previous study of typical commercial HPSs, we reported the discovery of considerable drift in time of “known” phase angles, with realistic waiting time to stability that was much longer than the warm-up time of 120 seconds set by the manufacturer’s control software.

This year we developed an empirical model for the warm-up drift of an HPS, which enables us to predict the warm-up time required to produce stable phase-angle values on a future measurement occasion. We found that two first-order exponential decay terms produced excellent representations of drift data collected to date in repeated calibration runs on the HPS device. The two decay terms can reasonably be associated with the power amplifier and step recovery diode of the HPS.

Based on the nonlinear decay model, large-sample uncertainties for the measurement “error” function in time can be applied to produce suitable statistical intervals for a waiting time appropriate to the tolerance requirements specified by the operator. A paper entitled “Modeling Warm-Up Drift in Commercial Harmonic Phase Standards,” by Jeff Jargon, Jolene Splett, Dom Vecchia, and Don DeGroot was presented at the Conference on Precision Electromagnetic Measurements in London, UK, on June 27 - July 2, 2004, and is being revised for the IEEE Transactions on Instrumentation and Measurement.

As wireless networks are pushed beyond the limits of linear network analysis, large signal descriptions are required to characterize nonlinear components, circuits, and systems. NVNAs are the instruments used to make the necessary measurements on various nonlinear devices used in digital wireless communications. Statistical intervals for predicting waiting times to stability of drifting NVNA calibration standards will ensure the improved accuracy of NVNA measurements.

4.5 Lifetime of Magnetically Trapped Neutrons

K. J. Coakley and G. L. Yang
Statistical Engineering Division, ITL

P. R. Huffman and A. K. Thompson
Ionizing Radiation Division, PL

L. van Buuren, S. N. Dzhosyuk, C. E. H. Mattoni, S. E. Maxwell, D. N. McKinsey,
L. Yang, and J. M. Doyle
Harvard University

R. Golub and E. Korobkina
Hahn-Meitner-Institut, Berlin

S. K. Lamoreauxa
Los Alamos National Laboratory

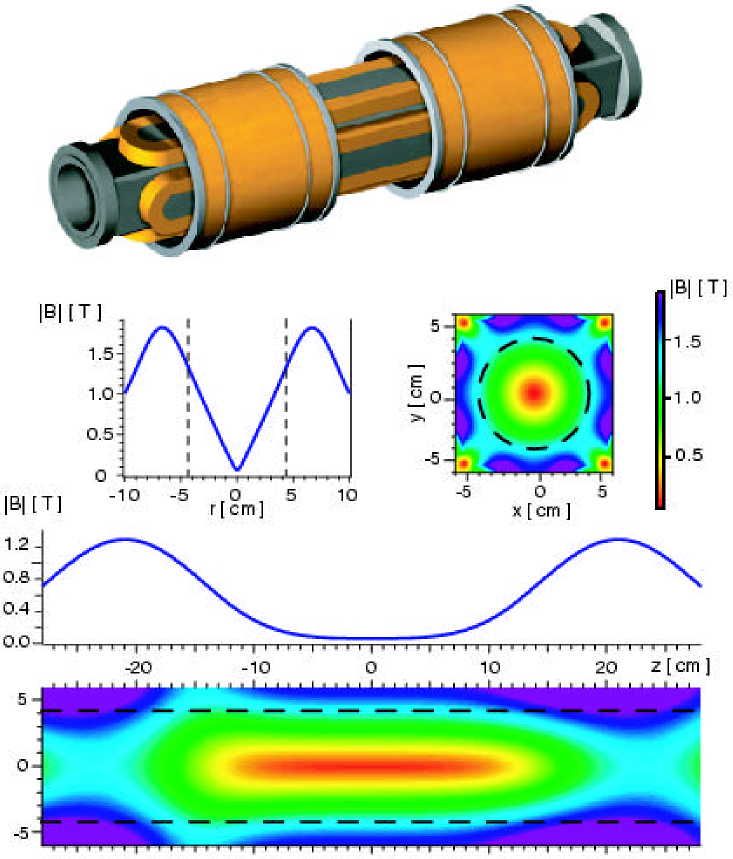


Figure 1. Magnet trap for confining ultracold neutrons. A rendering of the magnet coils and form is shown at the top. The two graphs depict the magnitude of the magnetic field as a function of *r* (along *z*=0) and *z* (along *x*=0). The two contour plots show the two-dimensional field profiles in the *x*-*y* plane (at *z*=0) and in the *x*-*z* plane (at *y*=0). The dashed lines denote the physical walls of the trap.

Along with other experimental data, the mean lifetime of the neutron allows one to test the consistency of the standard model of electroweak interactions. Further, the mean lifetime of the neutron is an important parameter in astrophysical theories. Statistical contributions include: birth-death stochastic models of neutron trapping process; likelihood models and estimation algorithms; optimal data collection strategies; methods for optimal redesign of apparatus; methods to account for background; and stochastic modeling of marginally trapped neutrons.

In 1999, a team of researchers from Harvard University, Los Alamos National Laboratory, University of Berlin, and NIST succeeded in producing and confining polarized Ultra Cold Neutrons (UCN) in a magnetic trap. In addition to the neutron lifetime experiment described here and other fundamental physics experiments, ultracold neutrons (UCN) have great potential in other major areas of research including neutron reflectometry and Quasi-Elastic neutron scattering. Neutron reflectometry is a technique which probes the composition and ordering of materials at surfaces and interfaces. Quasi-elastic scattering is a general term given to scattering events in which the energy change of the neutron is very small compared with the neutron's kinetic energy. Among the interesting cases for study using quasi-elastic scattering are large biological molecules and polymers. UCN offer a very interesting probe for the study of the dynamics of large molecules.

Data from the first generation neutron lifetime experiment using UCN yielded a neutron lifetime estimate of 660 s. The 68 percent confidence interval for this estimate is (490 s, 950 s) [1-3]. Along with other experimental data, the mean lifetime of the neutron allows one to test the consistency of the standard model of electroweak interactions. Further, the mean lifetime of the neutron is an important parameter in astrophysical theories. Although this proof-of-principle result is not as precise as the currently accepted value (885.7 s with a 1-sigma uncertainty of 0.8 s), a planned second generation experiment should yield a neutron lifetime more precise than the current value. Furthermore, systematic errors should be much lower than in other kinds of neutron lifetime experiments.

At the NIST Center for Neutron Research, ultracold neutrons are produced by inelastic scattering of cold neutrons from a reactor in superfluid ^{4}He. By creation of a single phonon in the superfluid, a cold neutron with wavelength near 0.89 nm can be scattered to a state of near rest. (The mean wavelength of a thermal ensemble of neutrons at 12 K is 0.89 nm (8.9 ?).) Very low energy neutrons are trapped in a potential field formed by the interaction of the neutron magnetic moment and a spatially varying magnetic field. The corresponding temperature of the trapped neutrons is less than 1 mK. When the trapped neutrons decay, they produce energetic charged particles that generate scintillations in the liquid helium. The scintillations are detectable with nearly 100 percent efficiency.

Statistical Analysis

Statistical contributions fall in two general areas. We have developed stochastic models for the experimental data as well as estimation procedures based on either binned data or arrival time data. Based on our stochastic models, we have studied a variety of strategies for estimation of the neutron lifetime. A primary consideration is how to efficiently estimate mean neutron lifetime with neutron decay data that are confounded with background noises. For instance, in one approach, we fit a model to the data from the primary experiment in which neutron decay signals are contaminated by background [4,5]. In another strategy, two separate experiments are performed. One experiment measures pure background signals (to be called the background-only experiment) and the other is the primary experiment of measuring neutron decays that contain unavoidable background signals. Neutron decay data are corrected for background with observations from the background-only experiment before the data are used for mean lifetime estimation. In yet another strategy, we estimate the mean neutron lifetime using the joint likelihood with the data from the primary experiment and the background-only experiment [7].

The primary experiment is composed of two stages of durations T_f and T_d, respectively. In the first stage, neutrons are generated and trapped magnetically, and in the second stage neutron decay signals as well as background noises are recorded. According to our birth-death stochastic model of the trapping process [4, 5], the expected number of trapped neutrons is $l\,t\,(1-\exp(-T_f/t))$, where l is the rate at which neutrons enter the trap, and t is the mean lifetime of the neutron. We developed a method to determine the optimal choice of the fill time T_f and the time spent observing decay events, T_d. The optimal values, found by simulations, minimize the asymptotic standard error of the lifetime estimate. For the case where a two-parameter exponential model is fit to background-corrected data, we determined the optimal ratio of "background-only" measurements to primary measurements for various models of the background as well as optimal values of T_f and T_d [6]. For the case where a more complex model is fit to joint likelihood using realizations of the background-only measurement and the primary measurement, we determined T_f, T_d, and R [7].

Based on our statistical analysis, a second generation version of the original experimental apparatus was redesigned so as to minimize the uncertainty associated with the lifetime estimate. In this study, candidate designs produced different background signals and different neutron intensities.

During the last year, SED staff focused on modeling the complex behavior of marginally trapped neutrons [9]. These neutrons have sufficient energy to escape the trap, but do not do so immediately. The decay of such marginally trapped neutrons can introduce systematic error into the lifetime estimate. Hence, they must be purged from the trap by varying the trapping potential in time.

Along with other experimental data, the mean lifetime of the neutron allows one to test the consistency of the standard model of electroweak interactions. Further, the mean lifetime of the neutron is an important parameter in astrophysical theories.

For more information, visit: http://www.doylegroup.harvard.edu/neutron/neutron.html.

Selected Refereed Publications

1. P. R. Huffman, C. R. Brome, J. S. Butterworth, K. J. Coakley, M. S. Dewey, S. N. Dzhosyuk, R. Golub, G. L. Greene, K. Habicht, S. K. Lamoreaux, C. E. H. Mattoni, D. N. McKinsey, F. E. Wietfeldt, and J. M. Doyle, "Magnetic Trapping of Neutrons," Nature, 403, 62-64 (2000).

2. 2. P. R. Huffman, C. R. Brome, J. S. Butterworth, K. J. Coakley, M. S. Dewey, S. N. Dzhosyuk, D. M. Gilliam, R. Golub, G. L. Greene, K. Habicht, S. K. Lamoreaux, C. E. H. Mattoni, D. N. McKinsey, F. E. Wietfeldt, and J. M. Doyle, "Progress Towards Magnetic Trapping of Ultracold Neutrons," Nuclear Instruments and Methods A, 440(3), 522-527 (2000).

3. C. R. Brome, J. S. Butterworth, K. J. Coakley, M. S. Dewey, S. N. Dzhosyuk, R. Golub, G. L. Greene, K. Habicht, P. R. Huffman, S. K. Lamoreaux, C. E. H. Mattoni, D. N. McKinsey, F. E. Wietfeldt, and J. M. Doyle, "Magnetic Trapping of Ultracold Neutrons," Physical Review C, 63, 055502 (2001).

4. K. J. Coakley, "Statistical Planning for a Neutron Lifetime Experiment Using Magnetically Trapped Neutrons," Nuclear Instruments and Methods A, 406, 451 (1998).

5. G. L. Yang and K. J. Coakley, "Likelihood Models for Two-Stage Neutron Lifetime Experiments," Physical Review C, 63, 014602 (2001).

6. K. J. Coakley, "Neutron Lifetime Experiments Using Magnetically Trapped Nutrons: Optimal Background Correction Strategies," Nuclear Instruments and Methods A, 469, 354 (2001).

7. K. J. Coakley and G. L. Yang, "Estimation of the Neutron Lifetime: Comparison of Methods Which Account for Background," Physical Review C, 65, 064612 (2002).

8. P. R. Huffman, K. J. Coakley, S. N. Dzhosyuk, R. Golub, E. Korobkina, S. K. Lamoreaux, C. E. H. Mattoni, D. N. McKinsey, A. K. Thompson, G. L. Yang, L. Yang, and J. M. Doyle, "Progress Towards Measurement of the Neutron Lifetime Using Magnetically Trapped Ultracold Neutrons," In H. Abele and D. Mund, eds., Quark-Mixing, CKM Unitarity, Proceedings of the Two-Day-Workshop "Quark Mixing, CKM-Unitarity", September 19-20, 2002, Heidelberg, Germany (Mattes Verlag Heidelberg, 2003) p. 97.

9. K. J. Coakley, J. M. Doyle, S. N. Dzhosyuk, L. Yang, and P. R. Huffman, "Chaotic Scattering of Marginally Trapped Neutrons," submitted to NIST Journal of Research.

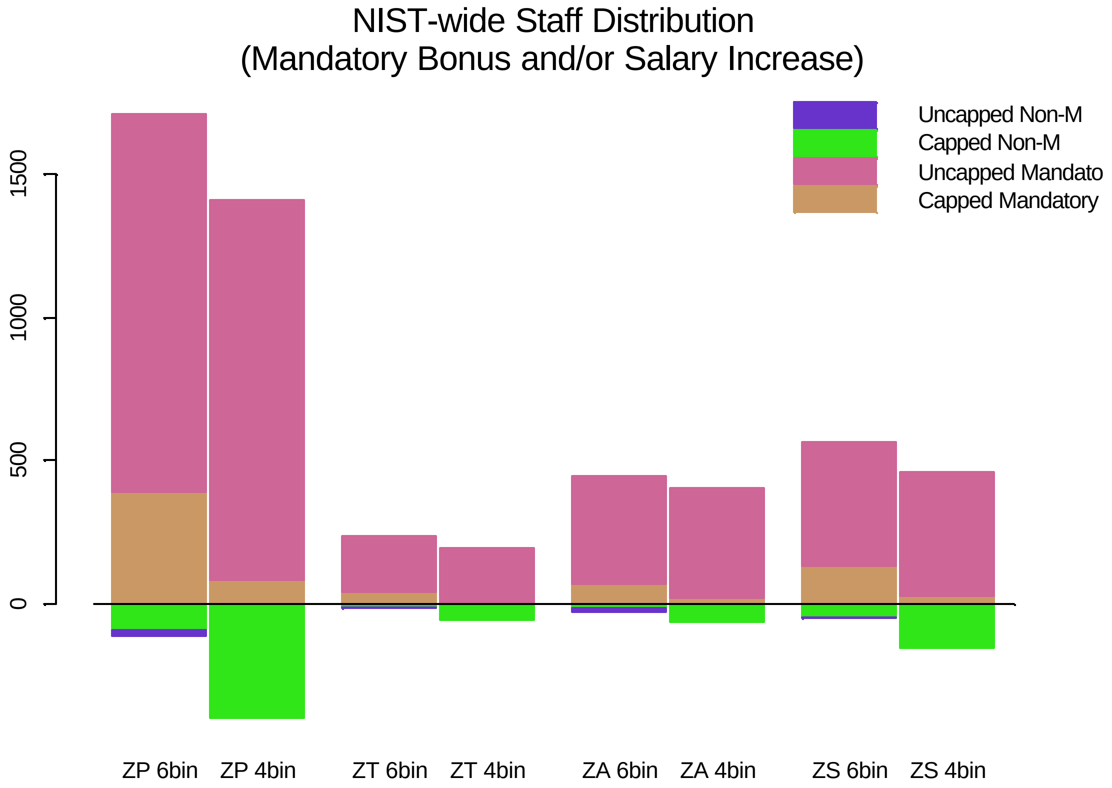
4.6 Pay-for-Performance System (PPS) Analysis

Hung-kung Liu and Nell Sedransk
Statistical Engineering Division, ITL

Harry S. Hertz
Baldrige National Quality Program, OD

Robert Kirkner
Human Resources Management Division, DACFO

Joe Kau
Applications System Division, DACFO



Comparisons of NIST staff distribution under different proposed systems.

The NIST Authorization Act for Fiscal Year 1987 established the NIST Personnel Management Demonstration Project to demonstrate an alternative personnel management system. It is now referred to as the NIST Alternative Personnel Management System (APMS). One of the key concepts of the APMS is pay for performance. The current NIST performance appraisal uses a 2-level rating system, a 100-point scoring system, and performance ranking within peer groups.

.

In recent years, the NIST Employee Survey and the Research Advisory Committee's Report to the Director have all identified a need for improvement to the NIST's current pay-for-performance system (PPS). The two major areas of concern are difficulties in understanding the method of scoring and determining pay increases, and the payout variations among employees with the same score within an OU or between OU's. To address these concerns, various proposals have been developed by the People Council, some OU directors, and the NIST Director.

SED, in collaboration with members of the People Council and the Office of Human Resources Management, analyzed graphically the 2001 and 2002 performance data to demonstrate the changes these different proposed systems will have on the current system, and to determine whether the intended improvements will be achieved by each proposed system. After several presentations to the Senior Management Board, a final proposal was selected that proposes four major changes to the current PPS:

(1) replace scoring and ranking with six performance levels;
(2) link performance pay increases to the six performance levels;
(3) convert pay increases to bonuses for pay-capped high performers; and
(4) place annual cost-of-living increase at risk for low performers.

We then performed a large-scale simulation study on the selected system when applied to 2 labs and 1 extramural OU using the 2003 performance data. Our simulation results clearly show a narrowing of the within OU dispersion in pay increases granted to employees. These results were summarized in the February 2004 staff briefings on the proposed changes to the pay-for-performance system.

Presently, SED is collaborating with staff of the Applications System Division in developing new PPS software to support different NIST user groups and their payroll management tasks.

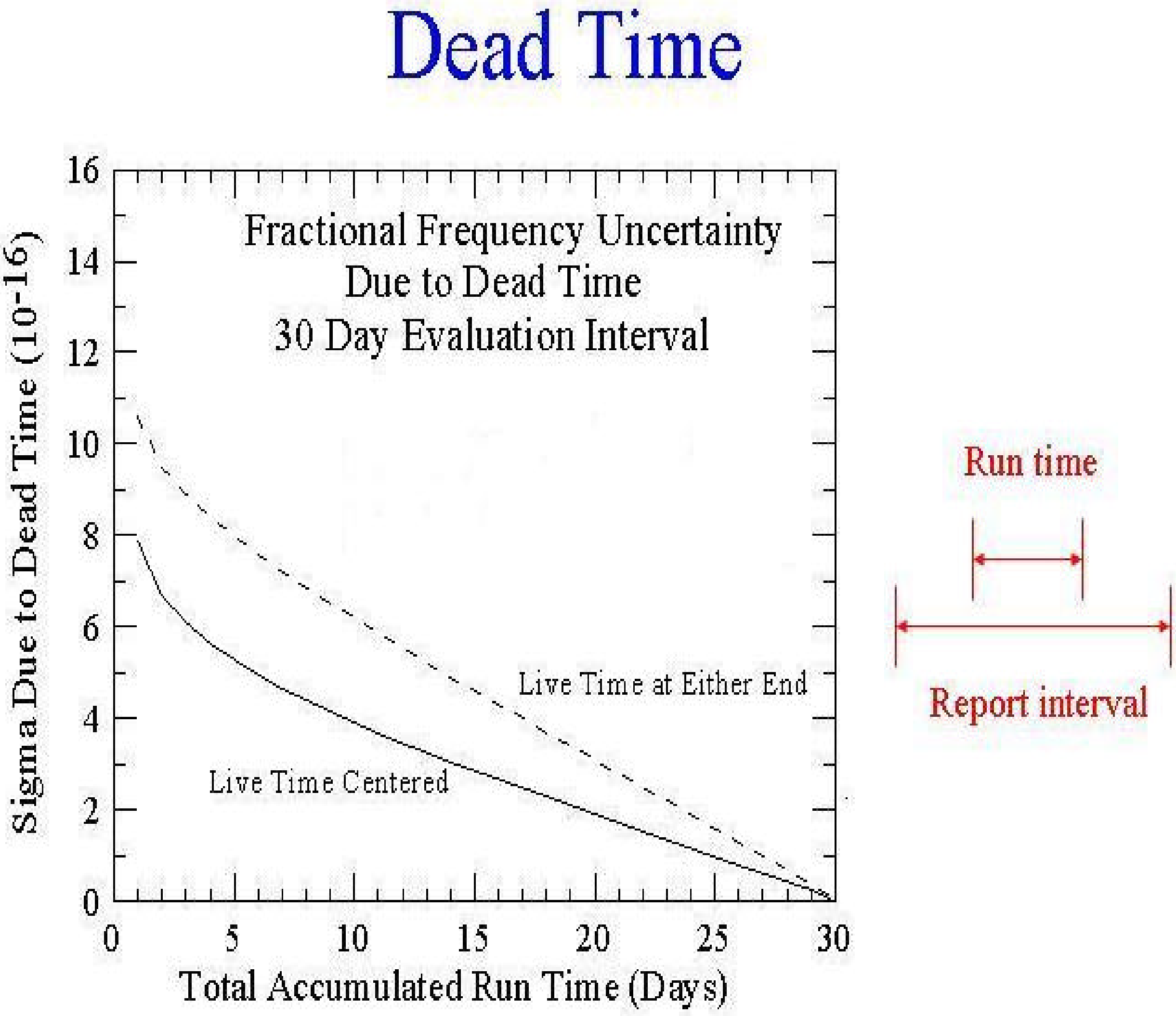
The new PPS has been approved and will be implemented with the 2005 rating period.

4.7 Estimating Clock Error Uncertainty

Sarah Streett
Statistical Engineering Division, ITL

Karen Kafadar
Guest researcher, Statistical Engineering Division, ITL

Tom Parker
Time and Frequency Division, PL



Clock error uncertainty as a function of the length of the reporting or calibration interval corresponding to a single continuous run time of the fountain. The dashed (solid) line represents a dead time (run time) period in the middle of the report interval.

The highly accurate cesium fountain, NIST-F1, is the current frequency standard at NIST. However, this primary standard can only be run intermittently, necessitating the use of a very stable, though inaccurate, secondary standard. In order to achieve an uninterrupted time measurement, the output from the secondary standard must be calibrated to the primary, after which model-based predictions are used for the frequency during the "dead time" intervals. The complicated noise models used for these time standards create challenges in estimating the resulting clock error uncertainty.

NIST-F1 provides the time standard through its measurements of the resonance frequency of cesium atoms. The resulting measurements have an uncertainty of less than 1 part in 10^{-15}. Due to the fountain's restrictive operating standards, it may be scheduled to operate for a period of only 20-30 days. It is during this time period that a secondary standard is calibrated to the fountain. The secondary standard actually consists of a group of 5 hydrogen masers and 4 commercial cesium oscillators. The hydrogen masers are very stable yet highly inaccurate and subject to frequency drifts. The commercial cesium standards are more stable in the long term but extremely noisy. Both are inferior to the fountain, yet can be operated on an almost continuous basis. To calibrate the secondary standards, the output of one of the hydrogen masers is compared with the fountain. The remaining secondary standards are then calibrated to the hydrogen maser using paired phase calculations. The final measurements from the secondary standards are based on an ensemble average of all 9 oscillators.

Although the fountain has scheduled operating periods, various factors can result in either erroneous measurements or periods of in-operation. These time intervals, together with the fountain's scheduled downtime, result in what are typically referred to as dead time intervals. As a result of dead time, it is necessary to have a model that provides accurate measurements of the phase deviations based on the output from the secondary standards. Because these standards run almost continuously, a great deal is known about the nature of the noise that affects their associated phase measurements. The typical noise model employed for these standards is a power-law model. A power-law model has a spectral density function given by $\sum h_a f^a$, where h_α is the intensity coefficient for the noise category. In the models used for the phase deviations, α ranges over the integers from -2 to 2. The complexity of the model for the phase deviations increases the difficulty of obtaining accurate measures of clock error uncertainty.

Our initial goal for this project is to provide a means for calculating the uncertainty in the phase deviations over a given time interval with periods of dead time interspersed. Currently methods exist for only the most simplistic cases. The second goal is to derive an improved statistical model for the phase deviations of the secondary standards. The resulting model must incorporate the different types of observed noise as well as accurately reflecting the underlying physical mechanisms of the clocks.

4.8 Statistics for Computational Measurements

Nien Fan Zhang, Charles Hagwood, Hung-kung Liu, Blaza Toman, James Yen
Statistical Engineering Division, ITL

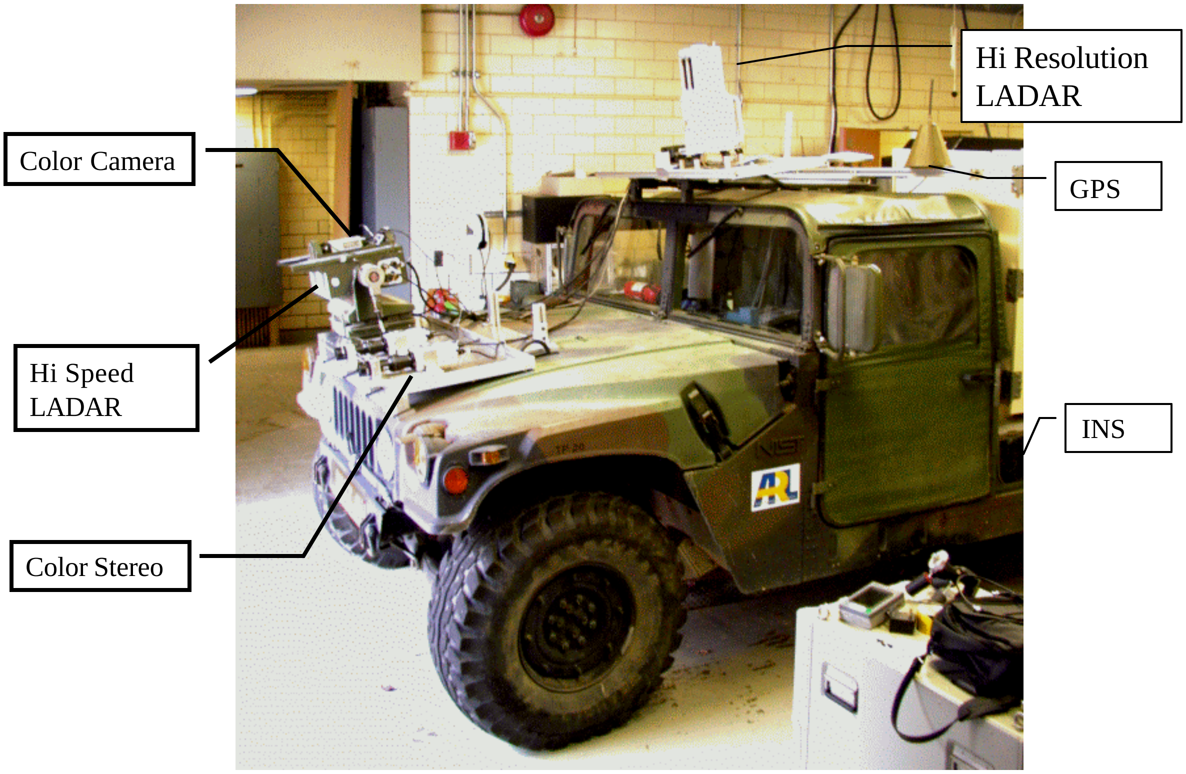


Figure: The NIST Unmanned Ground Vehicle with sensor-centric navigation

Because of advances in computing power, measurements at NIST that were once too expensive or physically impractical are now being generated on the computer. These computer-generated measurements are being used instead of experimental laboratory measurements in nanotechnology, computational chemistry, intelligent control of mobile systems, and research of building structural failures. Computer-generated measurements are the outputs of well-defined mathematical models based on theoretical principles and simulation algorithms. We call these measurements computational or virtual measurements.

A computational measurement is only as good as its accuracy. Therefor e, statistical oversight of virtual technology is essential for advancing this technology. Since most of these measurements are outside the realm of data-driven statistical methods, they present new challenges to statisticians. Statistical modeling and simulation are critical for obtaining accurate computational measurements. Advanced statistical methodologies are also needed to validate the corresponding computational measurements. More importantly, when computational measurements are used, the assessment of uncertainty for these computational measurements is vital to the measurement process. In 2004, a team of SED staff won a multi-year ITL competence award to address these issues with major emphasis on uncertainty for computational measurements. The Computation Measurements team this year has developed collaborations with individuals in the laboratories, CSTL, MEL, BFRL, and EEEL.

In CSTL, computational measurements are being used in quantum and kinetic chemistry. An algorithm, which uses quantum mechanics to compute Schrodinger's equation to determine the ground state, provides a computational measurement because no experimental data are used in the calculations. This generates substantial cost savings because experimentally measuring the heat required for the formation of certain molecules costs tens of thousands of dollars, whereas a computational measurement could be run on one's laptop computer for a few hundred dollars.

NIST's Manufacturing Engineering Laboratory, in collaboration with the U.S. Army, is developing software systems that control autonomous vehicles, i.e., vehicles that function without being manned or are remotely controlled. Such vehicles would make their decisions by incorporating information from an array of sensors. Testing and training unmanned vehicles physically requires a testing course and is very expensive. MEL engineers have created a "virtual world" that recreates a piece of the world as inputs and outputs for the vehicle sensors. This enables virtual testing of the vehicle's control algorithm, which does not know or care that its input sensory information is simulated rather than real.

In BFRL, physical tests of concrete formation are replaced by computer models in which the ingredients that make up concrete, such as cement paste, rock and sand, are modeled by spheres, ellipsoids, voids, and other shapes contained in a virtual matrix. Incorporating the pertinent principles of chemistry and physics into the model yields virtual measurements about concrete properties (mechanical, elastic, rheological, porosity).

The proposed work has the real potential to impact e-Metrology. At NIST, computational or virtual measurements have been advocated and used in many areas and in many laboratories. However, there has been no systematic statistical research on how to validate computational measurements and assess the corresponding uncertainties.

4.9 Failure of Complex Systems and Verification/Validation of Computational Models

James J. Filliben
Statistical Engineering Division, ITL

Jeffrey Fong
Mathematical and Computational Sciences Division, ITL

Emil Simiu
Materials and Construction Research Division, BFRL

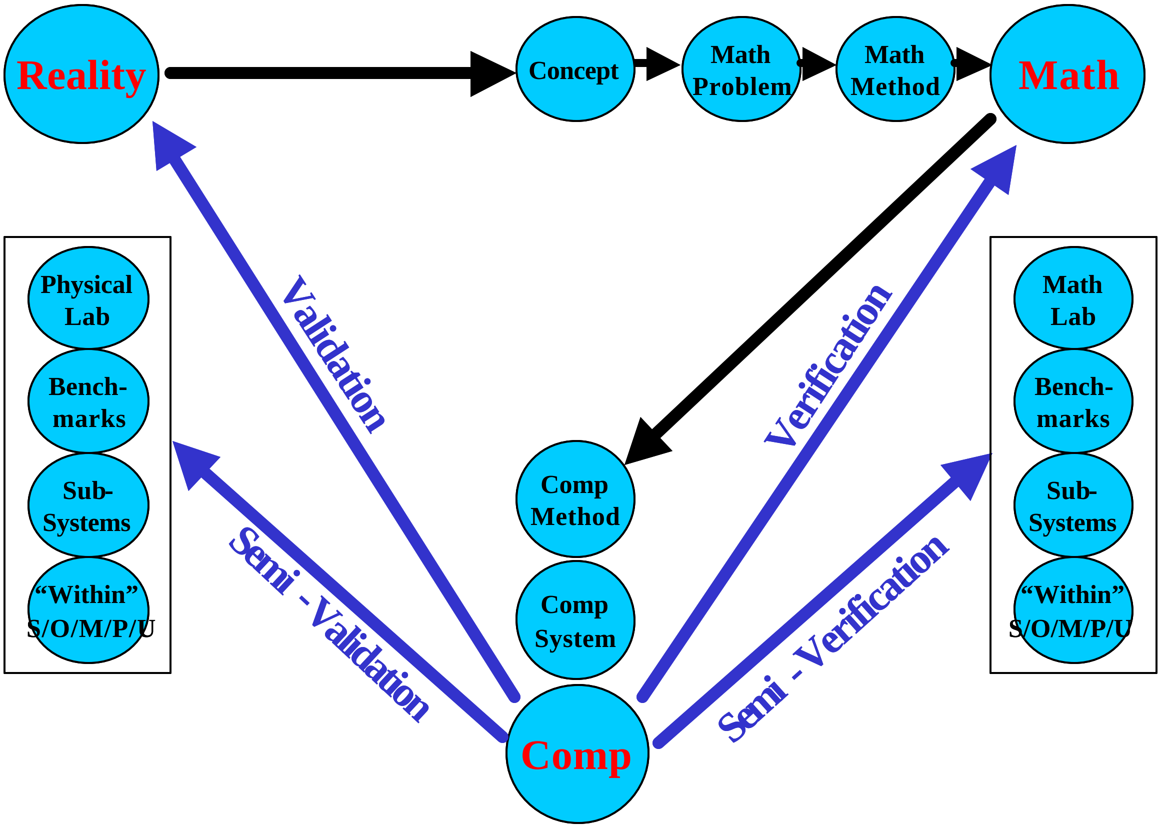


Figure 1. Verification and Validation of Virtual/Computational Systems

NIST's Building and Fire Research Laboratory (BFRL) provides state-of-the-art expertise in the characterization and modeling of structures. To that end, over the past decade there has been an increasing BFRL commitment toward the development and construction of computational models to serve as cyberspace surrogates for a variety of building component response characteristics. Long-standing classic BFRL examples of such computational models include FDS (the Fire Dynamics Simulator), CONTAM (the multi-zone indoor air quality and ventilation analysis simulator), and VCCTL (the Virtual Cement and Concrete Testing Laboratory). A more recent BFRL activity in this regard is the multi-stage finite element analysis (FEA) coding that was done in connection with the World Trade Center (WTC) collapse: 1) plane impact damage & fuel spillage; 2) fire spread; 3) thermal propagation through insulation from gaseous to steel and concrete; and 4) structural deformation/collapse.

Some of the above problems involve failure/collapse of systems. These types of problems fall under the general venue of an ongoing joint BFRL/ITL 5-year Competence Project: Failure of Complex Systems. All of the above problems involve the construction of computational models to either emulate a physical reality that we can observe, or to predict a reality that we cannot observe (due to economic or physical constraints). The quality assessment of these computational models falls under the general purview of V&V (verification and validation); that is, does the computational "black box" match the math (verification), and does the computational "black box" match reality (validation)?

These two areas: failure of complex systems and V&V are not identical, but are similar enough in goals, issues, and methodology that we report on them herein as a single article.

The Complex Systems Failure Analysis 5-Year Competence Project is a BFRL/ITL effort involving staff members from the BFRL/Fire Research Division (Howard Baum and Kuldeep Prasad), BFRL/Materials and Construction Research Division (Emil Simiu, Theresa McAllister, Dat Dutthinh, and Long Phan), ITL/Mathematical and Computational Sciences Division (Jeffrey Fong and Geoffrey McFadden), and ITL/Statistical Engineering Division (James Filliben).

This project is in its second year. Accomplishments in its first year focus primarily on one of the most visible "failures of a complex system" in modern times: The World Trade Center collapse. Although complex systems of all scientific types (e.g., physical, chemical, electronic, networks, biological, etc.) can fail, this project focuses primarily upon structural failures (where failure modes usually stem from material inhomogeneity, plastification, instabilities, fractures, fatigue, thermal weakening, excess loading, etc.) of building elements. In general, such structural failures can be of the following four types (see Figure 2): wind-structural engineering failures, fire-structural engineering failures, collision-structural engineering failures, and multi-mode structural engineering failures

1. Wind-structural engineering failures include, for example, the manifestation of energy transfer by deflection, pressure, and stress build-up in buildings induced by wind loads (e.g., hurricanes). Such wind-induced responses commonly result in non-linear failure mechanisms.

2. Fire-structural engineering failures include the heat transfer and subsequent failure of steel and concrete building components due to such components being in a temperature domain beyond their design specs. Several challenges present themselves in this fire-structural engineering area: for a given 1) building geometry, 2) fuel load, 3) office load, and 4) environmental condition, how does the fire spread and what is the thermo-spatial response at various points in time? How do the gaseous thermal conditions propagate through fireproofing to affect substrate (steel, concrete) temperatures? How do such substrate materials respond (deflection, deformation, collapse) in the presence of elevated temperatures? Fire-structural engineering played a major role in the analysis of the World Trade Center collapse. In each of the above three WTC cases (fire spread, thermal propagation, and structural deformation), the lack of WTC on-site data resulted in computational modeling becoming the critical tool in the engineering analysis. For the fire spread, BFRL's FDS (Fire Dynamics Simulator) was used to simulate fire spread across and between the one acre of space on a single WTC floor. For the thermal propagation, FEA code was written to simulate resulting substrate temperatures under various fireproofing conditions. For thermal deformation, FEA code was used to simulate steel/concrete deflection/failure at elevated temperatures.

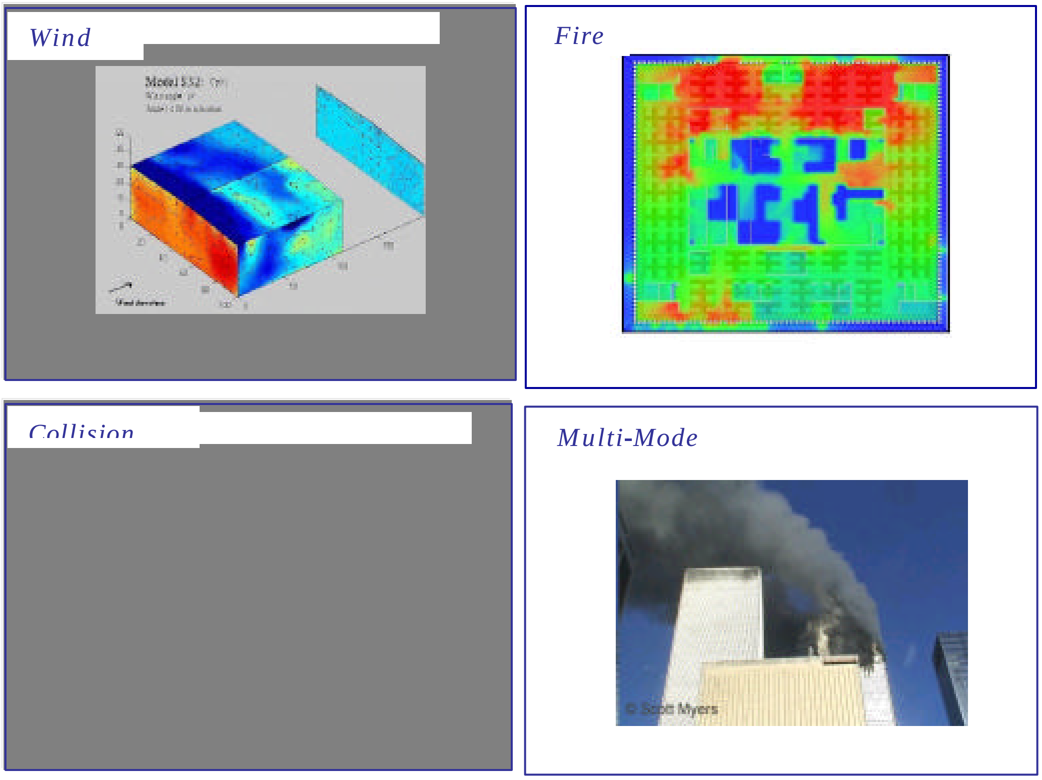


Figure 2. Engineering Failure Types: Wind/Structural (upper left), Fire/Structural (upper right), Collision/Structural (lower left), Multi-Mode Structural (lower right).

3. Collision-structural engineering failures include instantaneous energy-transfer deformation and failure of structural components due to a catastrophic impact event (external impact or internal explosion) in a structure. In this case, Homeland Security examples abound; for example, the impact of a projectile on a nuclear reactor, or the impact of the planes on the WTC buildings. Another example of collision/structural engineering would include the simulated car crashes that automotive companies worldwide carry out to assess vehicle/occupant safety. For this latter case, the existent data is a combination of physical crash tests and simulated FEA computational crash tests. For the first two examples, there is no (nuclear reactor) or little (WTC) physical data, and so heavy reliance is placed on FEA computational models. Constructing these models is difficult and time-consuming in itself; calibrating these computational models is yet another challenge; verifying and validating these computational models is an even more complicated challenge (from which we need to draw on principles and techniques from V&V as a discipline to provide structure and guidance).

4. Multi-mode structural engineering failures include all cases in which the final failure is contingent on two or more of the three above-described engineering failure types. The most striking example of this is the WTC collapse in which the wind-structural engineering failure was negligible, but the fire-structural and collision-structural engineering failure modes were significant, and very complicated. Being able to do reliable characterization and prediction is the ultimate goal of the Failure of Complex Systems Competence Project. A necessary prerequisite for this ultimate goal is to first achieve a high level of characterization and prediction expertise in each of the three components (wind, fire, collision) which are the fundamental building blocks of multi-mode structural engineering failure.

Note that of the above four general structural failure types, the Failure of Complex Systems Competence Project will formally focus in its remaining 3.5 years on structural failures of type 1 (wind-induced), type 2 (fire-induced), and type 4 (complex combinations of wind and fire). After immediate completion of the WTC study, the type 3 failures (collision-induced structural failures) will receive relatively less attention in the Competence Project.

Approach: It is clear that progress in improved characterization/ prediction of structural failures is intrinsically tied in with the quality of the underlying computational models. In an arena where "reality failure data" is limited (e.g., WTC) or non-existent (e.g., nuclear reactor projectile strike), the NIST scientist/engineer is increasingly becoming more dependent on conclusions drawn from our cyberspace representation of reality. This places enormous demands on the construction of such computational models, their calibration, their verification, and their validation.

All four of the above must be rigorously and successfully executed before the BFRL/NIST scientist/engineer can confidently declare fidelity to observed reality, and can confidently predict beyond observed reality. How is this done?

To answer this question, we refer to Figure 1, which is a representation of the 3-step process commonly carried out in constructing computational models:

1. **Reality**: this is observed (e.g., cantilever beam deflection & failure), partially observed (e.g., WTC) , or not observed at all (e.g., nuclear reactor projectile);
2. **Mathematics**: upon observing/imagining reality, the scientist/engineer/expert forms a concept of the failure process, and after considerable work forms a mathematical description and solution based on that concept;
3. **Computational**: if a mathematical solution is reached, the mathematics is converted into a computational solution--most commonly with the help of large-scale software platforms such as FEA (finite element analysis) or FDS (fire dynamics simulator).

The above (Figure 1) trichotomy serves as the generic framework for how computational modeling is done. This framework has had the benefit of 10+ years of extensive non-NIST interdisciplinary collaboration (AIAA and ASME), spearheaded primarily by Bill Oberkampf and colleagues at Sandia. In addition to the reality/mathematics/computational triangle, such historic collaboration over the last decade has resulted in a standardization of terms, the two most important of which are:

1. **Verification**: the computational solution matches the mathematics.
2. **Validation**: the computational solution matches reality.

Such terms are now universally agreed on across both structural mechanics and fluid mechanics arenas, as well as in DOD/military agencies. The question that remains is the general and the detailed roadmap as to how to carry out a rigorous V&V for a specific project.

NIST V&V Competence-Building: The catalyst for NIST attention to formal V&V was the WTC collapse with its extensive dependence on computational models (FDS and FEA). The question arose as to how one goes about rigorously (that is, acceptable from a scientific, engineering, and statistical point of view) ascertaining the validity of such computational models. To rapidly bring NIST/ITL up to speed in regard to existing V&V methods, Jeffrey Fong and Jim Filliben have done an intensive V&V immersion over the last 6 months.

In addition to attending to the extensive existent V&V literature, Fong and Filliben attended a DMSO (Defense Modeling and Simulation Office) conference headed by Simone Youngblood (who is the driving force behind DOD's V&V efforts). Fong & Filliben also attended and participated in a DOD-sponsored workshop: Foundations '04: A Workshop for VV&A (Verification, Validation, & Accreditation) in the 21st Century (Tempe, AZ, October 13-15). Immediately prior to that conference, they visited Southwest Research Institute (San Antonio) to learn of the V&V work of Ben Thacker (author of NESSUS: a probabilistic analysis tool for improving safety and reliability of complex systems) and Chris Freitas (whose fellow staff member showed them the SWRI on-site test results for the Columbia foam impact disaster). Immediately after the conference, Fong and Filliben visited Lawrence Livermore to give a talk and learn of Livermore's V&V work. Finally, on November 8-9, 2004, Jeffrey Fong's herculean efforts resulted in a NIST-DOD Workshop here in Gaithersburg: Verification and Validation of Computer Models for Design and Performance Evaluation of High-Consequence Engineering Systems--this workshop had an excellent representative attendance across government, DOD, and industry.

Role of NIST in V&V: The NIST workshop served as a catalyst to force the issue and focus on what NIST/SED/MACD should be doing in regard to the national V&V effort. There is a near-consensus opinion from the main players of the outside V&V community that they are looking to NIST to draw on its historic legacy, expertise, and excellence in standards to provide relevant V&V benchmarks to serve as references in the assessment of computational models. Ongoing SED/MACD efforts are currently being carried within NIST to determine the optimal nature, extent, (and funding) of such V&V benchmarking.

V&V Benchmarks: The first step in such NIST work will probably be with verification (as opposed to validation). The verification problem (which answers the question: Does the computational solution match the mathematics solution?) is inherently the easier of the two problems, and in itself does offer excellent insight as to the quality of the computational model. If NIST embraces this benchmark role, what benchmarks should NIST construct that would be best for the at-large V&V community? That is a question that yields different answers from different experts. At this time, Jeffrey Fong is leaning heavily toward the exploration of the "simple" cantilever--its deflection, its oscillation, its strain, its failure, etc.) The cantilever is an elemental structural engineering component. There is much to suggest that it would be a natural starting place for a suite of foundational benchmarks for verification purposes--and then for validation purposes.

The cantilever is an elemental component in the physical world at many levels:

1. **Macro**: At the macro (1 meter) level, how good is the FEA code in predicting deflections, deformations, and failure for WTC structural components? A necessary (but not

sufficient) step in the verification of such FEA code is the assessment of accuracy as compared to known mathematical solutions for a meter-sized steel cantilever.

2. **Micro**: At the micro (10^{-6} meter) level, Dario et al. point out the feasibility of MEMS (Micro Electro Mechanical Systems) and bioMEMS-based miniaturized sensors (including cantilevers) fabricated out of inorganic or organic materials to measure force and position for prostheses optimization. Other authors (e.g., Scherer) point out how the cilia of the human ear are in essence micro-cantilevers. Can computational models be constructed to serve as surrogates for predicting the frequency response properties of the cilia? How good are these computational models relative to the mathematical solutions?
3. **Nano**: At the nano (10^{-9} meter) level, a recent Science magazine article points out that the speed of the current Pentium 4 and Pentium 5 chips, and the future Pentium 6 chips depends directly on gate sizes which in turn are directly related to nano-cantilevers of the 50 nanometer to 10 nanometer range. Do the computational model predictions for these nano-cantilevers agree with the mathematics solutions?
4. **Ato**: At the ato (10^{-18} meter) level, Craighead at Cornell points out that tiny cantilevers can be used as weighing devices to measure viruses and other biological structures. Can these ato-cantilevers be computationally modeled? Do they agree with the mathematical models?

