Report of Activities:

Statistical Engineering Division

U. S. Department of Commerce
National Institute of Standards and Technology
Information Technology Laboratory
Gaithersburg, MD 20899 USA

February, 2002
Limited Edition

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

REPORT OF ACTIVITIES OF THE STATISTICAL ENGINEERING DIVISION

FEBRUARY 2002

Covering Period: January 2001 – December 2001

Publications: January 2000 – December 2001

Covers: With homage to Pollock and Gabo, Messy Data $\longrightarrow$ Model

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 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 role of SED is pervasive across NIST; SED staff actively collaborate with over 75% of the scientific Divisions at NIST and with some of the administrative offices as well. Collaborations are diverse in extent and in kind: SED staff provide the statistical basis for certification for Standard Reference Materials produced at NIST and for NIST calibrations; SED works with scientists engaged in designing and implementing international experiments to determine equivalence of standards; SED staff also engage in long-term, large-scale NIST research efforts that draw scientific expertise from several NIST laboratories.

As members of multidisciplinary research teams, SED staff collaborate in the definition of research objectives, formulate statistical strategies and develop statistical methods for process characterization and analysis of experimental data. The statistical expertise central to a multidisciplinary research project may lie in many subdisciplines (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, statistical computation and graphics, etc.). However, the SED objective is always 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 interactions with scientist colleagues.

As independent researchers, SED staff develop new statistical methodology for metrology and modeling, focusing on problems where innovation is critical to complex modeling and highly precise analysis. Advances in statistical methodology arise from fundamental issues in metrology, from scientific modeling problems unique to NIST, and from statistical requirements for incorporation into new standards. Fundamental research in probabilistic modeling, in design of experiments, in theory and methodology of inference, in computationally intensive statistical tools, in spatial statistics and in Bayesian inference and modeling not only expand the statistical methodology available to NIST scientists and engineers and to their customers, but also contribute in a fundamental way to the

discipline of statistics.

Within NIST, the role of SED extends to dissemination of statistical methodology through education and web products. Short courses and workshops are designed to equip scientists with sufficient understanding of basic statistical methodology to be competent data analysts for standard experiments, and to be astute customers when specialized statistical methodology is needed. As new methodology is developed, seminars and tutorials introduce scientists to software and to web products, some of which are also available on the external NIST web pages. Increasing attention is being given to discipline - specific tutorials for scientists and to the development as web products of templates for standard metrological calculations.

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; the third is in Boulder, Colorado. Visiting Faculty appointees and faculty collaborators from several universities are integral to SED activities.

This report provides technical summaries of some of the significant projects undertaken during the year 2001. The projects presented here have been selected to provide a sampling that is indicative of the spectrum of SED activities rather than a comprehensive list. The multidisciplinary collaborations summarized here are just that: joint work with intensive interaction with scientists and engineers. For descriptions of some of the many activities of SED that cannot be included here, consult the SED home page at:
http://www.itl.nist.gov/div898/.

Thank you for reading. Your comments are most welcome.

Nell Sedransk, Ph.D.
Chief, Statistical Engineering Division
100 Bureau Drive, Mail Stop 8980
National Institute of Standards and Technology
Gaithersburg, MD 20899-8980

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

2. Staff

Nell Sedransk, Chief
Stephany Bailey, Secretary

Metrology Statistics and Computation Group
Nien Fan Zhang, Manager
Mary Clark, Secretary
M. Carroll Croarkin*
Dipak Dey, Faculty (University of Connecticut)
Will Guthrie
Charles Hagwood
Alan Heckert
Walter Liggett
John Lu
Joan Rosenblatt*
Grace Yang, Faculty (University of Maryland)

Statistical Modeling and Analysis Group
James Filliben, Manager
Stephany Bailey, Secretary
Ivelisse Aviles
Dennis Leber
Stefan Leigh
Hung-kung Liu
Andrew Rukhin, Faculty (UMBC)
Blaza Toman
James Yen

Boulder Statistics Group
Dominic Vecchia, Manager
Lorna Buhse, Secretary
Duane Boes, Faculty (Colorado State University)
Kevin Coakley
Hariharan Iyer, Faculty (Colorado State University)
Jack Wang
Jolene Splett