Although good arguments have been presented from the outside that NIST 's benchmark efforts should be directed to artifacts other than the cantilever, there are opposing arguments (and the above scopes of application) that suggest that the cantilever may in fact be a good place for NIST to start. This is a matter of ongoing in-house discussion.

Benchmarks vs. Methodology: Other possible NIST contributions would include metrology-based statistical methodology for carrying out V&V; this is of course of considerable internal importance to NIST as reflected in Fong, Filliben, Sedransk, and Guthrie talks at the Gaithersburg V&V conference, but (as discussed earlier) the non-NIST V&V community consensus was that such methodology is of secondary importance to them at the moment, with the overriding priority being the actual selection and construction of relevant V&V benchmarks that the larger V&V community would use immediately for code assessment and comparison. Such benchmarks would of course be a critical component in a larger NIST/ITL computational modeling testbed, which arguably could be a high-priority future ITL capability/competence. In this context, appropriate statistical methodology will in fact be demonstrated by example in the construction, processing, and analysis of the resulting benchmarks.

Statistical Methodology: Although not an immediate deliverable, the issue of statistical methodology must nevertheless be in place for the successful construction and verification of benchmarks. It is our view that statistical contributions for V&V fall into the following three general areas:

1. **Unifying Approach**: This posits that for all V&V problems-- regardless of how different--a simple 5-step problem/ design/data/analysis/conclusion approach (as shown in Figure 3) serves well as a general, unifying framework for systematically addressing such problems.
2. **Problem Classification**: Just because an engineer is dealing with computational models does not necessarily mean they are doing only V&V. V&V is essentially a pairwise operation in which computational models are compared to mathematical models and to reality. Other operations (e.g., sensitivity analysis) are non-pairwise but are also of value in examining, characterizing, and gaining insight into the workings (and dominant factors) of a computational model. As it turns out, regardless of the engineering problem, we identify nine generic problems/activities (sensitivity analysis,

optimization, modeling, comparing, predicting, error analysis, calibration, verification, and validation) and encourage engineering specificity to choose/prioritize among them because different activities dictate different experiment designs, different statistical analyses, and different conclusion deliverables.

3. **Experiment Design Importance:** Although the above 5-step general approach involves both experiment design and statistical analysis, the former is very commonly under-appreciated and misunderstood by scientists/engineers when dealing with computational models.. The execution of high-quality computational models for realistic engineering systems is frequently time-consuming (days, even weeks) and complicated (involving many factors); hence this points out--just like with physical experiments--that computational experiments also have a need for statistically optimal designs: designs which are both efficient (in cost and time) and yet potent (in being able to generate valid scientific/engineering conclusions).

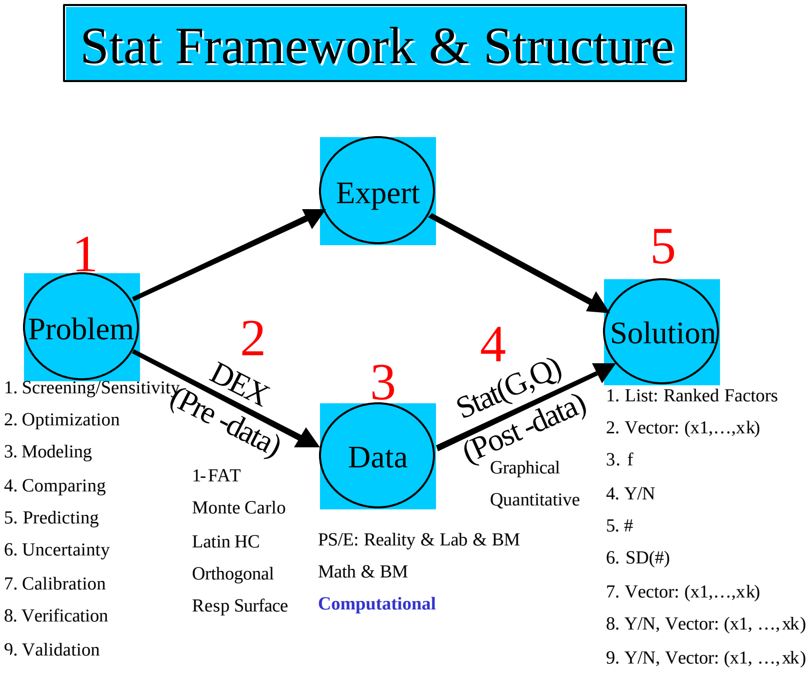


Figure 3. 5-Step General Problem Solving Approach

Many methodological details (both design and analysis) remain to be worked out. Again, statistical methodology is not seen by the larger V&V community as an end in itself. This implies that the exposition of such methodology will not be an overt end in itself, but rather will flow naturally from its application to the quality benchmarks.

e-Guide to the Design of Computational Experiments: On the other hand, a valuable overt forum for the demonstration and application of statistical methodology-- especially design methodology--to V&V and System Failure benchmarks is the planned e-Guide to the Design of Computational Experiments. This e-Guide will draw on metrology/methodology expertise from past SED/ITL WEB-projects (e.g., the NIST /Sematech e-Handbook of Stat Methods at http://www.itl.nist.gov/div898/handbook/; and the joint BFRL/FHWA/ITL discipline-specific project COST (Concrete Optimization Software Tool)--a concrete experiment design/analysis system at http://ciks.cbt.nist.gov/cost/). The e-Guide will go beyond both of the above products by 1) implementing significant enhancements 2) with greater interactive features and 3) with improved capabilities resulting from the recent improved understanding of the characteristics of computational experimentation.

Although analytic capabilities will of course be an important part of the e-Guide, the primary novel aspect of the proposed e-Guide is its focus on the much-neglected and little-appreciated design of computational experiments. Even more than the subsequent data analysis, the validity (and the domain of validity) of output from computational models depends critically on the executed experiment design. The e-Guide will provide both the framework and the specifics for optimal-information-content experiment designs to apply to each specific computational model. To do this, the e-Guide will first provide the unifying and simplifying high-level structure alluded to earlier--from problem-formulation, through design, through data collection, through analysis, through conclusion-formulation--that will be appropriate for all computational experiments. A second novel feature is the expert elicitation methodology whereby the e-Guide will—derived from experiment design principles and techniques--intelligently elicit from the scientist/engineer the specifics of the computational system at hand: system output(s), factors under study, admissible domain regions, scientific/engineering experiment objectives, etc.

Thirdly, within the limits of these computational specifics and time/cost constraints--the optimal experiment design (Orthogonal, Taguchi, Latin Hypercube, or other) will then be automatically constructed. Finally and most importantly, the e-Guide will be able to be used for computational systems at the component level up through the global level, and for computational experimentation problems of varied sizes and types, e.g., validation, sensitivity analysis,, optimization, uncertainty analysis, etc.

The e-Guide will benefit computational modelers at all levels: developer, user, and decision-maker.

1. In the near-term, the e-Guide's structure and optimal designs will serve a critical role in NIST V&V benchmark and Failure characterization;
2. in the intermediate term, the e-Guide will be a necessary component in a NIST configurable testbed for computational experimentation;
3. in the long-term, the e-Guide will lay a sound foundation for accurate (and efficient) quantification of high-consequence complex system validation, verification, and reliability.

Regarding funding for the e-Guide, a request has been submitted as an ATP/ITL Proposal, with operational completion in 2.5 years. The outcome is pending.

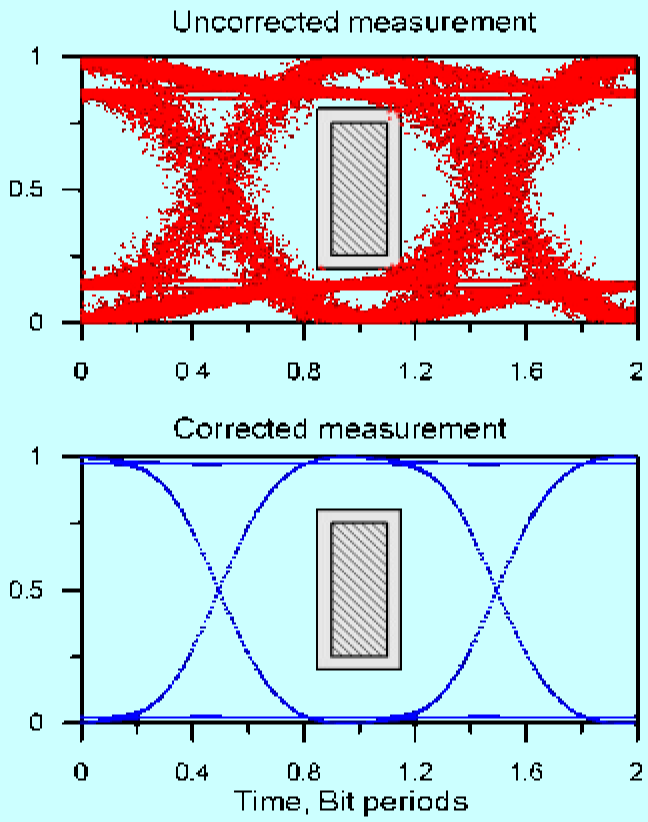
Both the formal Failure of Complex Systems Competence Project and the yet-to-be-formalized Verification/Validation of Computational Models Project reflect directly on the significant growth within NIST of computational experimentation. Progress in the production of V&V/Failure benchmarks and the development and application of associated methodology will yield immediate benefits (rigor, savings in cost/time, model-adequacy assessment, and accurate engineering predictions). These benefits will accrue not only for the leading-edge BFRL wind, fire and structural engineering problems addressed in the Failure of Complex Systems Competence Project, but also for the multitude of computational modeling projects throughout NIST (involving standards, measurement services, event modeling, and other leading-edge scientific/engineering modeling--especially in biotechnology, nanotechnology, and defense/safety arenas where reality is difficult--if not impossible--to observe). Dissemination of concomitant methodology via the e-Guide will share these methodologies (and the benchmarks behind the methodology) to the computational modeling community beyond NIST.

4.10 Time-Base Corrections in Waveform Calibrations

C. M. Wang
Statistical Engineering Division, ITL

P. D. Hale
Optoelectronics Division, EEEL

D. F. Williams
Electromagnetic Technology Division, EEEL



Simulated eye mask test of a long random bit sequence transmitted by a fiber optic transceiver. The forbidden region in the pass/fail test is shown in the hashed region surrounded by a 10 % guard band. Although the simulated transceiver meets specifications, the time-base errors in the oscilloscope measurement (top) cause the transceiver to fail the test as some samples fall in the guard band. The bottom graph shows the effect of correcting the oscilloscope errors, clearly passing the transceiver that would otherwise have been rejected.

We develop a method of correcting both random and systematic time-base errors using measurements of two quadrature sinusoids made simultaneously with a waveform of interest. The new time base is estimated from the sinusoids using a weighted error-in-variables regression approach that accounts for relative contributions of additive noise and timing error.

A waveform is a representation of a signal that varies with time. The most familiar waveform is the sine wave. Waveform measurements are required throughout the optical communications, computer, wireless communications, radar, and remote sensing industries. Waveform measurements are used to verify signal fidelity and standard compliance for the design and qualification of systems and components. High-speed sampling oscilloscopes are often used for displaying waveforms. Sampling oscilloscopes, however, suffer from several nonideal properties that must be characterized and compensated for. One of these effects is the timing error. At the *i*th sample, the timing error is the sum of a deterministic time-base distortion (TBD) h_i and a random jitter t_i. Thus the *i*th sample of the waveform of interest *g,* as a function of time, is given by

$$y_i = g(T_i + h_i + t_i) + e_i$$

where $T_i = (i-1)T_s$ is the target time of each sample, T_s is the target tim e interval between samples, ande_i is random additive noise. We assume the jitter and additive noise are independent zero-mean random variables with standard deviations s_t and s_e.

The TBD and jitter cause errors in the time in a waveform at which samples are acquired. These imperfections have tangible manufacturing costs. For example, the oscilloscope is often used for eye mask tests. Eye mask testing, shown in the accompanying figure, is a common pass/fail test for communications components. The eye diagram is obtained by plotting a long random bit sequence transmitted by a fiber optic transceiver as a series of short repetitive waveforms. The hashed region in the figure is the forbidden region for the pass/fail test. This region is surrounded by a 10 % guard band. Although the simulated transceiver meets specifications, the frequency response and timing errors in the oscilloscope measurement cause the transceiver to fail the test as some samples fall in the guard band. In a recent survey, a leading manufacturer of 10 Gbits/s Ethernet transceivers claims that it rejects 10 % of these components due to oscilloscope measurement inaccuracies.

The problem of estimating the TBD has been studied by many authors. Recent work has used a nonlinear least-squares approach that fits multiple measured sinusoids with multiple phases and frequencies to a distorted sinusoid model. This approach performs well at discontinuities in the TBD and allows simultaneous estimation of the harmonic distortion in the measured sinusoids. The distorted sinusoidal model, with harmonics number n_h, is given by

$$y_{ij} = a_j + \sum_{k=1}^{n_h}\left[b_{jk}\cos(2pkf_j t_{ij}) + g_{jk}\sin(2pkf_j t_{ij})\right] + e_{ij},$$

where f_j is the fundamental frequency of the *j*th measured waveform y_{ij} at the *i*th nominal time, $t_{ij} = T_i + h_i + t_{ij}$. The values of a_j, b_{jk}, g_{jk}, and h_i can be estimated by using a weighted least-squares approach. To obtain a solution using this approach, we typically measure a set of sinusoidal waveforms at two or three different frequencies. Each set includes two sinusoids that are approximately in quadrature at each frequency. Hence each set can have up to four or six waveforms.

The problem of estimating jitter and correcting for its effects has also been addressed by many authors. The typical approach is to obtain the signal variance of independent, repeated measurements and use the approximate model

$$\mathrm{var}(y_i) \approx s_e^2 + (g'(t_i))^2 s_t^2$$

to solve for s_t. Here $g'(t_i)$ is the derivative of the ideal signal evaluated at $t_i = T_i + h_i$. It is usually assumed that, on average, the jitter acts as a low-pass filter so that the average signal is the convolution of the ideal signal $g(t_i)$ and the probability density function $p(\cdot)$ of the jitter:

$$\mathrm{E}(g(t_i)) = \int g(t_i - t)p(t)dt \ .$$

The effects of the jitter can then be removed by deconvolution. This approach, however, has the following problems:

1. measurements must be repeated to find the mean and variance,
2. estimates of the jitter variance from the approximate model are generally biased,
3. $p(\cdot)$ must be known and be the same over the entire measured waveform,
4. the averaging process removes some of the inherent bandwidth of the measured signal, making the deconvolution subjective,
5. deconvolution is an “ill-posed” problem, so that in the presence of noise there is no unique solution.

Generally, it is desirable to avoid deconvolution, particularly in cases where the jitter is large, varies over the measurement time window, or has a non-Gaussian probability density.

In the present work, we are interested in the total time-base error, i.e., the sum of the TBD and jitter. We use data at the different frequencies to estimate the parameters of the distorted reference sinusoids and the new time base (h_i and t_{ij}) simultaneously. Our approach is to apply the so-called errors-in-variables or orthogonal distance regression to the distorted sinusoidal model. We take advantage of the parallel design of many equivalent-time sampling oscilloscopes. A result of the parallel architecture is that any jitter on the time-base delay generator is common to the sampling time of all the samplers in the oscilloscope mainframe. Consequently, we can generate reference sinusoids and the signal of interest simultaneously, estimate the time-base error from the sinusoidal signals, and then apply the estimate to the signal of interest and compensate for timing errors in its measurement.

We simultaneously estimate the systematic and random time-base errors of measured sinusoidal reference signals. Using the parallel sampling scheme, t allows us to use this estimate to correct the time-base errors in a simultaneously measured waveform by a factor of 10, effectively replacing the time base of the oscilloscope with a time base provided by the measured sinusoids. This allows us to correct the timing errors that might be present with long waveforms or large jitter, and lowers the noise floor significantly in most measurements

5 High-D Statistics and Informatics

The challenges in nanoscience, in bioscience, in measurement science and in information science are posed in high-dimensional terms. Data take the form of spectra and hyperspectra, of images either still or moving, and of 3–dimensional reconstructions. Obstacles to inference take many forms: exceeding computational limits, massive databases of inhomogeneous quality, multiple measurement scales (from nano to micro to meso to macro), and massive databases.

Consequently, the Statistical Engineering Division is increasingly in demand to provide sound statistical analysis for high-dimensional, multi-scale problems and statistical inferences from massive databases with quantifiable uncertainties. Along with participation in multidisciplinary research projects, the Statistical Engineering Division is also building new competence and developing new methodology for high-dimensional statistical metrology. As the problems explode in ambition and complexity, Bayesian statistics offers a natural paradigm for embedding and continually revising information.

Specific technological developments also drive the development of new statistical methods. Mass spectrometers are everywhere in the Science Laboratories at NIST and elsewhere; microsensor arrays for trace contaminant detection systems and microarrays for gene expression are almost as ubiquitous. The development of more generally applicable statistical metrology for *High-D Statistics and Informatics* will follow from the insight in defining and answering specific scientific questions.

5.1 Variation Linked to Sample Preparation in Protein Mass Spectrometry

Walter S. Liggett
Statistical Engineering Division, ITL

Peter E. Barker
Biotechnology Division, CSTL

O. John Semmes
Eastern Virginia Medical School

Lisa H. Cazares
Eastern Virginia Medical School

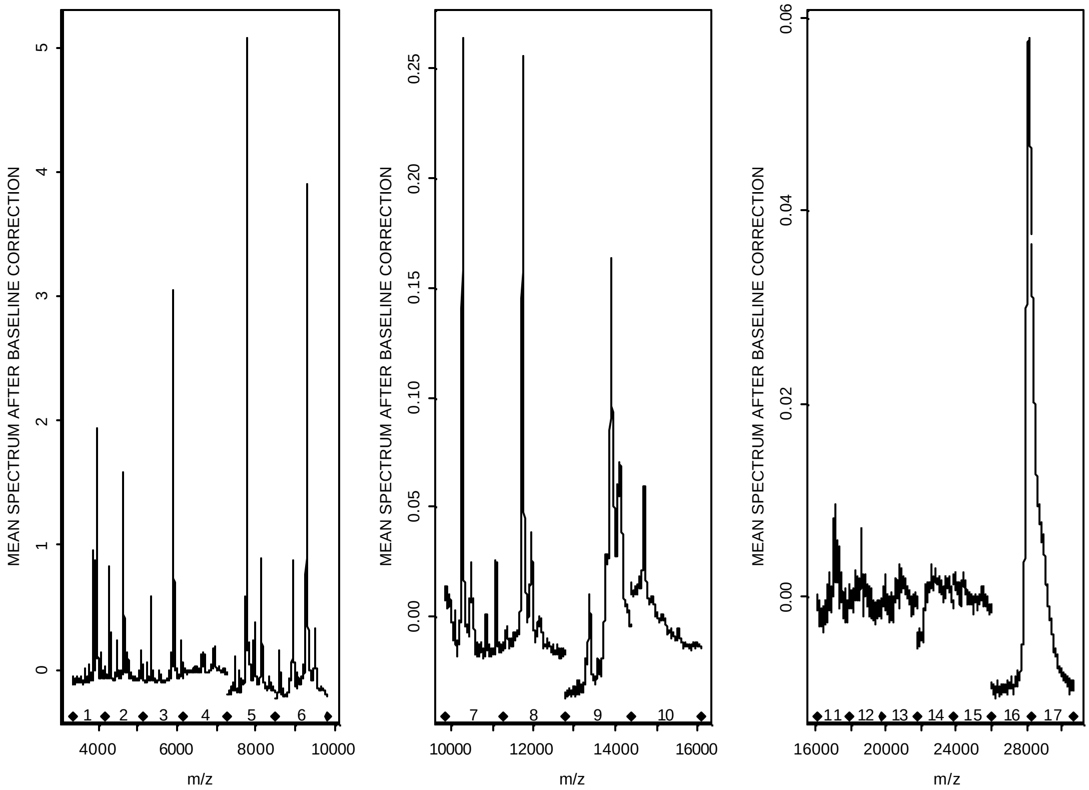


Figure 1. Mean intensity spectrum after baseline correction for each of the 17 intervals that are compared by functional CCA.

Background: Protein mass spectrometry when viewed as a measurement procedure presents performance issues that require identification of sources of measurement variation.

Methods: Sources of variation are identified through statistical analysis of repeated measurements of a human serum standard. Surface-enhanced laser desorption/ionization (SELDI) time-of-flight (TOF) mass spectrometry provided 88 spectra from 11 protein chips, each with 8 spots. The parts of these spectra in the mass-to-charge (m/z) interval (3300, 30700) are considered. The statistical approach involves functional canonical correlation analysis (CCA) applied to disjoint sections of the mass spectra for the purpose of finding long-distance correlation structure. Before this analysis, the spectra are normalized to remove spectrum-to-spectrum variation common to the entire interval. Examination of the relation between the CCA scores and the spectra at each m/z shows the spectral peaks responsible for high canonical correlation.

Results: We show that after normalization, the heights of some pairs of spectral peaks are correlated but others are not. Of the 17 spectral sections considered, we choose the seven pairs with highest canonical correlation for examination. Some pairs correspond to the singly- and doubly-charged ionization of the same protein. Other pairs in the seven may point to proteins with chemical similarities.

Conclusions: It seems likely that sources of variation in the sample preparation step are responsible for high correlations between proteins separated widely in m/z. Non-uniformity in the crystallization on the protein chip surface is a well-known source of long-distance correlation, but normalization should remove its effect. Thus, the remaining high correlations suggest other sources of variation in sample preparation.

A statistical approach to finding long-distance correlation in replicate mass spectra provides insight into sources of variation in the measurement procedure. The approach is functional canonical correlation analysis (CCA). The insight is into the sample preparation step of protein mass spectrometry.

The immediate purpose of our investigation is improvement of the measurement procedure. Insights into sources of variation should lead to proposals for reducing the effects of these sources. Our data are a batch of functions, specifically surface enhanced laser desorption/ionization time of flight (SELDI-TOF) mass spectra. The functions are all measurements of the same material, a human serum standard, not of different materials as in a case-control study. Thus, the variation in the batch of functions arises from the measurement procedure, and characterizing this variation is a step toward improving the measurement procedure.

The ultimate goal is development of biomarkers based on the measurement procedure. Improvement in the measurement procedure should help in reaching this goal. Formulation of biomarkers involves, in addition, case-control studies based on data drawn from the different classes that the biomarker is to distinguish. Statistical approaches to case-control studies differ from the approach in this paper.

Between two widely separated values of m/z, there might be strong correlation in the replicate-to-replicate spectral variation. It seems likely that such correlation is caused by the sources of variation that lie in the sample preparation step rather than in subsequent use of the mass spectrometer itself. In the case of SELDI-TOF mass spectrometry, the sample preparation step consists of applying the specimen as a coating on a protein chip with a surface that selectively binds to some proteins in the specimen but not to others. The unbound proteins are washed off, and the remaining proteins are co-crystallized with a matrix and introduced into the mass spectrometer. This sample preparation procedure contains at least one source of variation that can cause correlation between proteins with widely separated m/z values. One source is the non-uniformity of the crystallization, which has a scaling effect that is routinely eliminated from the spectra by normalization. Because of the possibility of other sources, characterization of long-distance correlation is important in assessing replicate spectral measurements.

To meet this need, we propose a method based on functional CCA applied to spectral segments from disjoint m/z intervals. CCA differs from inspection of point-by-point correlation maps in that CCA determines for each interval in a pair the combination of spectral intensities that gives the highest correlation for the pair. Data analysis for mass spectra usually begins with spectral peaks considered as the predetermined features of interest. In contrast, CCA constructs a weight function that defines a feature for each interval in the pair. These features are determined from the data rather than on the basis of predetermined peak shapes. Thus, CCA takes into account variation in peak shape due, for example, to instrument overload or common modifications of proteins. In this way, CCA exposes the full complexity of the long-distance correlation.

The spectra obtained from the Ciphergen system require further preprocessing despite the preprocessing already applied. In ideal terms, one might imagine an observed spectrum to be a superposition of peaks of various sizes. Different peaks would correspond to different proteins or a protein with different charges. Each peak would be centered at the m/z for the protein and charge, and the area under the peak would be proportional to the concentration of the protein in the spot from which the proteins were desorbed. That a spectrum obtained from the Ciphergen system does not conform to this model exactly can be remedied in part with further preprocessing. Because the observed spectra are not properly aligned with each other, we register the spectra. Because the spectra have an additive baseline, we make an adjustment to remove it. Because the constant of proportionality that relates the areas under the peaks to the concentrations at the spot on the protein chip varies from spectrum to spectrum, we normalize the spectra. These preprocessing steps have typically been applied to the analysis of SELDI-TOF mass spectra. A preprocessing step that we not apply is a square root or cube root transformation of the spectral intensities.

Lack of horizontal alignment among spectra is an issue because functional CCA is based on scores each defined by the integral of a weight function times the spectrum. The weight function is intended to emphasize or de-emphasize certain portions of each spectrum; for instance, the portion containing a spectral peak. Emphasis will not be applied to proper portions of a spectrum if the spectrum is not properly aligned. Spectral registration in this work is the same as the registration discussed in a previous paper. We represent the spectra after registration by a spline composed of fifth-order polynomials between the knots and having a continuous fourth derivative at every point. We denote registered spectra by $y_i(u), i = 1, \ldots, N$, where the independent variable u is the mass-to-charge ratio. To obtain $y_i(u)$, we interpolate spectrum i as originally observed with a cubic spline and evaluate this spline at $d_i + g_i u$. A previous paper discusses estimation of $d_i, g_i, i = 1, \ldots, N$. We have checked to see if this shift and scale form for the registration is valid beyond the interval previously considered and found no contradictory evidence.

We would like to apply functional CCA to functions consisting of the superposition of contributions from individual proteins as discussed above. Let these functions be denoted by $j_i(u)$. We assume that we in fact observe

$$y_i(u) = b_i j_i(u) + a_i(u),$$

where $a_i(u)$ is the slowly varying baseline and b_i is a scale factor that does not depend on u. Because we do not want the baseline $a_i(u)$ to influence the CCA results, we remove $a_i(u)$ from $y_i(u)$. Moreover, because spectrum-to-spectrum variation in the scale factor b_i can be expected, we want to distinguish this variation from other more interesting types of variation. For this reason, we reduce the dependence on b_i. Our approach to these preprocessing steps does not provide estimates of $j_i(u)$ but does provide modified spectra appropriate for functional CCA.

We do baseline correction separately for each of the intervals we use in our functional CCA. We denote these intervals by $\left[L_j, U_j\right]$. The baseline-corrected spectra are given by

$$y_{\mathrm{B}i}(u) = y_i(u) - \frac{1}{U_j - L_j}\int_{L_j}^{U_j} y_i(s)ds.$$

Underlying this correction is the assumption that $a_i(u)$ is essentially constant over $\left[L_j, U_j\right]$, which implies that $y_{\mathrm{B}i}(u)$ does not depend on $a_i(u)$. It is easy to show that $y_{\mathrm{B}i}(u)$ is proportional to b_i, a fact that is important in the normalization step described below.

The mean of $y_{\mathrm{B}i}(u)$ over i is shown in Figure 1. The 17 intervals $\left[L_j, U_j\right]$ are indicated. We note that the baseline-corrected spectra $y_{\mathrm{B}i}(u)$ are not always positive and are generally discontinuous from interval to interval. We account for this in our application of functional CCA. We also note that the y-axis scales for the three panels differ by a factor of 100.

Our normalization is based on $\bar{y}_{\mathrm{B}}(u)$, the mean of $y_{\mathrm{B}i}(u)$ over i. If the deviations of the individual spectra are small relative to their mean, then instead of dividing by an estimate of b_i, we can normalize by subtracting a quantity proportional to $\bar{y}_{\mathrm{B}}(u)$. We normalize by computing the spectral deviations

$$x_i(u) = y_{\mathrm{B}i}(u) - b_i\bar{y}_{\mathrm{B}}(u),$$

where the estimate of b_i is given by

$$b_i = \int y_{\mathrm{B}i}(s)\bar{y}_{\mathrm{B}}(s)ds \Big/ \int \bar{y}_{\mathrm{B}}^2(s)ds.$$

The integrals that define b_i are over the union of all the intervals $\left[L_j, U_j\right]$. Note that the spectral deviations $x_i(u)$ are not only normalized but also centered at the mean $\bar{y}_{\mathrm{B}}(u)$. Normalized spectra with the centering removed are given by $x_i(u) + \bar{y}_{\mathrm{B}}(u)$.

An elementary look at the spectra after preprocessing is given in Figure 2. This figure shows relations among three peaks, the largest peaks in intervals 7, 8, and 9. Horizontally, this figure shows spectral deviations $x_i(u)$ at the interval 8 peak, and vertically it shows the spectral deviations at the interval 7 peak and the interval 9 peak. We see that the deviations are uncorrelated for intervals 7 and 8, but correlated for intervals 8 and 9. Generalizing, we see that after normalization, some pairs of peaks are uncorrelated and others are correlated. This suggests the complexity of the correlation structure in which we are interested.

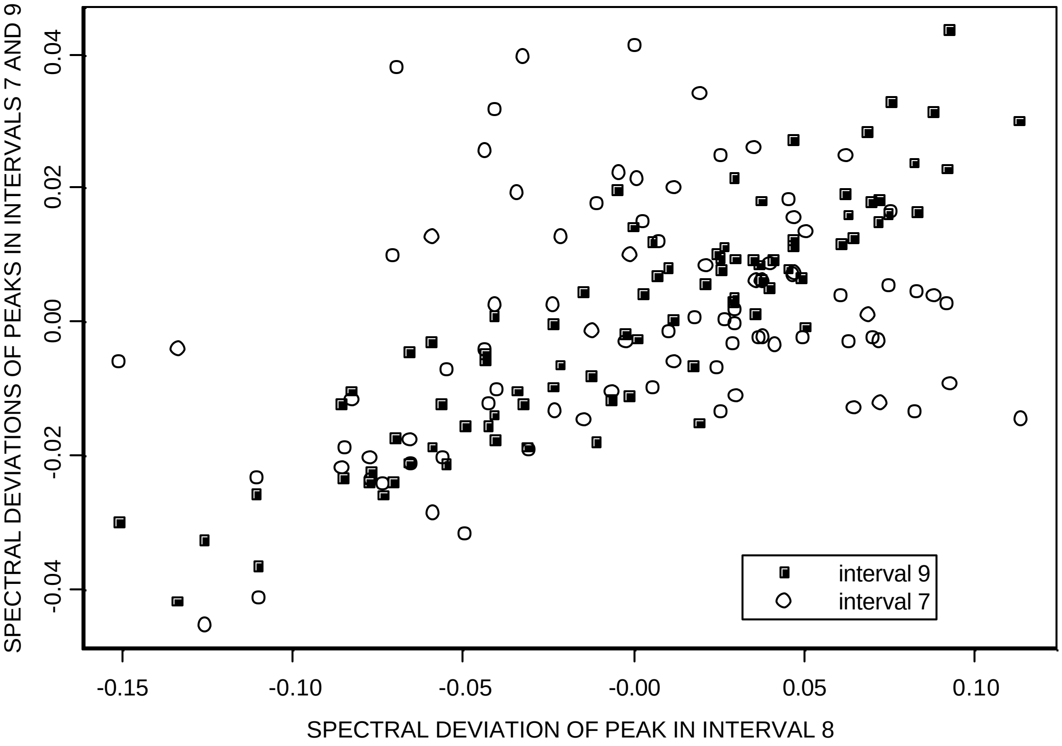


Figure 2. For the largest peaks in intervals 7 - 9, peak height scatter plots showing differing degrees of correlation.

We examine only the leading canonical correlation between disjoint intervals, although there are subsequent canonical correlations that one could consider. Central to functional CCA is the idea of a linear combination of the spectral values in an interval. Such a combination can be called a feature of the spectrum. For interval j, a weight function $x_j(s)$ specified for $L_j \le s \le U_j$ defines such a combination. Evaluation of this combination for a spectral deviation $x_i(u)$ is given by

$$\int_{L_j}^{U_j} x_j(s)x_i(s)ds\,,$$

which is the score for spectral deviation i. Corresponding to a weight function, we have a set of scores, one for each spectrum.

The leading canonical correlation in functional CCA can be specified in terms of two weight functions, $x_j(s)$ and $x_k(s)$, with scores that have maximum correlation subject to penalties on the smoothness of the weight functions. Were these weight functions given, then moments could be computed

$$s_{jk} = \frac{1}{N}\sum_i \left[\int_{L_j}^{U_j} x_j(s)x_i(s)ds\right]\left[\int_{L_k}^{U_k} x_k(s)x_i(s)ds\right].$$

The leading canonical correlation coefficient for intervals j and k is the value of s_{jk} obtained by maximizing over the weight functions subject to

$$s_{jj} + l_j \int_{L_j}^{U_j} \left\{D^2 x_j\right\}^2 ds = 1$$

and

$$s_{kk} + l_k \int_{L_k}^{U_k} \left\{D^2 x_k\right\}^2 ds = 1,$$

where the symbol D^2 denotes second derivative. The two constraints not only limit the sizes of the weight functions but also impose smoothness on the weight functions. The second terms in the two constraints limit the sizes of the second derivatives of the weight functions. Positive values for these terms, that is, for l_j and l_k, are necessary if functional CCA is to give reasonable results. Increasing the sizes of l_j and l_k increases the smoothness of the weight functions.

The baseline correction causes a problem in the use of this definition to compute the leading canonical correlation. The problem is that the constraints on the weight functions are not satisfied for the constant weight function. The reason is that the $x_i(u)$ are orthogonal to the constant weight function and consequently $s_{jj} = s_{kk} = 0$ for this weight function. We solve this problem by adding a small randomly chosen constant to each $x_i(u)$ in each interval. Thus, for the constant weight function, s_{jk} remains 0 while s_{jj} and s_{kk} do not, and the weight functions for the leading canonical correlation have mean close to 0.

To gain insight into variation in the measurement procedure, we examine pairs of intervals with high canonical correlation. The leading canonical correlation coefficient for each interval pair is shown in Figure 3. Note that this figure is symmetric because the correlation does not depend on the order of the intervals. Moreover, the values on the diagonal are 1. This figure directs our attention to seven interval pairs: (4—9), (5—6), (8—9), (9—10), (9—16), (9—17), and (16—17). As shown in the next section, the high canonical correlations associated with these pairs seem to have scientific explanations. Although other pairs also provide interesting insights, we will not discuss them.

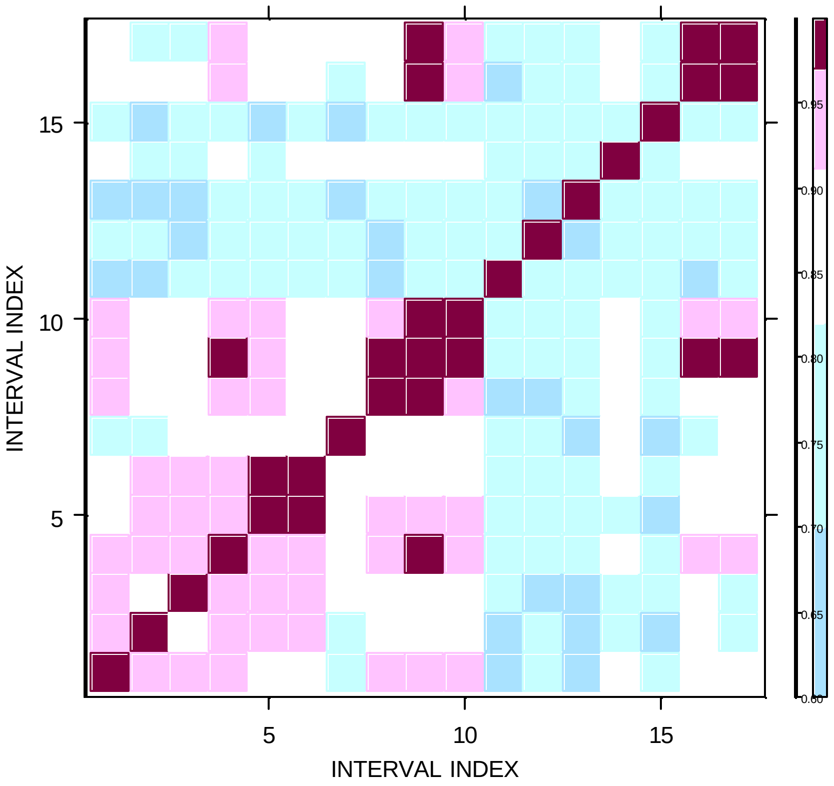


Figure 3. Leading canonical correlation coefficient for all pairs of intervals. Pairs with highest correlation are of greatest interest.

In trying to understand the overall pattern in Figure 3, one might start with the notion that high canonical correlation generally does not occur without distinct spectral peaks in both intervals. Figure 3 shows that the correlation is low for intervals 11 to 15. Figure 1 shows that these intervals contain only minor spectral peaks. It is interesting that interval 7 shows only low correlation despite its spectral peak. This corresponds to the lack of peak-height correlation shown in Figure 2. Interval 7 provides evidence that our normalization is effective in removing variation that affects all peaks proportionally. We note that the smallest canonical correlation coefficient in Figure 3 is larger than 0.6. The reason is that CCA always gives a positive coefficient because the algorithm is based on maximization.

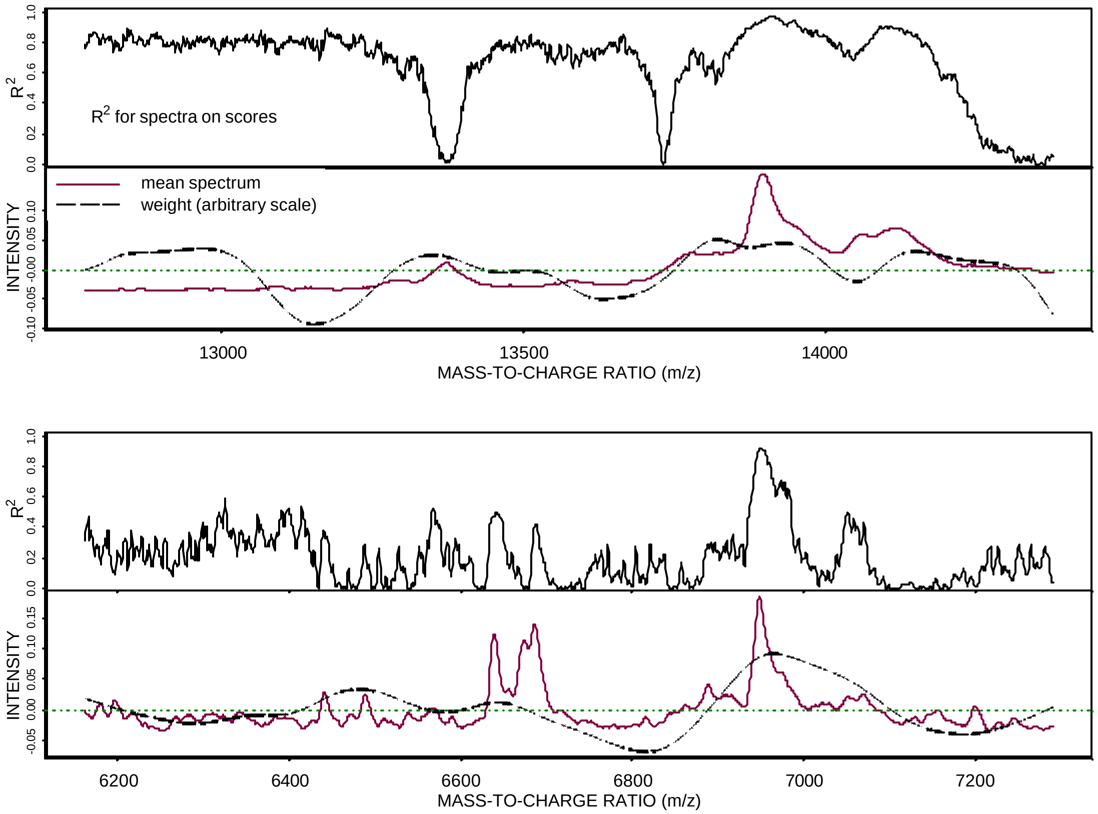


Figure 4. CCA results for intervals 4 (bottom) and 9 (top). High R^2 and high mean spectrum indicate highly correlated spectral peaks (e. g., m/z = 6950 and m/z = 13900). R^2 for spectra regressed on scores shows m/z values responsible for high canonical correlation. The weight functions convert spectra into scores.

To answer the question of why a pair of intervals has a high leading canonical correlation coefficient, we attempt to identify the parts of the intervals primarily responsible. As shown in Figure 4, we display for each interval, the mean spectrum, the weight function, and the R^2 values for the spectra at each point in m/z regressed on the CCA scores. The R^2 value at a particular m/z is the fraction of the variation of the spectra explained by the scores. The mean spectrum shows locations of spectral peaks in the interval. The weight function shows how the parts of the interval contribute to the scores for that interval. The scores underlie the canonical correlation coefficient that we want to explain. The R^2 values show how the spectra in various parts of the interval are related to the scores. Of initial interest in Figure 4 are m/z values for which the mean spectrum and the R^2 values are both high. High mean spectrum indicates high protein concentration, and high R^2 indicates a close relation to the scores that lead to the high canonical correlation coefficient.

Figure 4 provides the basis for examining the canonical correlation between intervals 4 (bottom) and 9 (top). High mean spectrum and high R^2 occurs for interval 4 at m/z = 6950 and for interval 9 at m/z = 13900. Apparently, these points correspond to the doubly- and singly-charged versions of the same protein. In mass spectrometry, evidence of peaks corresponding to differing amounts of charge is not unusual. More than this, Figure 4 indicates that the measurement-to-measurement spectral variations at these two m/z values are closely related. Identification of the underlying source of variation is an interesting question. That these variations occur despite the normalization suggests that this source of variation affects at least one but not all of the proteins in the specimen.

The R^2 curve for interval 9 (top) deserves further discussion. Variation of the spectral deviations $x_i(u)$ in this interval is dominated by a single component with the form $f_{i1}h_1(u)$. Because of the baseline correction applied, the integral of $h_1(u)$ over the interval must be close to 0. For this reason, the baseline correction spreads the variation of the spectral peaks near m/z = 14000 over the entire interval. Thus, the R^2 curve is high throughout the interval except where $h_1(u)$ is nearly 0. This accounts for the two places above m/z = 13500 where the R^2 curve dips nearly to 0. The dip below m/z = 13500 may be due in part to variation in the small peak evident in the mean spectrum. In contrast, the R^2 curve for interval 4 (bottom) seems to be affected by the generally uncorrelated variation in several peaks.

The weight functions shown in Figure 4 are not easily interpreted. The main reason is that the measurement-to-measurement variation in the spectral deviations $x_i(u)$ is correlated from one m/z value to another. We force the weight functions to integrate to 0 over the interval because the spectral deviations $x_i(u)$ do. Moreover, the weight functions are computed subject to a smoothness penalty. It is hard to understand how the algorithm meets these requirements when the spectrum-to-spectrum correlation is high.

This paper identifies correlation evident in replicate mass spectral measurements with the idea that such correlation points to sources of variation in the measurement procedure. One source of variation that leads to correlation is the well-known effect of nonuniform crystallization on the surface of the protein chip. We normalize the spectra to remove this effect and look for remaining correlation that points to sources of variation that affect some proteins proportionally more than others. We found two proteins, each of which varies enough after normalization that the correlation between the singly- and doubly-charged versions is obvious. We found pairs and groups of proteins with correlation after normalization that suggests a common source of variation. In both of these cases, the sources of variation seem to be in the sample preparation step. Finally, we found that the two proteins with the highest intensity have correlation in their peak distortion. This suggests that some source of variation, perhaps the nonuniform crystallization, sometimes causes the amount of material in the spectrometer to be so large that the detector is overloaded, thereby distorting the peak.

Sources of variation in the SELDI-TOF mass spectrometry procedure lead to spectral variation that is complicated in that neither the form of the variation nor the mechanism responsible for the variation is easily modeled. This has implications for the use of this procedure to obtain comparable measurements over time or from different laboratories. Because of the complexity of the form of the variation, detecting important differences in the execution of the measurement procedure is difficult. Similarly, because of the complexity of the mechanism, maintaining control of the sources of variation is difficult. These difficulties suggest that the most fruitful approach to measurement variation in SELDI-TOF mass spectrometry might be an effort to reduce the effects of sources of variation.

5.2 High-dimensional Analysis of Variance with Application to Polymer Mass Spectrometry

Z.Q. John Lu
Statistical Engineering Division, ITL

Charles M. Guttman
Polymers Division, MSEL

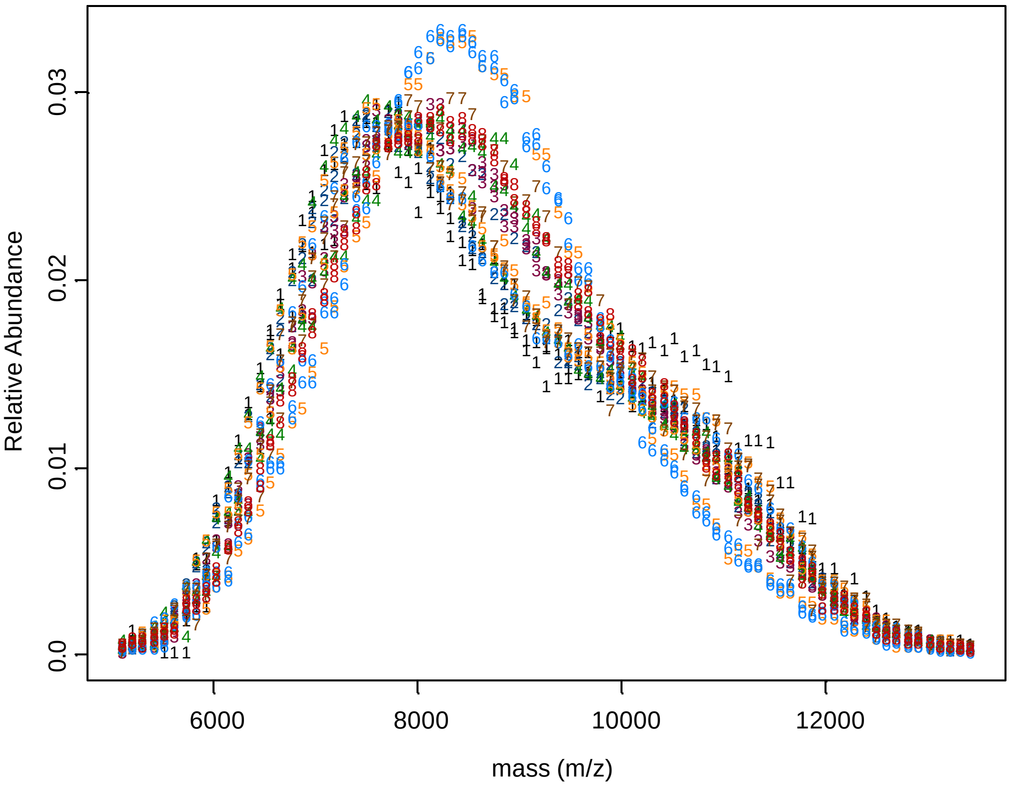


Figure 1. Measured molecular distribution curves from MALDI-TOF mass spectrometry. Eight sampling areas from a bottle of polymer material have been taken and there are three samples taken from each area. The curves for the three replicated samples taken from the same area are labeled by the same color and symbols (labeled 1,2,…,8).

Many high throughput measurements produce data that should be regarded as high-dimensional since the number of measured variables often outnumbers the number of samples or measurements. Examples may include MALDI-TOF mass spectrometry data in synthetic polymers, microarray gene expression, chemical spectra, spectral imaging, particle size distribution, etc. In all of these problems, there are great needs to extend statistical methods to problems of high-dimensional data in the experimental setup of statistical design. This summary describes some newly developed statistical methodology for the variability of high-dimensional data from designed experiments, such as one-way ANOVA data. The proposed high-dimensional analysis of variance (HANOVA) utilizes extensively the singular value decomposition as a structure discovery method, in analysis of both overall and between-setting variability of high-dimensional measurements.

As an example, we consider again the example of molecular mass distribution data from MALDI-TOF mass spectrometry experiments. Such data output of MALDI-TOF mass spectrometry can be put in the bivariate form of

(1) $$\{(x_{ij}, y_{ij}), i=1,...,m; j=1,...,r\},$$

from the simple design of measuring at *m* settings with *r* replications. Here we always use *m=8, r=3* in our analyzed data and the x_{ij}'s are the measured mass *(m/z)* vector based on the time-of-flight principle using a known reference sample and y_{ij} is the normalized intensity vector (corresponding to the sorted mass over charge values). Since each element in the vector y_{ij} is proportional to the number of ions of the measured mass of one type of molecules, the normalized intensity represents the composition or molecular mass (weight) distribution of molecules of different masses (lengths) in the sample, at the *i*th setting and *j*th replication. Though there are calibration errors associated with mass identification, this step is fairly well understood and the uncertainty is mostly negligible. The errors in intensity measurements, due to the intrinsic Poisson errors in the ions counting process, will be the main object of the analysis of variance.

Because the errors in x_{ij} due to mass resolution are negligible, for simplicity we may create mass spectral data with balanced support. That is, we want to find a common set of mass-to-charge values $x_{11},..., x_{sr}$ on which the intensity values y_{ij}'s are measured or interpolated. We regard this as an important part of the data normalization /preprocessing step, in that this will allow us to ignore x_{ij} and to treat y_{ij} as multivariate data vectors**,** just as in multivariate data analysis. One can then define the data matrix as:

(2) $$X = (y_{11},..., y_{1r}; y_{21},..., y_{2r};...; y_{m1},..., y_{mr})',$$

which have *n=mr* rows and *p*=number of resolved mass-to-charge values from mass spectrometer.

We may define the setting means as

$$\bar{y}_i = \frac{1}{r}\sum_{j=1}^{r} y_{ij},\ i=1,2,..., m.$$

The overall mean is defined as $\bar{y} = \frac{1}{mr}\sum_{i=1}^{m}\sum_{j=1}^{r} y_{ij} = \frac{1}{m}\sum_{i=1}^{m}\bar{y}_i.$

Define two versions of (column) mean-sweeped X:

$$X_w = (y_{11} - \bar{y}_1, \ldots, y_{1r} - \bar{y}_1; \ldots; y_{m1} - \bar{y}_m, \ldots, y_{mr} - \bar{y}_m)^T,$$

$$X_b = (\bar{y}_1 - \bar{y}, \ldots, \bar{y}_1 - \bar{y}; \ldots; \bar{y}_r - \bar{y}, \ldots, \bar{y}_r - \bar{y})^T.$$

Then

(3) $$X_c = X_w + X_b.$$

One can also prove the following ANOVA-type identity.

$$W_a = X_c^T X_c = \sum_{i=1}^{m} \sum_{j=1}^{r} \left(y_{ij} - \bar{y}_i\right)\left(y_{ij} - \bar{y}\right)^T$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{r} \left(y_{ij} - \bar{y}_i\right)\left(y_{ij} - \bar{y}_i\right)^T + \sum_{i=1}^{m} r\left(\bar{y}_i - \bar{y}\right)\left(\bar{y}_i - \bar{y}\right)^T$$

(4) $$= X_w^T X_w + X_b^T X_b \triangleq W_0 + W_b.$$

Thus, one may produce an ANOVA table.

Table 1. Analysis of variance table for functional data

Source of variation (covariance)	Sum of cross products (dispersion matrix)	Degree of freedom	Mean square	Expected value of mean square
Between-Settings	$\sum_{i=1}^{m} r\left(\bar{y}_i - \bar{y}\right)\left(\bar{y}_i - \bar{y}\right)^T \ (\triangleq w_b)$	*m-1*	S_b	$\Sigma + \Sigma_{i=1}^{m} r(m_i - \bar{m})(m_i - \bar{m})^T$ $/(m-1)$
Within-Settings	$\sum_{i=1}^{m} \sum_{j=1}^{r} \left(y_{ij} - \bar{y}_i\right)\left(y_{ij} - \bar{y}_i\right)^T \ (\triangleq w_0)$	*m(r-1)*	S_w	Σ
Total about the grand average	$\sum_{i=1}^{m} \sum_{j=1}^{r} \left(y_{ij} - \bar{y}_i\right)\left(y_{ij} - \bar{y}\right)^T \ (\triangleq w_a)$	*mr-1*		

However, arrangement of data as in Table 1 is not enough, for much more data reduction is needed in order to comprehend the variability in the dispersion matrices S_b and S_w, both of which are high-dimensional and may be singular due to the small sample size *(n=8*3=24)* and high-dimensionality of measurements *(p=81).*

Note that we may write X as given in (2) to *consist* of a sum of a signal matrix M (the mean structure) consisting of rows given by $m_i^T = (f_i(x_1), \ldots, f_i(x_p))$ and a noise matrix E (the variability component) consisting of rows e_{ij}, i.e.,

(5) $$X = M + E.$$

If there is no change in the spectral curve in each setting m_i, the rows of the signal matrix *M* should be the same, so *M* should have rank one. If there are changes in the spectral curves between the settings, the rank of the signal matrix is at least two. In general, the rank of the signal matrix *M* determines how many different spectral structures are contained, and we may write

$$M = Cb \quad (6)$$

where b is a $p \times d$ matrix, representing the unknown underlying basis of spectral curves, and C is a $n \times d$ matrix, representing the design matrix relating the $n = mr$ experiments to the unknown spectral curves of interest. The statistical problem is to recover the hidden structure or test on the hidden dimensionality *d*, which is much smaller than *p* or *n*, based on observed data matrix *X*.

In order to overcome the curse of dimensionality problem, we propose a singular value decomposition applied to both the overall deviation matrix *(X_c)* and the between-setting or within-setting deviation matrices X_b and X_w. Table 2 shows the results of this singular value decomposition applied to the data from Figure 1. Here we have shown only the first 10 leading singular values. From the last column of Table 2, it appears that there are probably two or three significant singular values from the between-setting variations (compare row 2 and row 3).

Table 2. Analysis of singular values for functional data

Source of variation	Leading 10 singular values of the deviation matrix (X_w, X_b, X_c)	Degree of freedom	Normalized singular values: or square roots of leading eigenvalues from the mean dispersion matrix (S_b, S_w, S_a)
Between-Settings	*0.077, 0.018, 0.007, 0.005, 0.005, 0.004, 0.003, 0.000*	*8-1=7*	*0.029, 0.007, 0.003, 0.002, 0.002, 0.001, 0.001*
Within settings	*0.022, 0.008, 0.007, 0.006, 0.005, 0.005, 0.005, 0.004, 0.004, 0.004*	*8(3-1)=16*	*0.005, 0.002, 0.001, 0.001, 0.001, 0.001, 0.001, 0.001, 0.001, 0.001*
Total about the grand average	*0.077, 0.027, 0.009, 0.007, 0.007, 0.006, 0.006, 0.005, 0.005, 0.005*	*8(3)-1=23*	*0.019, 0.007, 0.002, 0.002, 0.002, 0.001, 0.001, 0.001, 0.001, 0.001*

In summary, for high-dimensional data analysis, p is often larger than the sample size n. When this occurs, many standard tests from multivariate data analysis, such as MANOVA, do not apply directly. We argue that this curse of dimensionality problem can be overcome by examining the nature of the dimensionality in high-dimensional data using methods such as the singular value decomposition. A statistical paper to document the statistical methodology of high-dimensional analysis of variance is being written to address some of the statistical significance-testing issues.

5.3 Feasibility Study for the Development of a Motion Imagery Quality Metric

Ana Ivelisse Aviles,
Statistical Engineering Division, ITL

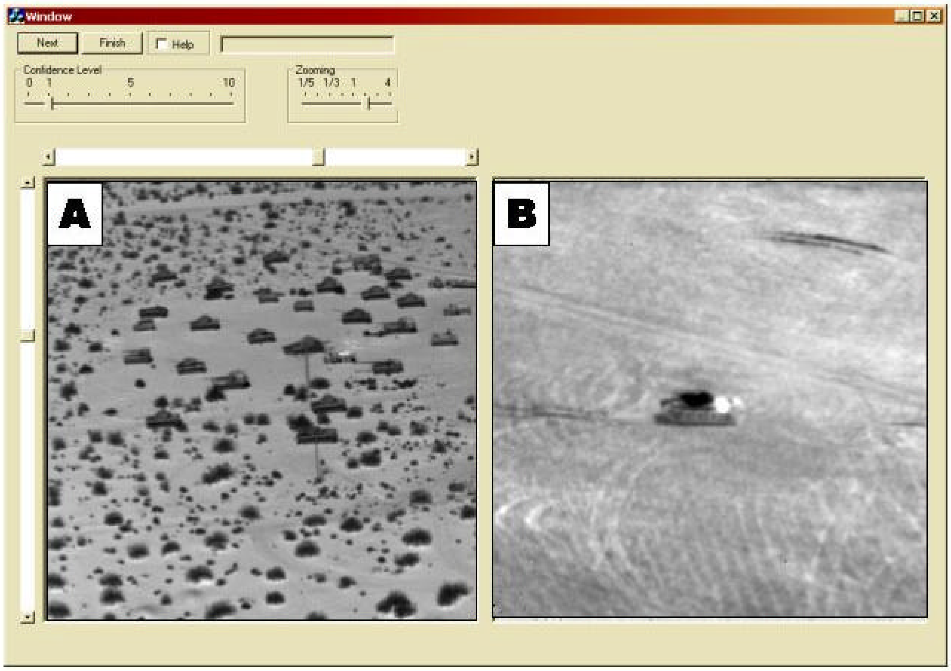
Charles Fenimore and John Roberts
Information Access Division, ITL

John Irvine, David Cannon, Steven Israel, Larry Simon, James Miller
Science Applications International Corporation (SAIC)

Paul Tighe
Booz, Allen, and Hamilton

Richard Behrens
Mitre Corporation

Charles Watts and Michelle Brennan
Moriarty and Associates



Format for Capturing Paired Comparison Quality Ratings.

The National Imagery Interpretability Rating Scale (NIIRS) is an approach embraced by the Intelligence Community for still imagery. Each NIIRS level indicates the types of exploitation tasks an image can support based on the expert judgments of experienced analysts. Development of a NIIRS for a specific imaging modality rests on a perception-based approach. Accurate methods for predicting NIIRS from the sensor parameters and image acquisition conditions have been developed empirically and substantially increase the utility of NIIRS.

The motion imagery community would benefit from the availability of standard measures for assessing image interpretability. NIIRS has served as a community standard for still imagery, but no comparable scale exists for motion imagery. Several considerations unique to motion imagery indicate that the standard methodology employed in the past for NIIRS development may not be applicable or, at a minimum, require modifications. Traditional methods for NIIRS development rely on a close linkage between perceived image quality, as captured by specific image interpretation tasks, and the sensor parameters associated with image acquisition. The dynamic nature of motion imagery suggests that this type of linkage may not exist or may be modulated by other factors. The goal of this study was to understand the interplay of motion and perceived image quality. The findings of this study provide a first step in developing a quality metric motion imagery.

In exploring avenues for development of a quality metric for motion imagery, a clearer understanding of the factors that affect the perceived quality of motion imagery is needed. If perceived quality is highly dependent on scene content (such as target motion), then predicting image quality from the sensor and acquisition parameters is not possible. Although scene complexity does not appear to be a major factor affecting perceived quality of still imagery, it could be important for motion imagery. In particular, an interaction between target motion and scene complexity has been hypothesized for motion imagery. Another area of concern is the effect of the motion of the sensor platform, since change in camera position affects obscuration, masking, and perception of three-dimensional information.

The study was conducted to understand the effects of specific factors on perceived image interpretability for motion imagery. These factors are:

1. Target motion: Other studies indicate that moving targets exhibit greater salience that can enhance target detection and recognition
2. Camera motion: The parallax effect and changing viewing geometry assist the analyst, particularly when viewing partially occluded targets
3. Scene complexity: It has been hypothesized that both target and camera motion exhibit greater effects on perceived interpretability when the scenes are more complex.

Imagery

The image matrix was populated with existing holdings at the National Geospatial-Intelligence Agency's Persistent Surveillance Office (NGA/IXA) and special collections by the National Institute of Standards and Technology (NIST). The imagery used in this evaluation was High Definition Television (HDTV) data collected from a 720x1280 progressive scan camera system. While the ultimate development of a motion imagery quality metric must embrace a range of camera systems and imaging conditions, this effort focuses on understanding specific effects related to perceived image quality. Consequently, the relatively limited range of image conditions effectively controls for a number of factors that might otherwise confound the effects of interest in this study. The image set was well characterized in terms of target motion, camera motion, and scene complexity. Each clip was rated from 1 (low) to 5 (high) with respect to each of these factors. The ratings are subjective, based on the following definitions:

- Target Motion: The targets (usually vehicle or people in the scene) are moving with respect to the background and/or the raster
- Camera Motion: The camera is moving with respect to the background
- Scene Complexity: High complexity scenes include diverse clutter, multiple independent motions, higher spatial frequency information, target confusers, partial obscuration, or other features that make it difficult for an observer to detect and track the targets

Table 1. Characteristics Represented in Each GSD Bin

Target Motion	**Scene Complexity**	**Camera Motion**
Low	Low	Low
Low	High	Low
High	Low	Low
High	High	Low
High	High	High

In addition, the ground sample distance (GSD) was estimated via mensuration of known objects in the scene and, where possible, validated by comparison to metadata. From a full database of several hundred motion imagery clips, a set of 35 clips was selected for the evaluation. These clips were grouped into bins of similar GSD, where each grouping spanned five combinations of conditions (Table 1). The unbalanced design arose from the limitations of the available imagery. From each clip, a high-quality still image was generated using 5 consecutive frames (using a super-resolution technique). Thus, the full set of imagery consisted of 35 video clips of approximately 5 seconds in length and 35 corresponding still images.

Approach

The general approach was a small, focused evaluation that addresses the fundamental issues related to the development of a measure of interpretability for motion imagery. Imagery analysts were asked to provide NIIRS ratings and perform pairwise comparisons for a set of 35 motion imagery clips. For each pairwise comparison, the analyst was asked to indicate the relative image interpretability for the two clips using a ratio scale. In addition, analysts were asked to rate each video clip and corresponding still images using the current Visible NIIRS.

Analysis of these ratings allowed us to examine a number of critical issues:

1. Relationships between perceived image quality and target motion.
2. Relationships between perceived image quality and scene complexity.
3. Interactions between scene complexity and target motion that could affect perceived image quality.
4. Relationships between perceived image quality and camera motion.
5. Relationships between perceived image quality for motion imagery and the Visible NIIRS.
6. Consistency of the perceived relative quality levels across analysts.
7. Internal consistencies of the ratings from each analyst.
8. Relationships between perceived image quality and sensor/acquisition parameters.

Evaluation Design and Execution

Twelve image analysts (IAs) participated in the evaluation. All of the analysts had experience with operational exploitation of imagery and were NIIRS certified. Experience levels spanned a range from junior analysts to exceedingly experienced ones. Following the initial introduction, each IA worked through the evaluation at his/her own pace, taking breaks as needed. All imagery was viewed on calibrated monitors under controlled lighting conditions. To facilitate display of motion imagery for paired comparisons, the set-up used two PCs, each with a high-end color monitor. All responses were recorded in hardcopy. At the end of the evaluation, each IA completed an exit questionnaire to provide subjective feedback. The four steps in the evaluation were:

> Step 1: Visible NIIRS ratings of still images that were extracted from each motion imagery clip
> Step 2: Visible NIIRS ratings of the motion imagery clips
> Step 3: Paired comparisons of the motion imagery clip to a single frame from the clip sequence
> Step 4: Paired comparisons between various pairs of motion imagery clips.

Results of the Evaluation

Throughout the evaluation, target motion has a significant effect on perceived image quality, in terms of both NIIRS ratings and paired comparisons. Motion imagery clips in which the targets are moving are consistently rated higher. This result is not surprising, since motion increases target salience. It is interesting to note, however, that the effects due to camera motion were not statistically significant and there are only weak indications of an interaction effect involving target motion and scene complexity.

Steps 1 and 2 of the evaluation demonstrate that trained IAs are capable of providing consistent NIIRS ratings for motion imagery. On average, the NIIRS ratings for the motion imagery clips are slightly (about 0.25 NIIRS units) higher than for the corresponding still image (Figure 1a). Both the NIIRS ratings and the paired comparisons indicate that image interpretability is inversely related to log10(GSD). While the relationship is linear, the slope is much lower than expected (Figure 1b). Historically, a doubling or halving of GSD produces a one NIIRS unit shift. These ratings exhibit about a half NIIRS unit shift when GSD varies by a factor of two. This flatter relationship may be due to changes in the image associated with softcopy display or because the color imagery provides better target contrast than for panchromatic imagery.

Visible NIIRS ratings for the motion imagery clips are slightly, but statistically significantly, higher than for the corresponding NIIRS ratings of still images. The paired comparisons suggest that the perceived interpretability of motion imagery is considerably higher than for still images, but the Visible NIIRS is not sensitive to all the factors influencing the perceived interpretability of motion imagery. Figure 2 illustrates this point. On the left side, the bars represent the values of t-statistics to test for significant differences between motion imagery and still imagery. The blue bars are computed from the NIIRS ratings, i.e., the t-statistic arises from the paired test of NIIRS ratings for still images versus NIIRS ratings for the corresponding video clips. The red bars are the t-statistics computed from the paired comparisons of stills to video clips (step 3) and test for a significant difference from zero. Note that the red bars show a much stronger difference, indicating the motion imagery has much higher interpretability than the corresponding still frames.

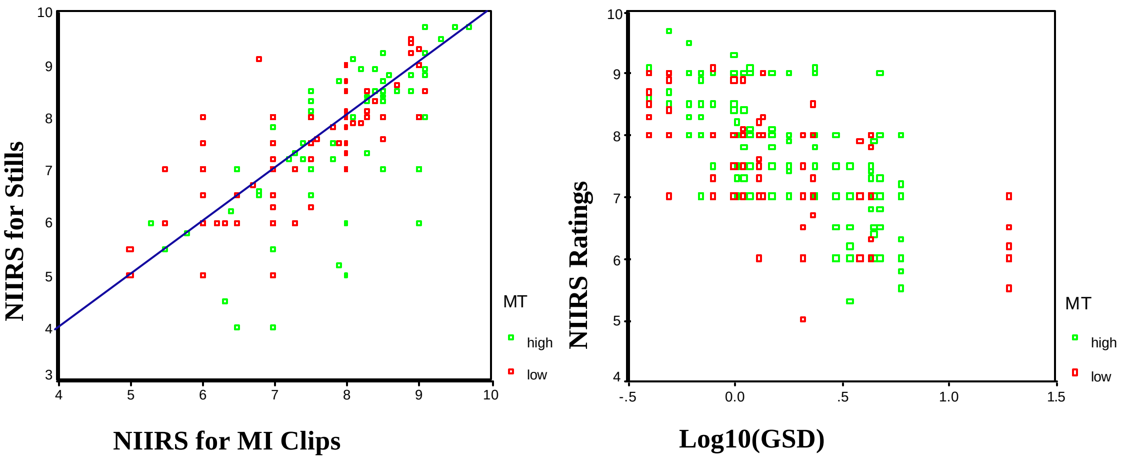


Figure 1. (a) The Graph on the Left Depicts the NIIRS Ratings for the Motion Imagery Clips Compared to the Visible NIIRS Ratings for the Corresponding Still Images; (b) The Graph on the Right Shows the Relationship Between the NIIRS Rating of the Motions Imagery Clip and the Estimated GSD for the Same Clip

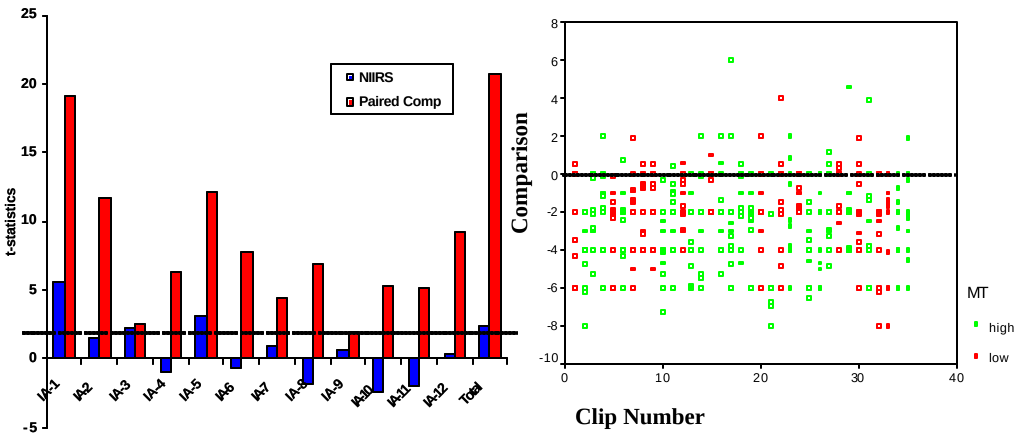


Figure 2. (a) The t-statistics for Testing a Difference Due to Motion and (b) Raw Ratings from the Paired Comparisons of Motion Imagery Clips to Still Images

Conclusions and Future Directions

This study indicates that target motion is a significant and consistent factor affecting image quality. One of the implications of this finding is that traditional NIIRS development methods will not translate effectively into the motion imagery domain. The study also raises questions about other factors that influence image interpretability. To address these issues, we propose a series of small, focused evaluations, each intended to answer a single basic question about interpretability (Table 2) for FY2005. Finally, the analysis of the NIIRS ratings and the paired comparisons indicate that Visible NIIRS does not capture the full range of factors inherent in the perceived quality of motion imagery. A new scale, therefore, is needed to address the quality and interpretability aspects of motion imagery.

Table 2. Proposed Focused Evaluations for Investigation of Motion Image Quality

Evaluation	**Questions Addressed by Evaluation**
Frame rate study	How does image interpretability vary with frame rate?
Color study	Assess interactions between color and motion
Resolution & viewing geometry	How do low grazing angles affect relationship between interpretability and GSD?
Criteria satisfaction	Can IAs consistently rate criteria relative to MI markers? How does target motion affect criteria ratings? Are motion-sensitive tasks rated differently?

Currently, no NIIRS or similar quality metric exists for motion imagery. If a Motion Imagery NIIRS (MI NIIRS) or similar quality metric were developed, it would be a useful tool for a number of applications:

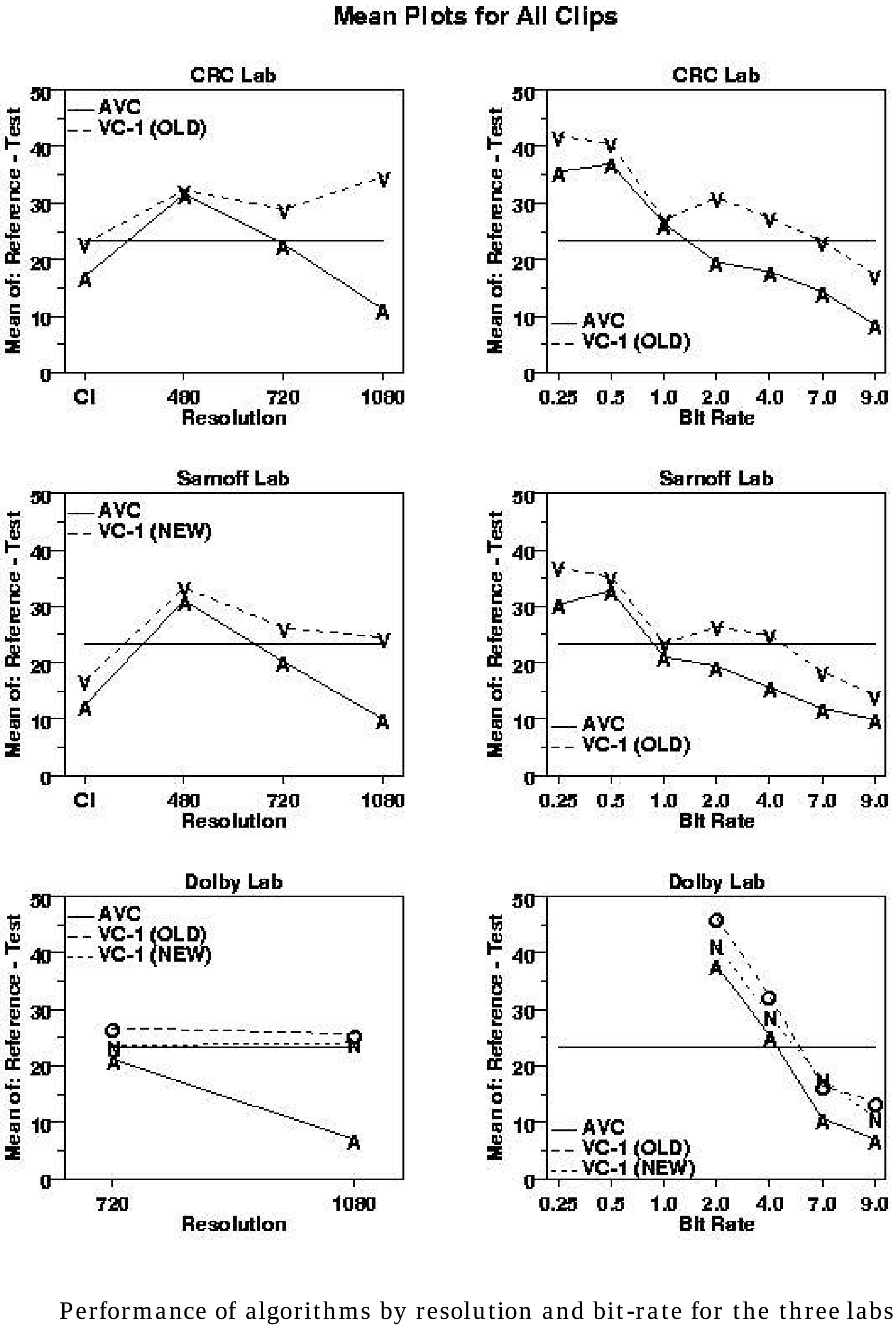
- Sensor and System Requirements: By expressing the requirements for new Intelligence, Surveillance, and Reconnaissance (ISR) systems in terms of NIIRS, the system performance can be related directly to fulfillment of specific military missions. Design and trade studies can be performed.
- Tasking and Collection Management: By linking sensor parameters to MI NIIRS through an Image Quality Equation, collection managers can assess how best to task imaging assets to satisfy critical needs.
- Evaluation of Image Processing and Image Compression: An MI NIIRS would quantify the benefits from image processing to enhance the imagery. Conversely, an MI NIIRS can quantify the loss associated with image compression or other modifications to the image chain. This could be a significant concern since the capability to acquire digital motion imagery is likely to exceed the communications capacity in many tactical settings.

This study has started to explore the implications for development of a "NIIRS-like" scale for motion imagery. A significant finding of the study is that traditional NIIRS development methods will not translate effectively into the motion imagery domain. A series of small, focused evaluations, each intended to answer a single basic question about interpretability, has been proposed for the next year.

5.4 Video Quality Assessment for MPEG and ATSC

Alan Heckert, Stefan Leigh
Statistical Engineering Division, ITL

Charles Fenimore
Information Access Division, ITL



Performance of algorithms by resolution and bit-rate for the three labs

The Motion Imagery Quality Project of the Information Access Division and SED have collaborated in designing and analyzing tests intended to quantify video quality. Two rounds of subjective video quality tests for the assessment of the compression efficiency of a new generation of video codecs have been undertaken. These tests examine emerging open standards from the Motion Pictures Experts Group (MPEG) and the Society of Motion Picture and Television Engineers (SMPTE).

The gold standard in video picture quality testing is the use of human assessors for subjective ratings of video clips. While there are well-established approaches for many aspects of such tests, it has not been uncommon for the industry to analyze resulting data in questionable or inappropriate ways.

At MPEG, NIST was one of four leaders in Verification Testing for the new Advanced Video Codec (AVC). The subjective tests showed the new technology delivers a 50% gain in efficiency compared to the older MPEG-2. The report of this collaboration is being used by the industry to guide investment decisions in AVC and competing technology. Because the results were very positive for AVC, there is AVC product on the market today. In that collaboration a year ago, IAD and SED provided an analysis of the data and compared our results to those of collaborators using methods developed over time within the video picture testing community itself.

In the Advanced Television Standards committee (ATSC) we have continued this effort by analyzing another round of tests comparing AVC with new technology from SMPTE, the Video Codec 1 (VC-1). Collaboration with the Communications Research Center (CRC) of Canada was undertaken to compare the performance of these two codecs. We have reported to the ATSC Test Chair that we are able to determine the significance of the differences between the two codecs. The report came in for substantial criticism from one of the new codec proponents.

ANALYSIS: Data consisted of matched opinion scores obtained from 20-30 human subjects, each at 3 participating laboratories (CRC, Sarnoff, Dolby), using clips compressed by 3 algorithms (AVC, VC-1 “old”, and VC-1 “new”) at 4 distinct resolutions and 6 different bit rates. All paired scores represented the same subject rating the identical clip untreated (no compression) and treated. Ratings were scored as integers between 0 and 100. All analyses were performed on raw differences or differences of means: reference clip minus test (treated) clip result. Analyses consisted of graphical organization of summarized and non-summarized data, representing f(Reference-Test) against resolution and/or bit rate, segregated into different plot panels for Lab, clip, or aggregated Lab or clip. Most visuals directly contrast the performances of the algorithms being compared (AVC versus VC-1). (See Figures.) The graphics largely support the superiority of the existing MPEG algorithm over the developmental SMPTE one, although areas of crossover in performance are potentially rich in information for the designers of the algorithms. For formal support of trends observed in the visuals, nonparametric tests of slippage between score clusters or means were employed on a level-by-level basis. We eschewed a one-size-fits-all linear modeling (ANOVA) approach as sweeping under the rug many interesting frame-specific patterns visible in the data.

The introduction and adoption of statistical tools that are new to the television and motion picture industry is a slow process that requires extensive ongoing cross-fertilization. We expect to continue these kinds of efforts within MPEG and to document our experiences in a suitable (IEEE) journal.

5.5 Cryogenic Detection of Weakly Interacting Particles

K. J. Coakley
Statistical Engineering Division, ITL

D. N. McKinsey
Yale University

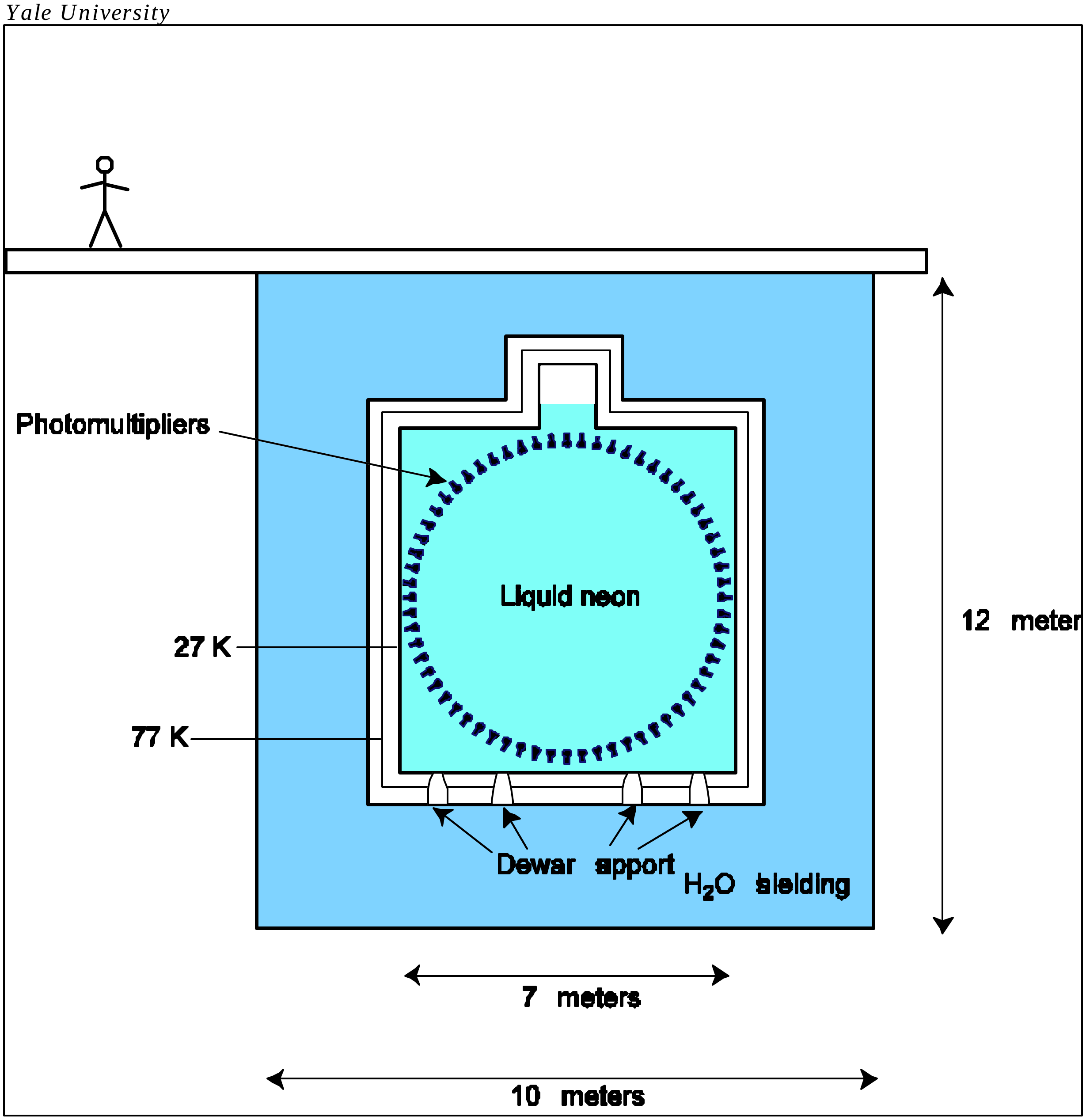


Figure 1. Diagram of the proposed CLEAN experiment.

In the proposed experiment CLEAN (Cryogenic Low Energy Astrophysics with Noble gases), the low energy spectrum of solar neutrinos, supernova neutrinos and other weakly interacting particles would be detected. Statistical efforts include development of event reconstruction algorithms, background discrimination, and experimental planning.

The study of neutrinos plays a prominent role in astrophysics and particle physics. Though they are emitted in vast numbers by stars and can easily be made in modern particle accelerators, neutrinos are difficult to detect because they have no charge and only interact through the weak force (which explains radioactive decay and related phenomenon). Recent experiments demonstrate that solar neutrinos oscillate between different mass states as they travel from the Sun to Earth. The CLEAN instrument should provide invaluable data for rigorously testing competing theories of the neutrino and of the Sun. CLEAN should be sensitive to weakly interacting massive particles (WIMPS). Astrophysical evidence on a variety of distance scales clearly shows that a large fraction of the mass of the universe cannot be accounted. This matter is dark because it does not appear to emit or absorb any electromagnetic radiation. The existence of WIMPS is a very plausible explanation of this dark matter. Data from CLEAN should improve theoretical understanding of the supernova collapse mechanism.

In CLEAN, the unwanted background signal can be orders of magnitude more intense than the signal of interest. Thus, we need powerful statistical methods for background discrimination. Low energy neutrinos would be detected based on scintillation light produced by neutrino-electron scattering, or neutrino-WIMP scattering, in a large cryostat filled with liquid neon. Such events of interest would occur uniformly throughout the cryostat. For a spherical cryostat geometry, the probability distribution function (pdf) for the radial location r of an event of interest would be proportional to r^2. On average, the number of scintillation photons produced by an event would be proportional to the energy deposited by the neutrino. The scintillation photons Rayleigh scatter as they propagate in the neon. Thus, the scintillation photons do not travel in straight line trajectories. Further, in our detection model, each scintillation photon is shifted to lower energy and re-emitted by detectors before ultimate detection. The background signal is mainly due to gamma rays, i.e., photons, produced by radioactive decay of isotopes found in the materials from which the outer spherical walls and photomultiplier tubes are constructed. As these background gamma rays propagate inward, they deposit energy when they Compton scatter or are absorbed. Like events of interest, background gammas produce scintillation light. Due to attenuation, the probability that a gamma penetrates an inner fiducial volume occupying a fraction p of the total detection volume (defined by $r < p^{1/3}R$) decreases as p decreases. Because of the attenuation of background gamma rays, the instrument is said to be self-shielding. Thus, if one can accurately estimate the radial position of an event, one can potentially discriminate background events from events of interest with high confidence.

Current and future statistical work includes: stochastic modeling of scattering and transport of gamma ray and scintillation photons, statistical planning, development and testing of statistical background discrimination methods, development of empirical models for calibration of statistical estimates of event location, development of sampling schemes for training data for calibration, energy spectrum estimation, quantification of detector efficiency and false detection rate, and uncertainty analysis.

We estimate the location of an ionizing radiation event that produces multiple photons within a spherical detection volume of radius R based on count data recorded by detectors mounted on the boundary of this detection volume, using a calibrated spatial Maximum Likelihood method. The scattering length for the photons is l_S. The detectors cover approximately 75 % of the total area of the detection volume boundary. In the "shift" detection model, photons are absorbed and isotropically re-emitted by the detectors. In the "no-shift" detection model, photons are not re-emitted. CASE A: $l_S = \infty$ and "shift" detection model. CASE B: $l_S = 0.1R$ and "no-shift" detection model. CASE C: $l_S = 0.1R$ and "shift" detection model. For each plot, we show the 0.1 quantile of the estimate (dashed horizontal line), $r_p / R = p^{1/3}$ where $p = 0.1$ (solid horizontal line), and the line of equality corresponding to a perfect estimate.

Presentations

D.N. McKinsey and K.J. Coakley, "CLEAN", Mini-Symposium on Underground Science, April 2002 Meeting of the American Physical Society, Albuquerque, NM.

K.J. Coakley and D.N. McKinsey, "Event Location Estimation and Background Discrimination in a Proposed Low Energy Neutrino Experiment," poster session, April 2002 Meeting of the American Physical Society, Albuquerque, NM.

Daniel McKinsey, Walter Lippincott, James Nikkel, Andrew Hime, Mark Boulay, Jeff Lidgard, Kevin Coakley, and Edward Kearns, "Neutrino and WIMP detection with CLEAN", April 2004 Meeting of the American Physical Society, Denver, CO.

K. J. Coakley, "Classification Problems in Neutrino Physics" 2004 Meeting of the International Federation of Classification Societies. Chicago, IL. July 15-18, 2004.

Publications

C.J. Horowitz, K.J. Coakley, and D.N. McKinsey, "Supernova Observation via Neutrino-Nucleus Elastic Scattering in the CLEAN Detector," Physical Review D, 68(2), 23005, 2003.

K.J. Coakley and D.N. McKinsey, "Spatial Methods for Event Reconstruction in CLEAN," Nuclear Instruments and Methods in Physics Research A, 522, pp. 504-520, 2004.

D.N. McKinsey and K.J. Coakley, "Neutrino Detection in CLEAN," to appear in Astroparticle Physics.

The CLEAN instrument has high scientific potential because of its low energy sensitivity. Further, development of measurement technology related to CLEAN (particularly noble gas purification, low-background light detection, the use of light detectors at low temperature, and statistical methods for background discrimination) should have a broad impact in nuclear and particle physics as well as the detection of fast neutrons in homeland security applications.

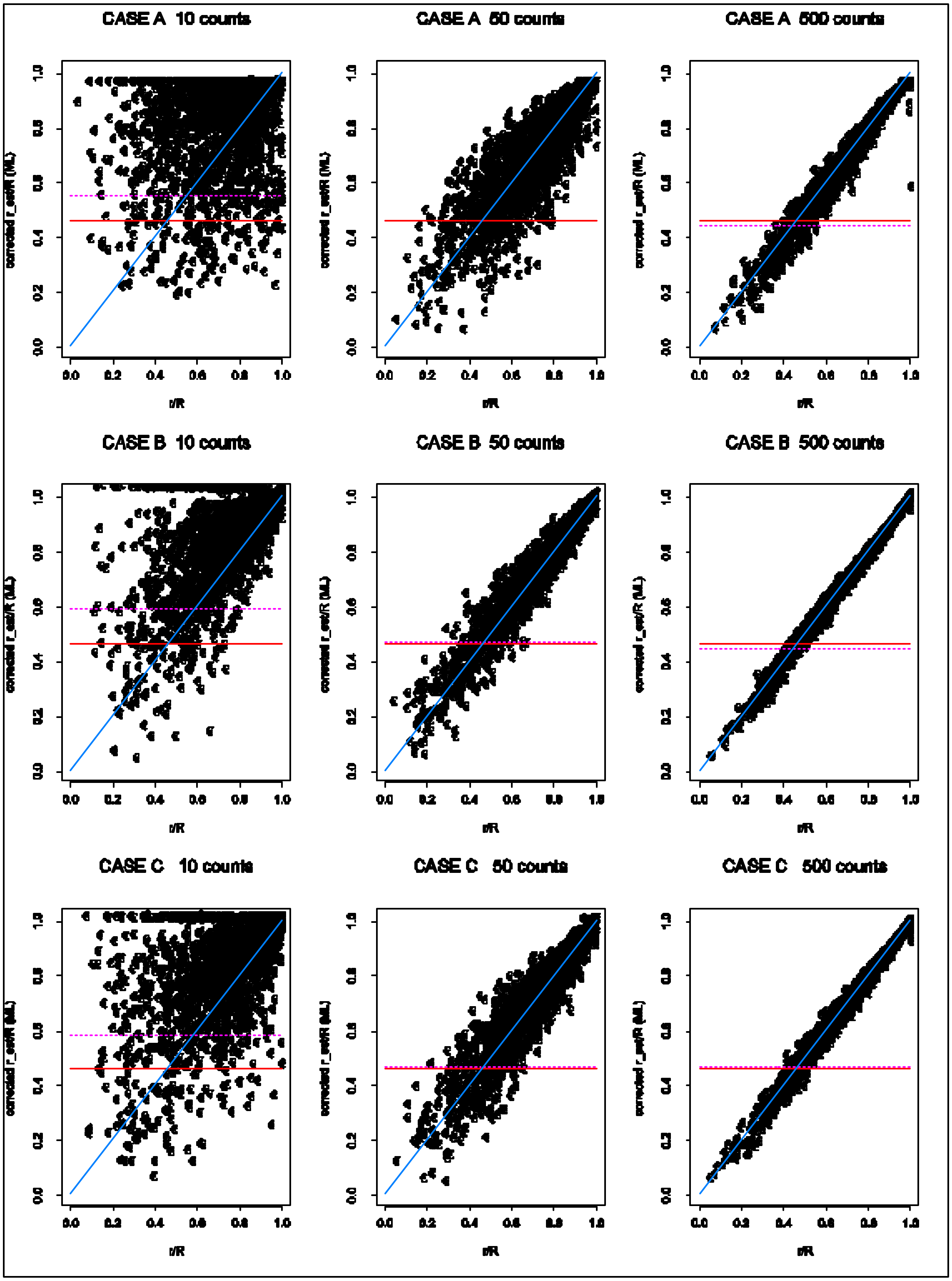


Figure 2. For three cases, we display the 0.1 quantile of the estimate (dashed horizontal line), $r_p / R = p^{1/3}$ where $p = 0.1$ (solid horizontal line), and the line of equality corresponding to a perfect estimate (diagonal line).

5.6 Functional Data Analysis Methods for Remote Sensing

Kevin Coakley and Jolene Splett
Statistical Engineering Division, ITL

David Walker
Electromagnetics Division, EEEL

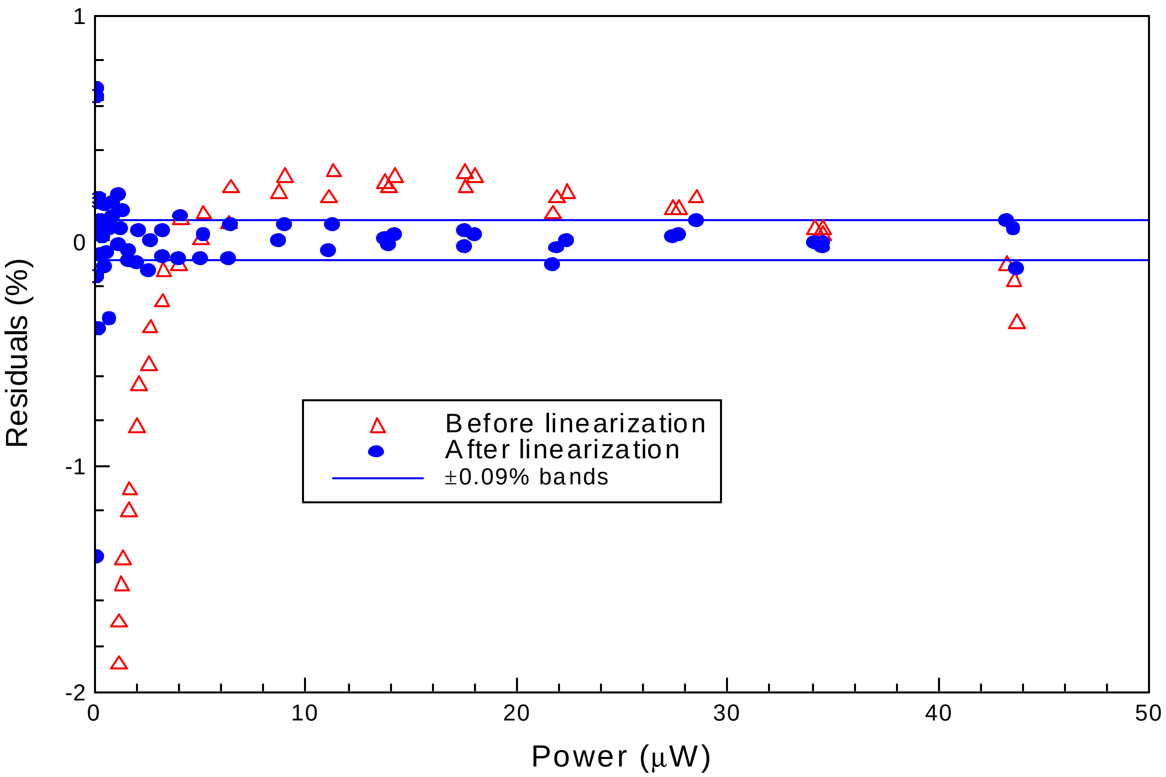


Figure 1. Residuals, $(P_U - \hat{a}P_L)/P_U$, before and after power meter linearization, where $\hat{a}$ was determined by least squares.

Microwave radiometers are critical components of satellite remote sensing measurement systems that provide information about the earth's oceans, land masses, atmosphere and near-space environment. In many microwave radiometers, the brightness temperature of a scene is inferred from the observed voltage of a tunnel diode detector. Typically, linear calibration models are used to estimate the power and hence the brightness temperature from the observed voltage of the tunnel diode. Since the voltage-power transfer function of a tunnel diode is nonlinear, this linear approach introduces systematic error. Thus, quantification of systematic calibration errors is critical in order to evaluate the accuracy of next generation remote sensing systems, including the Advanced Technology Microwave Sounder (ATMS) on the National Polar-orbiting Operational Environmental Satellite System (NPOESS). The NPOESS program is managed by the tri-agency Integrated Program Office (IPO), employing personnel from the Department of Commerce (DoC), Department of Defense (DoD) and the National Aeronautics and Space Administration (NASA).

Analysis is complicated by the fact that power meters are nonlinear. In collaboration with the Electromagnetics Division of NIST, SED staff developed functional data analysis methods to: (1) linearize power measurements; (2) model an observed voltage-power curve; and (3) quantify systematic calibration errors associated with using a linear calibration method. We anticipate that future efforts will focus on development of nonlinear calibration models so as to reduce systematic calibration errors.

In a calibration experiment, we measured many pairs of powers (P_L, P_U). Due to the additive bias and nonlinearities in the power meter, the observed ratio for the *i*th pair $P_U(i)/P_L(i)$ differed, on average, from the true power ratio $r_{true} (\approx 1.73)$. Based on measured power P_m (either P_L or P_U) we predicted true power P as

$$\hat{P} = f(P_m) = SP_0 \left(\frac{P_m}{P_0} - \Delta \right)^{g\left(\frac{P_m}{P_0} - \Delta\right)}$$

where

$$g(x) = a_0 + \sum_{k=1}^{5} a_k x^k \ .$$

The model parameters formed a seven dimensional vector $(\Delta, a_0, a_1, a_2, a_3, a_4, a_5)$; P_0 is a reference power and S is a scale factor. We set P_0 to the maximum measured power in a calibration data set. We modeled the additive bias as the product ΔP_0 where Δ is a dimensionless model parameter to be determined. We estimated the model parameters by minimizing

$$MSPE = \sum_i (\hat{r}(i) - r_{true})^2$$

where

$$\hat{r}(i) = \frac{f(P_U(i))}{f(P_L(i))} \;.$$

We determined S by requiring that the maximum predicted power, $f(P_0)$, equal $P_0(1-\Delta)$. Figure 1 displays fractional residuals before and after power meter linearization.

We modeled the measured voltage-power transfer function as a cubic B-spline where the number of knots was selected by cross-validation. Based on measurement of $V(P)$, our linear interpolation model to predict power P is

$$\hat{P} = P_1 + \frac{V(P) - V(P_1)}{V(P_2) - V(P_1)}(P_2 - P_1)\,,$$

where $V(P_1), V(P_2), P_1,$ and P_2 are measured in a calibration experiment. We simulated noise-free realizations of the voltage-power transfer function according to our regression spline model. For particular values of P we show the fractional systematic calibration error $(\hat{P} - P)/P$ in Figure 2. The dashed curve represents values of power where we took calibration data. The approximate ± 1 standard error bands were estimated using a nonparametric bootstrap resampling scheme.

We developed functional data analysis methods to linearize power meter measurements and estimate systematic calibration errors in next generation remote sensing systems that utilize tunnel diode detectors in microwave radiometers. The results of this study were documented in the paper, "Nonlinear Modeling of Tunnel Diode Detectors," and appeared in proceedings of the 2004 IEEE International Geoscience and Remote Sensing Symposium held September 20-24, 2004 in Anchorage, Alaska.

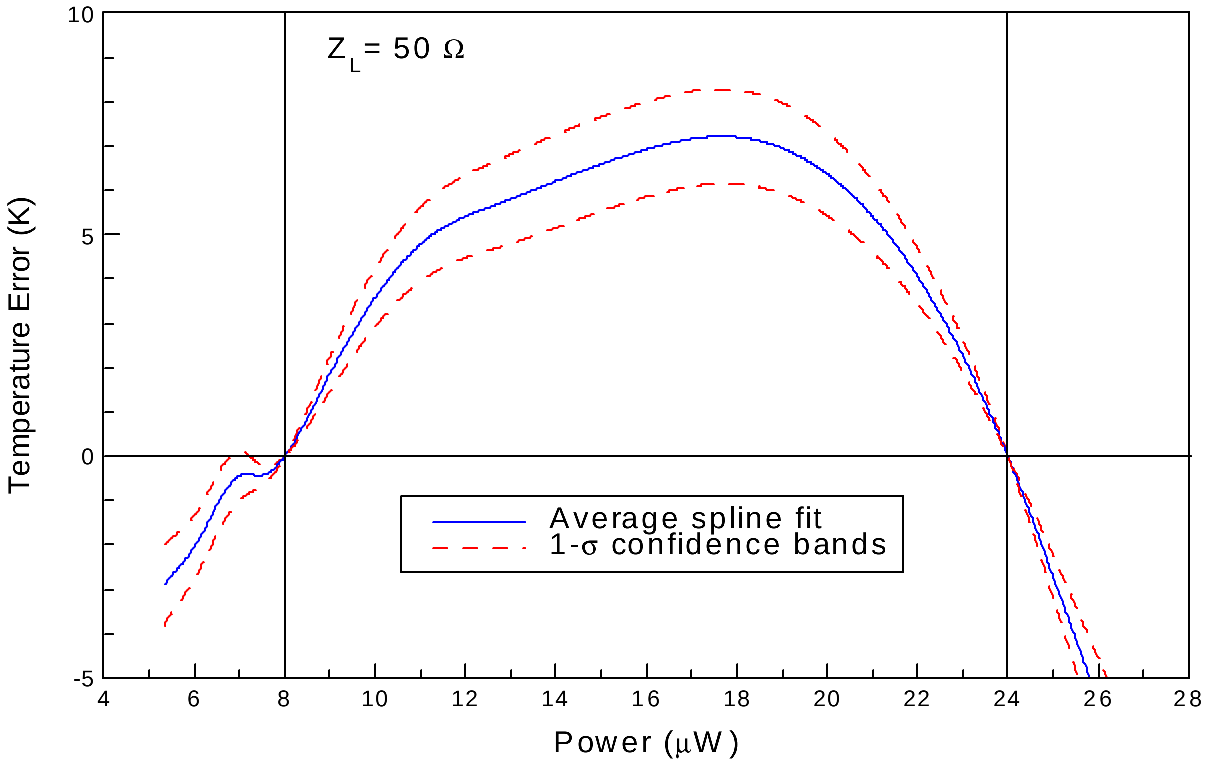


Figure 2. Estimate of systematic error associated with linear interpolation method based on spline model fit for voltage versus power curve.

5.7 Three Dimensional Chemical Imaging at the Nanoscale

Donald Malec, Juan Soto, Abderahman Cheniour
Statistical Engineering Division, ITL

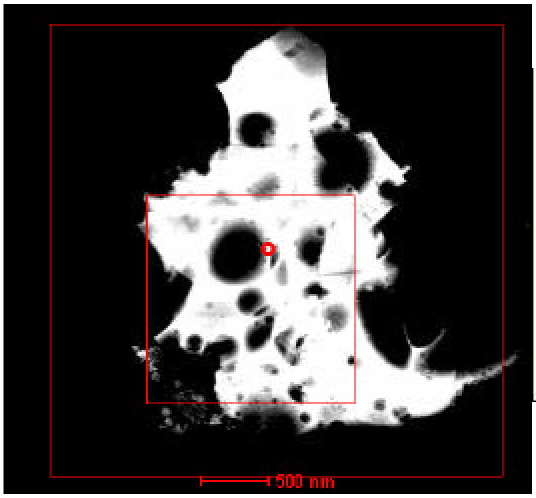
Eric Steel
CSTL

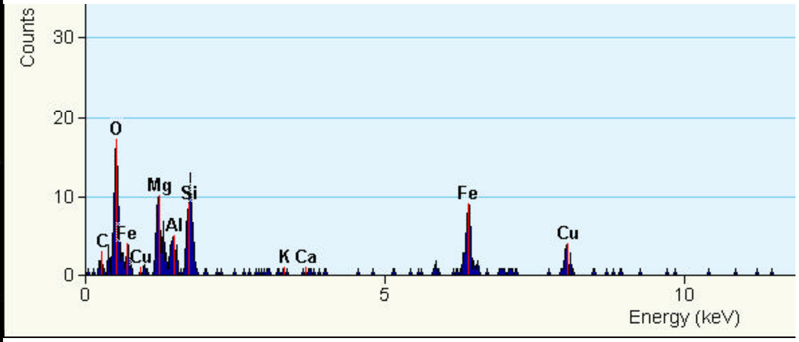
John Henry Scott
Surface and Microanalysis Science Division, CSTL

Zachary Levine
Electron and Optical Physics Division, PL

John Bonevich
Metallurgy Division, PTL

Judith Devaney, John G. Hagedorn, William L. George
Mathematical and Computational Sciences Division, ITL





First Figure: Dark-field STEM image of a sample with overlays showing the area where hyperspectral data were acquired (inner red square)

Second Figure: A picture/plot of a typical spectrum from the dataset (i.e., a random pixel's x-ray data.)

A quantitative understanding of the distribution of chemical species in three dimensions, including the internal structure and surfaces of micro- and nanoscale systems, is critical to the development of successful commercial products in nanotechnology. Current nanoscale-chemical 3D measurement tools are in their infancy and must overcome critical measurement barriers to be practical. This project will develop measurement approaches to attain three-dimensional chemical images at nanoscale resolution. These approaches will be broadly applicable to nanoscale technologies from microelectronics to pharmaceuticals and subcellular biomedical applications.

As one part of this wide-ranging project, the use of Bayesian approaches to determine structure and composition will be explored and developed. Potential areas of application include the use of Markov Random Field models of voxels for estimating 3D structure, Bayesian methods for picking likely 3D structures based on 2D images, and prior knowledge of composition (without structure). Spatial identification of atomic and molecular composition based on measurement of electron diffraction, X-ray spectra and electron energy-loss spectra may also be amenable to Bayesian methods. Lastly, adaptive design with the aim of obtaining accurate descriptions of the distribution of chemical species using a reduced number of measurements is a possibility. The Bayesian part of this project will build on basic 2D and 3D tomographic imaging techniques, knowledge of the dynamics of electron beams and their interactions with materials, multivariate statistical analysis of multiple spectra data, Bayesian decision theory and Bayesian computation.

It is widely known that image reconstruction techniques are computationally intensive. These techniques require considerable workspace and often force algorithm designers and software developers to improve their design and implementation. To address these matters, we have compiled strategies that should manage these issues. They include:

- a study of programming language constructs in Fortran 90 that may be exploited to improve upon the Fortran 77 implementation
- the use of profiling tools to identify the critical components (bottlenecks) in the algorithm; such information is useful in algorithm component redesign and lead to a more efficient implementation
- a study of the computer architecture for the target platform (a Sun Sparc 60 workstation with a single processor)
- a study of Fortran compiler options available on the target platform that lead to optimized binaries (executables)
- the use of timing tools to aid in assessing execution time and comparing implementations

Ongoing work is underway to test and evaluate the algorithm on sample data.

To achieve these aims, an understanding of the non-statistical tomographic methods of reconstruction that have been developed over the past twenty years has begun, with the aim of incorporating the salient features into a statistical framework. In parallel work, the development of Bayesian methods for ascertaining the unknown size and distribution of known 3D structures (e.g., cubes) given 2D projections has begun. Additional modeling to include identification of compounds and structure from spectra may be included later. The statistical aspects of this project are extensively linked to the subject of electron microscopy and cannot be easily abstracted as a "statistical problem". As such, extensive collaboration with the other team members is essential.

5.8 Statistical Data Mining Tools

Juan Soto, Z.Q. John Lu
Statistical Engineering Division, ITL

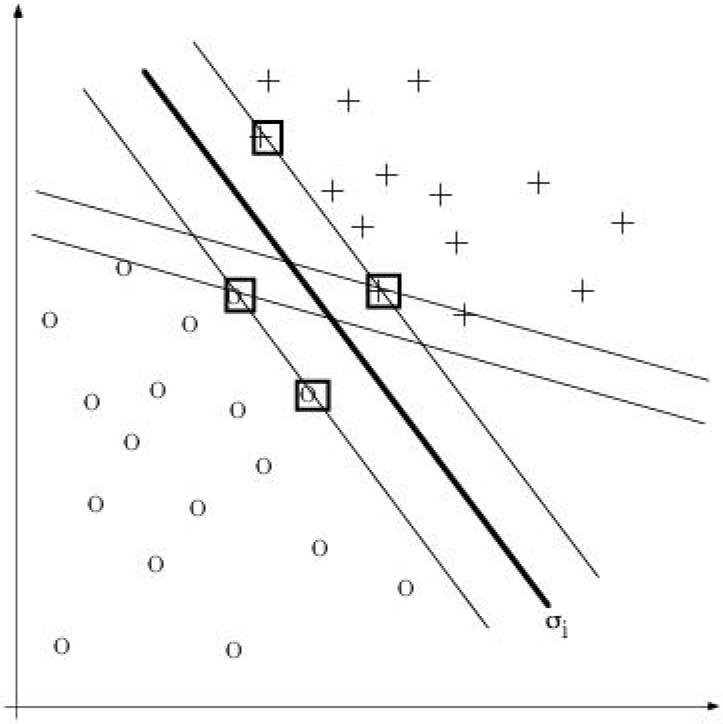


Figure: A graphic image depicting a support vector machine in a linear classification.

Collaborative research between members of the Statistical Engineering Division (SED) and members of the Process Measurements Division (Chemical Sciences and Technology Laboratory) has required that SED staff investigate various statistical tools for data mining. These tools include some very powerful statistical classification/prediction methods for high-dimensional data. This article briefly summarize this ongoing effort with the goal of bringing attention to a wide array of methods in a statistical toolkit that is already easily available to NIST scientists who may need them. Most of these functions have a user-friendly interface in the open source environment R and widely available commercial product S-plus.

The neural network (NN) function (in library *nnet*) fits a single hidden layer**,** possibly with skip layer connections. NN can be used for both classification and prediction problems. It utilizes a log-linear or a logistic model to solve classification problems and a nonlinear regression function to solve prediction problems. The implementation is based on a quasi-Newton optimizer. NN is fairly fast and robust in many problems.

The support vector machine (svm function in library libsvm) fits to the data a classifier function in terms of a linear combination of a positive definite kernel. Options for the choice of kernels include linear, polynomial, radial, or sigmoid. The special feature of support vector machine methodology is that its fit to data is done through a special fitting criterion. For classification problems, it controls classification errors while maximizing the class (margin) separation. Svm is primarily used for classification. Developments for probabilistic modeling and prediction are also possible, but there is less advantage over other approaches. Its speed is independent of input dimension, but is dependent on the number of training examples. Because of the limitations in the choice of kernels, svm is less flexible than NN, but there is rapid progress in this area in which capability and features of svm are continuously being improved and expanded.

advantage over other approaches. Its speed is independent of input dimension, but is

The K-nearest neighbor (KNN) method is a much older technique in statistics and pattern recognition literature. But it works very well in some problems and is much simpler in concept and implementation. It is based on nonparametric estimates of the posterior probabilities of a given class or label. It finds the k nearest samples in the training data and assigns a class label by majority rule, or estimates the posterior probabilities by the proportions of the classes among the k samples. For example, a Euclidean distance metric is used here though other similarity metrics may be used as well. KNN is mainly used for classification. For prediction, analogous methods exist, called kernel nonparametric regression. The accuracy of the KNN methods depends on identifying good "analogies" in the training data. The choice of k depends on the quality of data (level of noise) as well as the amount of data available at decision time. The computational speed is adversely affected by the dimension of the input data, as well as the size of the training data. However, these drawbacks may be overcome via preprocessing, e.g., dimension reduction or choosing subsets of the training data.

Additional data mining methods, which will likely be considered in the future by us, include recursive partitioning and tree methods, random forests, Bayesian methods, and bagging and boosting methods. Typically, these methods:

- are computationally intensive but are often robust and perform well in many noisy situations

- do not involve pre-specified functions, in contrast to NNs or SVMs, but the predictive function or classifier is very much data dependent; the model complexity is controlled by some model selection criteria as in the case of tree-type methods

- can be used for classification and probabilistic modeling, as well as prediction

- fairly fast on moderate-sized data sets, but may be adversely affected by the dimensionality of input predictors, and some dimensionality reduction or data preprocessing may be needed.

The hugely popular book, *Modern Applied Statistics with S*, 4th edition, by W. N. Venables and B. D. Ripley (2002) gives some detailed account of some of these methods. The NN functions are available in S+ under the "nnet" library. The SVM functions "libsvm" may be downloaded (http://www.stats.ox.ac.uk/pub/MASS04/Winlibs/libsvm.zip). The KNN functions are available in S+ under the "class" library. Tree based methods may be found in library "rpart." The analogous R functions may be found at http://www.r-project.org.

Researchers who are interested in utilizing these tools in conducting their own data mining project are encouraged to download and familiarize themselves with these tools. Alternatively, you may contact members of the SED to engage in collaborative research. It is important to note that these are a few among many tools developed specifically for data mining. The reader is cautioned that one should not be left with the impression that these are the only preferable techniques since no single technique works well on all problems. Depending on the nature of problems, one technique may work noticeably better than others. A good data miner always tries various techniques in order to find the appropriate technique for the given problem, and uses the right metrics for prediction performance evaluation.

6 Collaborative Research

The Statistical Engineering Division's collaboration with NIST colleagues focuses on NIST's highest priorities and initiatives, particularly in the emerging areas of science and technology with major agency-wide research in Nanotechnology, in BioSystems and Health and in Homeland Security. Work on these high priority projects spans all the Science Laboratories at NIST, and almost all of the Divisions in each Laboratory. While the goal of collaborative research is ensuring that NIST scientific results are based on a statistically solid foundation, the collaborative process also paves the way for an expanded role for statistics as smaller exploratory research projects develop into major research programs.

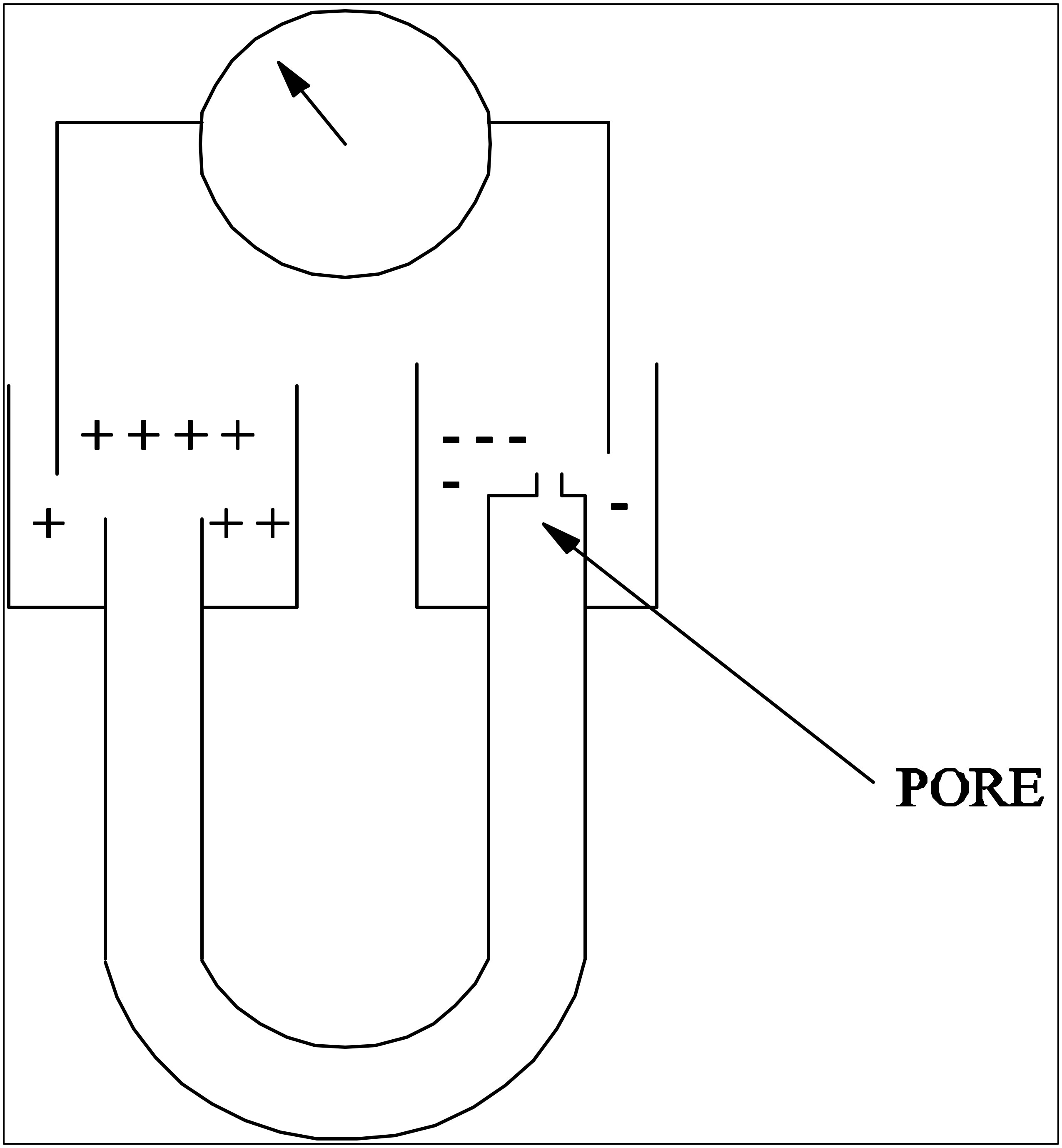
Decisions to create e-Tools for Statistics are often based on collaborative work that identifies metrologists' needs; illustrative web-based case studies and example data sets are drawn from specific joint projects that illuminate statistical principles, methods and modeling techniques.

Collaborative Research also underpins NIST's Core Mission activities and the delivery of measurement services to industrial customers and partners. The Standard Reference Materials program provides certified materials to be used in high-precision calibration and testing – certification of the attached uncertainty is the responsibility of the Statistical Engineering Division. Conventional and new statistical tools take the forms of experiment designs, data analyses, specialized algorithms and software.

6.1 Nanotechnology

6.1.1 First Entrance of DNA into a Nanopore

Charles Hagwood
Statistical Engineering Division, ITL



A horizontal bilayer apparatus used in nanopore translocation.

John Kasianowicz (NIST) and team discovered in 1996 that sequencing of DNA could in principle be done by observing the blockage of ions flowing through a nanopore as single-stranded DNA traverses the pore. The mechanism by which this is done is illustrated in the figure below. Two solutions, one of positive and a second of negative ions, are separated by a membrane. On this membrane, a small hole or pore is drilled, only large enough to allow one single strand of DNA passage at a time, thus initiating a flux of positive ions through the pore to the oppositely charged ions on the other side, and vice versa. Negatively charged DNA is driven through the channel, called translocation, and mostly blocks this flow of ions. The current blockades were found to be sensitive to the properties of the DNA, such as the size of its bases. The translocation process can be broken down into two parts: (1) the three-dimensional problem of the DNA finding the pore, and (2) the one-dimensional problem of the DNA treading the pore.

The solution of the three-dimensional problem of the polymer finding the pore is sought in our analysis. A proper model for this problem is the motion of a chain of particles (monomers) in a box. In polymer dynamics, there are several models for describing the interaction between monomers. We use the Zimm model. In the Zimm model, the motion of the polymer is described by a stochastic differential equation. Let $r_{1,}(t)\ldots$, $r_n(t)$ denote the positions, at time t, of the monomers in R^3. The equation of motion for the chain is given by

$$dr(t) = -k_{sp}HAr(t)dt + s\sqrt{H}\,dB(t)$$

where $r(t)=(r_1(t),\ldots, r_n(t))$, $B(t)=(B_1(t),\ldots,B_n(t))$, k_{sp} is a spring constant, H the hydrodynamic interaction matrix and A the connectivity matrix. The vector B(t) represents the diffusion part of the equation and its components $B_1(t),\ldots,B_n(t)$ are independent 3-dimensional Brownian motions.

The transition probability for the motion of DNA in a box with all sides reflecting, except one absorbing side on which the pore resides, is derived from the Fokker-Planck equation. From this, the probability that the DNA polymer is found at the entrance to the pore is computed. The first entrance time is also computed.

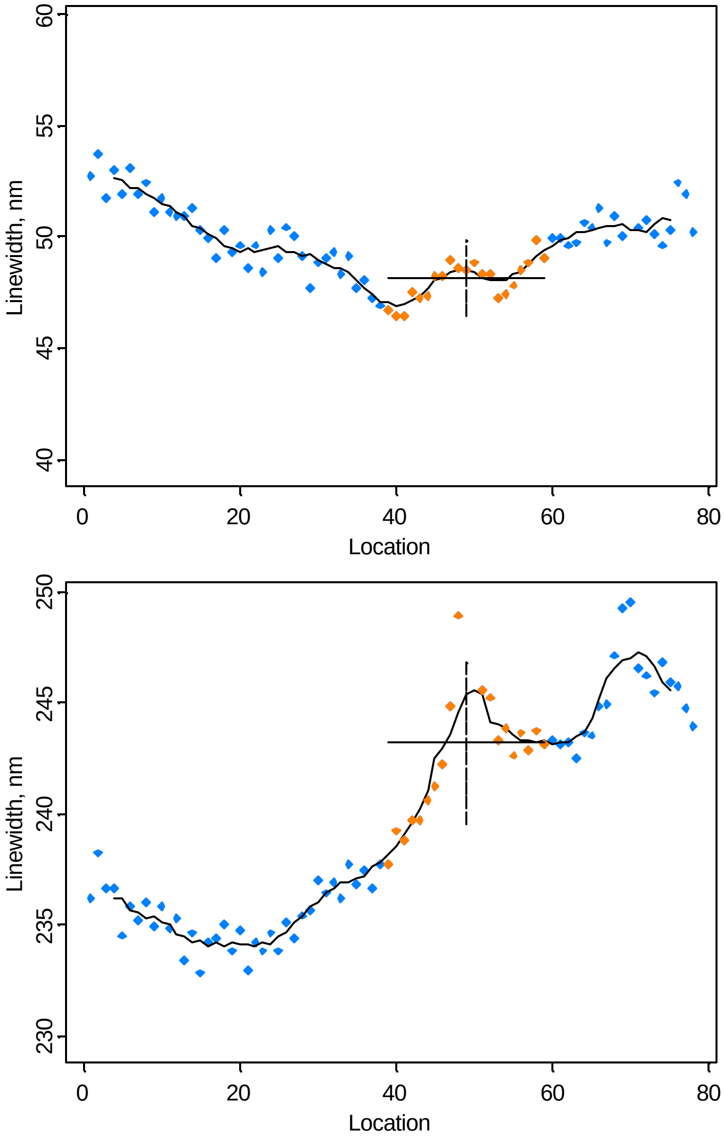
The fascinating work of John Kasianowicz promises to bring about new discoveries in cell biology, gene transfer and sequencing and properties of biopolymers. Working out the transport properties and dynamics of polymers in nanopores is extremely important to the growing fields of nanobiology, nanochemistry, and nanophysics.

6.1.2 Development of a Single-Crystal Reference Material for Calibration of Semiconductor Linewidth Measurements

William Guthrie
Statistical Engineering Division, ITL

Michael Cresswell and Richard Allen
Semiconductor Electronics Division, Electronics and Electrical Engineering Laboratory

Ronald Dixson
Precision Engineering Division, Manufacturing Engineering Laboratory



AFM data with associated moving average fits for two reference features on a chip used for AFM calibration. The data points shown in orange correspond to the HRTEM window that is of interest. The data points shown in blue, outside the window of interest, were only used to help estimate the uncertainty.

As the integrated circuits (IC's) used in computers and other electronic devices become more complex, the semiconductor industry requires increasingly accurate metrological tools for characterization of its manufacturing processes and products. The International Technology Roadmap for Semiconductors, an industry-led assessment of semiconductor technology requirements, notes a critical need to be able to manufacture IC's that have conducting lines with linewidths that are within 10 nm of specifications. To help meet these goals, staff of the Semiconductor Electronics Division, the Precision Engineering Division, and the Statistical Engineering Division have developed a prototype single-crystal critical-dimension reference material for calibrating metrology instruments used to measure linewidth. The reference material is a silicon chip mounted in a 200-mm silicon carrier-wafer. Six reference features on each chip range in size from about 50 nm to 200 nm, with typical expanded uncertainties of between ±1 nm and ±2 nm. The calibrated linewidths of the reference features are determined using atomic force microscopy (AFM) referenced to high-resolution transmission-electron microscope (HRTEM) images that reveal the cross-sectional counts of silicon lattice planes, whose spacing is traceable to the SI meter.

One of the interesting features of the data used to calibrate the AFM is the fact that the AFM makes measurements along the reference lines, while the HRTEM only measures the linewidth at a single point. This would not be a difficulty if the line were completely uniform or the location of the HRTEM slice through the sample were exactly known. However, some non-uniformity is currently unavoidable, even when using special etching techniques, and the location of the HRTEM slice can only be fixed within a 0.5 *m* m window. As a result, the AFM data must be averaged within the window of interest and an estimate of the uncertainty of the AFM linewidth that accounts for the random variation in the AFM data and the variability due to the non-uniformity of the line is also needed.

In order to estimate the AFM linewidth and its uncertainty, the first step was to smooth the data to describe the deterministic non-uniformity in the data and to reduce the effects of outliers in the data. AFM scans of two lines with the associated moving average fits are shown in the figure on the preceding page. The data points shown in orange correspond to the HRTEM window that is of interest. The data points shown in blue, outside the window of interest, were only used to help estimate the uncertainty. In order to reduce the effect of outliers that occasionally affected the measurement process due to surface contamination, the fitted values from the moving average, rather than the raw data values, were combined to obtain the AFM linewidth. The data were combined using a weighted average with weights inversely proportional to the distance of each point from the center of the HRTEM window. This type of weight function was chosen because the procedures used to obtain the HRTEM measurement were designed so that the slice of the reference feature that the HRTEM imaged is more likely to be in the center of the window than at an edge. The AFM linewidth is indicated in the plots of the AFM data by a horizontal line segment in the HRTEM window. The uncertainty of the AFM linewidth was obtained using bootstrap methods with an allowance for the reproducibility of AFM measurements when samples are mounted and remounted in the AFM. The AFM reproducibility was based on long-term experience with AFM data and was incorporated as a Type B standard uncertainty using the methods outlined in the ISO Guide to the Expression of Uncertainty in Measurement. The effect of the line's non-uniformity on the AFM linewidth, which was the largest source of uncertainty, was estimated by computing a bootstrap upper bound on the range of the non-uniformity within the HRTEM window and also converted to a Type B standard uncertainty. Because the estimates for each component of uncertainty were based on over 70 data points, the number of effective degrees of freedom for the combined standard uncertainty is assumed to be infinite. The expanded uncertainty of the AFM linewidth is shown in each plot as a vertical dashed line.

The top plot in the figure on the next page shows the calibration data relating the AFM linewidths to the HRTEM linewidths. Data from two chips, each with six reference features, were used for the calibration. After centering the data, a straight-line model was fit to the data using weighted least squares since the uncertainties of the AFM linewidths are not equal. The output from the fit of the model is given in the table below. The fact that the slope does not significantly differ from unity indicates that the scale of the AFM, which was previously calibrated with an independent standard, is on target and does not need to be corrected using the information from the single-crystal reference material, as was expected.

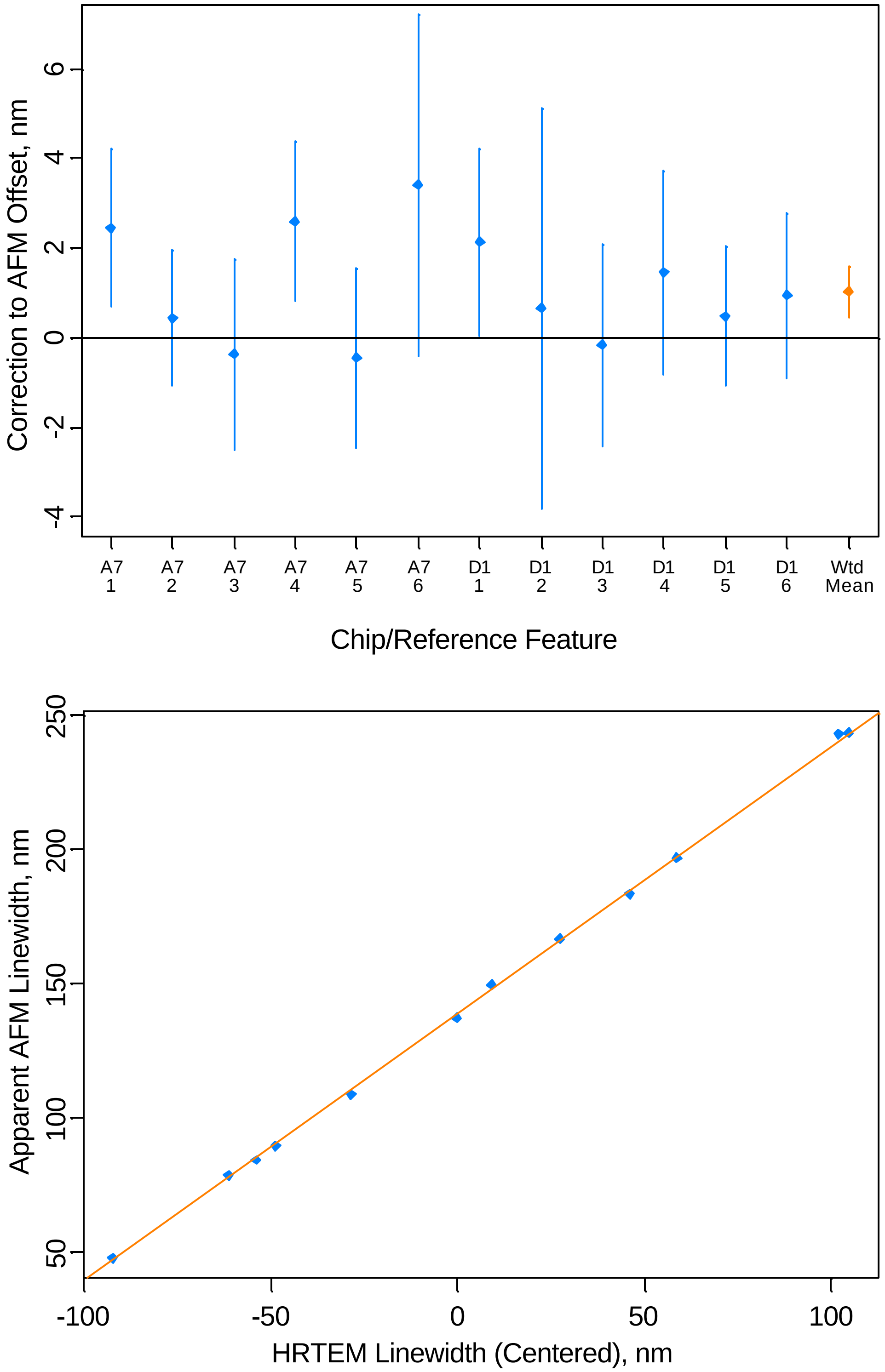
Output from the fit of a straight-line regression model to the calibration data.

```
                  Value   Std. Error    t value    Pr(>|t|)
Intercept      138.8825       0.3250   427.3745      0.0000
Slope            0.9960       0.0053   187.8507      0.0000

Residual standard error: 1.469 on 10 degrees of freedom
Multiple R-Squared: 0.9997
F-statistic: 35290 on 1 and 10 degrees of freedom, the p-value is 0

Correlation of Coefficients:
              Intercept     Slope
Intercept         1           0
Slope             0           1
```

The standard previously used to calibrate the offset of the AFM is not as precise as the scale calibrant, however, so an offset correction was expected. The offset correction was estimated using a weighted average of individual offset corrections computed for each reference feature. This approach allows for explicit inclusion of the uncertainty in the HRTEM linewidths, although the uncertainty from the HRTEM linewidths is believed to be negligible relative to the combined uncertainty of the apparent AFM linewidths. The estimate of the correction to the AFM offset is 1.03 nm ± 0.58 nm (expanded uncertainty). The individually-estimated offsets and the mean offset are shown in the bottom plot in the figure on the next page. The error bars on each estimate of the offset are expanded uncertainties.

With the offset of the AFM established, the linewidths of reference features on chips manufactured at the same time as the calibration chips and measured under the same conditions were then estimated. The effect of any non-uniformity of the lines on the chips to be distributed is reduced, however, because the center of each reference feature, the location selected for estimation of the linewidth, does not need to be as large as the HRTEM window.

Reference materials like these will improve metrology for semiconductor manufacturing by reducing the disagreement between metrology tools within a company and between results from different laboratories. These materials will also provide an absolute assessment of linewidth that is critical for controlling circuit properties. A workshop to obtain industry feedback on these prototype reference materials is planned for February 2005.



A straight-line calibration curve fit to the AFM/HRTEM data (top), and individual and mean corrections to the existing AFM offset (bottom).

6.1.3 Statistical Methods for Microanalysis and Nuclear Forensics

K. J. Coakley and A. M. Leifer
Statistical Engineering Division, ITL

D. S. Simons
Surface and Microanalysis Division, CSTL

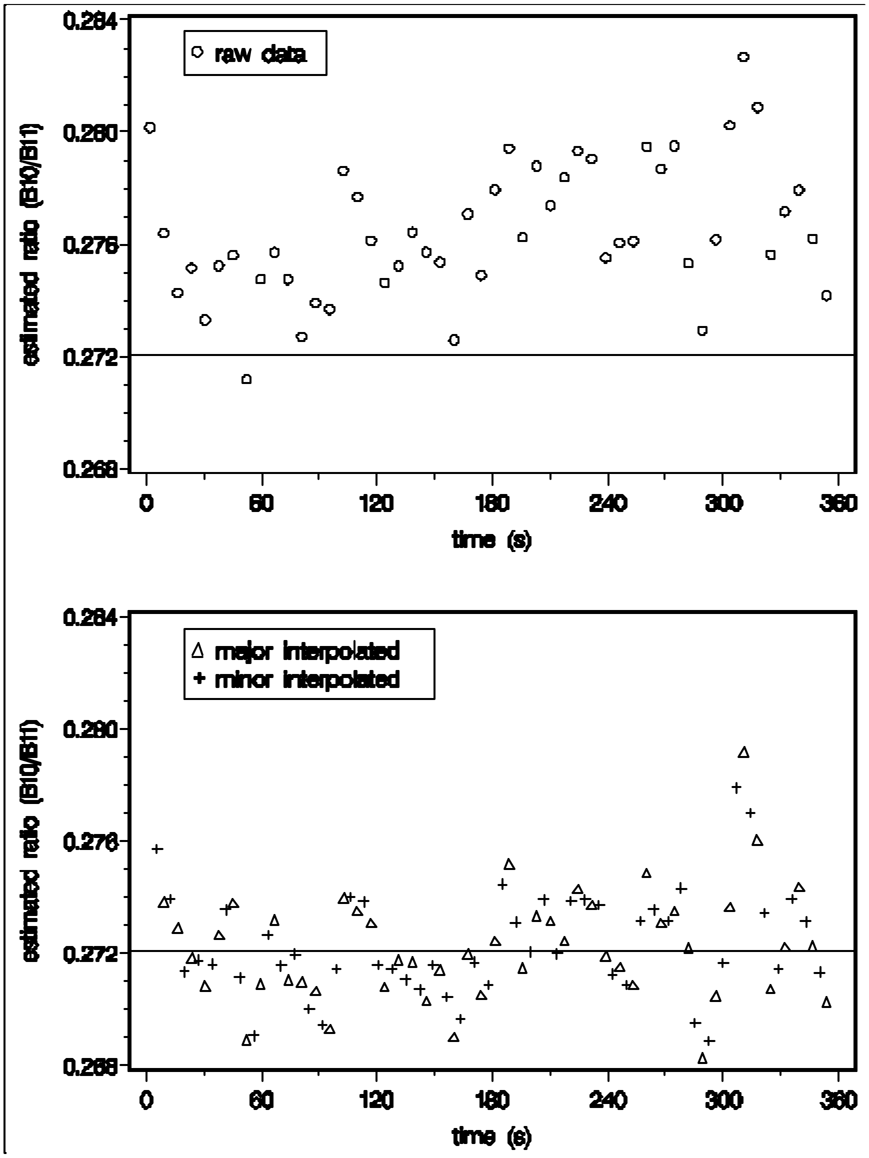


Figure 1. In a Secondary Ion Mass Spectrometry measurement, isotopes of boron are measured sequentially. Instrumental drift introduces systematic error into the estimate of the ratio of B10 and B11 isotopes. Top: Estimates of ratio of B10 to B11 isotopes computed from raw data. Bottom: Estimates of ratio of B10 to B11 isotopes computed from interpolation schemes which correct for drift. We plot the average of the interpolation method estimates of the isotopic ratio as a reference line.

Secondary ion mass spectrometry (SIMS) is a specialized analytical method that can be used to perform localized isotopic ratio measurements on a micrometer scale. Such measurements have broad applicability in areas of geology, astronomy, and biology. A specific application area of recent interest is nuclear forensics, whereby SIMS has been applied to the search for evidence of uranium enrichment activities through the measurements of the relative abundances of U-235 and U-238 in micrometer-sized particles. We developed statistical methods to correct SIMS measurements for the effects of instrumental drift.

In SIMS measurement systems, the count rate of isotopes may vary in time as a particle is consumed during the analysis. Since only one isotope at a time is measured in conventional ion counting systems, this drift can introduce systematic error into the estimate of the ratio of any two isotopes. Hence, correcting the SIMS instrument for drift is critical to the accurate determination of isotopic ratios and their associated random uncertainties.

In each of two interpolation schemes, we align one isotope time series with respect to the other. In the major-minor interpolation scheme, the minor time series is fixed and the major time series is interpolated. In the minor-major interpolation scheme, the major time series is fixed and the minor time series is interpolated. In simulation studies, we show that the average of the estimates obtained from both interpolation schemes is superior to the estimate computed either from the unaligned data or from a single interpolation scheme for the case where the count rate varies in time.

We present a formula for the approximate standard deviation of the isotopic ratio estimated from the aligned data. Our formula accounts for the effect of interpolation on the variability of the data. We also present an approximate hypothesis test procedure to detect and quantify possible systematic temporal variation in the isotopic ratio time series data.

Publication and Presentations

K. J. Coakley, D. S. Simons, and A. M. Leifer, “Secondary ion mass spectrometry measurements of isotopic ratios: correction for time varying count rate,” The International Journal of Mass Spectrometry 240 (2005) pp. 107-120.

D. S. Simons (CSTL), K. J. Coakley, A. M. Leifer, “Application of Time Interpolation to SIMS Isotopic Ratio Measurements,” 17th Annual SIMS Workshop, Westminster, CO, May 19, 2004 and American Vacuum Society, 51st International Symposium, Anaheim, CA, November 15, 2004.

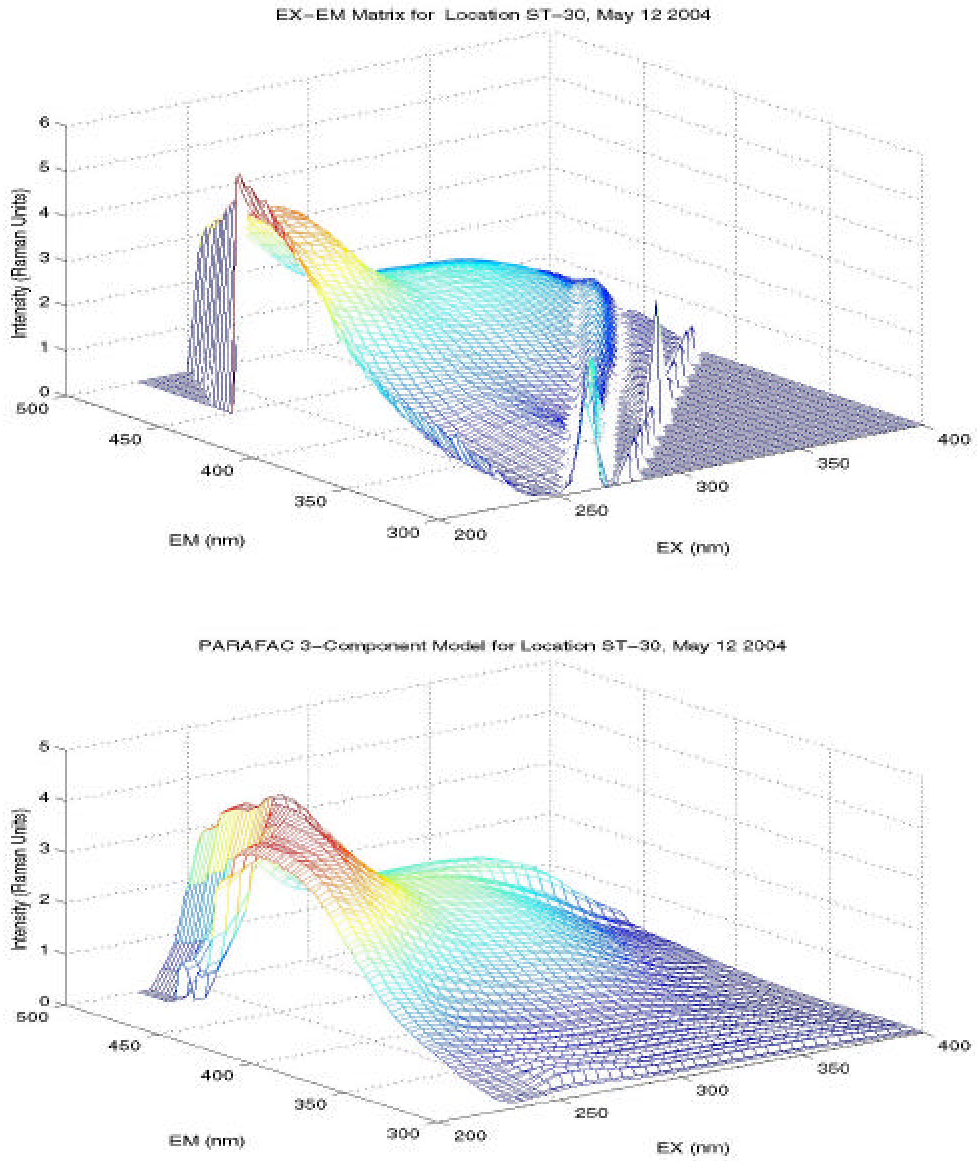
Our statistical methods for drift correction and quantification of systematic and random uncertainty in SIMS measurements should have broad applicability in geology, astronomy, biology, and nuclear forensics.

6.2 BioSystems & Healthcare

6.2.1 Excitation-Emission Matrix (EEM) Fluorescence Spectroscopy for Water Purity

James Yen, Stefan Leigh, Alan Heckert, Andrew Rukhin
Statistical Engineering Division, ITL

David Holbrook
Surface and Microanalysis Science Division, CSTL



Depicted are the raw data (top) and the estimated PARAFAC model (bottom) for an EEM data taken from a watershed in the Mid-Atlantic region.

Naturally occurring bodies of water, such as are used for bathing, drinking, and irrigation, are known to contain agglomerated microscopic particles of dissolved organic molecules (DOMs, or humics). Although these complex mixtures of organic polymers are known to play an influential role in aquatic ecosystems, the biology, chemistry, speciation, transport, and toxicity of such systems, the details of their dispersion, their interactions, and their propagation remain poorly understood. The humics of immediate interest frequently incorporate residues of pesticides, herbicides, and other hydrocarbons. Due to their complexity, the exact chemical characterization of DOMs would be an arduous process, involving large volume sampling and many stages of analysis.

Excitation-Emission Matrix (EEM) Fluorescence Spectroscopy is a simple technique for measuring and monitoring the heterogeneous composition of organic material contained in aqueous samples. The method entails collection of the emission spectrum at multiple excitation wavelengths, producing a topographic or contour map where the fluorescence peaks can be used to help identify the constituents of the sample. Advantages of the technique include small sample volume, minimal sample preparation, and quick analysis time. EEM devices are operationally robust and could be stationed in environmentally sensitive bodies of water to provide a remote monitoring capability.

Initial SED work focuses on two sets of issues: (1) how to determine if two EEM matrices are essentially different from one another, both in terms of qualitative and quantitative information that they convey, and (2) automation of the task of determining where and how two matrices considered not identical do in fact differ.

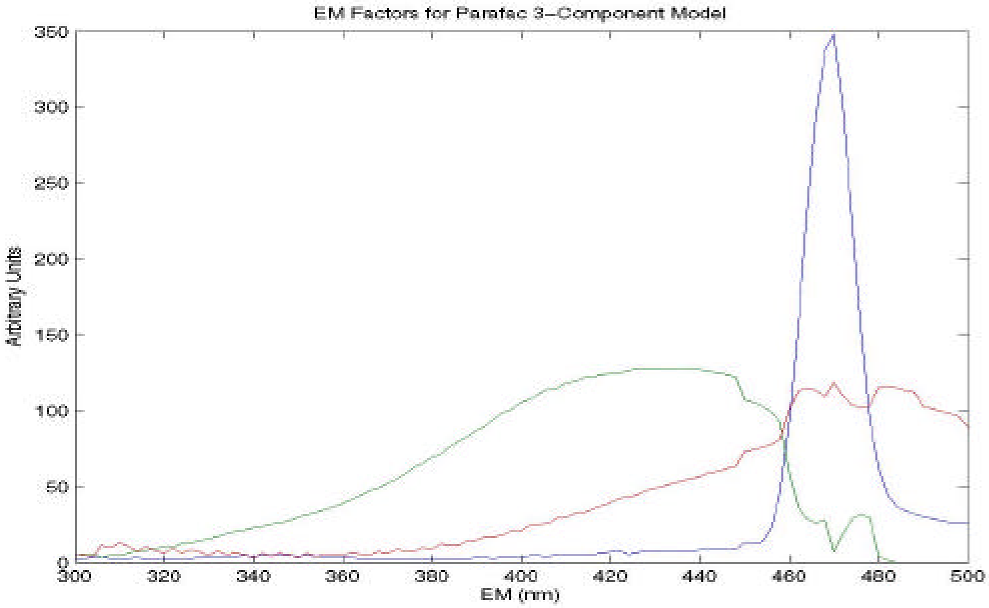
A first simple tool for assessing differences between successive EM topographical snapshots is a graphic that plots correlations or comovements between complete Excitation spectra against the corresponding Emission Spectrum wavelengths. It has the virtue of representing 3D contour differences in the form of an easily interpretable 2D graphic. For a more formal test of differences, at each individual wavelength, perform a hypothesis test that the correlation (or comovement) of the corresponding data from different EEM matrices falls into some pre-fixed range of interest. The spectrum of resulting p-values can then be combined into a single test of equivalence using any number of standard combining techniques, such as Fisher's method. An application of this approach to calibration of EEM against representative samples from the International Humic Substance Society samples is currently being documented for submission to the Journal of Environmental Quality.

Previous work in DOM analysis has partitioned the spectral matrix into regions that contain groupings of molecules such as aromatic proteins or humic-like organics. More elaborate approaches to the analysis of EEM data for environmental application have only just begun to appear in the literature. Stedman et al (Marine Chemistry, 2003) make use of parallel factor analysis (PARAFAC), documented by Bro (Chemometrics and Intelligent Laboratory Systems, 1999), who also provides a free PARAFAC tool in Matlab (Anderson and Bro, Chemometrics and Intelligent Laboratory Systems, 2000). We illustrate PARAFAC on water sample data taken from a watershed in the mid-Atlantic region. We do a 4-way analysis with the 4 dimensions being Emission wavelength (EM), Excitation wavelength (EX), Location, and Date. We will look at samples taken on 5 different dates (coded as 1=May 5, 2=May 12, 3=May 21, 4=May 28, 5=June 2) at 4 different locations at the watershed. These data are modeled with a PARAFAC 3-Component model. The top figure on the previous page is a perspective plot of the EEM matrix at one location (ST-30, coded as Location 3) on May

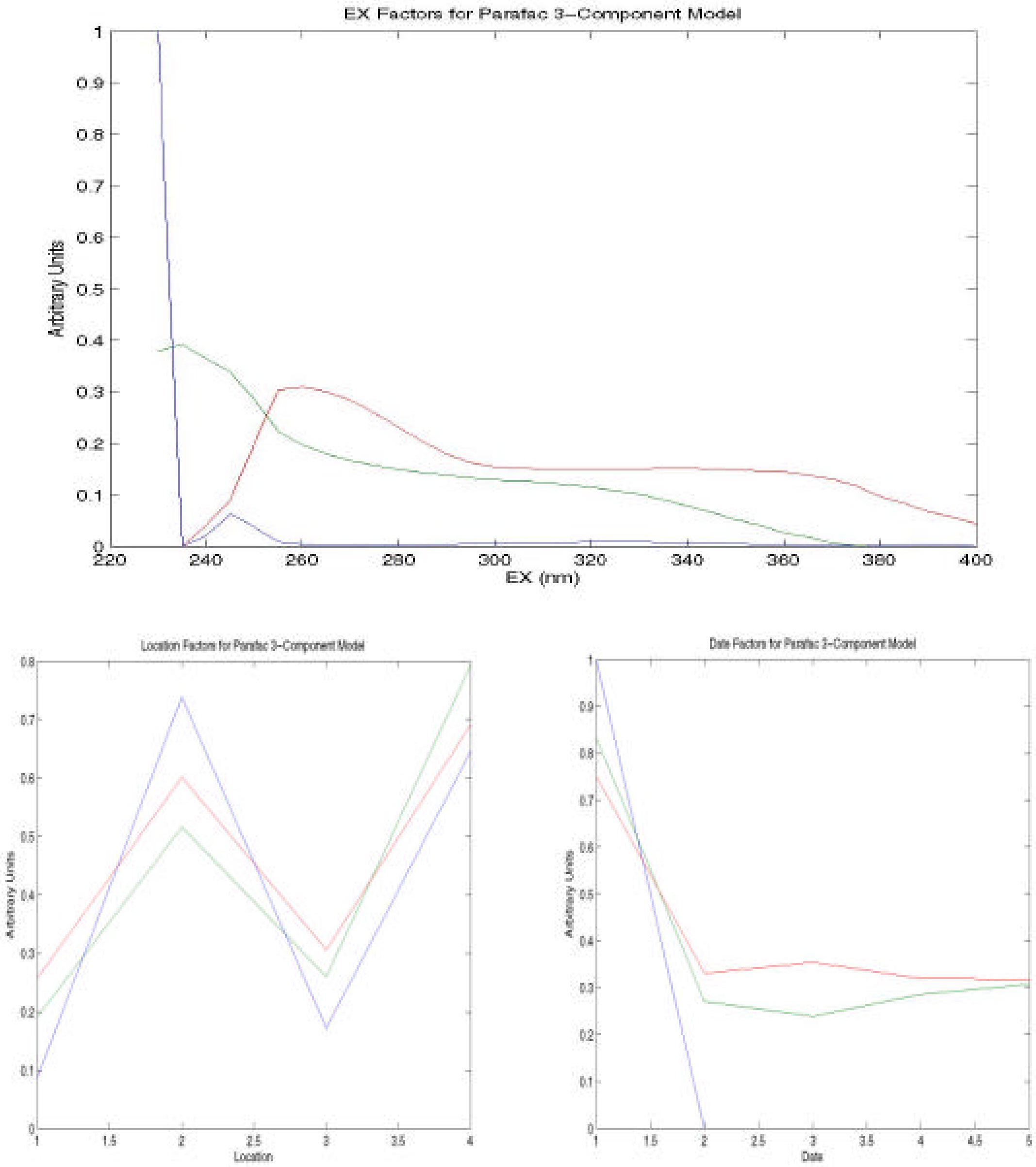
12. The figure below that is a perspective plot of the PARAFAC 3-Component model of this matrix. The PARAFAC model captures the main features of the data, except that it seems to round the main hump too much and it does not capture the diagonal ridges near the opposite corner. Such diagonal elements violate the conditions of the quadrilinear model used by first-order PARAFAC. The ridges are probably scattering noise that has not been fully accounted for, rather than being indicative of the underlying chemical composition.

The number of components is chosen beforehand; ideally, the components will correspond to different chemicals with distinctive spectral signals. However, correlated components may lead to convergence problems, necessitating a choice of fewer components. For each of the four factors (EM, EX, Date, and Location) in the watershed data, we show below a graph depicting the loadings of the 3 estimated components. Note that the factor loading units are all on an arbitrary scale because the factor loadings of a component are only meaningful relative to the loadings of the other components for that factor.

Each estimated component has the same color in all 4 factor graphs. All 3 components tend to be concentrated at small EX and large EM, which agrees with the perspective plots. However, the sharpest peaks are from the blue component, which is concentrated on the low edge of the EX range. This blue component is high at Date 1 (May 5, 2004) and basically disappears for the other dates. This corresponds to the data from May 5 having a huge spike there that is not present on other days. That may indicate a contamination or, more likely, incomplete suppression of scattering noise. Thus, PARAFAC can be a useful tool for data cleaning as well as for data mining. The green and red components combine to form the dominant features of pictured EEM matrices; these components tend to be at a higher level for Locations 2 and 4, and at a lower level for Locations 1 and 3. That they are at a higher level for Date 1 may be an artifact of the huge peak at Date 1. Scientists can determine if the red and green components correspond to particular chemical compounds, and how important is the difference in locations. A series of articles documenting the analysis of water data using PARAFAC and other methods is planned.



Depicted are the estimated EM factor loadings for the 3 estimated components of the PARAFAC model.



Depicted above are the estimated EX, Location, and Date factor loadings for the 3 components of the PARAFAC model.

Dissolved organic matters (DOMs) are known to be influential actors in aquatic ecosystems, but their compositions and short-term and long-term effects on water sources are only beginning to be appreciated and understood. The complexity of DOMs stands as a major roadblock to fully understanding their form and function. EEM holds the promise of being able to construct a meaningful taxonomy and catalogue of DOMS as an important first step in appreciating and counteracting their potentially deleterious effects.

6.2.2 Stress-Strain Behavior of Rat Pulmonary Arteries

Jolene Splett and Dom Vecchia
Statistical Engineering Division, ITL

Liz Drexler
Materials Reliability Division, MSEL

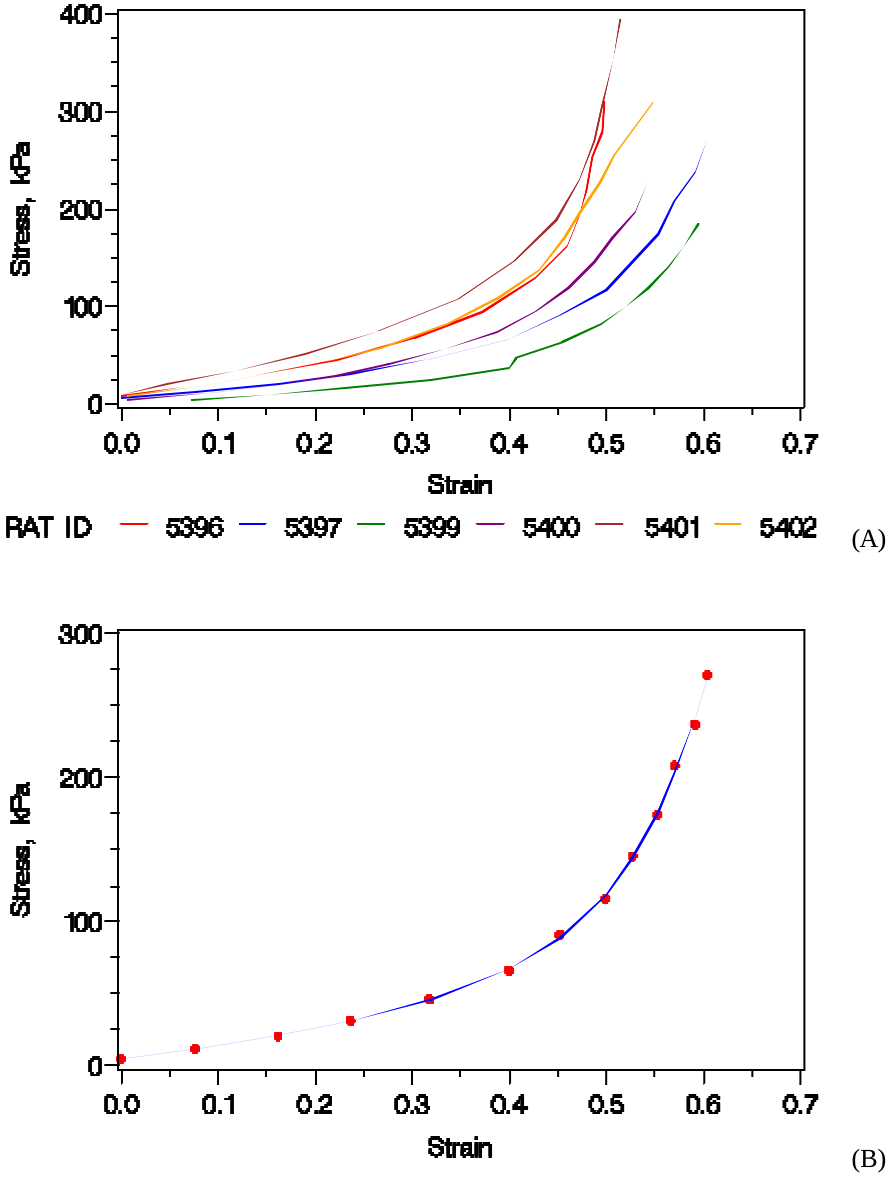


Figure 1. (A) Stress versus strain for six normotensive rats. The samples were taken from the right ventricle and were measured in the longitudinal direction. (B) Stress versus strain for rat number 5397 with fitted hyperbola.

It is well known in the medical community that pulmonary arteries "remodel," or stiffen, in the presence of pulmonary hypertension, but detailed biomechanical studies of the proximal arteries are lacking. Characterizing the biomechanical features of the proximal arteries under normal and hypertensive conditions is an important step in understanding pulmonary vascular dynamics.

Rats were chosen for the initial study because they can be bred to be genetically identical, thus minimizing variation among rats and increasing the probability that observed differences are due to the disease. A bubble inflation technique is used to compare the stress-strain behavior of diseased versus healthy proximal pulmonary arteries in rats. Each test on a single tissue sample provides a stress-strain curve based on increasing pressures.

Limited tests were conducted to quantify the difference in mechanical properties of normotensive and hypertensive rat pulmonary arteries. Stress-strain measurements were obtained for comparison of the properties at multiple orientations of the trunk, right, and left pulmonary arteries from normal (control) and treated rats.

We are currently examining models to describe the length versus pressure relationship that can be used to predict the initial length (or diameter) of the bubble before inflation. The true length is difficult to measure because of small wrinkles and bumps in the surface of the tissue. The length before inflation is crucial to subsequent calculations of both stress and strain.

Assuming the initial length is correct, preliminary analyses indicate that a hyperbolic model fits the stress-strain data quite well. The form of the model is chosen such that one of the parameters can be associated with the onset of strain stiffening of the arterial material with increasing load. Estimated values of this parameter may be useful in analyses for various comparisons between normotensive and hypertensive experimental rats.

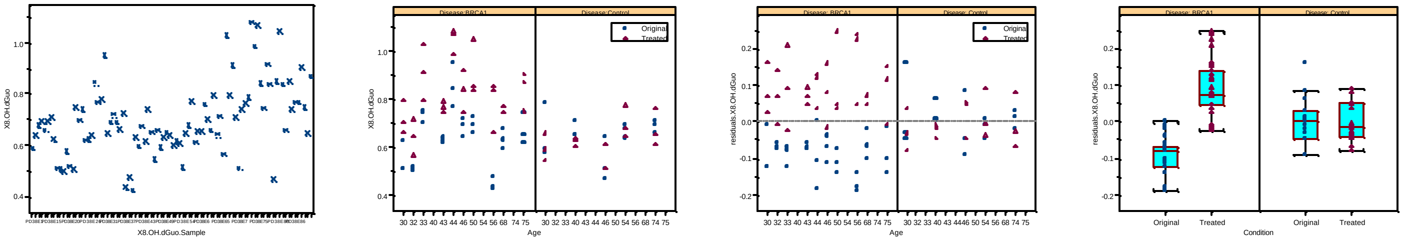
Eventually a constitutive model that fully describes the viscoelastic non-linear characteristics of arterial tissue will be developed. Understanding biomechanical properties of proximal pulmonary arteries will facilitate the development of novel diagnostic techniques with respect to pediatric pulmonary hypertension.

6.2.3 Displaying the Effect of Radiation on Cancer Prone Cells

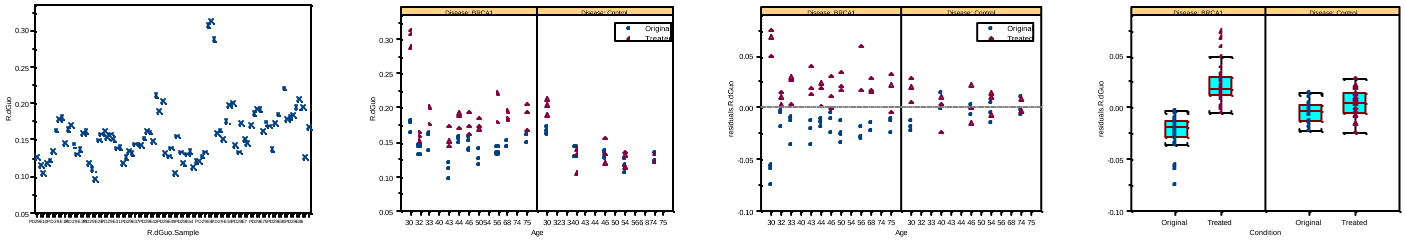
Dennis Leber
Statistical Engineering Division, ITL

Miral Dizdar, Henry Rodriguez
Biotechnology Division, CSTL

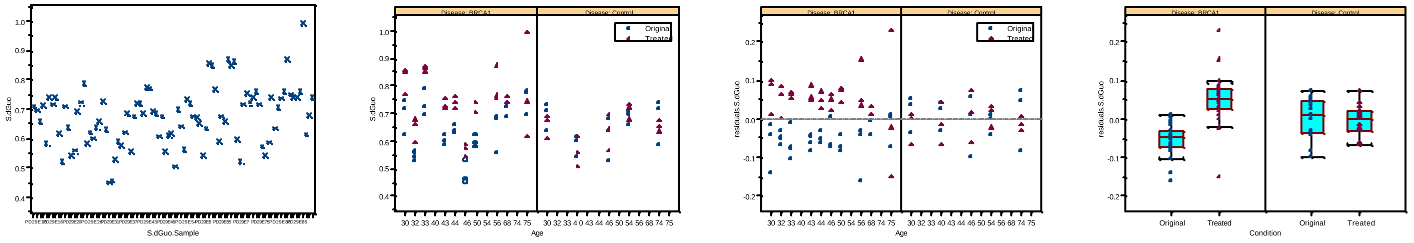
8-OH-dGuo Measurements



R-dGuo Measurements



S-dGuo Measurements



Measurements taken during an experiment to assess the effect of radiation on cancer prone cells. From left to right in each measurement type: 1.) Raw data are observed; 2.) Raw data separated into BRCA1 cells (left panel) and control cells (right panel) plotted by age and color-coded by radiation exposure (red); 3.) Residuals (age effect removed) separated into BRCA1 cells (left panel) and control cells (right panel) plotted by age and color-coded by radiation exposure (red); 4.) Residuals (age effect removed) separated into BRCA1 cells (left panel) and control cells (right panel) plotted by and color-coded by radiation exposure (red).

In ongoing efforts in the scientific and medical communities to understand, treat, prevent and potentially find a cure for cancer, Miral Dizdar and Henry Rodriguez of the DNA Technologies Group at NIST have examined the effect of radiation on cells with a genetic structure, BRCA1, known to potentially lead to breast cancer in women at an early age.

Samples of cells from individuals of various ages containing the genetic makeup BRCA1 were examined in this study, as well as cells from a control group. The cells from each individual within each group were divided into six subsamples. Since these subsamples were taken from cells from the same individual, it is assumed that the subsamples are identical from the genetic perspective.

As normal, healthy cells from the human body are damaged, human nature repairs the damage done. In this study, several genetic measurements, 8-OH-dGuo, R-dGuo, and S-dGuo, were utilized to assess damages and repair of the cells. Measurements were taken on three of the subsamples for each individual in the study. The remaining three subsamples from each individual were subjected to a dose of radiation, allowed sufficient time to repair itself, and then measurements were taken.

At first view, meaningful relationship in the data is hard to depict as seen in the leftmost graphs above for each measurement type. A first step to uncover structure in the data was to graphically view the data according to the group with which it is associated, BRCA1 or Control, as illustrated by the left and right panels, respectively, of the second graphs above for each measurement type. Additionally, the data are plotted by Age and color-coded to display weather or not the subsample was exposed to radiation.

Since individuals vary greatly, it is no surprise that the data show great variation from one individual, or age, to another within each group. Since this variation is of no interest to the study, an Analysis of Variance (ANOVA) model was used to detect and remove the effects due to the differences in individuals. The signal remaining in the data after removing these effects is now due to remaining factors such as the effect due to exposure to radiation and the effect of the BRCA1 genetic makeup. Using the same graphical separation of groups and color-coding of radiation exposure as previously used, the residual data are plotted and shown in the third graphs above for each measurement type. The graphs of the residual data clearly display the effect the radiation had on the BRCA1 cells, but did not have on the Control group cells. Since the effect due to individual has been removed, it would now be appropriate to combine the data across age. The result is displayed in the final graphs above for each measurement type. It can now easily be seen that the measurements of the cells exposed to radiation for the BRCA1 cells (left panel) are significantly higher than those not exposed to radiation. The cells from the control group (right panel) do not display this difference. Quantitative ANOVA models verify the significant difference just observed.

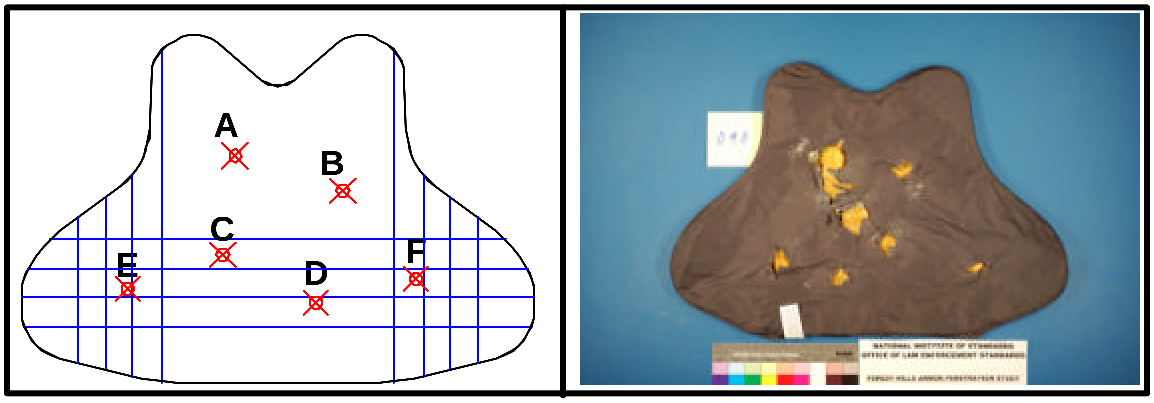
Although very early in the research, these findings may lead to alternative treatment plans for individuals with the BRCA1 genetic makeup that do not include exposure to radiation. In fact, since it is seen that cells containing the BRCA1 genetic makeup do not completely repair damages done as a result of exposure to radiation, special precautions to avoid radiation altogether may be in order for individuals with this genetic makeup.

6.3 Homeland Security

6.3.1 Experimental Design for Body Armor Failure Investigation

Dennis Leber
Statistical Engineering Division, ITL

Kirk Rice, Michael Riley
Office of Law Enforcement Standards, EEEL



Shot zones/locations on a ballistic panel

Ballistic panel after testing

Shot zone/location, as defined by the above graphic, was identified as a factor of interest in the body armor failure study. The violent impact of the bullet with the body armor produces the drastic deformations shown.

On June 23, 2003, an officer from a municipal police department in a suburb east of Pittsburgh, working undercover as a member of the state attorney general's Drug Enforcement Task Force, was shot in the abdomen while attempting to arrest a drug suspect. The officer was wearing ballistic-resistant body armor; the bullet completely penetrated the six-month-old armor and lodged near the officer's kidney. The officer survived the incident. Two additional armor penetrating incidents have been reported, one fatal. In all three cases, the armor was of the same model from the same leading body armor manufacturer. The material used, Zylon, has only been in use in the manufacturing of ballistic-resistant body armor since 1998.

NIST's Office of Law Enforcement Standards, sponsored by the National Institute of Justice (NIJ), in conjunction with SED and other key NIST divisions are investigating the June 23rd armor failure. Additionally, in response to US Attorney General Ashcroft's Bulletproof Vest Safety Initiative, the research group is examining all Zylon-based bullet resistant vests (both new and used), as well as the upgrade kits offered by manufacturers to retrofit the Zylon-based vests. Furthermore, this research will be expanded in the future to include all materials used to manufacture body armor as well as the test methodologies used to ensure their reliability.

In an attempt to determine the cause for the June 23rd body armor failure, the focus of the experimental testing was the central question of:

> Why was the vest able to pass the NIJ standard testing, but failed on the street?

Experimental Factors

To explore the question at hand, major differences between the NIJ standard testing in the laboratory and the incident on the street were explored. Extensive characterization testing was done on unfired ammunition recovered at the crime scene, similar ammunition purchased in the area the crime took place, the recovered crime weapon, the failed vest, and vests identical in model to the failed vest. Noteworthy results include a crime scene bullet velocity slightly lower than that used in the NIJ certification standard and vest fabric strength significantly lower than when certified. The differences (factors) of interest, and appropriate levels for these factors chosen to be included in the study were as follows:

Experimental Factor	**NIJ Standard Laboratory**	**Forest Hills Street**
Material Condition	New	Weakened
Bullet Type	Remington	Magtech
Bullet Velocity	1055 ft/sec	975 ft/sec
Barrel/Twist	SAAMI	Hi Point
Angle of Incidence	0º	45º

In addition to the above five factors, a sixth factor, Vest Zone, was included in the experimental design. Vest Zone is a three-level factor that refers to the areas of the vest that have:

1.) No additional stitching;
2.) Additional stitching in one direction, the vertical or horizontal direction, and;
3.) Additional stitching in both the horizontal and vertical directions.

Observations

The five two-level factors and the three-level factor were organized into a $2^5 * 3^1$ Full Factorial Design, where each level of every factor was combined with each level of all other factors, for a total of 96 unique observations. Each of the 96 unique observations was replicated, resulting in a total of 192 observations taken for the experiment.

At the request of NIJ, the manufacturer of the armor in question provided 100 identical body armor panels representing the armor model in question. Approximately half of these panels were subjected to extreme, but carefully monitored, temperature and humidity to weaken the material strength to the levels of that found within the failed vest. Thirty-two randomly selected panels of armor, 16 new and 16 weakened were used in the investigative study. Six shots were placed on each panel of armor. The six shots within each armor panel were arranged in such a manner that no two shots were under identical conditions. Exactly two shots were placed within each of the Vest Zones and within each vest exactly three observations were made at each level of the two-level factors (Bullet Type, Bullet Velocity, Barrel/Twist, and Angle of Incidence).

Across both the 16 new and 16 weakened vests, each condition was observed in each zone exactly twice (creating the replicated shot under each condition). The conditions assigned to the 16 new vests were repeated exactly as for the 16 weakened vests, creating the ability for a paired comparison.

Observation Order

Forty-eight observations were made throughout each of four days. Twenty-four observations were taken each morning using a single defined barrel with the remaining twenty-four observations taken in the afternoon using the second barrel. The order of barrels was arranged in such a manner to reduce the need for the barrel to be changed.

Each new vest was paired with a weakened vest as described above. Each vest was observed at least once, but no more than twice each day. The new/weakened vest pair was observed sequentially with the order randomly predetermined.

Eight shots were taken against a given clay block backing fixture before the block is replaced/repaired as described in the NIJ Standard. The eight shots contained exactly four New Vests and four Weakened Vests; four Magtech Bullets and four Remington Bullets; four 1055 ft/sec Velocity shots and four 975 ft/sec Velocity shots; four 0° Angle shots and four 45° Angle shots; with all Vest Zones represented and no paired location appearing more than once. The order of these observations that met the constraint requirements for a given clay block were randomly selected.

Six clay blocks (two blocks used three times) were used each day. Given the balanced factor constraints on each clay block, an overall daily factor balance was achieved.

Testing

Prior to the observations being made in both the morning and afternoon, shots were fired through the defined barrel verifying the velocities; 975 ft/sec and 1055 ft/sec, with each bullet type; Remington and Magtech, could accurately be achieved. Upon verifying the velocity accuracies, observations were taken as defined. The four-day test plan follows:

Day	Time	Barrel	Observation Number	Panel ID	Position	Zone	Material Condition	Bullet	Velocity ft/sec	Angle
ONE	AM	Hi Point	1	103	A	I	New	Remington	1055	0º
			2	058	A	I	Weakened	Remington	1055	0º
			3	041	F	III	Weakened	Magtech	975	45º
			4	035	F	III	New	Magtech	975	45º
			5	090	E	III	New	Remington	975	0º
			6	092	E	III	Weakened	Remington	975	0º
			7	080	C	II	New	Magtech	1055	45º
			8	056	C	II	Weakened	Magtech	1055	45º
			9	070	A	I	New	Remington	975	45º
			10	064	A	I	Weakened	Remington	975	45º
			11	057	D	II	Weakened	Remington	1055	0º
			12	069	D	II	New	Remington	1055	0º
			13	096	B	I	Weakened	Magtech	1055	45º
			14	068	B	I	New	Magtech	1055	45º
			15	058	E	III	Weakened	Magtech	975	0º
			16	103	E	III	New	Magtech	975	0º
			17	065	D	II	Weakened	Remington	1055	45º
			18	099	D	II	New	Remington	1055	45º
			19	077	B	I	New	Magtech	1055	0º
			20	054	B	I	Weakened	Magtech	1055	0º
			21	091	E	III	Weakened	Magtech	975	45º
			22	098	E	III	New	Magtech	975	45º
			23	055	A	I	Weakened	Remington	975	0º
			24	085	A	I	New	Remington	975	0º
	PM	SAAMI	25	048	B	I	Weakened	Magtech	1055	0º
			26	038	B	I	New	Magtech	1055	0º
			27	087	A	I	Weakened	Remington	975	45º
			28	037	A	I	New	Remington	975	45º
			29	029	E	III	New	Magtech	975	45º
			30	045	E	III	Weakened	Magtech	975	45º
			31	092	D	II	Weakened	Remington	1055	0º
			32	090	D	II	New	Remington	1055	0º
			33	056	B	I	Weakened	Magtech	1055	45º
			34	080	B	I	New	Magtech	1055	45º
			35	052	D	II	Weakened	Remington	1055	45º
			36	025	D	II	New	Remington	1055	45º
			37	026	E	III	New	Magtech	975	0º
			38	060	E	III	Weakened	Magtech	975	0º
			39	035	A	I	New	Remington	975	0º
			40	041	A	I	Weakened	Remington	975	0º
			41	055	F	III	Weakened	Magtech	975	45º
			42	085	F	III	New	Magtech	975	45º
			43	069	E	III	New	Remington	975	0º
			44	057	E	III	Weakened	Remington	975	0º
			45	026	A	I	New	Remington	1055	0º
			46	060	A	I	Weakened	Remington	1055	0º
			47	068	C	II	New	Magtech	1055	45º
			48	096	C	II	Weakened	Magtech	1055	45º

Day	Time	Barrel	Observation Number	Panel ID	Position	Zone	Material Condition	Bullet	Velocity ft/sec	Angle
TWO	AM	SAAMI	49	035	E	III	New	Magtech	1055	0º
			50	041	E	III	Weakened	Magtech	1055	0º
			51	090	F	III	New	Remington	1055	45º
			52	092	F	III	Weakened	Remington	1055	45º
			53	026	C	II	New	Remington	975	45º
			54	060	C	II	Weakened	Remington	975	45º
			55	096	A	I	Weakened	Magtech	975	0º
			56	068	A	I	New	Magtech	975	0º
			57	029	C	II	New	Remington	975	0º
			58	045	C	II	Weakened	Remington	975	0º
			59	037	E	III	New	Magtech	1055	45º
			60	087	E	III	Weakened	Magtech	1055	45º
			61	052	F	III	Weakened	Remington	1055	0º
			62	025	F	III	New	Remington	1055	0º
			63	077	A	I	New	Magtech	975	45º
			64	054	A	I	Weakened	Magtech	975	45º
			65	064	F	III	Weakened	Magtech	975	0º
			66	070	F	III	New	Magtech	975	0º
			67	054	C	II	Weakened	Magtech	1055	0º
			68	077	C	II	New	Magtech	1055	0º
			69	065	E	III	Weakened	Remington	975	45º
			70	099	E	III	New	Remington	975	45º
			71	045	A	I	Weakened	Remington	1055	45º
			72	029	A	I	New	Remington	1055	45º
	PM	Hi Point	73	038	C	II	New	Magtech	1055	0º
			74	048	C	II	Weakened	Magtech	1055	0º
			75	052	E	III	Weakened	Remington	975	45º
			76	025	E	III	New	Remington	975	45º
			77	091	A	I	Weakened	Remington	1055	45º
			78	098	A	I	New	Remington	1055	45º
			79	037	F	III	New	Magtech	975	0º
			80	087	F	III	Weakened	Magtech	975	0º
			81	080	A	I	New	Magtech	975	0º
			82	056	A	I	Weakened	Magtech	975	0º
			83	069	F	III	New	Remington	1055	45º
			84	057	F	III	Weakened	Remington	1055	45º
			85	058	C	II	Weakened	Remington	975	45º
			86	103	C	II	New	Remington	975	45º
			87	085	E	III	New	Magtech	1055	0º
			88	055	E	III	Weakened	Magtech	1055	0º
			89	038	A	I	New	Magtech	975	45º
			90	048	A	I	Weakened	Magtech	975	45º
			91	065	F	III	Weakened	Remington	1055	0º
			92	099	F	III	New	Remington	1055	0º
			93	091	C	II	Weakened	Remington	975	0º
			94	098	C	II	New	Remington	975	0º
			95	064	E	III	Weakened	Magtech	1055	45º
			96	070	E	III	New	Magtech	1055	45º

Day	Time	Barrel	Observation Number	Panel ID	Position	Zone	Material Condition	Bullet	Velocity ft/sec	Angle
THREE	AM	Hi Point	97	035	D	II	New	Magtech	975	0º
			98	041	D	II	Weakened	Magtech	975	0º
			99	026	B	I	New	Remington	975	45º
			100	060	B	I	Weakened	Remington	975	45º
			101	056	E	III	Weakened	Remington	1055	0º
			102	080	E	III	New	Remington	1055	0º
			103	052	A	I	Weakened	Magtech	1055	45º
			104	025	A	I	New	Magtech	1055	45º
			105	055	C	II	Weakened	Remington	1055	45º
			106	085	C	II	New	Remington	1055	45º
			107	054	D	II	Weakened	Remington	975	45º
			108	077	D	II	New	Remington	975	45º
			109	029	F	III	New	Magtech	1055	0º
			110	045	F	III	Weakened	Magtech	1055	0º
			111	065	B	I	Weakened	Magtech	975	0º
			112	099	B	I	New	Magtech	975	0º
			113	048	E	III	Weakened	Remington	1055	45º
			114	038	E	III	New	Remington	1055	45º
			115	090	A	I	New	Magtech	1055	0º
			116	092	A	I	Weakened	Magtech	1055	0º
			117	087	D	II	Weakened	Magtech	975	45º
			118	037	D	II	New	Magtech	975	45º
			119	029	B	I	New	Remington	975	0º
			120	045	B	I	Weakened	Remington	975	0º
	PM	SAAMI	121	064	D	II	Weakened	Magtech	975	45º
			122	070	D	II	New	Magtech	975	45º
			123	069	A	I	New	Magtech	1055	0º
			124	057	A	I	Weakened	Magtech	1055	0º
			125	058	B	I	Weakened	Remington	975	45º
			126	103	B	I	New	Remington	975	45º
			127	068	E	III	New	Remington	1055	0º
			128	096	E	III	Weakened	Remington	1055	0º
			129	055	D	II	Weakened	Magtech	975	0º
			130	085	D	II	New	Magtech	975	0º
			131	098	B	I	New	Remington	975	0º
			132	091	B	I	Weakened	Remington	975	0º
			133	077	E	III	New	Remington	1055	45º
			134	054	E	III	Weakened	Remington	1055	45º
			135	065	A	I	Weakened	Magtech	1055	45º
			136	099	A	I	New	Magtech	1055	45º
			137	038	D	II	New	Remington	975	45º
			138	048	D	II	Weakened	Remington	975	45º
			139	052	B	I	Weakened	Magtech	975	0º
			140	025	B	I	New	Magtech	975	0º
			141	041	C	II	Weakened	Remington	1055	45º
			142	035	C	II	New	Remington	1055	45º
			143	098	F	III	New	Magtech	1055	0º
			144	091	F	III	Weakened	Magtech	1055	0º

Day	Time	Barrel	Observation Number	Panel ID	Position	Zone	Material Condition	Bullet	Velocity ft/sec	Angle
FOUR	AM	SAAMI	145	069	C	II	New	Magtech	975	45º
			146	057	C	II	Weakened	Magtech	975	45º
			147	064	B	I	Weakened	Remington	1055	0º
			148	070	B	I	New	Remington	1055	0º
			149	098	D	II	New	Magtech	1055	45º
			150	091	D	II	Weakened	Magtech	1055	45º
			151	048	F	III	Weakened	Remington	975	0º
			152	038	F	III	New	Remington	975	0º
			153	055	B	I	Weakened	Remington	1055	45º
			154	085	B	I	New	Remington	1055	45º
			155	056	F	III	Weakened	Remington	975	45º
			156	080	F	III	New	Remington	975	45º
			157	058	D	II	Weakened	Magtech	1055	0º
			158	103	D	II	New	Magtech	1055	0º
			159	065	C	II	Weakened	Magtech	975	0º
			160	099	C	II	New	Magtech	975	0º
			161	090	B	I	New	Magtech	975	45º
			162	092	B	I	Weakened	Magtech	975	45º
			163	037	C	II	New	Remington	1055	0º
			164	087	C	II	Weakened	Remington	1055	0º
			165	103	F	III	New	Magtech	1055	45º
			166	058	F	III	Weakened	Magtech	1055	45º
			167	080	D	II	New	Remington	975	0º
			168	056	D	II	Weakened	Remington	975	0º
	PM	Hi Point	169	096	D	II	Weakened	Remington	975	0º
			170	068	D	II	New	Remington	975	0º
			171	060	F	III	Weakened	Magtech	1055	45º
			172	026	F	III	New	Magtech	1055	45º
			173	064	C	II	Weakened	Remington	1055	0º
			174	070	C	II	New	Remington	1055	0º
			175	069	B	I	New	Magtech	975	45º
			176	057	B	I	Weakened	Magtech	975	45º
			177	090	C	II	New	Magtech	975	45º
			178	092	C	II	Weakened	Magtech	975	45º
			179	054	F	III	Weakened	Remington	975	0º
			180	077	F	III	New	Remington	975	0º
			181	087	B	I	Weakened	Remington	1055	0º
			182	037	B	I	New	Remington	1055	0º
			183	045	D	II	Weakened	Magtech	1055	45º
			184	029	D	II	New	Magtech	1055	45º
			185	060	D	II	Weakened	Magtech	1055	0º
			186	026	D	II	New	Magtech	1055	0º
			187	096	F	III	Weakened	Remington	975	45º
			188	068	F	III	New	Remington	975	45º
			189	041	B	I	Weakened	Remington	1055	45º
			190	035	B	I	New	Remington	1055	45º
			191	025	C	II	New	Magtech	975	0º
			192	052	C	II	Weakened	Magtech	975	0º

Recorded Data

The data recorded for each defined observation included:

- Complete or Partial Penetration
- Bullet Velocity Readings from each of two sets of Chronographs
- Back Face Signature Measurement
- Laboratory Temperature
- Laboratory Humidity
- Clay Block ID used

Velocity Error

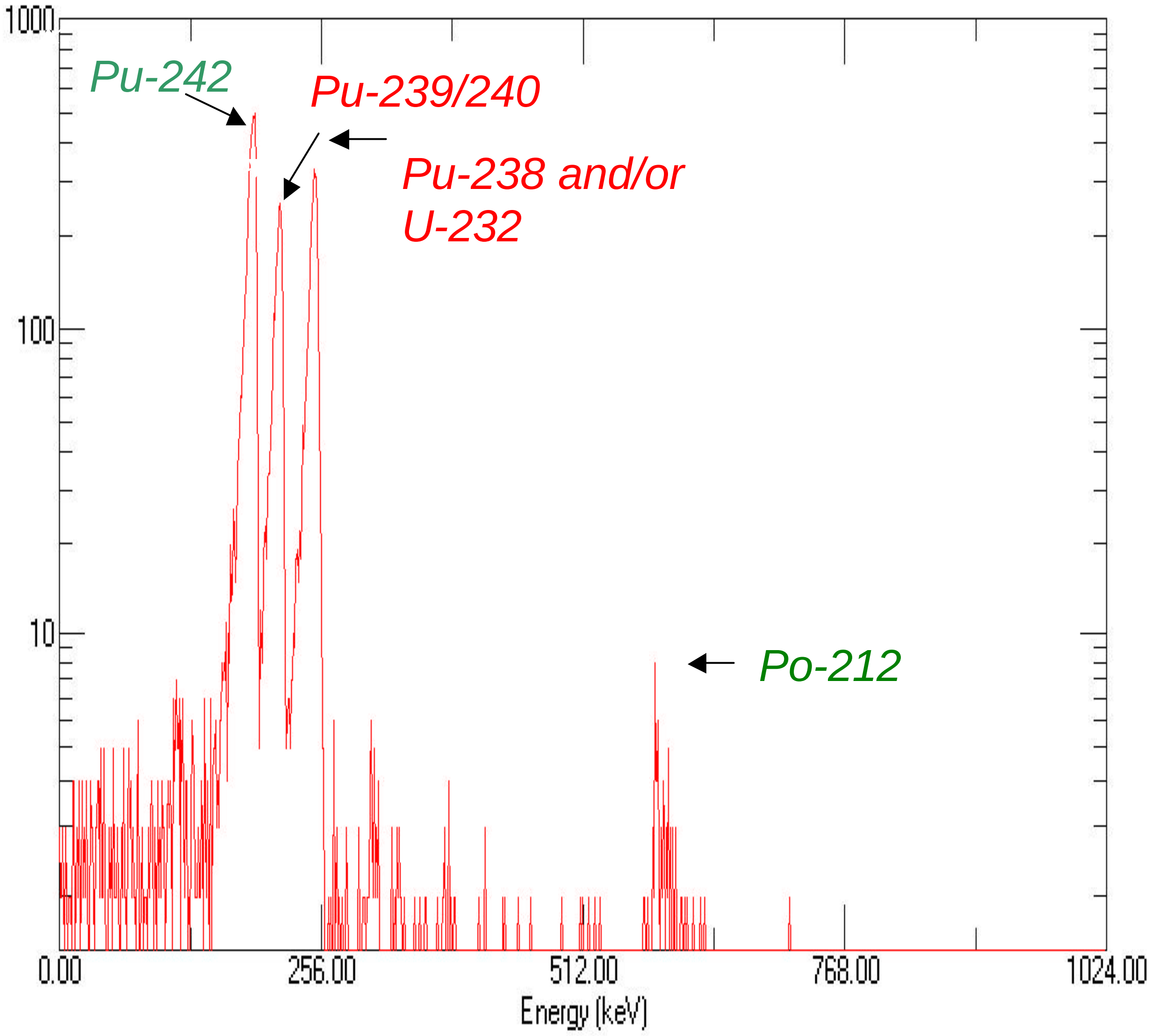
In the event that the desired velocity was not achieved due to the velocity variation introduced as a result of hand loading, the fair shot criterion as described in the NIJ standard was to be used. If the shot is deemed unacceptable, an additional shot elsewhere within the desired zone was taken after all other shots on the vest had been observed. No shots were deemed unacceptable throughout the testing.

Unfortunately, at the time of writing, the results from the experiment have yet to be approved by the funding agency, NIJ, for release to the public and therefore cannot be discussed in detail. The experiment did lead to several additional questions to be explored, for example: Were the assumptions relating to vest material strength based on observations of the rear panel correctly extrapolated to the failed front panel? Did the artificial weakening of the samples in the laboratory accurately simulate that which occurred in the officer's vest at the molecular level? These and many more questions, future experiments, and potentially life-saving conclusions will address the issues surrounding the safety of Zylon-based bullet-resistant vests. The extensive knowledge gained throughout the study has been and will continue to be used in the further refining of the measurement and analytical methods used in testing body armor as well as improvements to the National Institute of Justice's certification standard, ensuring the continuing commitment to safety for those who wear NIJ certified ballistic-resistant body armor.

6.3.2 Troubleshooting Plutonium Contamination in NIST Radionuclide Shellfish Measurements via Experiment Design

James J. Filliben
Statistical Engineering Division, ITL

Kenneth Inn, Robert Radford
Radiation Physics Division, PL



Low-level Plutonium-(239, 240 and 238) Contamination Detected in the Clean Lab

The NIST Radiation Physics Division (RPD) provides state-of-the-art measurements for a variety of radionuclides over a wide range of concentrations. NIST/RPD has particular expertise at low-dose oceanic (shellfish, seaweed, etc.) measurements, and the outside environmental/pollution community looks to this Division for definitive measurements, standards and calibration of such low-dose artifacts.

On occasion, low-concentration measurements of one radionuclide are contaminated by trace residues of another radionuclide. This write-up describes one such case, whereby it was determined that trace concentrations of plutonium had crept into a particular NIST shellfish alpha spectrometry measurement process. In order to maintain the world-class metrological standards symbolic of NIST/RPD, it was necessary that the source of plutonium contamination be identified and removed. Statistically designed experiments were critical in the uncovering of the contaminant source, and in doing so with a minimal number of runs/time/$.

As part of its ongoing program of monitoring ocean pollution worldwide, the NIST Radiation Physics Division regularly carries out high-accuracy measurements of radionuclides in shellfish, seaweed, and other oceanic artifacts. Increases in such radionuclide concentrations would suggest a possible worsening of pollution across space and/or time. In this regard, a question of interest is what procedural safeguards does NIST/RPD have in place to assure that such measurements are accurate.

One such safeguard is the incorporation and use of controls. In the context of these marine measurements, this translates into the execution of the measurement protocol on "blank" samples--that is, samples with zero concentration for all of the various radionuclides. This zero-base reading is valuable all by itself; an additional value-added is achieved, however, when the physical chemists artificially augment the blank with a "spike"--a known concentration of a single known radionuclide. This permits the proper calibration and interpretation of the resulting alpha spectrum.

In early 2003, in connection with NIST measurement of shellfish samples, the routine spectral examination (see Figure 1) of the spiked blank revealed a serious anomaly: not only did the spike (Pu-242) show up on the spectrum (as it should), but 2 additional plutonium peaks also appeared (unexpectedly)--one at Pu-239/240 and the other at Pu-238. There was no physical basis for these 2 additional peaks; they were, in fact, a clear indication that plutonium contamination existed somewhere in the shellfish measurement process itself.

Two courses of action were possible:

1. Leave the process unchanged, estimate the amount of contamination, and correct for it during the statistical analysis stage; or
2. Experiment with the process, determine the source of the contamination, and remove it.

The second option was clearly preferable. In this regard, over the course of the following year, a series of "best guess trial and error" tests were run by a RPD member with the hope of discovering the specific source of the contamination. This effort was unsuccessful.

In the early summer of 2004, an alternative approach was employed by RPD--the use of orthogonal fractional factorial designs--to systematically uncover the cause for the contamination. This entire effort took only 12 weeks--and was successful in resolving the problem. Details of this approach follow.

The first step in the experimental design structured approach is to carefully enumerate and scrutinize the components of the shellfish measurement process. The following is a summary of the 8-step procedure that was used for processing a (shellfish or blank) sample for alpha spectrometry:

1. Ashing
   Ashing is the process of burning the shellfish (or blank) sample. This is done to remove the organics. This ashing could be done at 2 different workstations: workstation 1 (belonging to staff member I), and workstation 2 (belonging to staff member H).

2. Dissolution
   Dissolution is the dissolving of the ash in acid. This is done to liquify the sample for further processing. The dissolution could be done by either of 2 NIST staff: I or H.

3. Chemistry
   Chemistry is the process of separating the plutonium (Pu) from the ash. This is done by putting the dissolved samples into an anion exchange column. By varying the acid concentration, the type of acid, and the oxidation state, one can cause the deposition of a radionuclide (plutonium, uranium, etc.) on the column. For this experiment, 2 different chemistry approaches were used: the above-described, or none.

4. Glassware
   The samples were at this time put into glassware. With the thought that the glassware could be contaminated due to the glassware storage area being contaminated and/or the wash solution contaminated, 2 different glassware wash procedures were considered: washing/storage by staff member I, and washing/storage by staff member H.

5. Hood
   The sample is further processed in a class 100 clean environment: (a max of 100 particles per cubic meter with each particle < .1 micron). In order to achieve this environment, 2 separate hoods were utilized: one at staff member I's workstation, and one at staff member H's workstation.

6. Reagents
   The addition of reagents facilitates the ultimate measurement of plutonium. To address the possibility that reagent bottles may be contaminated, two different reagents sources were used: one belonging to staff member I and the other belonging to staff member H.

7. Tracer
   The "tracer" refers to the spiking of the blank by Pu-242. To address the possibility that the Pu-242 was contaminated by other plutonium radionuclides, 2 tracers were used: one belonging to staff member I, and the other belonging to staff member H.

8. Sample Preparation
   Sample preparation refers to the final step of preparing the sample for (plutonium) counting. To determine if further plutonium contamination was caused by the sample prep method, 2 different sample prep procedures were used: one was electrodeposition, and the other was selective precipitation and filtering ("co-precipitation").

In terms of design essentials, this problem may be classified as a (k = 8 factor, n < 20 run) experiment. In terms of design goals, this is a classical screening experiment with its usual objectives:

1. Determine the important factors out of the 8 that maximize the response (= plutonium contamination); and
2. Determine the "worst" settings--those settings of the 8 factors that yield the highest plutonium contaminations.

In terms of factor levels, the natural binary settings for each of the 8 factors suggests taking advantage of the efficiency and power of 2-level orthogonal fractional factorial experiment designs. Based on k = 8 and n < 20, NIST/RPD selected a 2**(8-4) orthogonal fractional factorial design, which effectively and efficiently examines the relative importance of 8 factors with only 16 experimental runs.

The 8 factors and their 2 coded settings are given below in table 1:

	Factor	-1	1
1	Glassware	Staff I	Staff H
2	Reagent	Staff I	Staff H
3	Sample Prep	Co-precipitation	Electrodeposition
4	Tracer	Staff I	Staff H
5	Dissolution	Without	With
6	Hood	Staff I	Staff H
7	Chemistry	Without	With
8	Ashing	Without	With

Table 1. Factors for the Plutonium Contamination Study

Note that out of the 8 factors, 4 are "operator" in nature; that is 4 factors (1, 2, 4, and 6) have a human component involved in the operation (e.g., cleaning the glassware, mixing the reagent, adding the tracer, and doing the dissolution, respectively). Although such "human operator factors" are not physical in nature, it is a well-known fact that many measurement processes have human components, and these human components are oftentimes vulnerable to affecting/contaminating the final result. In this study, the same 2 staff members (staff member "I" and staff member "H") shared responsibility in 4 steps of the measurement process protocol.

Ken Inn of the RPD has been a strong local advocate of statistically designed experiments for many years. Ken has successfully incorporated many such designs for solving specific radiometric physics problems over the years; this "plutonium contamination detective work" problem was the latest. The final design (a 2^{8-4}) is a resolution 4 which of course means that none of the 8 estimated main effects will be confounded with any of the $\binom{8}{2}$ (= 28) 2-factor interactions. This is an excellent design for the "detective" problem at hand. The design is given in Table 2 below:

2^{8-4}	Glassware	Reagent	Sample Prep	Tracer	Dissolution	Hood	Chemistry	Ashing	Response
Run	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**Y**
1	-1	-1	-1	-1	-1	-1	-1	-1	0
2	1	-1	-1	-1	-1	1	1	1	0.003312
3	-1	1	-1	-1	1	-1	1	1	0.000037
4	1	1	-1	-1	1	1	-1	-1	0
5	-1	-1	1	-1	1	1	1	-1	0.000065
6	1	-1	1	-1	1	-1	-1	1	0.000133
7	-1	1	1	-1	-1	1	-1	1	0.000046
8	1	1	1	-1	-1	-1	1	-1	0.00003
9	-1	-1	-1	1	1	1	-1	1	0
10	1	-1	-1	1	1	-1	1	-1	0.000287
11	-1	1	-1	1	1	1	1	-1	0.000133
12	1	1	-1	1	1	-1	-1	1	0.000048
13	-1	-1	1	1	-1	-1	1	1	0.000133
14	1	-1	1	1	-1	1	-1	-1	0.005749
15	-1	1	1	1	1	-1	-1	-1	0.000015
16	1	1	1	1	1	1	1	1	0.00247

Table 2. 2^{8-4} Orthogonal Fractional Factorial Design (k = 8 factors, n = 16 runs)

Upon executing the design, the response (= total amount of plutonium contamination) is given in the last column of Table 2. Note that some (3) of the 16 runs yielded zero contamination.

With respect to the analysis, the design and data were inputted into the DEXPLOT.DP macro in NIST/SED's Dataplot statistical software system; this macro automates the recommended 10-step graphical procedure for analyzing orthogonal 2^{k-p} factorial designs (see "An EDA Approach to Experimental Design" in the "Advanced Topics" section of the Process Improvement chapter (i.e., Chapter 5) of the NIST/Sematech e-Handbook of Engineering Statistics). Three of those 10 steps are presented below to provide a clear answer to the problem of determining the plutonium contamination source.

Figure 2 is an ordered data plot: the vertical axis is the ordered response Y (= contamination); the horizontal axis is the "carry along" settings of all 8 factors. Note that there are 3 large contamination values. Note also the settings of the 8 factors that yield these 3 large values. Note finally that 2 of the 8 factors, namely (X1 = Glassware and X6 = Hood) have identical settings for all 3 of these extreme runs; in particular, when the contamination Y is large, X1 = (+ + +) and X6 = (+ + +).

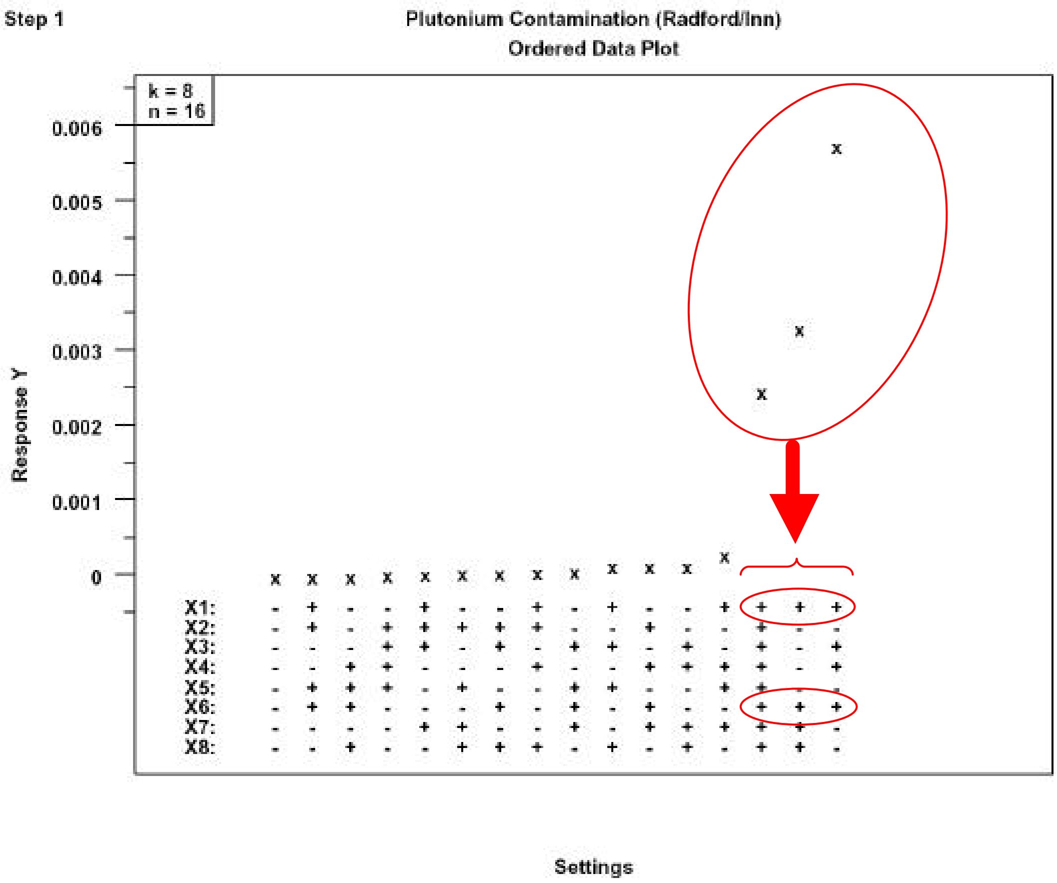


Figure 2. Ordered Data Plot. Vertical Axis = Plutonium Amount. Horizontal Axis = Experimental Runs (1 to 16)

The conclusions based on this EDA plot are as follows:

1. X1 (Glassware) and X6 (Hood) are the 2 most important factors;
2. X1 (Glassware) = + (that is, staff member H) and X6 (Hood) = + (that is, staff member H) are leading candidates for the plutonium contamination source.

Figure 3 (below) is a main effects plot, which shows the mean of each of the 2 settings for each of the 8 factors. Since the design is orthogonal, the difference in the means is identically the least squares estimate for the factor effect. The main effects plot reaffirms the above conclusions; namely, that

1. X_1 (Glassware) and X_6 (Hood) are the 2 most important factors;
2. X_1 (Glassware) = + (that is, staff member H) and X_6 (Hood) = + (that is, staff member H) yield the highest average contamination.

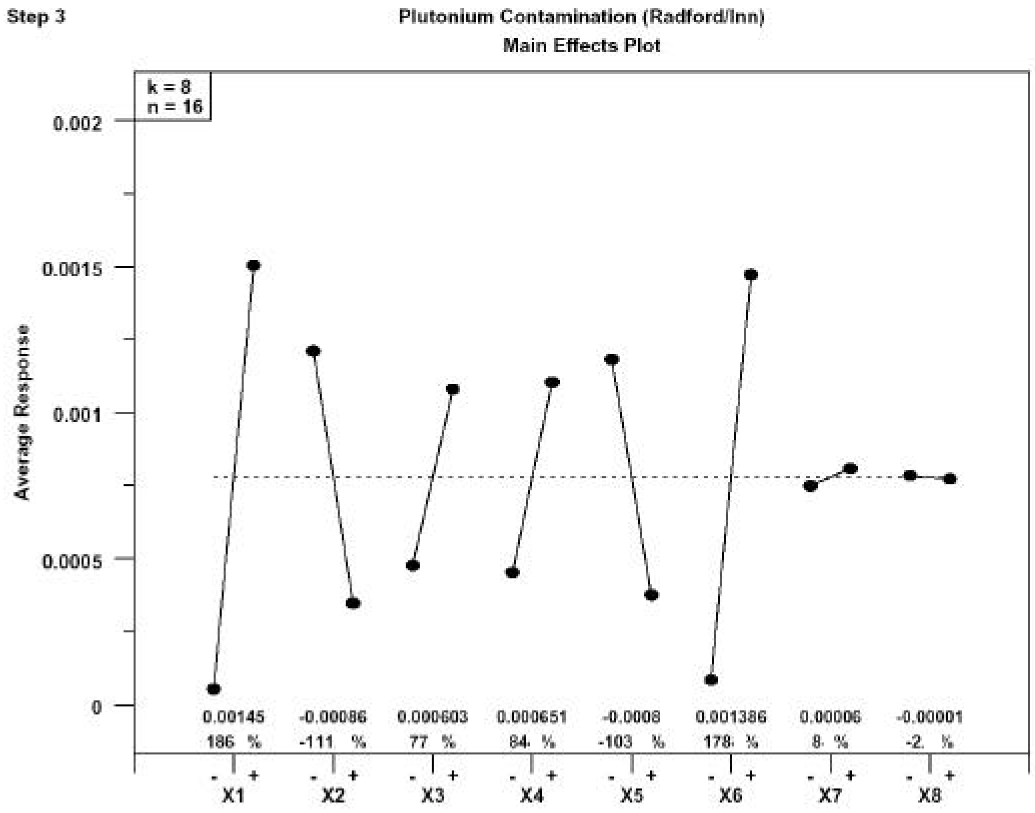


Figure 3. Main Effects Plot. Vertical axis = Average Plutonium Contamination. Horizontal Axis = 8 Factors and 2 Settings per Factor

A final analysis technique of interest is the half-normal probability plot of |estimated effects| (see Figure 4 below). Note how the 3 largest effects have separated themselves from the remaining 13 effects. This indicates that the 3 most important factors are

1. X1
2. X6
3. X3X4, X1X6, X2X5, X7X8 (confounded)

From other considerations, we may deduce that the 2-term interactions X3X4, X2X5, and X7X8 make relatively minor contributions to this third effect, and that the bulk of this effect comes from the X1X6 interaction. Given that, we make a final conclusion that the plutonium contamination from this system is being driven by

1. X1 (Glassware)
2. X6 (Hood)
3. X1X6 (the Glassware*Hood interaction)

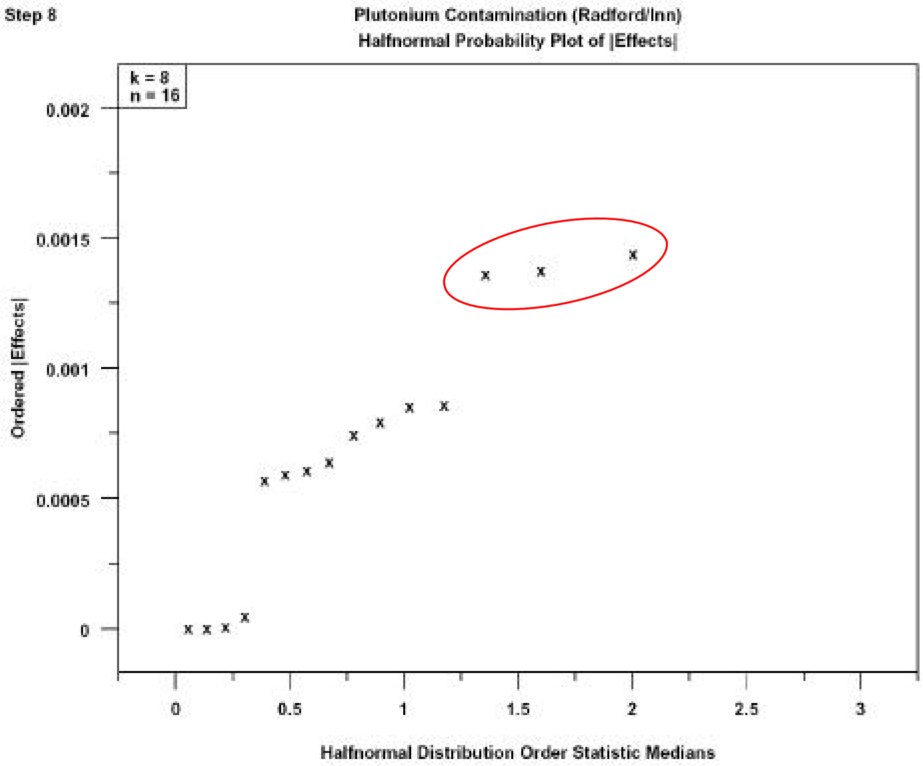


Figure 4. Half-normal Plot of |Effect Estimates|.
3 Factors Dominate: 1, 6, and 16 Interaction

Thus in summary, this analysis shows that the plutonium contamination has been narrowed down to 2 sources: the cleansing of the glassware and the hood. For item 1 (glassware cleaning), stricter cleaning and storage room protocols are being put in place. With respect to item 2 (hood), it is believed that the filter within the hood is the culprit--this filter is currently being replaced ($3000).

The use of orthogonal fractional factorial designs to simultaneously and efficiently study a large number of factors to provide information about dominant factors and best/worst settings played a critical role in the RPD's effort to troubleshoot and maintain high-accuracy contamination-free shellfish measurements. This approach has potentially many similar applications. The presentation of this case study by one of the collaborators (Radford) to the radiometric community at the recent Cincinnati meeting led one of the Los Alamos management-level attendees to declare what a superb presentation/method it was, and how he/she could think of many applications back at Los Alamos that might benefit from the same structured orthogonal experiment design approach. In any event, the NIST RPD oceanic radionuclides measurement program specifically benefitted by identifying the major sources of pollution contamination in a minimal amount of time and effort. This will serve to maintain NIST/RPD's leadership role in continuing to provide the precise and bias-free measurements so vital to environmental and pollution communities worldwide.

6.3.3 Statistical Design and Signal Processing of Microhotplate Sensor Arrays in Detection of Chemical Warfare Simulants

Z.Q. John Lu, Juan Soto, Nell Sedransk
Statistical Engineering Division, ITL

Jon Evju, Zvi Boger, Steve Semancik
Process Measurement Division, CSTL

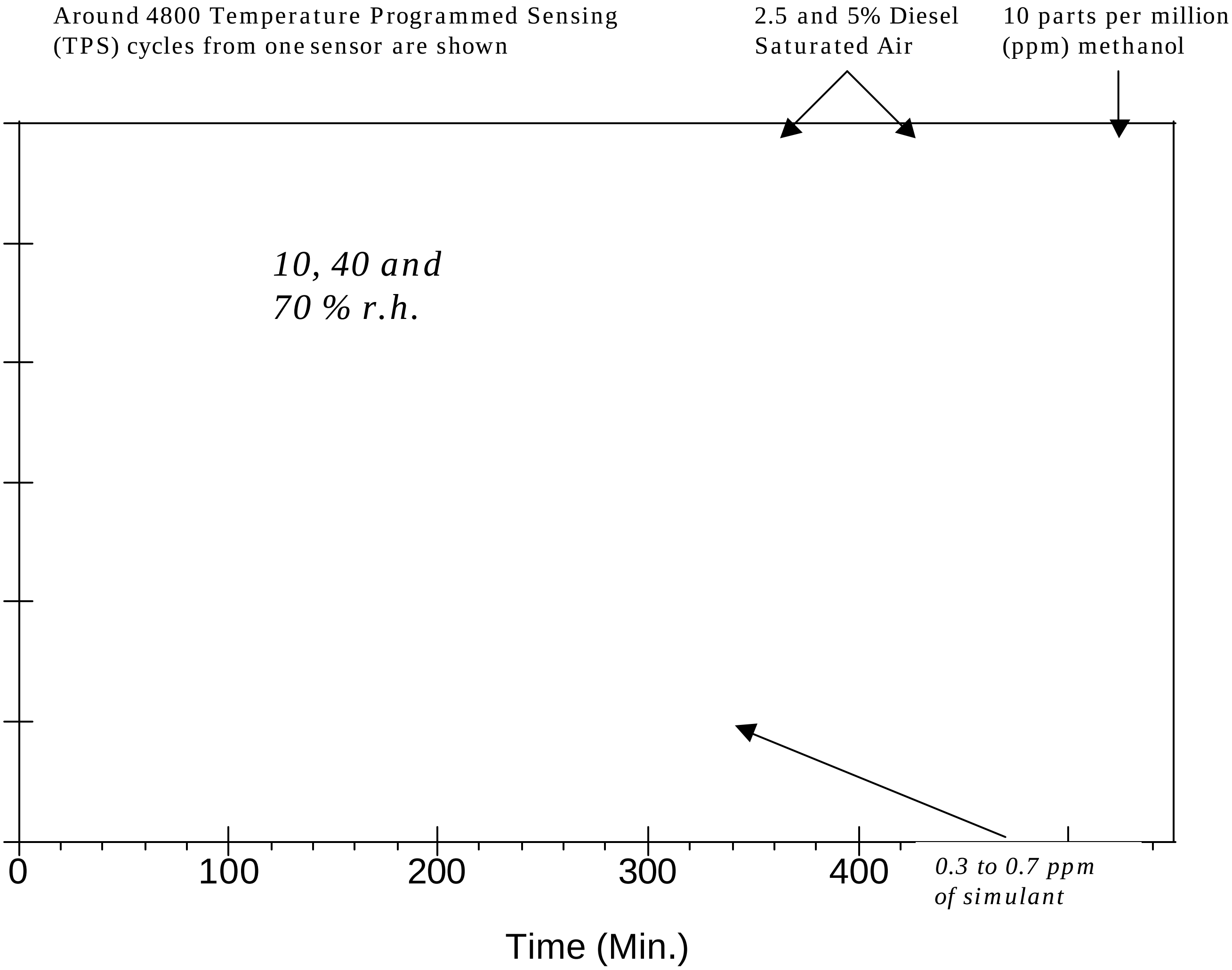
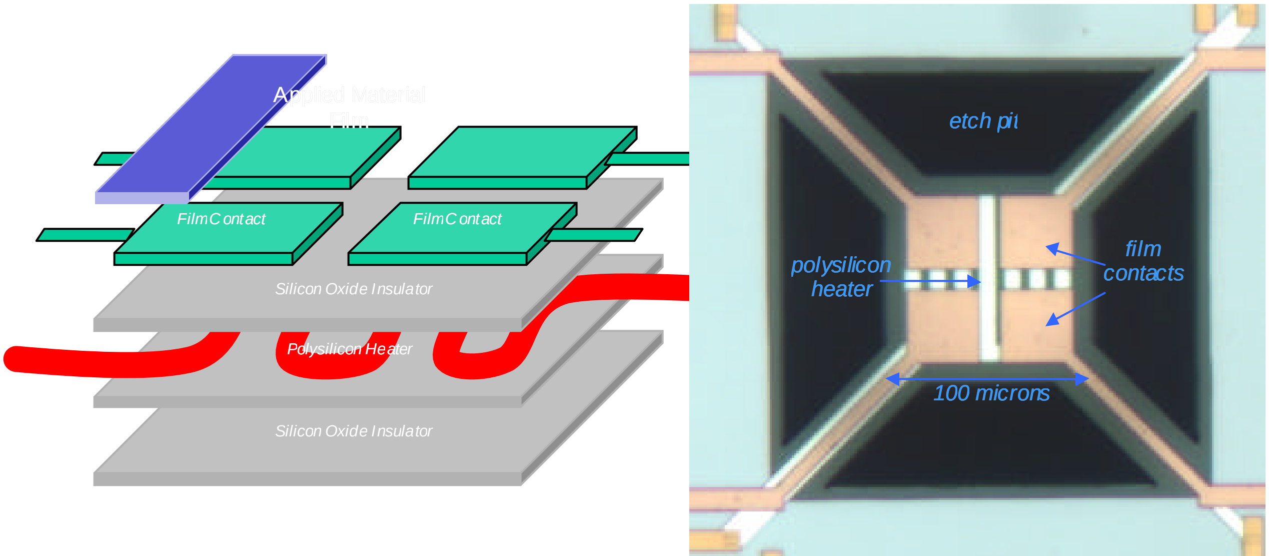


Figure 1. Design introducing interferences and simulant to TPS sensing cycles during one 540-minute-run of 36 steps (experimental settings)

Microhotplate sensor arrays are low-cost and extra-sensitive monitoring devices that can be used to detect certain chemical analytes at extremely low levels of concentration, such as parts per million (ppm) or parts per billion (ppb). This interdisciplinary project, funded in part by the Department of Defense, aims to demonstrate for the first time the feasibility of this NIST-patented technology for detection of chemical warfare simulants at low concentration and under strong interference factors such as water, diesel, etc. The Statistical Engineering Division has made key contributions in this project, which include: (a) the design of experiments with the collection of high throughput data under multi-factor influences and inputs and high-dimensional signal outputs, and (b) the development of a statistical methodology for signal processing and statistical data mining: finding weak fingerprints (needle) from high-interference signal data (haystacks).

MicroElectroMechanical Systems (MEMS) design-based microhotplates are miniaturized platforms that are used to collect relevant chemical signals in a parallel and high-throughput fashion so that a particular chemical analyte of interest may be distinguished from other environmental interferences via its kinetic reactions to certain thin films while heated at some very high temperature ramping cycles. Figure 2 depicts a 2x2 design of this device.



<u>Features include:</u> Thermal isolation

- Lateral dimensions 30 to 200 mm; mass 250 ng
- Capable of heating rates of 10^6 °C/s
- Time constants ~2 ms• 20 °C to 500 °C normal operation range
- Both films and hotplates are independently controlled and measured
- Tool for controlled materials preparation and gas analyte sensing

Figure 2. Diagram of the microhotplate-sensing device

Traditional conductometric sensors have but a single response to their environment – conductance change. As a result, they are fundamentally non-selective, since any reducing agent present in a given analyte stream will interact with the metal oxide film, resulting in a conductance change. Rapid thermal control of microhotplates allows a sensing mode not available to traditional conductometric sensors: Temperature Programmed Sensing (TPS).

High-speed temperature modulation during conductometric measurements results in the collection of a series of kinetic transients rather than thermodynamic equilibria. In this way, an array of four microsensors programmed to sample 20 temperatures in 13.5 seconds can be treated as 80 different virtual sensors, each with a time resolution equal to the program time. Larger numbers of sensors, virtual or otherwise, expand response orthogonality, and thus selectivity, for a given analyte system. Through careful selection of sensor materials and temperature programs, maximum selectivity and sensitivity can be achieved.

Our goal in this DoD funded project is to demonstrate the feasibility of using the microhotplate technology for detection of some chemical warfare simulants at very low concentration with various interferences being introduced during the experimental setup. Figure 1 shows a design for one run at 36 experimental settings with about 4800 temperature cycles. Figure 3 shows an example of signal data from 4 sensors under the TPS mode. It is seen that the two types of films, SnO_2 and TiO_2 can give rise to quite different kinetic reactions for the given simulant at various experimental conditions. However, it is clear that within each experimental setting (denoted by the same color), the response cycles are quite consistent and compact, indicating very good reproducibility of the sensors at a given condition.

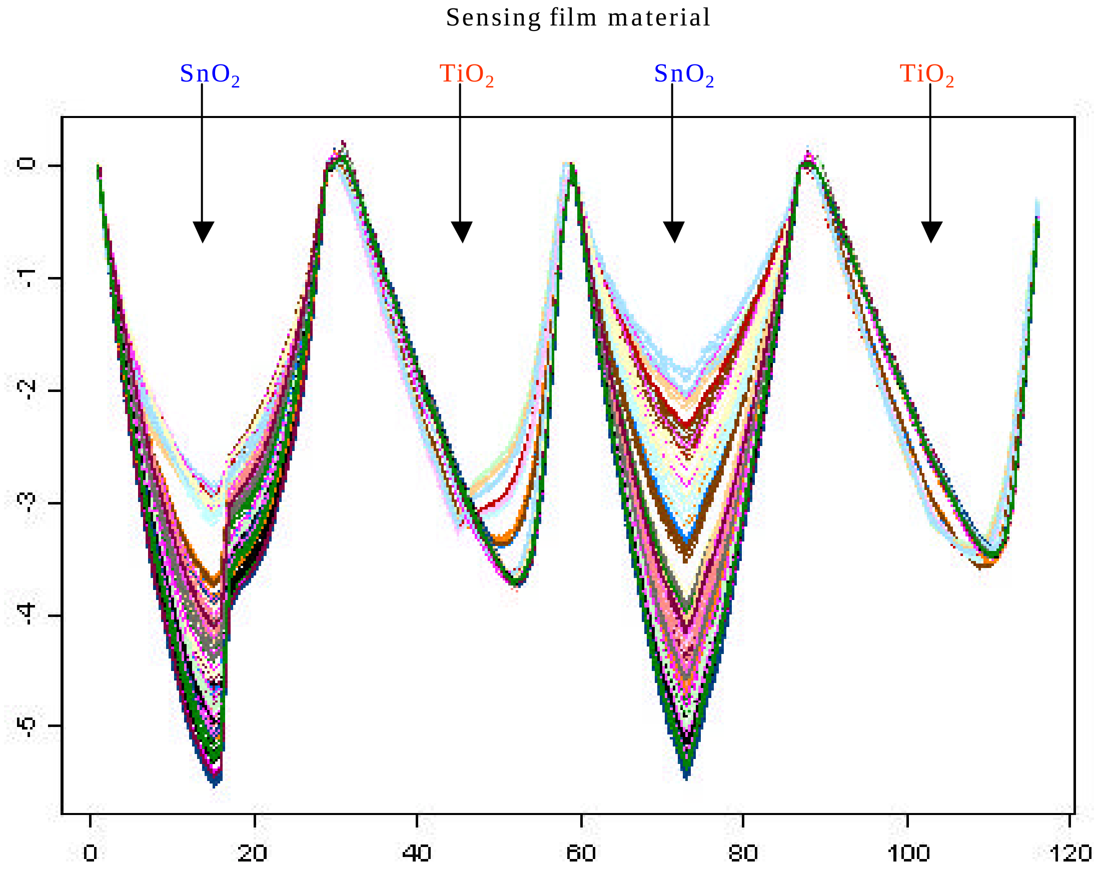


Figure 3. An example of sensor signals (conductivity response cycles) from one experiment from TPS temperature cycles: 480^{o}C? 420^{o}C? …? 367^{o}C? 420^{o}C? 480^{o}C

To validate this claim, Figure 4 shows the prediction results using a very simple statistical data mining algorithm. The training data consist of the cycles for the time interval 8-12 minutes. The rest of the data within the 15-minute setting are tested based on a fitted nonlinear classifier or posterior probability model. Only data from one experimental run from all 36 settings are used. Two conclusions can be drawn. First, sensor signals from a given short time period can be used for training. These yield prediction results corresponding to the experimental inputs with very good accuracy. Second, there is a short transient period when the sensors are switched to different experimental settings. The initial signals from sensors are vulnerable to residual effects from previous settings. These nonstationary effects may be detected or calibrated for in the future.

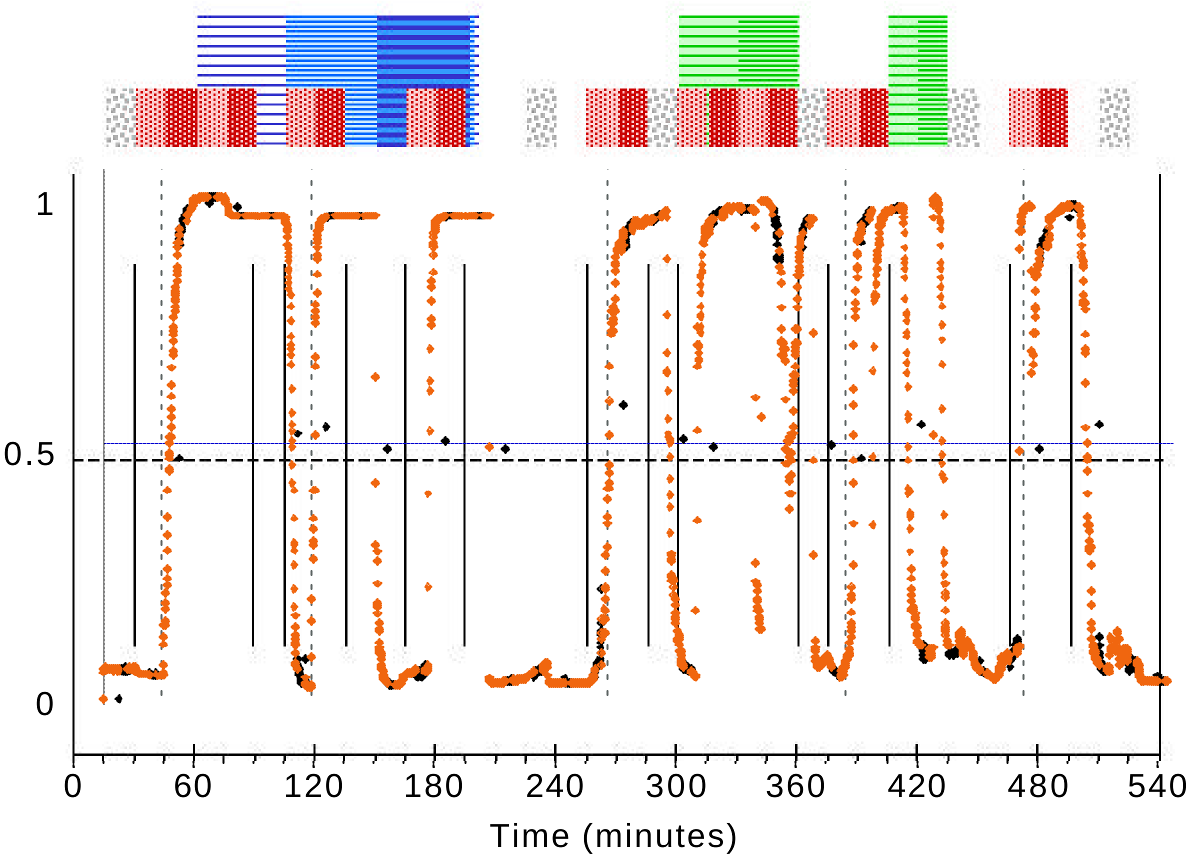


Figure 4. Prediction results from a statistical data mining algorithm. The black diamonds denote posterior probabilities for the presence of a simulant during the training phase (time period), whereas the orange diamonds denote posterior probabilities for the presence of a simulant during the testing phase (time period).

*In summary, statistical science provides the much-needed experimental design and decision-theoretic support***,** *as well as pattern recognition algorithms that can be implemented easily in an S-PLUS or GNU R environment. This project illustrates SED expansion in the challenging experimental design and signal processing areas of combinatorial and high-throughput experiments. The statistical data mining efforts using neural networks, support vector machines, and k-nearest neighbor methods are applicable to many other problem areas, e.g., bioinformatics.*

6.4 Web Products and Software

6.4.1 Rewrite of MASSCOMP Calibration Software

Alan Heckert, Hung-Kung Liu
Statistical Engineering Division, ITL

Jalilian Firouzeh
University of Maryland

Zeina Jabbour
Manufacturing Metrology Division, MEL

Standards: K20 – National Standard
K4 - Check Standard
K79 – Research Standard

The mass of an object, a measure of quantity of matter, is an intrinsic property of the object. Precise measurement of mass is a cornerstone for trade and commerce and therefore serves a vital role in science and manufacturing. To ensure equity and equivalence in trade and manufacturing at the national and international levels, uniform standards for mass are needed. Using weighing designs, mass standards are calibrated at NIST by comparison measurements that relate the mass of a client's standard to the NIST standard kilogram. The NIST standard kilogram is in turn related to the internationally defined unit for mass, the IPK.

The current mass calibration software was initially written in Fortran 66 in the 1970's. Although this code has served for several decades, it now has a number of limitations.

- The code is unstructured (e.g., Fortran 66 did not support IF-ELSEIF-ELSE constructs and it uses a limited number of large subroutines). This makes the code difficult to read and therefore difficult to update. As a result, a number of desired code updates have not been implemented.
- The code has a rigid data input format.
- The code only runs in a command line and does not support any graphical output.

For these reasons, we are currently rewriting the mass calibration software. The primary goals for this project are

- Rewrite the mass calibration software in a modern language to provide a more structured and modularized code. This will result in a more maintainable code.
- Provide a more flexible data input model.
- Implement a large number of proposed updates to the program. Some of these updates are relatively minor while others are significant.
- Implement a graphical user interface (GUI) for Windows. The mass calibration software is used by a number of non-NIST users of varying computer skills.

We chose Fortran 90 as the language for the computational engine. Fortran 90 provides modern control structures and convenient methods for modularizing the code. Of particular note, it supports sophisticated array syntax that allows simplification of the mathematical and statistical computations. We chose Visual Basic as the language for developing the graphical user interface (GUI). Visual Basic provides easy-to-implement user-interface features while maintaining compatibility with a Fortran 90 computational engine. For the GUI, the main program is written in Visual Basic while the Fortran 90 modules are implemented as a dynamic link library (DLL).

Accomplishments for the mass calibration software rewrite over the last year include

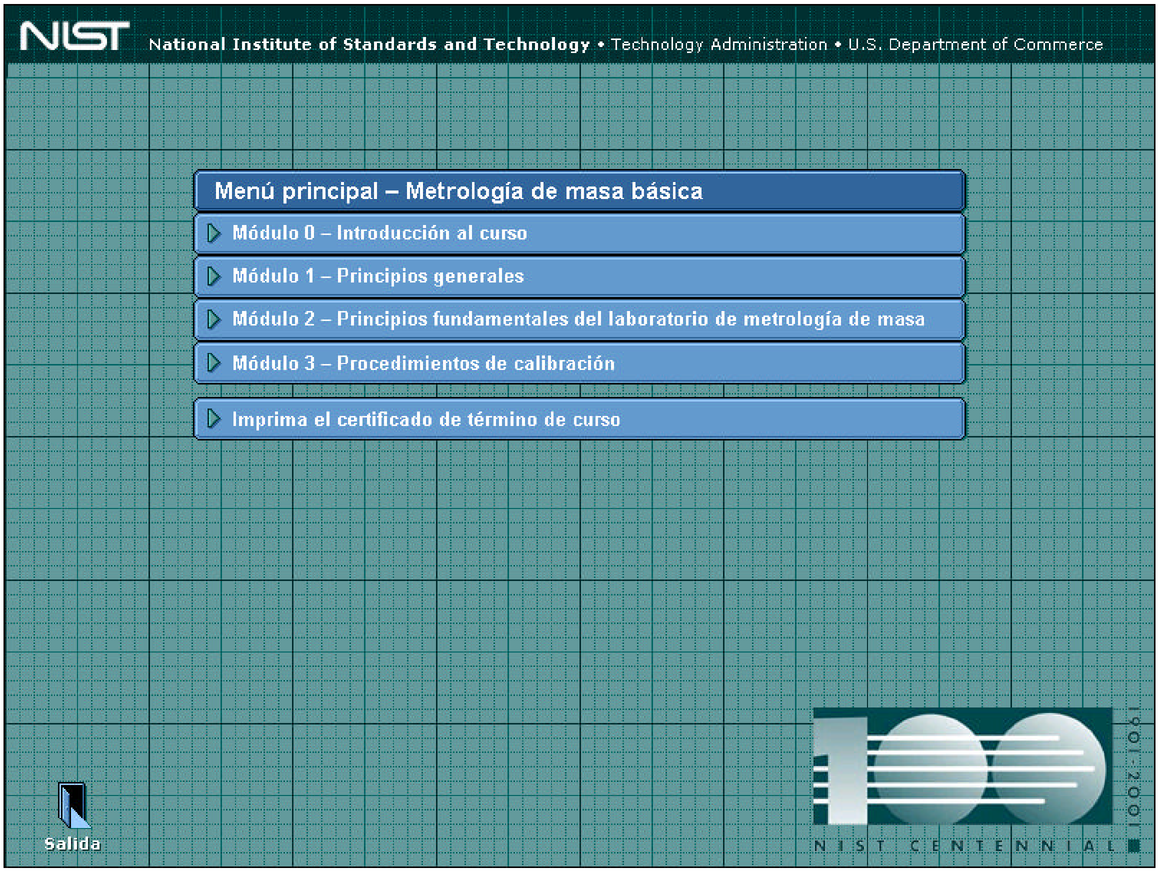
- The Fortran 90 rewrite of the current code is essentially complete. The large subroutines were split into smaller and easier to understand modules. Extensive documentation was added to the Fortran subroutines. Computations were simplified by using modern control structures and array syntax. All of these steps lay a strong foundation for future enhancements to the code.
- A keyword data input format was designed and implemented. This also served as a basis for designing the input module in the GUI.
- Most common weighing designs are now built into the code.

Future efforts include refining and developing the GUI and enhancing the functionality of the mass calibration software. The rewritten mass calibration software should provide a solid foundation for a program that is easier to maintain and update and that will be easier for non-NIST customers to use.

6.4.2 The Basic Mass Metrology CD-ROM – Spanish Version *(CD-ROM de metrología básica de masa del NIST)*

A. Ivelisse Avilés
Statistical Engineering Division, ITL

Georgia L. Harris
Weights & Measures Division,TS


NIST
National Institute of Standards and Technology • Technology Administration • U.S. Department of Commerce
Menú principal – Metrología de masa básica
Módulo 0 – Introducción al curso
Módulo 1 – Principios generales
Módulo 2 – Principios fundamentales del laboratorio de metrología de masa
Módulo 3 – Procedimientos de calibración
Imprima el certificado de término de curso
Salida
100
1901-2001
NIST CENTENNIAL

The multimedia CD-ROM metrology course is divided into modules. At the end of the course, a participation certificate can be printed.

The National Institute of Standards and Technology (NIST) trains many state and industrial laboratory metrologists who verify the accuracy of standards used to test the measuring equipment used in commercial transactions. To alleviate a training backlog of officials from government and industrial mass calibration laboratories, NIST released an electronic mass measurement training course in October 2003. Although NIST receives numerous requests to provide training in Spanish, it has been difficult to get enough students in one class to make it cost-effective to offer the course in Spanish. Thus, funding was obtained to translate the entire project and provide a version of the Basic Mass Metrology course in Spanish. This version of the CD-ROM was just released. The Statistical Engineering Division has been involved in the review assessment of technical (statistical) validity, with appropriate technical and linguistic editing and/or revision, as required, in both English and Spanish.

The free multimedia CD-ROM covers NIST's basic one-week mass metrology course. It includes interactive activities, knowledge quizzes, examples, video demonstrations, and specialty graphics and photos for specific products. The CD-ROM is designed to introduce mass metrology to newcomers to the field; offer supplementary training for those who have recently attended a metrology course and want to review their knowledge before entering the laboratory environment; and act as a refresher for long-time laboratory staff unfamiliar with the latest measuring techniques.

The course is divided into the following modules:
Module 1 – General Principles
Module 2 – Basics in Mass Metrology Lab
Module 3 – Calibration Procedures

By the time metrologists finish this course, they will be able to:

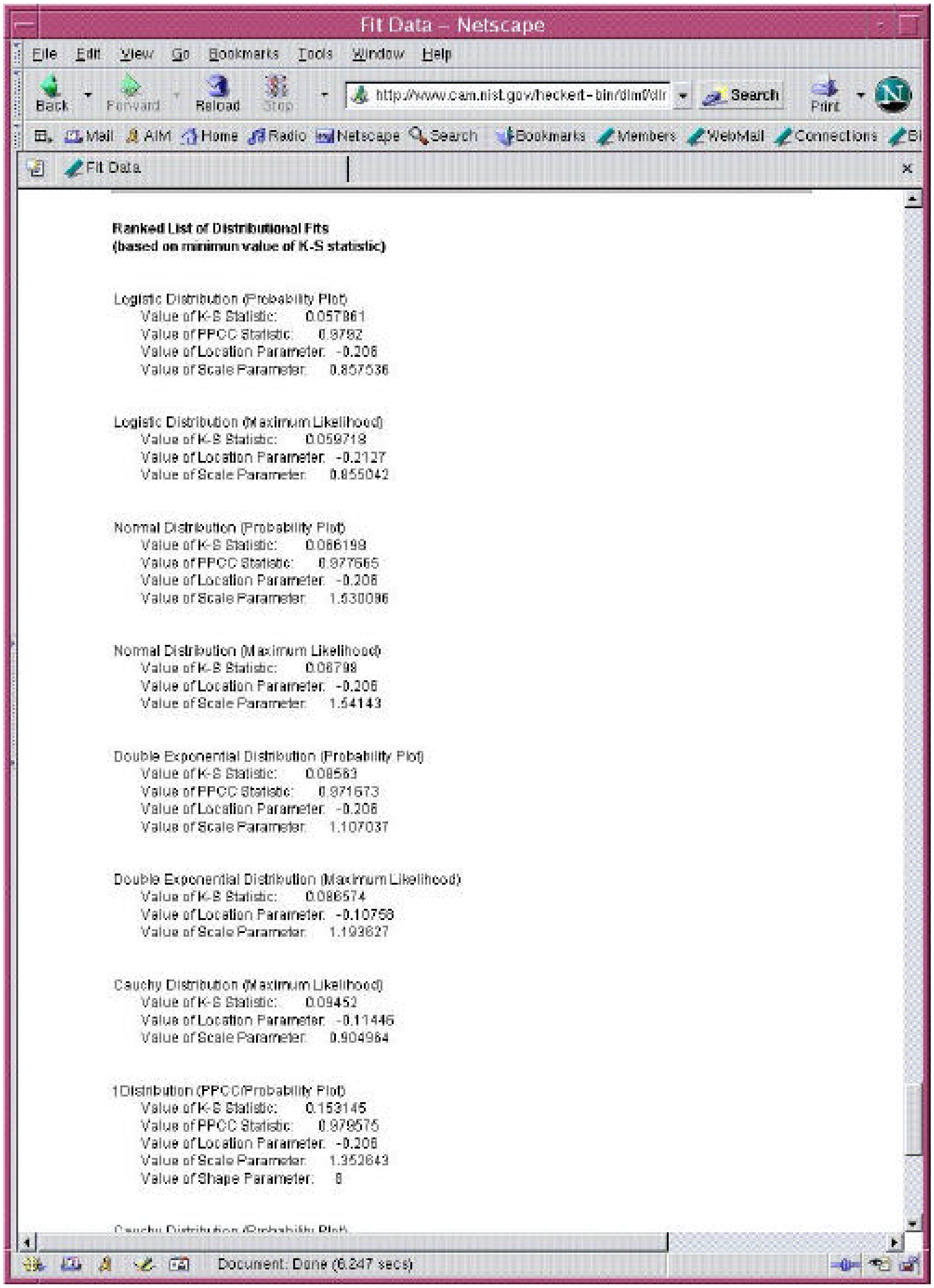
- Perform basic math and statistics in metrology calculations
- Explain the measurement assurance process and the factors that influence measurement uncertainty
- Develop good laboratory practices
- Create control charts, calculate measurement uncertainties, and interpret the measurement results
- Explain basic mass metrology theory used in making accurate mass measurements
- Select an appropriate standard operating procedure for a mass measurement
- Correctly find the correction and uncertainty for the unknown weight being calibrated

The CD-ROM was developed so that metrologists can have a ready reference on mass calibration that can be reviewed at any time. The course material has been translated into Spanish for even wider distribution and metrology support. It provides a good introduction for new metrologists while providing supplemental training for the experienced metrologist. The information is divided into modules, lessons, and topics to help them quickly locate information as necessary for the job. The Basic Mass Metrology CD-ROM (NIST Special Publication 1001) is available from NIST at (301) 975-4004 or by e-mail at owm@nist.gov

.6.4.3 e-FITS

Alan Heckert
Statistical Engineering Division, ITL

Ken Inn
Ionizing Radiation Division, PL



Fit Data – Netscape
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.cam.nist.gov/heckert-bin/dlmf/dlr
Search
Print
Mail AIM Home Radio Netscape Search Bookmarks Members WebMail Connections Bi
Fit Data

```
Ranked List of Distributional Fits
(based on minimun value of K-S statistic)


Logistic Distribution (Probability Plot)
      Value of K-S Statistic:       0.057861
      Value of PPCC Statistic:       0.9792
      Value of Location Parameter:  -0.206
      Value of Scale Parameter:      0.857536


Logistic Distribution (Maximum Likelihood)
      Value of K-S Statistic:       0.059718
      Value of Location Parameter:  -0.2127
      Value of Scale Parameter:      0.855042


Normal Distribution (Probability Plot)
      Value of K-S Statistic:       0.066199
      Value of PPCC Statistic:       0.977665
      Value of Location Parameter:  -0.206
      Value of Scale Parameter:      1.530086


Normal Distribution (Maximum Likelihood)
      Value of K-S Statistic:       0.06798
      Value of Location Parameter:  -0.206
      Value of Scale Parameter:      1.54143


Double Exponential Distribution (Probability Plot)
      Value of K-S Statistic:       0.08563
      Value of PPCC Statistic:       0.971673
      Value of Location Parameter:  -0.206
      Value of Scale Parameter:      1.107037


Double Exponential Distribution (Maximum Likelihood)
      Value of K-S Statistic:       0.086574
      Value of Location Parameter:  -0.10758
      Value of Scale Parameter:      1.193627


Cauchy Distribution (Maximum Likelihood)
      Value of K-S Statistic:       0.09452
      Value of Location Parameter:  -0.11446
      Value of Scale Parameter:      0.904964


1Distribution (PPCC/Probability Plot)
      Value of K-S Statistic:       0.153145
      Value of PPCC Statistic:       0.979575
      Value of Location Parameter:  -0.206
      Value of Scale Parameter:      1.352643
      Value of Shape Parameter:      8


Cauchy Distribution (Probability Plot)
```

Figure 1: A ranked list of distributional fits generated by e-FITS.

The e-FITS software is a Web-based tool used to generate graphs, tables, critical values, random numbers, and distributional fits for a large number of probability distributions. The motivation for e-FITS is to support the statistics chapter in the Digital Library of Mathematical Functions (DLMF) and to provide a convenient way to perform distributional modeling. Although e-FITS complements the DLMF, it is a distinct product from the DLMF.

The statistics chapter of the DLMF documents 22 univariate continuous distributions and seven discrete distributions. e-FITS generates the following for these distributions.

- Graphs of common probability functions (probability density, percent point, hazard, cumulative hazard, survival, and inverse survival).
- Tables and single values for each of the above functions.
- Random numbers for the distribution.
- Generate a fit for the distribution with user-supplied data. Fits will be generated using either the PPCC/probability plot or with maximum likelihood methods of moment estimates (or both). Diagnostic plots and Anderson-Darling or Kolmogorov-Smirnov goodness of fit tests are generated to assess the adequacy of the fit.
- For selected distributions, generate tables of critical values.

The e-FITS software is implemented using forms and CGI scripts. The user of e-FITS does not need to install or learn a statistical software program. The user fills out a Web form and a CGI script utilizes Dataplot to generate the requested graph, table, or fit. Ken Inn of the Ionizing Radiation Division provided support for the e-FITS project over the last year. His group is particularly interested in distributional fitting and in finding uncertainty intervals for percentiles of the fitted distributions.

Accomplishments for e-FITS over the last year include

- 28 (up from 14) continuous distributions are operational with two additional distributions in progress. Nine (up from zero) discrete distributions are operational. Fitting algorithms are still being developed for a few of these distributions.
- Forms were added for fitting a number of distributions and returning a ranked list of best fit. One form fits five common symmetric distributions while another fits twelve common asymmetric distributions. Forms were also added to generate distributional graphs of the user-supplied data (e.g., histograms, kernel density plots).
- The fitting forms are being expanded to include pre-binned data and censored data for a number of the supported distributions.
- A number of updates were made to the underlying Dataplot software to better support e-FITS. Maximum likelihood estimation was enhanced to support additional distributions, to support censored data (normal, exponential, gamma, lognormal, Weibull, Gumbel), and to provide better confidence intervals for estimated parameters and percentiles for selected distributions. The PPCC and probability plots were enhanced to support censored data. We are currently developing a command to provide bootstrap based confidence intervals for estimated parameters and percentiles for a wide range of distributions.

Future efforts include adding additional univariate distributions, support for selected multivariate distributions, and using Java to dynamically generate tables and graphs for selected distributions.

The Web-based e-FITS software will be a useful educational and reference tool. It will also provide a useful tool for NIST staff and industry to perform distributional modeling.

7 Education

SED provides education and training in a variety of ways: (1) short courses on both campuses at NIST, (2) Web based, (3) professional society short courses and workshops, (4) SED sponsored seminars, (5) case studies, and (6) talks by NIST staff.

SED has reinstituted the series of statistics short courses known as Statistics for Scientists and Engineers. These courses target NIST staff, although they typically draw attendees from outside of NIST as well, both from other U. S. government agencies and private corporate sources. The courses are of varying duration and depth, but are designed to cover statistics, probability, data analysis, and statistical computing topics deemed to be relevant to NIST scientific staff at a level appropriate for NIST staff, from technician to senior Ph.D. level.

Each course typically covers one major area or aspect of statistics, with an emphasis on applications to NIST scientific and engineering problems. The principal objective of each course is to help researchers recognize opportunities for the use of particular statistical methods and to offer practical guidance in their applications.

7.1 Short Courses

Hands-on Workshop on Estimating and Reporting Measurement Uncertainty
Will Guthrie, Hung-kung Liu, Adriana Hornikova
Measurement Science Conference, Anaheim CA
January 12/13, 2004

This workshop on uncertainty describes the statistical framework and methods needed to develop uncertainty statements based on the ``ISO Guide to the Expression of Uncertainty in Measurement.'' Methods for uncertainty estimation are illustrated with practical examples from different metrological areas. The workshop includes hands-on examples to be analyzed by the participants. The examples are handled using both propagation of uncertainty formulas and the Kragten spreadsheet, an easy-to-use computational tool for propagation of uncertainty.

Introduction to Markov Chains and Markov Chain Monte Carlo for Scientists and Engineers
Andrew Rukhin, Stefan Leigh, Van Molino
August 16/17, 2004

This two-day introductory course devotes one day to an overview of the basic theory of finite Markov chains and their applications, and one day to their application to Markov Chain Monte Carlo.

Day one topics include: probability concepts review, Markov chain basics (order, time homogeneity, transition matrices), visualization and application of chains, operations to extend applicability of first-order time homogeneous chains, Markovian taxonomy (classification of states), canonical representation, periodicity, irreducibility, limiting/stationary distributions, equilibrium, reversibility, and absorbing chain taxonomy and properties.

Day two on MCMC follows the introductory exposition of the course text, FINITE MARKOV CHAINS AND ALGORITHMIC APPLICATIONS (Haggstrom, Cambridge Univ. Press, 2002). The concepts are motivated at length through a study of interspecies competition measures on a closed eco-community that can be modeled by simulating progressions of zero-one co-occurrence matrices with constrained marginal sums (Cobb and Chen, AMERICAN MATHEMATICAL MONTHLY, April 2003). Coding issues, standard generic variations, convergence issues (and proofs), the Propp-Wilson Algorithm, sandwiching, and simulated annealing are covered at a gentle pace.

Introduction to Nonparametric Regression

Will Guthrie NIST
September 8, 2004

This course introduces several popular non-parametric regression methods for use by scientists and engineers. The course starts with spline models with pre-specified knots, which can be fit with regular regression software, and then moves on to cover smoothing splines, LOESS, and kernel smoothers. The properties of the different methods are compared and methods for assessing model fit, with emphasis on graphical residual analysis, are discussed. Detailed NIST examples on nuclear tank calibration, thermal expansion of copper, and ozone spectroscopy are used to illustrate and compare the different methods.

NIST-SEMATECH e-Handbook of Statistical Methods and DATAPLOT: an Interactive Demonstration

Alan Heckert, Dennis Leber
September 29, 2004

The NIST/SEMATECH e-Handbook of Statistical Methods is a Web-based book for which the goal is to help scientists and engineers incorporate statistical methods into their work as efficiently as possible. It is hoped that the e-Handbook will serve as a useful educational tool that will help users of statistical methods and consumers of statistical information better understand statistical procedures and their underlying assumptions, and more clearly interpret scientific and engineering results stated in statistical terms. The e-Handbook has been integrated with the Dataplot statistical software.

The e-Handbook demo is presented in two parts. The first part is an introduction to the e-Handbook. We cover the basic chapters (exploratory data analysis, measurement process characterization, production process characterization, process modeling, process improvement, process or product monitoring and control, product and process comparison, and process reliability), and demonstrate running the case studies with the Dataplot software. The second part of the demo consists of a brief introduction to the Dataplot software. Students are invited to bring sample data sets on a floppy drive or CD-R.

Statistical Analysis of Incomplete Data for Scientists and Engineers

Grace Yang, Hung-kung Liu
December 13/15, 2004

This two half-day course introduces scientists to modern techniques for handling imputational issues for partially observed and missing data. Censoring mechanisms are discussed. The Kaplan-Meier estimator for the reliability function for right-censored data is covered. The EM algorithm is motivated as a computational device for Kaplan-Meier. More general types of censoring and regression analysis with censored data are discussed. A partially observed Poisson process is used for modeling missing data, and maximum likelihood estimators of parameters of interest and their uncertainties are derived. NIST case studies, including electromigration in microelectronics reliability, deadtime in phase Doppler interferometry recordings, partially observed neutron lifetimes, software reliability, and quality assurance for software embedded systems are used to motivate the specific imputation approaches covered.

COURSES TAUGHT BY SED STAFF IN 2004

January

- Hands-on Workshop on Estimating and Reporting Measurement Uncertainty
  Measurement Science Conference, Anaheim, CA
  Will Guthrie, Hung-kung Liu, Adriana Hornikova

- Estadística para Experimentos (Statistics for Experiments)
  University of Puerto Rico, Mayagüez, PR
  Ivelisse Aviles (taught in Spanish)

May

- Symposium on Statistical Methods for Analyzing Color Differences
  NIST (at the ASTM E12 - Color and Appearance Meeting)
  Ivelisse Aviles

August

- Introduction to Markov Chains: Markov Chain Monte Carlo for Scientists and Engineers
  NIST
  Andrew Rukhin, Stefan Leigh, Van Molino

September

- Introduction to Nonparametric Regression
  NIST
  Will Guthrie

- NIST/SEMATECH e-Handbook of Statistical Methods and DATAPLOT: An Interactive Demonstration
  NIST
  Alan Heckert and Dennis Leber

December

- Statistical Analysis of Incomplete Data for Scientists and Engineers
  NIST
  Grace Yang and Hung-kung Liu

COURSES PROJECTED FOR 2005

Hands-on Workshop on Estimating and Reporting Measurement Uncertainty
Will Guthrie, Hung-kung Liu, and Adriana Hornikova

Experimental Design for Calibration and Interlaboratory Studies
Dennis Leber and Ivelisse Aviles

Hands-on Workshop on Estimating Uncertainties for Chemical Analysis
Thomas Vetter (CSTL/ACD) and Will Guthrie

Regression Models
Will Guthrie

Introduction to Nonparametric Regression Analysis
Will Guthrie

Using WinBUGS for Bayesian Analysis of Physical Science and Engineering Data
Will Guthrie and Blaza Toman

Analysis for Incomplete Measurements
Hung-kung Liu & Grace Yang

Introduction to e-Handbook & DATAPLOT
Alan Heckert

Introduction to ANOVA
Dennis Leber & Stefan Leigh

Introduction to Markov Chain Monte Carlo
Andrew Rukhin & Stefan Leigh

Introduction to Principal Components Analysis
John Lu and Stefan Leigh

7.2 SED Seminar Series

Dale Newbury and David Bright (NIST), *Finding the Unexpected in X-Ray Spectrum Images: Developing Tools to Aid in Searching the Cube*, March 2004.

George Ostrouchov (Oak Ridge National Laboratory), *Data Intensive Analysis and Visualization Projects at ONRL*, May 2004.

Karol Marsina (Geological Survey of the Slovak Republic), *Geochemical Mappings Surveys*, June 2004.

Fern Hunt (NIST), *Visualizing the Frequency Patterns of DNA Sequences*, September 2004.

Seungseok Oh (Purdue University), *Nonlinear Multigrid Inversion for Bayesian Optical Diffusion Tomography*, September 2004.

Adriana Hornikova (NIST), *A Survey of Design, Analysis and Reporting of Results*, October 2004.

G. P. Patil (Pennsylvania State University), *Surveillance Geoinformatics of Hotspot Detection, Priortization, and Early Warning*, October 2004.

William Notz (Ohio State University), *The Design of Computer Experiments to Determine Optimum and Robust Control Variables*, November 2004.

7.3 Summer Students

Brian Cordes and Charles Hagwood
Statistical Engineering Division, ITL

Summer Students (left to right): Firouzeh Jalilian, Daniel Cogut, Van Molino, Abderahman Cheniour, Brian Cordes

Each year, for the last 4 years, the division has run a summer students program. Students are hired to spend approximately two months working on a project under the direction of one of our staff members. Then, they make a final presentation of their results to the division staff. Our students are chosen from a variety of resources. Our three main resources are: (1) The Department of Commerce's Post-Secondary Internship Program (PSIP), (2) The NIST Summer Undergraduate Research Fellowship (SURF) Program, and (3) The Statistical Engineering Division's Minority Internship Program.

The purposes of the PSIP are to use work experience to integrate academic theory and workplace requirements, to expose interns to the federal workplace and federal career opportunities, and to develop professional networks. Basic eligibility requirements are enrollment as an undergraduate or graduate student in a two- or four-year accredited educational institution, as well as U.S. citizenship. Interns receive stipends as well as paid round-trip transportation expenses between their homes/schools and NIST. Assistance with temporary housing arrangements is also provided. Interns are not employees of the Department of Commerce; rather, they are affiliated with one of four sponsoring organizations with which the Department of Commerce collaborates to recruit interns. The four sponsoring organizations are the American Indian Science and Engineering Society (AISES), the Hispanic Association of Colleges and Universities National Internship Program (HACU/HNIP), Minority Access, Inc. (MAI) and the Oak Ridge Associated Universities (ORAU). In addition to these programs, we welcome and consider applications from all students who express an interest in working in the Statistical Engineering Division.

The NIST SURF program is a partnership, supported by NIST, NSF, and participating colleges/universities. The program is open to students majoring in science, mathematics and engineering, whose universities are participating institutions. Applications for participation in the SURF program are only accepted from colleges and universities, and not from individual students. Students are selected by the participating universities from a pool of applicants. Students have to be undergraduates at a U.S. university or college with a scientific major, a G.P.A of 3.0/4.0 or better, intend to pursue a Ph.D., and must be covered by a health insurance plan (either through school or family). Students attend weekly seminars of outside speakers and partake of other science seminars available at NIST. They present their own research results to NIST scientists and other SURFers at the end of the summer. They receive a $4,000 stipend for the summer (for 12 weeks, prorated for shorter periods if their school year does not permit the full 12 weeks).

The SED Minority Internship is a recent program initiated to recruit minority summer students. The program began in 2001 and has been quite successful. The purpose of the program is to interest more minorities in the field of statistics and probability. An announcement is sent out annually to Historically Black Colleges and Universities, to Hispanic Serving Institutions, and to other colleges and universities.

2004 Summer Students

1. Abderahman Cheniour, I.S.T.I.L (Institut des Sciences et des Techniques de L'ingenieur de Lyon) at the University Claude Bernard Lyon - Mentors Juan Soto and Don Malec
2. Daniel Cogut, The College of William and Mary (SURF) - Mentor Andrew Rukhin
3. Brian Cordes, Worcester Polytechnic Institute (SURF) - Mentor Ivelisse Aviles
4. Firouzeh Jalilian, University of Maryland College Park (SURF) - Mentor Alan Heckert
5. Van Molino, Princeton University (SURF) - Mentors Stefan Leigh and Andrew Rukhin

ABSTRACTS

Abderahman Cheniour
Sensitivity Testing for Bayesian 2D and 3D Tomography

The primary objective of the 3D chemical study is to understand the 3D spatial distribution of chemical species in materials at the nano-scale level. The information available for doing this is in the form of non-destructive, transmission electron microscopy measurements. To achieve this objective, the dual aim of the study to extend Bayesian methods of tomographic reconstruction from 2 dimensions to 3 dimensions and to utilize all electron microscope information pertinent to chemical species identification, such as the electron energy loss spectra and X-ray spectra. After developing a new 3D Bayesian algorithm, the next step is to evaluate its properties under a variety of conditions. This project specifically evaluates how well a proposed Bayesian method can reconstruct images (both 2D and 3D) using known inputs. The results of this testing will be useful in improving the proposed method and for assessing its computational efficiency.

Daniel Cogut
Fusion of Biometric Algorithms

Biometric identification plays an important role in security today. Biometric technology allows for recognition of an individual's identity based on certain characteristics, called signatures. These characteristics can include facial features, fingerprints, and vocal expressions. A signature of an unknown individual, called a probe, is compared to a database of known individual's signatures, called a gallery. Then a biometric algorithm produces similarity scores between the probe and each signature in the gallery and ranks them. Despite the fact that many biometric algorithms are now available for commercial use, there is no optimal algorithm that is widely accepted. Thus it is practical to create aggregations or fusions of various algorithms. In this project, the similarity scores produced by four biometric algorithms were used, with data from the Face Recognition Technology (FERET) program. The method of fusion was implemented by means of weighted averaging of the ranks of the similarity scores.

Brian Cordes
Developing a Graphical Tool to Determine the Estimation Capacity of an Experimental Design

In scientific research, the quality of the design of an experiment has a great deal of impact on the quality of the experimental results. Experimental design is concerned with exactly this:

finding the most efficient method for obtaining the most amount of information about a process. By using a certain class of experimental designs called assembled designs, Aviles (2001) has focused on determining optimal designs for robustness experiments. Assembled designs are very useful to study a process in which batches are produced and samples are made. In this case, the batch-to-batch (between-batch) variance and the sample-to-sample (within-batch) variance are known as the variance components. The goal of robustness experiments is to find the settings of the factors so that the process mean is on target and the variance is minimized. Location effects are effects on the response mean and dispersion effects are effects on the response variance. The model used in these scenarios is a linear mixed-effects model with two variance components (between-batch variation and within-batch variation).

The concept of the Precision Plot has been developed to help the scientist determine the expected effectiveness of a design. This is accomplished by identifying how large the dispersion effects and variance components must be in order for the design to detect them. That is, if smaller effects must be detected, then the experimental budget must be increased (i.e., need a larger design) or the level of confidence in our results is decreased (for this design). Similarly, resources could be saved by running a smaller experiment if it is determined that the estimation capacity of a design is greater than what is practically needed. The work done this summer includes both refining the Precision Plot method and improving the presentation of the outputted results. Precision Plots are calculated using normal theory and asymptotic approximations. Research done in the future will include establishing the adequacy of the approximation when sample sizes are much smaller.

Firouzeh Jalilian
Developing a Graphical User Interface for the Mass Calibration Program

The Statistical Engineering Division and the Manufacturing and Metrology Division are rewriting the Masscomp program that is used to assign mass values to weights submitted to NIST for calibration. The current production software was written in the 1970s using the programming language Fortran 66. The code is being modernized to Fortran 90. However, the program could only be run in a shell such as MS DOS. My project has been to develop a Graphical User Interface for this calibration program in order to make it easier to use. Different approaches were investigated and it was decided that the programming language Visual Basic would be most suitable for developing this interface. In addition to learning Visual Basic (VB) and Fortran 90, I also had to learn Fortran/Visual Basic Mixed-Language programming. The project involved creating a Fortran Dynamically Linked Library (DLL) and exporting its routines into VB. These routines could be called by VB during execution of the graphical user interface. The next stage of the project is to add a feature that plots the results at the end of the program. Once completed, this graphical user interface will make it significantly easier and faster to use the Masscomp program. In addition, Visual Basic proved to be a versatile tool for developing a graphical user interface.

Van Molino
Markov Chains and MCMC Methods

The concept of a Markov chain was introduced in 1906 by A. A. Markov. The usefulness of the Markov chain as a modeling tool was quickly realized and by the 1960's so much work had been done that many considered the field to be an unlikely source of new research. However, when Markov chain Monte Carlo (MCMC) algorithms were introduced, the field enjoyed a spectacular revival and exciting new work is being done using MCMC in many different disciplines ranging from mathematical logic to ecology.

Because of the wide applicability, my advisors and I have developed a two-day course on these topics that will be taught to NIST scientists in late August. The first day will cover Markov chain theory, which is both powerful in its own right and should be understood in order to utilize MCMC algorithms intelligently. The second day will cover MCMC methods and some of the interesting problems to which these methods have been applied. This talk will survey some basic Markov chain theory, emphasizing the theory's versatility by applying it to several simple problems in different fields of study. MCMC methods will be introduced briefly. To illustrate the power of these simulation tools, this talk will discuss how this machinery can be applied to the satisfiability problem of mathematical logic, a famous problem shown to be NP-complete by Stephen Cook in 1971.

Accolades of Appreciation

"I'd like to thank everyone in the SED for all their help and hospitality throughout this great summer program. The SED was a great place to work because everyone was both easy and fun to work with. Whenever I needed help with something, whoever I asked was always willing to take time out of their schedule to make sure that I resolved my issues. Being able to work with researchers one-on-one was really a great opportunity for me because of my interest in pursuing an advanced degree. I definitely gained a solid amount of valuable experience and information relating to the field of research. Most of all I'd like to thank my own advisor Ivelisse Aviles for her devotion to my project and my summer experience. Whatever I needed she always helped me find. I hope everyone in the division continues to prosper with their work and thanks again for a great summer!"

- Brian Cordes

"This has been my second summer working with the Statistical Engineering Division at NIST and, once again, I have nothing but good things to say about the experience. Everyone in this division that I have ever interacted with has been great. They are always willing to take time out of their busy days to explain the research in which they are involved or even to discuss things like my academic interests or my future career plans. It is clear that Dr. Sedransk takes great interest in making sure that the students placed in this division have an enlightening and enjoyable summer.

Last summer I learned that members of the SED not only perform world-class research in the field of statistics, but that they also help scientists all around NIST design good experiments and analyze their data effectively. This summer I experienced some of this second aspect of the SED when I was allowed the opportunity to help develop and teach a course along with Dr. Andrew Rukhin and Stefan Leigh. Both Andrew and Stefan have been very helpful throughout the summer and have given me excellent guidance. I would especially like to thank Stefan for all the time he spent with me making sure I understood and enjoyed my work. Once again, I appreciate what you have all done for me and wish you success in your future research. I hope we have the opportunity to interact more in the future."

- Van Molino

8 Conferences and Standards Activities

8.1 International Organization for Standardization (ISO)

Nien Fan Zhang
Statistical Engineering Division, ITL

ISO/TC69 on Applications of Statistical Methods is an international standards group that develops generic statistical standards. The technical committee (TC) has four active subcommittees that develop documents in the following subject areas:

SC1-Vocabulary,
SC4-Process Control,
SC5-Acceptance Sampling, and
SC6-Measurement Methods.

Within each subcommittee, there are working groups for different areas. Each member country of the TC has a Technical Advisory Group (TAG) that sends a delegation to the international meetings; develops strategies and positions for advancing the interests of national industry via the standards arena; and coordinates the dissemination and critiquing of standards under development. The Statistical Engineering Division (SED) has a long history of supporting the development of international standards, particularly those that impact measurement science. SED participates at ISO/TC69 and its SC 6 on Measurement Methods and SC 4 on Process Control. Nien Fan Zhang of the SED is a member of US TAG.

The document ISO/TC69/SC 6 TS 21749: "Measurement uncertainty for metrological applications - Repeated measurements and nested experiments" has been approved and is being published by ISO as a Technical Specification of ISO. The document is intended for metrological and scientific laboratories that are capable of collecting data to evaluate both short-term and long-term sources of error in the measurement process and have the capability of performing statistical analyses. The document was originally initiated by Carroll Croarkin of SED. Nien Fan Zhang is the current project leader.

During the 2004 TC69 meeting held in Stockholm, ISO/TC 69/SC 6 resolved to develop a stronger working relationship with the Joint Committee for Guides in Metrology (JCGM) of the International Bureau for Weights and Measures (BIPM) regarding works related to the evaluation and use of uncertainty of measurement. Two draft documents prepared by JCGM: Revision of International Vocabulary of Basic and General Terms in Metrology (VIM) and Guides to the Expression of Uncertainty in Measurement Supplements 1: Numerical Methods for the Propagation of Distribution were sent to ISO/ TC69 for comments. For the second draft, SED supplied comments on GUM Supplement to TC 69/SC 6. In another resolution, SC 6 resolved to change the title of SC 6/WG 7 to “Statistical methods to support measurement uncertainty evaluation.”

During the SC 6 meetings, Nien Fan Zhang was elected to be the Convenor of TC69/SC 6/WG 4, which is for Statistical Aspects of the Preparation and Use of Reference Materials.

8.2 ASA/IMS Spring Research at NIST

Will Guthrie
Statistical Engineering Division, ITL

Karen Kafadar and Dana Franklin
Mathematics Department, University of Colorado - Denver

Tom Loughin
Statistics Department, Kansas State University

11th Spring Research Conference on Statistics in Industry and Technology

“Statistics on Data Streams for Scientific Research and Implementation”

May 19-21, 2004
National Institute of Standards and Technology (NIST)
Gaithersburg, MD

Front cover of the conference program and photos of the student scholarship winners, conference attendees discussing one of the plenary talks, and James J. Filliben closing the conference with a talk titled, “*The World Trade Center Collapse: The Critical Role of Statistics in the Diagnosis of a Failure of an Engineering Marvel.*”

The Spring Research Conference (SRC) is an annual conference sponsored jointly by the Institute of Mathematical Statistics and the Section on Physical and Engineering Sciences of the American Statistical Association. It provides a forum for promoting statistics in engineering, technology, industry, information and physical sciences. The SRC attracts an international group of statisticians, engineers and physical scientists from universities, industry, and government.

The 11th annual Spring Research Conference on Statistics in Industry and Technology was held May 19-21, 2004, in Gaithersburg, Maryland and was hosted by the NIST Statistical Engineering Division. The theme of the conference was "Statistics on Data Streams for Scientific Research and Implementation."

Edward Wegman (George Mason Univ.) opened the conference with the keynote address, "*Visualization of Internet Traffic Data.*" He emphasized that new data types historically have inspired new developments in statistics, and massive data sets and streams are new data types that present such opportunities. Subsequent plenary speakers at the conference who described other problem areas where large data sets are demanding new ways of visualizing and analyzing data included: Robert Jacobsen, Lawrence Berkeley Laboratory (high-energy physics), Donna Stroup, Centers for Disease Control (public health), Vijay Nair, University of Michigan (internet tomography), and James Filliben, NIST (computational modeling of the WTC). Invited sessions were also organized around these themes; e.g., internet traffic modeling (organized by James Landwehr), physics (Kevin Coakley), clock synchronization (William Notz), data visualization (Barbara Bailey), business applications (Bonnie Ray), computer experiments (Michael Trosset and Lisa Moore), and data mining (David Banks).

The conference also had an excellent contributed program with a wide range of talks and active student participation. The twenty-seven contributed talks included "*Dynamic Profiling of Online Auctions Using Curve Clustering*" by Wolfgang Jank and Galit Shmueli, which described an application of functional data analysis to eBay auction data to display different types of bidding behavior, "*Empirical Bayes Estimation of Crash Reduction Factors*" by Peter Hovey and Mashrur Chowdhury, which described the use of empirical Bayes methods for reducing the impact of regression to the mean when evaluating crash reduction factors computed for roadway improvements made at sites chosen by crash prevalence, and "*Brain Activity at Nine Million Bytes per Second: Synchronizing Video Data and Physiological Recordings to Reveal Brain Function*" by Kary Myers, which explored the use of covariates such as blood pressure and expired carbon dioxide to "clean" video data of the brain activity of laboratory animals of physiological fluctuations unrelated to the brain stimulus of interest.

Among the student participants at the conference were six winners of scholarships that supported student attendance. This year's winners, pictured from left to right in the first row of the upper-right photograph on the preceding page, were Deng Huang (Ohio State), Shiling Ruan (Ohio State), Kary Myers (Carnegie Mellon), Gavin Richards (Ohio State), and Tirthankar Dasgupta (Georgia Tech). David Hutchison (Johns Hopkins) is not pictured.

Active participation by NIST staff in conferences like the SRC helps maintain an awareness of current trends in statistical research and fosters recognition of the Division outside of NIST.

8.3 NIST Quality System

Adriana Hornikova, Stefan Leigh, Nien Fan Zhang
Statistical Engineering Division, ITL

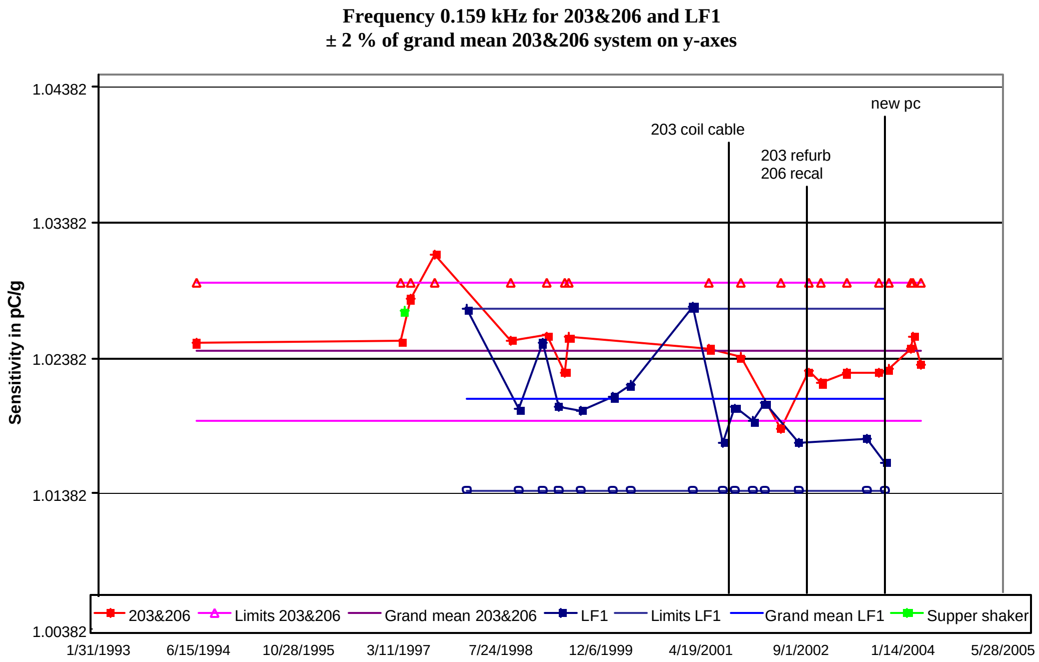
David Evans
Manufacturing Metrology Division, MEL

NIST simultaneously with other National Metrology Laboratories has undertaken the implementation of a formal agency-wide Quality System for its measurement services to document compliance with ISO/IEC 17025. The NIST Measurement Services Advisory Committee formed a Working Group under the leadership of Willie May to draft a NIST Quality Manual detailing the requirements and procedures for documentation and assessment of calibration services. Nien Fan Zhang served on the Working Group and held special responsibility for statements of requirements dealing with uncertainty computation and certification. A second group of measurement scientists constituted as a Task Force, convened to address the Quality System requirements for Standard Reference Materials (SRMs), with Nien Fan Zhang appointed to the working group and responsible for statistical issues. The completed NIST manuals were adopted formally and the ambitious assessment process was implemented immediately for NIST measurement services. At the Regional Metrology Organization meetings, quality assessments for NIST measurement services have been approved for international acceptance, effectively certifying acceptance of NIST's stated measurement capabilities for the specific listed measurement services.

At the practical level, the Statistical Engineering Division also provided statistical expertise to metrologists to establish formal process control and quality monitoring procedures and/or to characterize and document measurement system performance to meet the reporting guidelines for the NIST Quality System and ISO 17025.

For vibration measurements made in the Acoustics and Vibration Laboratory, the quality assessment process included the statistical analysis and validation of measurement uncertainties associated with several different calibration systems (all using sinusoidal excitation) that are used for vibration measurements at NIST. These systems use both primary and secondary methods to generate measurements of the sensitivity of accelerometers in terms of electrical output per unit of applied acceleration. An additional high-precision calibration system, the "Super Shaker," is not typically used to provide calibrations for NIST customers, but has been used as a benchmark device for high-precision internal calibration and for calibrations with special requirements such as international intercomparisons in support of the MRA.

NIST calibration records for the past decade provided the data for examining the performance of the several systems in terms of stability, sensitivity of sensors used in the measurement devices, and recalibrations over that time. Graphical tools for displaying analytical results also show control limits and interventions (recalibration, etc.).



Control limits of ± 2 % are shown in magenta and in black for two of the measurement systems with their calibration values plotted in red and blue, respectively. A single point in green for the supershaker serves to confirm the calibration of the red measurement device.

8.4 Conferences

- On January 12-13, Will Guthrie and Hung-kung Liu gave a "Hands-on Workshop on Estimating and Reporting Measurement Uncertainty" to metrologists at the 2004 *Measurement Science Conference* in Anaheim, CA. Adriana Hornikova served as a teaching assistant. The workshop, attended by sixteen participants from industry and government, described the statistical framework and methods needed to develop uncertainty statements based on the "ISO Guide to the Expression of Uncertainty in Measurement." Methods for uncertainty estimation were illustrated with many practical examples from different metrological areas, including mass, temperature, and chemical measurements. The workshop also included hands-on examples to be analyzed by the participants. The hands-on examples wer e done using both propagation of uncertainty formulas and the Kragten spreadsheet, an easy-to-use computational tool for propagation of uncertainty. Jack Wang also attended the conference.

- Nell Sedransk, Nien Fan Zhang, and Will Guthrie participated in the *National Nanotechnology Initiative (NNI) Workshop* on Instrumentation and Metrology for Nanotechnology, held January 27-29, 2004 at NIST in Gaithersburg, Maryland. The purpose of this workshop was to convene a wide range of scientists and engineers working in the fields of instrumentation, metrology, and nanotechnology to identify needs and opportunities for nanometrology research. This workshop was one of a series of workshops being sponsored by the NNI member agencies to address the NNI's "grand challenges." Workshop attendees participated in open discussions of grand challenges in nanotechnology and contributed to the drafting of the final report from the workshop.

- The article “The Recognition Problem of Biometrics” by Andrew Rukhin was published in the issue of *Chance* magazine dedicated to statistical issues in counter-terrorism This paper discusses the use of statistical measures of dependence for biometric algorithms and techniques for their fusion. Dr. Rukhin was invited to present these results at the *Spring Biometrics Meeting* in Pittsburgh in March 2004.

- Kevin Coakley attended the *International Conference on Precision Measurements with Slow Neutrons* held at NIST in Gaithersburg, MD on April 5-7, 2004. The conference served as a forum for highlighting progress in the field of fundamental physics using cold and ultracold neutrons, and development of new neutron sources. “Measuring the neutron with magnetically trapped neutrons” and “Development of a long wavelength neutron monochromator for superthermal production of ultracold neutrons” were talks that were presented based on the work of a collaborative research team, which included Coakley and G.L. Yang of SED.

- On May 18, 2004, Ivelisse Aviles offered a Symposium on Statistical Methods for Analyzing Color Differences at the *ASTM E12, Color and Appearance Meeting*, which was held at NIST on May 17-20, 2004. Other presenters included Maria Nadal, Cameron Miller, and Ted Early from the Optical Technology Division, Physics Laboratory. The objective of this meeting was to advance the development of standards for characterizing color and appearance attributes.

- The 11th annual *Spring Research Conference on Statistics in Industry and Technology* was held May 19-21, 2004, in Gaithersburg, Maryland and was hosted by the Statistical Engineering Division. The theme of the conference was “Statistics on Data Streams for Scientific Research and Implementation.” All SED staff attended the conference and served different roles: Will Guthrie was the local chairperson; Kevin Coakley organized and chaired a session on “Design and Analysis for Fundamental Physics and Metrology”; Ivelisse Aviles was the discussant of a session on “Data Mining”; and Jim Filliben closed the conference with a talk entitled, “The World Trade Center Collapse: The Critical Role of Statistics in the Diagnosis of a Failure of an Engineering Marvel.”

- The 36th *Symposium on the Interface: Computing Science and Statistics* was held in Baltimore, May 26-29, 2004. John Lu organized a session “Statistical and Metrological Issues in Proteomics Using Time-of-Flight Mass Spectrometry.” Because the presenting author could not attend, Adriana Hornikova (SED) presented the first paper, “Exploring Bioinformatics in Serum Proteomic Analysis for Early Detection of Prostate Cancer.” Walter Liggett (SED) presented the second paper, “Data-Driven and Peak-Based Feature Selection in Serum Protein Mass Spectrometry.” John Lu (SED) presented the third paper, “SVD-based Functional ANOVA for Measurement Evaluation of MALDI-TOF Mass Spectrometry.” Zhen Zhang (Johns Hopkins) presented the fourth paper, “Bioinformatics for Clinical Proteomics: Usage and Abuseage.”

- Kevin Coakley, Sarah Streett, and Jolene Splett attended the short course “Applied Spatial Statistics,” presented by Dr. Jay Ver Hoef on June 16, 2004, and the *Graybill Conference: Spatial Statistics* on June 17-18, 2004. Both were held in Fort Collins, CO.

- Blaza Toman and Will Guthrie attended the *NCSLI Conference* in Salt Lake City from July 11-15, 2004. They participated in the organization and presentation of a panel discussion “GUM and Supplements: Conventional vs. Bayesian Methods” with colleagues from the Navy Calibration Laboratories, Duke University, and ITRI, the Taiwanese Metrology Institute. The panel session was very well attended and drew many questions and comments. Blaza also organized an invited session on Bayesian methods in metrology, where Will presented a paper titled “MCMC in STRD” and Blaza presented the paper “A Bayesian Approach to the Estimation of a Key Comparison Reference Value”.

- Kevin Coakley attended the 2004 *Meeting of the International Federation of Classification Societies* held at Chicago, IL on July 15-18, 2004 and presented an invited talk “Classification Problems in Neutrino Physics”. The conference hosted leading researchers working in the fields of classification, clustering, and data mining.

- In July 2004, Jack Wang, John Lu, and Nien Fan Zhang of the Statistical Engineering Division (SED) participated in the 2004 *Summer Topical Meeting of the American Society for Precision Engineering* (ASPE) at State College PA. The topic of the meeting was "Uncertainty Analysis in Measurement and Design". The meeting focused on measurement uncertainty in engineering and manufacturing. Many of the presentations dealt with standards issues for geometric objects. The first SED presentation described a simulation procedure for uncertainty evaluation in measurement results. The proposed method is relevant since a draft of the supplement (Numerical Methods for the Propagation of Distributions) to the ISO Guide to the Expression of Uncertainty in Measurement (GUM) by the Joint Committee for Guides in Metrology (JCGM) of the International Bureau of Weights and Measures (BIPM) is being circulated for final comments. The second SED presentation discussed the Bayesian statistics approach and how Bayesian statistics provides a unifying approach to uncertainty analysis. The presentation also discussed a full Bayesian approach to combining means from multiple methods. The third SED presentation described a statistical analysis of Key Comparisons -- a special type of interlaboratory study with linear trends.

- In August, eight staff members from the Statistical Engineering Division participated in the 2004 *Joint Statistical Meetings* (JSM) in Toronto. Presentations by SED staff included "Statistical Process Control on Biochemical QC Data" by Nien Fan Zhang, "Generalized Prediction Intervals" by Jack Wang, "An Update on the NIST Statistical Reference Datasets for MCMC" by Hung-kung Liu, "A Survey of Design, Analysis, and Reporting of Results in Key Comparisons" by Adriana Hornikova, and "Ranking the Sources of Numerical Error in MCMC Computations" by Will Guthrie. Ivelisse Aviles chaired a session on the design and analysis of experiments for complex biological experiments run on 96 well plates. Several staff members were also involved in Committee activities, including Ivelisse Aviles who is currently serving as Awards Chair for the Section on Physical and Engineering Sciences (SPES) and as ASA Representative to the American Association for the Advancement of Science (AAAS), Section P - Industrial Sciences. Adriana Hornikova, Will Guthrie, and Sarah Streett served on the Awards Committee for SPES. In addition, Nell Sedransk completed her term on the Deming Lectureship Committee and Will Guthrie completed his term as Secretary of the Quality and Productivity (Q&P) Section. Highlights of the conference included the Deming Lecture, "Deming and Bell Labs" by Colin Mallows, the Presidential Address, "Bayesians, Frequentists, and Scientists" by Brad Efron, and the President's Invited Address, "The Romance of Hidden Components" by David Donoho. Two SED members, Jack Wang and Will Guthrie, were also recognized by the Section on Physical and Engineering Sciences for their presentations at the 2003 Joint Statistical Meetings.

- James Yen collaborated in a geochemical investigation, whose results were presented by Andrew Grosz of the U.S. Geological Survey at the 32nd *International Geological Congress* at Florence, Italy on August 22, 2004. The one-hour talk, "Arsenic in Surficial Sediments of North America," co-authored by scientists at the U.S, Canadian, and Mexican Geological Surveys as well as NIST, was part of the Geological Congress's Workshop on Global Geochemical Baselines. The study analyzed the distribution of arsenic and other elements in North American soils and stream sediments. The data analysis uncovered patterns and anomalies that can be interpreted in terms of both the underlying geology of the regions and anthropogenic land-use effects.

- Jeffrey Fong (Mathematical and Computational Sciences Division) and Jim Filliben participated in a DOD-sponsored workshop, *Foundations '04: A Workshop for VV&A (Verification, Validation, & Accreditation) in the 21st Century* in Tempe, AZ on October 13-15. Prior to that conference, they visited Southwest Research Institute (San Antonio) to learn of the V&V work of Ben Thacker (author of "NESSUS: a probabilistic analysis tool for improving safety and reliability of complex systems") and Chris Freitas (whose fellow staff member showed them the SWRI on-site test results for the Columbia foam impact disaster). Immediately after the conference, Fong and Filliben visited Lawrence Livermore to give a talk and learn of Livermore's V&V work.

- During October 24-25, 2004, Nien Fan Zhang attended the *INFORMS annual meeting* in Denver, Colorado. INFORMS stands for the Institute for Operations Research and the Management Sciences. This society with 12,000 members is playing an important part in making decisions. Models that are based on the core foundations of operations research are assisting governments around the world in tackling issues of utmost importance, such as those related to intelligence, emergency systems and the environment. Nien Fan Zhang was invited to give a presentation entitled "Statistical Process Monitoring for Autocorrelated Data" in the invited session, Process Monitoring and Diagnosis for Autocorrelated data. In the session there were four speakers from academia and government, including an associate editor of *Technometrics*.

- On October 25-29, 2004, John Lu attended a one-week meeting that is part of the NSF-funded *UCLA/IPAM* (University of California at Los Angeles/Institute of Pure and Applied Mathematics) workshop on multiscale structures in the analysis of high-dimensional data, which ran from September 7 - December 17, 2004. The *UCLA/IPAM* workshop is one of a series of workshops on mathematical, computation al and statistical issues related to multiscale geometry, machine learning, and statistical theory. For example, the workshop contains a big segment on recent developments in 3-D image analysis and low-dimensional structure embedding, which can be traced back to multidimensional scaling in psychometrics in the 1930s. Some exciting developments on support vector machines (SVM) were also presented, such as kernel methods for canonical correlation analysis and dimension reduction, a Laplacian operator approach to geometric structure in high-dimensional data analysis, and SVM methods for unlabeled and partly labeled data.

- On November 8-9, 2004, a NIST/DOD Workshop entitled *Verification and Validation of Computer Models for Design and Performance Evaluation of High-Consequence Engineering Systems* was held in Gaithersburg. This workshop was well-attended with excellent representation from government, DOD, and industry. Among the many presentations at the workshop, SED was well represented with talks by Nell Sedransk, Will Guthrie, and Jim Filliben. The workshop demonstrated clearly that V&V as a discipline has many potential directions, with diverse technical needs/desires by the various members of the large V&V community. Feedback from this workshop served (in a very constructive way) to enumerate and focus the many challenges and opportunities existent for NISTSED/MACD to beneficially accelerate the national V&V effort.

- Will Guthrie and Dean Ripple (CSTL Process Measurements Division) were invited to give a seminar on the use of guardbands in thermocouple calibration by the members of the *ASTM Committee E20 on Temperature Measurement*. The talk, titled "Validating the Temperature-EMF Response of Thermocouples with Confidence", was given on November 9, 2004 as part of an ASTM Committee Week held in Washington, DC. The talk was followed by a lively discussion of calibration issues facing thermocouple manufacturers and their suppliers. Based on this introduction to the use of guardbands, the Committee agreed to investigate further the impact that guardbanding would have on thermocouple calibration procedures to determine if current thermocouple calibration standards should be updated to mandate the use of guardbands.

- Jolene Splett attended the *Second Symposium on Pendulum Impact Machines: Procedures and Specimens* on November 10, 2004 in Washington D.C., and presented the paper, “Analysis of Charpy Impact Verification Data: 1993-2003”, which was co-authored by Chris McCowan of the Materials Reliability Division (MSEL). Jolene and Chris also contributed to a second conference paper, “International Comparison of Impact Reference Materials,” which was the result of a joint effort by four national measurement institutes.

- John Lu attended the *Critical Assessment of Microarray Data Analysis* (CAMDA 2004) workshop at Duke University on November 10-12, 2004. This is the fifth international conference of this series organized by Duke University on microarray data analysis. The participants included about 80-120 researchers and students from all over the world from both academia and industry. The goal of the workshop was to discuss the most useful and important statistical and computational techniques for some pre-selected microarray data sets released early every year, given in about 20 papers chosen by the scientific committee in July. This year's data was on the gene expression levels during the four-stage development cycles of Plasmodium falciparum, measured by the P. falciparum specific DNA microarray using long oligonucleotides. The dataset is essentially a large spatial-temporal dataset, for which analysis techniques using spatial correlation, time series, Bayesian methods, and graphical network modeling were employed.

- On November 22-23, 2004, Ivelisse Aviles attended the *MR Workshop on Translational Research in Cancer - Tumor Response*, in Bethesda MD. NIST's interest for partnership in working with the medical imaging community in standardization and infrastructure of imaging databases was acknowledged. Dr. Aviles briefly introduced NIST and the response was very positive since NIST can help standardization, development of novel statistical design and analysis techniques, and infrastructure on the imaging databases.

9 Staff and Professional Activities

9.1 New Staff

9.1.1 Abderahman Cheniour

BIOGRAPHICAL SKETCH

Abderahman Cheniour has been a guest researcher in the Statistical Engineering Division at NIST since June 2004. He carried out his internship at the end of his studies to obtain his M.S. degree from the ISTIL (Institut des Sciences et Techniques de l'Ingénieur de Lyon) at the French University Claude Bernard in Lyon. His M.S. is in the area of modeling and computer science and he also has a background in statistics. This is the first time that he has visited the USA.

STATISTICAL RESEARCH

The main project that he is working on is the implementation of a Bayesian algorithm for Image Reconstruction. The primary objective of the 3D chemical study is to understand the 3D spatial distribution of chemical species in materials at the nanoscale level. The information available for doing this is in the form of nondestructive, transmission electron microscopy measurements. To achieve this objective, the dual aim of the study is to extend Bayesian methods of tomographic reconstruction from 2D to 3D and to utilize all electron microscope information pertinent to chemical species identification, such as the electron energy loss spectra and X-ray spectra. After developing a new 3D Bayesian algorithm, the next step is to evaluate its properties under a variety of conditions. This project specifically evaluates how well a proposed Bayesian method can reconstruct images (both 2D and 3D) using known inputs. The results of this testing will be useful in improving the proposed method and for assessing its computational efficiency.

9.1.2 Sarah Streett

BIOGRAPHICAL SKETCH

Sarah Streett has been a mathematical statistician in the Statistical Engineering Division at NIST in Boulder, Colorado since January 2004. She received her Ph.D. degree in Statistics from Colorado State University in August 2000, an M.S. degree in Mathematics from Northern Arizona University in May 1994 and a B.A. in Mathematics with emphasis in Computer Science from Hendrix College in May 1992.

Prior to joining SED, Sarah was a postdoctoral researcher at the National Center for Atmospheric Research in Boulder, Colorado. While at NCAR, she gained valuable experience in working with large data sets and interdisciplinary collaboration. Her research interests include time series, stochastic processes and spatial statistics.

STATISTICAL RESEARCH

Sarah's main area of statistical research is with the Time and Frequency Division. She is currently working on a project involving clock error uncertainty. She continues to build collaborative relationships with members of the Time and Frequency Division.

9.2 Publications

9.2.1 Publications in Print

1. C. F. Ferraris, V. A. Hackley, and A. I. Aviles (2004), "Measurement of Particle Size Distribution in Portland Cement Powder: Analysis of ASTM Round-Robin Studies," *Cement, Concrete and Aggregate Journal*, 26(2).
2. J. M. Irvine, C. Fenimore, D. Cannon, J. Roberts, S. A. Israel, L. Simon, C. Watts, J. D. Miller, A. I. Aviles, P. F. Tighe, R. J. Behrens (2004), "Feasibility Study for the Development of a Motion Imagery Quality Metric," *Proceedings of the 33rd Applied Imagery and Pattern Recognition Workshop: Image and Data Fusion*, IEEE Computer Society, Washington, 13-15 October 2004.
3. K. J. Coakley, D. S. Simons, A. M. Leifer (2004), "Secondary Ion Mass Spectrometry Measurements of Isotopic Ratios: Correction for Time Varying Count Rate," *The International Journal of Mass Spectrometry*.
4. D. K. Walker, K. J. Coakley, J. D. Splett (2004), "Nonlinear Modeling of Tunnel Diode Detectors," *Proceedings of the 2004 IEEE Geosciences and Remote Sensing Society Symposium*, Anchorage, Alaska, September 20-24, 2004.
5. K. J. Coakley, D. J. McKinsey (2004), "Spatial Methods for Event Reconstruction in CLEAN," *Nuclear Instruments and Methods in Physics Research A*, 522, 504-520.
6. D. Dey, Y. Wang (2004), "Wavelet modeling of priors on triangles," *Journal of Multivariate Analysis*.
7. A. Micheas, D. Dey (2004), "Modeling shape distribution and inferences for assessing differences in shapes," *Journal of Multivariate Analysis*.
8. A. Rupp, D. Dey, B. Zumbo (2004), "To Bayes or not to Bayes: Application of Bayesian methodology to item response modeling," *Structural Equations Modeling*.
9. M. H. Chen, D. Dey, J. Ibrahim (2004), "Bayesian criterion based model assessment for categorical data," *Biometrika*.
10. G. A. Klouda, J. J. Filliben, H. J. Parish, J. C. Chow, J. G. Watson, R. A. Cary (2004), "Reference Material 8785: Air Particulate Matter on Filter Media," *Proceedings of the Symposium on Air Quality Measurement Methods and Technology 2004*, Research Triangle Park, NC, April 2004, CD-ROM ISBN 0-923204-62-8.
11. A. Cope, K. Gurley, S. Hamid, J. P. Pinelli, C. Subramanian, L. Zhang, E. Simiu, J. J. Filliben (2004), "Hurricane Damage Prediction Model for Residential Structures," *Journal of Structural Engineering* 130, 1685.
12. S. J. Wetzel, C. M. Guttman, K. M. Flynn, J. J. Filliben (2004), "The Optimization of MALDI-TOF-MS for Synthetic Polymer Characterization by Factorial Design," *Proceedings of the 52nd ASMS Conference on Mass Spectrometry and Allied Topics*, May 2004, Nashville, TN.
13. W. F. Guthrie, H.-K. Liu, D. Malec, G. Yang (2004), "MCMC in StRD," *Proceedings of the 2004 NCSLI Conference.*
14. H. Liu, W. F. Guthrie, D. Malec, G. L. Yang (2004), "MCMC in StRD," *Proceedings of the 2003 Quality and productivity Conference.*
15. Z. Q. Lu, W. F. Guthrie (2004), "Bayesian Methods for Statistical Inference on the Common Mean from Multiple Data Sources," *Proceedings of ASPE 2004 Summer Topic Meeting - Uncertainty Analysis in Measurement and Design*, 95-99.
16. W. F. Guthrie, C. D. Simon, F. W. Wang (2004), "Cell Seeding into Calcium Phosphate Cement," *Journal of Biomedical Materials Research*, 68A, 628-639.
17. W. Liggett, P. E. Barker, O. J. Semmes, L. H. Cazares (2004), "Measurement Reproducibility in the Early Stages of Biomarker Development," *Journal Disease Markers*, 20, 295-307.

18. H. Liu, C. R. Hagwood, N.F. Zhang (2004), "Comparisons of Bayesian approaches to combining results from multiple methods," *Proceedings of the 2004 Measurement Science Conference*.
19. H. Iyer, C. Wang, T. Matthew (2004), "Models and confidence intervals for true values in interlaboratory trials," *Journal of the American Statistical Association*, 99, 1060-1071.
20. H. Iyer, C. Wang, D. F. Vecchia (2004), "Consistency tests for key comparison data," *Metrologia*, 41 (4), 223-230.
21. S. Leigh, D. L. Poster, M. M. Schantz, S. A. Wise (2004), "Standard Reference Materials (SRMs) for the Calibration and Validation of Analytical Methods for PCBs (as Aroclor Mixtures)," NIST *Journal of Research*, 109, 245-266.
22. R. B. Marinenko, S. Leigh (2004), "Heterogeneity Evaluation of Research Materials for Microanalysis Standards Certification," *Microscopy and Microanalysis*, 10, 491-506.
23. R. B. Marinenko, J. R. Sieber, L. L. Yu, T. A. Butler, S. Leigh (2004), "A New NIST SRM for Microanalysis and X-ray Fluorescence, TiAl(NbW) Alloy," *Proceedings of Microscopy and Microanalysis Conf*erence, Savannah Ga.
24. S. Leigh (2004), Review of *Statistics for the Quality Control Chemistry Laboratory* by Eamonn Mullins, *Journal Anal Bioanal Chem*.
25. H. Liu, "High-dimensional empirical linear prediction," *Advanced Mathematical Tools in Metrology III*, 79-90.
26. N.F. Zhang, H. Liu, N. Sedransk, W. E. Strawderman (2004), "Uncertainty Analysis of Interlaboratory Studies with Linear Trends, *Proceedings of ASPE 2004 Summer Topic Meeting - Uncertainty Analysis in Measurement and Design*, 100-105.
27. H. Liu, G. N. Stenbakken (2004), "Empirical modeling methods using partial data," *IEEE Transactions on Instrumentation and Measurement*, 53, 271-276.
28. N.F. Zhang, H. Liu, N. Sedransk, W. E. Strawderman (2004), "Statistical Analysis of Key Comparisons with Linear Trends," *Metrologia*, 41, 231-237.
29. Z. Q. Lu (2004), "Functional ANOVA with Applications to MALDI-TOF Mass Spectrometry in Polymers," *Proceedings of the 36st Symposium on the Interface: Computational Biology and Bioinformatics*.
30. Z. Q. Lu (2004), Review of *Nonlinear Time Series: Nonparametric and Parametric Methods* by J. Fan, Q. Yao, *Technometrics*, Vol.46, No.1, 114-115.
31. A. L. Rukhin (2004), "Gamma-Distribution Order Statistics, Maximal Multinomial Frequency and Randomization Designs," *Journal of Statistical Planning and Inference*.
32. A. L. Rukhin, I. Malioutov (2004), "Fusion of Biometric Algorithms in the Recognition Problem," *Pattern Recognition Letters*.
33. A. L. Rukhin, A. Osmoukhina (2004), "Nonparametric Measures of Dependence for Biometric Data Studies," *Journal of Statistical Planning and Inference*.
34. A. L. Rukhin (2004), "The Recognition Problem of Biometrics," *Chance*, 17, 30-34.
35. A. L. Rukhin (2004), "Limiting Distributions in Sequential Occupancy Problems," *Sequential Analysis*, 23, 141-158.
36. T. P. Ryan (2004), Review of *Semiparametric Regression* by David Ruppert, Matt Wand and Ray Carroll, *Journal of Quality Technology*, 36(2), 242-243.
37. T. P. Ryan (2004), Review of *Planning, Construction, and Statistical Analysis of Comparative Experiments* by Francis G. Giesbrecht and Marcia L. Gumpertz, *Journal of Quality Technology*, 36(4), 454-457.
38. T. P. Ryan (2004), Review of *Handbook of Statistics 22* by R. Khattree and C. R. Rao, eds., *Journal of Quality Technology*, 36(3), 339-341.
39. L. F. Goodrich, T. C. Stauffer, J. D. Splett, D. F. Vecchia (2004), "Unexpected Effect of Field Angle in Magnetoresistance Measurements of High-Purity Nb," *Proceedings of the Applied Superconductivity Conference*, Jacksonville, Florida.

40. J. A. Jargon, J. D. Splett, D. F. Vecchia, D. C. DeGroot (2004), “Modeling Warm-Up Drift in Commercial Harmonic Phase Standards,” *Conference on Precision Electromagnetic Measurements Digest*, London, England, 612-613.
41. L. F. Goodrich, T. C. Stauffer, J. D. Splett, D. F. Vecchia (2004), Measuring Residual Resistivity Ratio of High-Purity Nb,” *Advances in Cryogenic Engineering: Transactions of the International Cryogenic Materials Conference*, Vol. 50, pp. 41-48.
42. A. Katsis, B. Toman (2004), “A Bayesian Double Sampling Scheme for Classifying Binomial Data,” *The Mathematical Scientist*.
43. B. Toman (2004), “Bayesian Approach to the Estimation of a Key Comparison Reference Value,” *NCSLI*.
44. G. L. Yang, S. Y. He, K. T. Fang, J. F. Widmann (2004), “Estimation of Poisson intensity in the presence of dead time,” *Journal of the American Statistical Association*.
45. A. E. Grosz, J. N. Grossman, J. H. Yen, et al (2004), “Arsenic in Surficial Sediments of North America,” *Global Geochemical Baselines Workshop (DWO 16) of the 32nd International Geological Congress*, Florence, Italy.
46. P. D. Over, J. H. Yen (2004), “An Introduction to DUC 2004 Intrinsic Evaluation of Generic New Text Summarization Systems,” *Document Understanding Workshop at the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting*.
47. J. M. Brown Thomas, J. H. Yen, M. M. Schantz, B. J. Porter, K. S. Sharpless (2004), “Determination of Caffeine, Theobromine, and Theophylline in Standard Reference Material 2384, Baking Chocolate Using Reversed-phase Liquid Chromotography,” *Journal of Agriculture and Food Chemistry*.
48. N.F. Zhang, P. Winkel (2004), “Statistical Process Control on Biochemical and Hematological Quality Contorl Data,” *Proceedings of Section on Quality and Productivity of Americal Statistical Society*.
49. N.F. Zhang, P. Winkel (2004), “Serial Correlation of Quality Control Data on the Use of Proper Control Charts,” *Scandinavian Journal of Clinical and Laboratory Investigation*, 64, 195-204.

9.2.2 NIST Technical Reports

1. K. J. Coakley, J. D. Splett, M. D. Janezic, R. K. Kaiser (2005), “Relative Permittivity and Loss Tangent Measurement Using the NIST 60 mm Cylindrical Cavity,” *NIST Special Publication*.
2. C. R. Hagwood, K. L. Stricklett, O. Petersons (2004), “Operating Characteristics of the Proposed Sampling Plans for Testing Distribution Transformers,” *NIST Technical Note 1456*.
3. W. J. Rossiter, B. Toman, M. E. McKnight, I. Emenanjo (2004), “Ultrasonic Extraction/Anodic Stripping Voltammetry for Determining Lead in Dust: Analyses of Field-Sampled Wipes,” *NISTIR 7109*.
4. W. J. Rossiter, B. Toman, M. E. McKnight, I. Emenanjo, M. B. Anaraki (2004), “Ultrasonic Extraction/Anodic Stripping Voltammetry for Determining Lead in Dust: A Laboratory Evaluation,” *NISTIR 6998*.
5. D. Williams, P. D. Hale, T. S. Clement, C. Wang (2004), “Uncertainty of the NIST Electrooptic Sampling System,” *NIST Technical Note 1535*.
6. S. Laskowski, M. Autry, J. Cugini, B. Killam, J. H. Yen (2004), “Improving the Usability and Accessibility of Voting Systems and Products,” *NIST Special Publication 500-256*.

9.2.3 Publications in Process

1. A. I. Aviles, B. E. Ankenman, J. C. Pinheiro (submitted), "Robustness Experiments with Two Variance Components," *Journal of Quality Technology*.
2. D. J. McKinsey, K. J. Coakley (to appear), "Neutrino Detection in CLEAN," *Astroparticle Physics*.
3. S. N. Dzhosyuk, K. J. Coakley, J. M. Doyle, R. Golub, E. Korobkina, S. K. Lamoreaux, A. K. Thompson, G. L. Yang, L. Yang, P. R. Huffman (submitted), "Determination of the Neutron Mean Lifetime Using Magnetically Trapped Neutrons," *NIST Journal of Research*.
4. K. J. Coakley, J. M. Doyle, S. N. Dzhosyuk, L. Yang, P. R. Huffman (submitted), "Chaotic Scattering and Escape Times of Marginally Trapped Ultracold Neutrons," *NIST Journal of Research*.
5. J. J. Filliben, R. R. Zarr (submitted), "Collaborative Thermal Conductivity Measurements of Fibrous Glass and Expanded Polystyrene Reference Materials," *Proceedings of International Thermal Conductivity Conference*.
6. J. J. Filliben, R. R. Zarr (submitted), "An International Study of Guarded Hot Plate Laboratories Using Fibrous Glass and Expanded Polystyrene Reference Materials," *ASTM Special Technical Publication*.
7. G. A. Klouda, J. J. Filliben, H. J. Parish, J. C. Chow, J. G. Watson, R. A. Cary (to appear), "Reference Material 8785: Air Particulate Matter on Filter Media," *Aerosol Science and Technology*.
8. F. H. Sadek, R. Wilcox, E. Simiu, J. J. Filliben (to appear), "Wind Speeds in ASCE 7 Standard Peak-Gust Map: Assessment. Closure," *Journal of Structural Engineering.*
9. S. J. Wetzel, C. M. Guttman, K. M. Flynn, J. J. Filliben (submitted), "Significant Parameters in the Optimization of MALDI-TOF-MS for Synthetic Polymers," *Analytical Chemistry*.
10. A. Hornikova, W. F. Guthrie (2005), "A Survey of Key Comparisons," *Proceedings of the 2005 Measurement Science Conference*.
11. A. Hornikova, W. F. Guthrie (2005), "Troubleshooting Key Comparisons (A Survey of Design, Analysis, and Reporting of Results in Key Comparisons)," *Proceedings of the 2004 Joint Statistical Meeting.*
12. C. R. Hagwood, W. F. Guthrie (2005), "Combining Data in Small Multiple Method Studies," *Technometrics*.
13. C. Wang, H. Iyer (submitted), "Propagation of uncertainties in measurements using generalized inference," *Metrologia*.
14. C. Wang, H. Iyer (submitted), "Detection of influential observations in the determination of the weighted-mean KCRV," *Metrologia*.
15. S. Leigh, C. Beauchamps (to appear), "Best Measurement Practice Guide: Using Magnetic Methods for the Determination of Nonmagnetic Coating Thickness on Magnetic Substrates," *US DOC Publication*.
16. W. Liggett, C. Buckley (2005), "System Performance and Natural Language Expression of Information Needs," *Information Retrieval* 8(1), 101-128.
17. D. Malec, B. Toman (submitted), "A Bayesian Approach to Gas Chromatography Calibration and Prediction for Multiple Laboratory Experiments with Corrupted Material," *Journal of the American Statistical Association*.
18. A. L. Rukhin (submitted), "Nonparametric Inference for Balanced Randomization Designs," *Journal of Statistical Planning and Inference.*
19. A. L. Rukhin (submitted), "Recognition Problem in Biometric Data Studies: Nonparametric Dependence Characteristics and Aggregated Algorithms," *Statistical Methods in Counter-Terrorism*, A. Wilson and D. Olwell, eds.
20. T. P. Ryan, W. H. Woodall (2005), "The Most-Cited Statistical Papers," *Journal of Applied Statistics*.

21. C. McCowan, G. Roebben, Y. Yamaguchi, S. Lefranois, J. D. Splett, S. Takagi, A. Lamberty (2005), “International Comparison of Impact Reference Materials,” *Journal of ASTM International*.
22. J. D. Splett, C. Wang (submitted), “Uncertainty in reference values for the Charpy V-notch verification program,” *ASTM Journal of Testing and Evaluation*.
23. J. D. Splett, C. McCowan (2005), “Analysis of Charpy Impact Verification Data: 1993-2003,” *Journal of ASTM International*.
24. B. Toman (submitted), “A Bayesian Approach to Assessing Uncertainty and Calculating a Reference Value in Key Comparison Experiments,” *Technometrics*.
25. L. A. Muth, C. Wang, T. Conn (submitted), “Robust separation of background and target signals in radar cross section measurements,” *IEEE Transactions on Instrumentation and Measurement*.
26. P. D. Hale, C. Wang, D. Williams, K. A. Remley (submitted), “Compensation of random and systematic timing errors in sampling oscilloscopes,” *IEEE Transactions on Instrumentation and Measurement*.
27. G. L. Yang, J. F. Widmann, S. He, K. T. Fang (to appear), “Estimation of Poisson intensity in the presence of dead time,” *Journal of the American Statistical Association*.
28. F. W. Wang, S. Hirayama, J. H. Yen, S. Takagi (2005), “Physical Properties of Composite Bone Grafts Consisting of Calcium Phosphate Cement and Chondroitin Sulfate,” *Proceedings of Biomaterials Society Meeting*, Memphis.
29. N.F. Zhang (submitted), “The Batched Moving Averages of a Stationary Process and Their Applications,” *Communication in Statistics: Theory and Methods*.

9.2.4 Working Papers

1. A. I. Aviles, B. Cordes, “Developing a Graphical Tool to Determine the Estimation Capacity of an Experimental Design.”
2. D. Dey, J. Liu, J. Soto, B. Toman, “Prior Elicitation from Expert Opinion: An Interactive Approach.”
3. A. Hornikova, S. Leigh, “Report on the CCPR S2.”
4. A. Hornikova, N.F. Zhang, “Statistical Approaches to Evaluate the Uncertainty for NIST SRM 1508a.”
5. F. R. Guenther, S. Leigh, A. L. Rukhin, “Errors in Variables as Applied to Gas Standard Calibration.”
6. W. R. Kelly, B. S. MacDonald, S. Leigh, “A 'Designer' Calibration Standard Method for Determination of Sufur in Fossil Fuels: User Prepared NIST Traceable CRMs with Known Concentrations and Uncertainties.”
7. N. Sedransk, A. L. Rukhin, “Statistics in Metrology: International Key Comparisons and Interlaboratory Studies.”
8. A. L. Rukhin, Z. Volkovich, “Testing Randomness via Aperiodic Words.”
9. R. A. Davis, W. T. Dunsmuir, S. B. Streett, “Maximum Likelihood Estimation for an Observation Driven Model for Poisson Counts.”
10. S. D. Phillips, B. Toman, W. T. Estler, “Uncertainty due to Finite Resolution Measurements.”
11. B. Toman, “Linear Statistical Models with Type B Uncertainty: A Bayesian View of Annex H.3 and H.5 of the Guide to the Expression of Uncertainty in Measurement.”
12. S. Hirayama, S. Takagi, J. H. Yen, F. W. Wang, “Physical Properties of Moldable, Resorbable, Composite Bone Graft.”
13. C. Khatri, G. Du, E. S. Wu, J. H. Yen, F. W. Wang, “Focal Adh Focal Adhesions of Osteoblast-like Cells on Films of Poly(d,l-lactide)/Poly(vinyl alcohol) Blends.”

14. L. C. Sander, K. S. Sharpless, M. B. Satterfield, K. W. Phinney, J. H. Yen, S. A. Wise, et al. "Determination of Ephedrine Alkaloids in Dietary Supplement Standard Reference Materials."

9.3 Talks

1. A.I. Aviles, Statistical Methods for Analyzing the Particle Size Distribution of Portland Cement, ASTM C01 Meeting on Fineness, Washington, DC, December 2004.
2. A.I. Aviles, Statistics for Experiments, University of Puerto Rico, Mayagüez, PR, January 2004.
3. A.I. Aviles, Symposium on Statistical Methods for Analyzing Color Differences, ASTM E12 - Color and Appearance Meeting, Gaithersburg, MD, May 2004.
4. A.I. Aviles, Careers, Choices & Opportunities, NSF/CISE Maryland REU-Sites, First Annual Meeting, NIST, Gaithersburg, MD, June 2004.
5. K.J. Coakley, Classification Problems in Neutrino Physics, invited talk at the 2004 Meeting of the International Federation of Classification Societies, Chicago, IL, July 2004.
6. K.J. Coakley, Nonlinear Modeling of Tunnel Diodes, (with D.K. Walker and J.D. Splett), poster session at the 2004 IEEE Geoscience and Remote Sensing Society Symposium (IGARSS), Anchourage, AK, September 2004.
7. K.J. Coakley (with D.S. Simons (CSTL), K.J. Coakley, A. Leiffer), Application of Time Interpolation to SIMS Isotopic RatioMeasurements, Americal Vacuum Society 51st International Symposium, Anaheim, CA, November 2004.
8. K.J. Coakley, (with D. McKinsey, W. Lippincott, J.Nikkel, A. Hime, M. Boulay, J. Lidgard, E. Kearns), Neutrino and WIMP Detection with CLEAN, 2004 Meeting of American Physical Society, Denver, C0, May 2004.
9. K.J. Coakley, (with D.S. Simons and A.M. Leiffer), Application of Time Interpolation to SIMS Isotopic Ratio Measurements, 17th Annual Secondary Ion Mass Spectrometry Workshop, Westminister, C0, May 2004.
10. K.J. Coakley (with K. Ehara, N. Fukushima, K. Worachotekamjorn), Analysis and Improvement of the Temporal Response of the Aerosol Mass Analyzer, European Aerosol Conference, Budapest, September 2004.
11. J.J. Filliben, World Trade Center Analysis of Structural Failure, NRC Panel Informal Assessment of Division Activities, NIST, Gaithersburg, MD, April 20, 2004.
12. J.J. Filliben, Statistical Analysis of Influential Parameters, NIST WTC Working Group, NIST, Gaithersburg, MD, April 27, 2004.
13. J.J. Filliben, The World Trade Center Collapse: The Critical Role of Statistics, Spring Research Conference, NIST, Gaithersburg, MD, May 21, 2004.
14. J.J. Filliben, Design of Numerical Experiments, a part of Jeffrey Fong's Talk: A Metrology-based Uncertainty Analysis Approach to V&V of Computer Models of High-Consequence Engineering Systems, Southwest Research Institute, October 11, 2004.
15. J.J. Filliben, Design of Numerical Experiments, a part of Jeffrey Fong's Talk: A Metrology-based Uncertainty Analysis Approach to V&V of Computer Models of High-Consequence Engineering Systems, Foundations '04: A Workshop for V&V in the 21st Century, Tempe, AZ, October 13, 2004.
16. J.J. Filliben, Design of Numerical Experiments, a part of Jeffrey Fong's Talk: A Metrology-based Uncertainty Analysis Approach to V&V of Computer Models of High-Consequence Engineering Systems, Lawrence Livermore Laboratories, Livermore, CA, October 18, 2004.
17. J.J. Filliben, World Trade Center Analysis, SED/ITL Program Review, NIST, Gaithersburg, MD, November 2, 2004.

18. J.J. Filliben, A Structured Roadmap for Verification and Validation--Highlighting the Critical Role of Experiment Design, 2004 Workshop of Verification and Validation of High-Consequence Engineering Systems, NIST, Gaithersburg, MD, November 8, 2004.
19. J.J. Filliben, Basketballs, Funnels, and Designed Experiments, Adventures in Science, NIST, Gaithersburg, MD, November 20, 2004.
20. J.J. Filliben, Statistical Approaches in the NIST World Trade Center Collapse Analysis, University of Maryland, December 2, 2004.
21. W.F. Guthrie, A Bayesian Approach to Combining Results From Multiple Methods, Workshop on the Verification and Validation of Computer Models for the Design and Performance Evaluation of High-Consequence Engineering Systems. NIST, November 2004.
22. W.F. Guthrie, Hands-on Workshop on Estimating and Reporting Measurement Uncertainty, Anaheim, California, January 2004.
23. W.F. Guthrie, Introduction to Nonparametric Regression, NIST Administration Building, Lecture Room D, Gaithersburg, Maryland, September 2004.
24. W.F. Guthrie, MCMC in StRD, Salt Lake City, Utah, July 2004.
25. W.F. Guthrie, Ranking the Sources of Numerical Error in MCMC Computations, Toronto, August 2004.
26. W.F. Guthrie, Validating the Temperature-EMF Response of Thermocouples with Confidence, Washington, DC, November 2004.
27. C. Hagwood, An application of stochastic differential equations to sizing ultrafine particles, 2nd International Work Shop in Applied Probability, Pireaus, Greece, March 2004.
28. C. Hagwood, Dynamic Calibration, Talk at George Mason University, Farifax, Virginia, February 2004.
29. C. Hagwood, An Application of Stochastic Differential Equations to Sizing Ultrafine Particles, Talk at Department of Statistics, University of Delaware, May 2004.
30. A. Hornikova, A Survey of Design, Analysis, and Reporting of Results in Key Comparisons, Room 152, NIST North, Gaithersburg, Maryland, October 2004.
31. D.D. Leber, NIST-SEMATECH e-Handbook of Statistical Methods and DATAPLOT: An Interactive Demonstration ACSL Conference Room, Gaithersburg, Maryland, September 2004.
32. H.K. Liu, Comparisons of Bayesian approaches to combining results from multiple methods, Measurement Science Conference, Anaheim, CA, January 2004.
33. H.K. Liu, Weighing Designs for Mass Calibration, Spring Research Conference, Gaithersburg, MD, May 2004.
34. H.K. Liu, An update on the StRD for MCMC, Joint Statisitical Meetings, August 2004.
35. H.K. Liu, G. L. Yang, Statistical Analysis of Incomplete Data for Scientists & Engineers, NIST Administration Building, LRB & LRE, Gaithersburg, Maryland, December 2004.
36. W. Liggett, Data-Driven and Peak-Based Feature Selection in Serum Protein Mass Spectrometry, invited talk in Interface 2004: Computational Biology and Bioinformatics, 36th Symposium on the Interface, Baltimore, Maryland, May 2004.
37. Z.Q. Lu, Statistics and Standards for High-Throughput Measurements, Invited talk ICSA 2004 Applied Statistics Symposium in Bio-tech Research and Computing Intensive Methodologies, in San Diego, California, June 2004.
38. Z.Q. Lu, Bayesian Statistics in Uncertainty Analysis, Combining Multiple Methods, and Inter-laboratory Studies, State College, Pennsylvania, June 2004.
39. Z.Q. Lu, Statistical and Metrological Issues in Proteomics using Time-of-flight Mass Spectrometry, invited talk in Interface 2004: Computational Biology and Bioinformatics, 36th Symposium on the Interface, Baltimore, Maryland, May 2004.

40. T.P. Ryan, The Role of Modern Experimental Design in Quality Improvement, Southeastern Quality Conference, Atlanta, Georgia, October 2004.
41. A.L. Rukhin, Statistics in Metrology: International Key Comparisons and Interlaboratory Studies, 11th Spring Research Conference on Statistics in Industry and Technology, NIST, Gaithersburg, MD, May 2004.
42. A.L. Rukhin, Balanced Randomization Designs and Classical Probability Distributions, NIST North, Room 145, Gaithersburg, Maryland, February 2004.
43. A.L. Ruhkin, Nonparametric Dependence Characteristics in Recognition Problem of Biometrics, ENAR-IMS Spring Meeting, Pittsburgh, PA, March 2004.
44. A.L. Rukhin, Introduction to Markov Chains: Markov Chain Monte Carlo for Scientists & Engineers, Lecture Room B, Administration Building, Gaithersburg, Maryland, August 2004.
45. A.L. Rukhin, Issues of Trend and Linkage in Interlaboratory Studies, International Workshop "Longevity, Aging and Degradation Models in Reliability, Medicine and Biology," St.-Petersburg, Russia, June 2004.
46. N. Sedransk, Potential Roles for Statistics in NanoScience, NNI Interagency Workshop: Instrumentation and Metrology for Nanotechnology - Grand Challenges, January 2004.
47. N. Sedransk, Role of Statistics in International Metrology: Role of SED at NIST, presentation at BIPM Visit to NIST, April 2004.
48. N. Sedransk, Statistics in Metrology: International Key Comparisons and Interlaboratory Studies, Spring Research Conference, NIST, Gaithersburg, MD
49. N. Sedransk, A Statistical Metrologist Looks at Computational System Models, V&V Conference at NIST, November 2004.
50. N. Sedransk, Statistical Engineering and SRM Certification, ASTM E01 at NIST, November 2004.
51. J.D. Splett, Analysis of Charpy Impact Verification Data: 1993-2003, Second Symposium on Pendulum Impact Machines: Procedures and Specimens, Washington, DC, November 2004.
52. B. Toman, Bayesian Approach to Calculating a Key Comparison Reference Value 2004 NCSL Workshop and Symposium, Salt Lake City, Utah, July 2004.
53. C. Wang, Consistency tests for key comparison data, Measurement Science Conference, Anaheim, California, January 2004.
54. C. Wang, Generalized prediction intervals, Joint Statistical Meetings, Toronto, Canada, August 2004.
55. C. Wang, Structural inference for uncertainty evaluation with application to calibration experiments, ASPE Summer Topical Meeting, State College, Pennsylvania, June 2004.
56. C. Wang, Structural approach for propagating uncertainties, International Microwave Symposium, Fort Worth, Texas, June 2004.
57. G.L. Yang, Statistical Analysis of Incomplete Data for Scientists & Engineers, Administration Building, Gaithersburg, Maryland, December 2004.
58. N.F. Zhang, Statistical Analysis of Interlaboratory Studies with Linear Trends, ASPE Summer Topic Meeting, Penn State University, PA, July 2004; University of Science and Technology of China, Hefei, China, November 2004; and Department of Mathematical Statistics, East China Normal University, Shanghai, China, November 2004.
59. N.F. Zhang, Statistical Process Monitoring for Autocorrelated Data, INFORMS Annual Meeting, Denver, Colorado, October 2004.
60. N.F. Zhang, Statistical Process Control on Biochemical and Hematological Quality Control Data, Joint Statistical Meetings, Toronto, Canada, August 2004.

9.4 Professional Activities

9.4.1 NIST Committee Activities

1. A. I. Aviles, Member, ITL Awards Committee.
2. A. I. Aviles, Member of NIST Employees Concerned with Disabilities (ECD).
3. A. I. Aviles, Advisor to the ITL coordinators, SURF/NSF program.
4. A. I. Aviles, Chair, Yellow Book Team; Editor, SED Report of Activities.
5. S. Bailey, Member, ITL Diversity Committee.
6. Carroll Croarkin, Member, US TAG for ISO TC69.
7. W. F. Guthrie, ITL Representative to the Washington Editorial Review Board.
8. S. Leigh, ITL Representative to the Washington Editorial Review Board.
9. S. Leigh, SED liaison for NIST/NRC postdoctoral associateship program.
10. W. Liggett, Member, NIST Institutional Review Board.
11. N. Sedransk, Member of Measurement Services Group.
12. N. Sedransk, Member of MSAG Task Force on SRM Business Practices.
13. N. Sedransk, WSS Planning Committee on Statistics for Homeland Defense and Security.
14. N. Sedransk, Leader of Task Force on Statistical Methodology for Key Comparisons.
15. C. Wang, EEEL MCOM Technical Subcommittee on relative permittivity and loss tangent SRM..
16. N.F. Zhang, Member, EEEL, MCOM subcommittee on AC-DC Difference of Voltage.
17. N.F. Zhang, Member of Task Force for reviewing NIST Quality Manual, QM-1.
18. N.F. Zhang, Convenor of ISO TC69/SC6/WG4.
19. N.F. Zhang, EEEL MCOM Technical Subcommittee on NIST Measurement Service for DC Standard Resistor.
20. N.F. Zhang**,** Invited Panel Discussion on Statistical Process Control in 23th Decision Science Institute Meeting.

9.4.2 Standards Committee Memberships

1. A. I. Aviles, Member, ASTM International, E12 - Color and Appearance.
2. K. J. Coakley, Telecommunications Industry Association International Electrotechnical Commission, TIA/EIC, Working Group 4, TC-86.
3. N.F. Zhang, Member, ASC Z1 Subcommittee on Statistics.
4. N.F. Zhang, US TAG member on ISO TC69.
5. N.F. Zhang, Project Leader of ISO/CD/TS 21749 of ISO TC69/SC6.

9.4.3 Other Professional Society Activities

1. A. I. Aviles, Chair, SPES Awards Committee.
2. A. I. Aviles, ASA Representative to AAAS, Section P- Industrial Sciences.
3. A. I. Aviles, Member, American Council of the Blind (ACB).
4. Carroll Croarkin, Chair, Mary Natrella Scholarship Committee for ASA.
5. D. Dey, Member, editor selection committee of the Institute of Mathematical Statistics.
6. D. Dey, Member, archive committee of the American Statistical Association.
7. K. J. Coakley, NIST representative to National Institute of Statistical Sciences (NISS).
8. Z. Q. Lu, Organizer of an Invited Session at Interface 2004: Computational Biology and Bioinformatics, May 26-29, 2004.

9. T. P. Ryan, Chair of the Committee to Nominate Fellows of the Quality and Productivity Section of ASA.
10. T. P. Ryan, Member, Editorial Review Board.
11. A. L. Rukhin, Monitored Gordon Research Conference on Statistics in Chemistry, Mount Holyoke College.
12. N. Sedransk, Member, ASA Subcommittee on Publications Marketing.
13. N. Sedransk, Member, ASA-ENVR Committee on Fellows.
14. N. Sedransk, Vice Chair, ASA Publications Committee.

9.5 Professional Journals

9.5.1 Editorships

1. D. Dey, Associate Editor of *Journal of Statistical Planning and Inference.*
2. W. Liggett, Board of Editors of the *NIST Journal of Research*.
3. T. P. Ryan, Book Review Editor of the *Journal of Quality Technology* (JQT).
4. A. L. Rukhin, Associate Editor of *Statistics and Probability Letters.*
5. A. L. Rukhin, Associate Editor of *Mathematical Methods of Statistics.*
6. A. L. Rukhin, Associate Editor of *Applicationes Mathematicae.*
7. A. L. Rukhin, Coordinating Editor of *Journal of Statistical Planning and Inference.*

9.5.2 Refereeing

1. A. I. Aviles, *Technometrics*.
2. A. I. Aviles, *Statistics and Probability Letters*.
3. A. I. Aviles, *Washington Editorial Review Board* (WERB).
4. K. J. Coakley, *Boulder Editorial Review Board*.
5. K. J. Coakley, *Biometrics*.
6. J. J. Filliben, *Washington Editorial Review Board* (WERB).
7. C. Hagwood, *Statistics and Probability Letters*.
8. C. Hagwood, *Washington Editorial Review Board* (WERB).
9. S. Leigh, *Washington Editorial Review Board* (WERB).
10. S. Leigh, *IEEE Transactions*.
11. S. Leigh, *Analytical and Bioanalytical Chemistry*.
12. Z. Q. Lu, *IEEE Transactions on Signal Processing.*
13. T. P. Ryan, *IIE Transactions*.
14. T. P. Ryan, *JQT*.
15. T. P. Ryan, *Statistics in Medicine.*
16. T. P. Ryan, *Technometrics*.
17. A. L. Rukhin, *Technometrics*.
18. A. L. Rukhin, Journal *of Computational and Graphical Statistics*.
19. A. L. Rukhin, *British Journal of Mathematical and Statistical Psychology*.
20. N. Sedransk, *Metrologia*.
21. J. Soto, *Journal of Statistical Planning and Inference.*
22. W. Strawderman, *Annals of Statistics.*
23. W. Strawderman, *Bernoulli*.
24. W. Strawderman, *JSPI.*
25. W. Strawderman, *Statistics and Probability Letters*.
26. W. Strawderman, *Statistical Science.*
27. W. Strawderman, *Journal of Multivariate Analysis*.
28. B. Toman, *Technometrics*.

29. B. Toman, *Washington Editorial Review Board* (WERB).
30. J. Wang, *Psychometrika*.
31. G. L. Yang, *Journal of Statistical Planning and Inference*.
32. G. L. Yang, *Journal of Nonparametric Analysis*.
33. G. L. Yang, John Wiley Publ. Co.
34. N.F. Zhang, *Technometrics*.
35. N.F. Zhang, *Metrologia*.
36. N.F. Zhang, *Mathematical Methods of Statistics*.
37. N.F. Zhang, *International Journal of Production Research*.

9.6 Review Panels

1. C. Hagwood, National Science Foundation.
2. N. Sedransk, National Science Foundation.
3. G. L. Yang, National Institute of Health.

9.7 Honors

A. I. Aviles received the SROP Alumni Achievement Award by the Committee on Institutional Cooperation (CIC). It was presented at the SROP (Summer Research Opportunity Program) Conference held at the University of IOWA on July 10, 2004.

W. F. Guthrie and **J. Wang**: Honorable Mention - Outstanding Presentation Award for Contributed Paper at JSM 2003 by the Section on Physical and Engineering Sciences.

The Silver Medal Award in the category of Scientific and Engineering Achievement was awarded to **W. F. Guthrie** and **N. F. Zhang** for development and provision to industry of a long-sought two-dimensional (2-D) grid standard (SRM 5001). The Silver Medal Award is bestowed for "exceptional performance characterized by noteworthy or superlative contributions that have a direct and lasting impact within the Department." This award recognizes scientific/engineering or technological breakthroughs that resolve long-standing problems, radically advance the state of the art, significantly impact the Department of Commerce or the economy, or significantly advance the understanding, knowledge, or mastery of a given discipline. The Silver Medal Award is the second highest honor awarded by the Department of Commerce.

ITL Outstanding Contribution Award to **C. Hagwood** for his work in managing student programs, including the minority student program in the Statistical Engineering Division.

A. Hornikova: ITL Thank you Award for her energetic and creative support with the publication of the Statistical Engineering Division Brochure and Report of Activities 2003.

Outstanding achievement in poetry to **A. Hornikova** by the International Society of Poets, on August 15, 2004.

The Bronze Medal Award in the category of Scientific and Engineering Acievement was awarded to **S. Leigh** for exemplary leadership in expert application and dissemination of statistical metrology for the chemical, physical and information sciences. The Bronze Medal award is the highest honorary recognition available for Institute presentation. The award, approved by the Director, is given for significant contributions affecting major programs, scientific accomplishments within the Institute, and superior performance of assigned taks for at least five consecutive years.

Accolades of Appreciation

- *"This summer has been a great experience for me. The SURF program is excellent and allowed me the opportunity to participate in some of the world-class research that takes place here at NIST. At first, I was a little disappointed with my placement in the Statistical Engineering Division. I thought there would be little overlap between my academic strengths and interests and what I would be doing over the summer. I couldn't have been more wrong and my placement in SED turned out to be a blessing in disguise. Not only did I get to work on two very interesting and challenging projects with two excellent advisors (thanks Dr. Hagwood and Dr. Rukhin), but I also experienced a bit of the consulting nature of SED. I was allowed to tag along with Stefan Leigh as he helped various scientists all around NIST handle the statistical problems that arise in their research. All in all, I had an extremely enjoyable and educational summer. Thanks SED!"*

  -Van Molino

- *"My internship experience this summer at NIST is one of the most memorable experiences of my life. I learned a lot about research at NIST and the importance of statistics. Before my internship, I had no idea why research at NIST is important. After my internship, I have a greater appreciation of the importance of standards and accurate measurements. I realize that research at NIST is improving our everyday lives. In the beginning of my internship, when I found out about being placed in the SED, Statistical Engineering Division, I did not know what to expect. I had only taken one course in statistics, so I didn't have a strong background or interest in statistics. At the end of my internship at NIST, I am very glad that they placed me in the SED. I learned a lot about statistics, and I understand that many statisticians in the SED are applying their knowledge to help solve real-world problems. Although I probably will not dedicate myself to the field of statistics as a career, I am very impressed with the work that statisticians in the SED have done. My experience in the SED has given me a new perspective about statistics. Overall, my internship experience this summer at NIST was very valuable and enjoyable."*

  - Chiu Yeung