**Guest Researcher*

3. Project Summaries

3.1 Bayesian Metrology

3.1.1 Bayesian Metrology – Overview

Nell Sedransk
Statistical Engineering Division, ITL

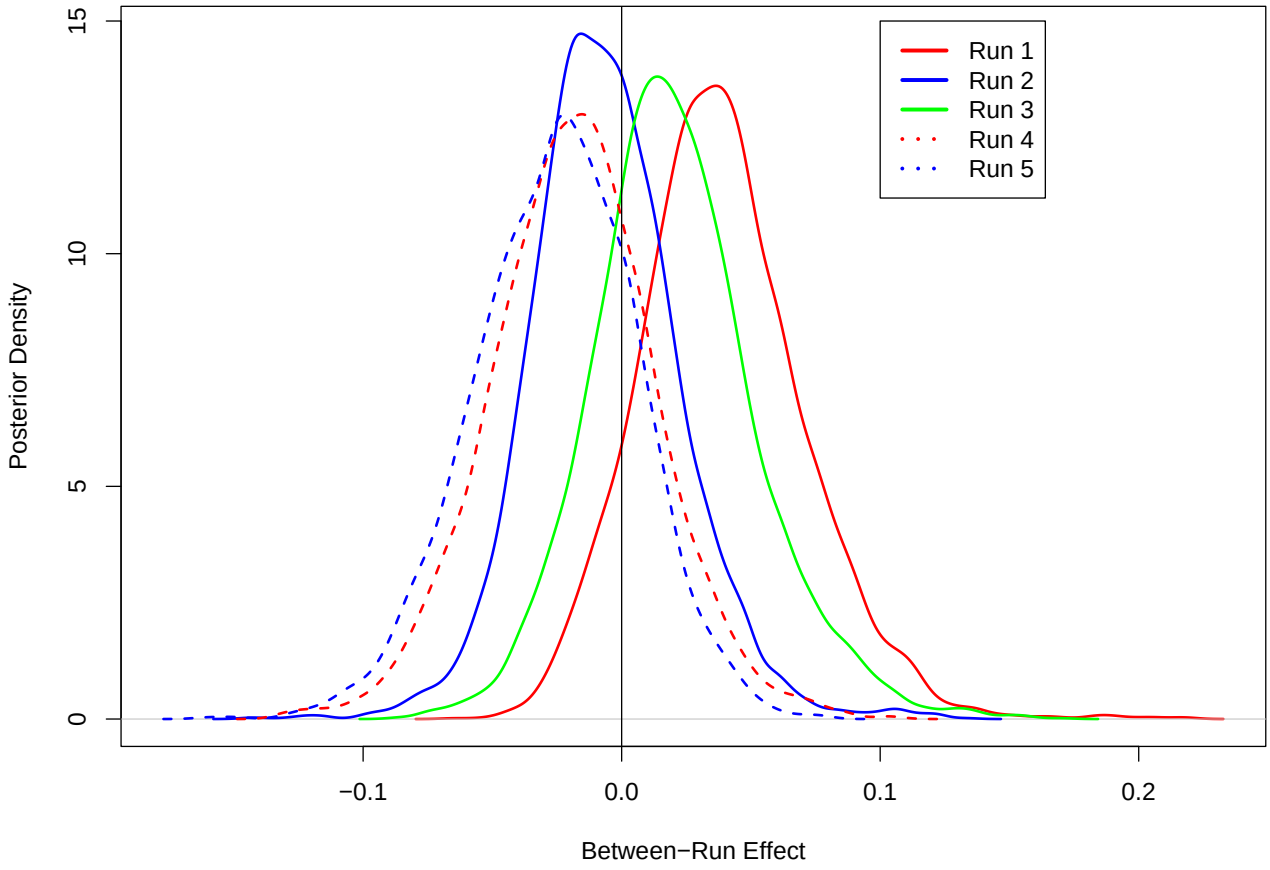


Figure 1: Random effects due to runs.

Bayesian Metrology is emerging as an invaluable statistical methodology for the modern challenges of science which demand efficient designs of experiments to characterize highly complex systems. Bayesian methodology summarizes data in terms of information and the quality of that information. Scientific investigations are often dynamic, accomplished through a sequence of experiments, each built upon the one before. Bayesian statistics follows this paradigm. With each new experiment, Bayesian inference starts with existing information from previous experiments, from observation of a process in operation, from expert opinion, or by applying physical laws. From this starting point, efficient experiments yield new data for statistical modeling, analysis and inference that explicitly and systematically incorporate existing information.

The Bayesian Metrology Project is a five-year effort to develop and integrate Bayesian statistical methodology into the design and analysis of NIST research. Applications for Bayesian metrology cover a broad spectrum, including both frequently encountered traditional NIST analyses and one-of-a-kind highly complex specialized problems for which standard mathematical / statistical tools do not apply. In some cases available Bayesian methodology can be applied; other cases require the development of new methodology for more specialized metrological analyses. During the past year, Bayesian methods have been applied in collaborations with every NIST Laboratory.

The first objective is to develop within SED a cadre of Bayesian statisticians, both a core with primary expertise in Bayesian inference and a broad group with skill in the application of Bayesian methodology. The research objective is to develop and implement Bayesian metrology tools with specific application to NIST measurement problems. The wider objective is to extend the utilization of Bayesian methodology to scientists at NIST through web products, NIST courses and tutorials.

Development of new Bayesian methodology initially focused on traceability, analysis of Standard Reference Material experiments, calibration and inspections processes. More recent work addresses interlaboratory intercomparisons (especially the international Key Comparisons among National Metrological Institutes), elicitation of prior information for uncertainties (second moments), and development of nonparametric Bayesian models using empirical distributions.

To date, this project has been highly successful with 20 professional presentations and 12 publications in professional statistical journals. The goal of integrating Bayesian methodology into NIST practice is also being met; thus far, 13 Standard Reference Materials have been analyzed and certified using Bayesian methodology. (Examples appear elsewhere in this book.) Introductory lectures at NIST have been very well-received. A series of NIST courses in Bayesian modeling begins in February 2002. For a more detailed description of the Bayesian Metrology Project and for examples of Bayesian analyses of NIST experiments, consult: www.itl.nist.gov/div898/bayesian/homepage.htm.

The NIST-wide impact of Bayesian methods has already been demonstrated via successful application into core NIST tasks and into some complex NIST research. Impetus comes from scientific and metrology communities within and outside NIST to adopt Bayesian modeling techniques in order to embed simultaneously in the models empirical and stochastic elements in combination with physical laws.

Figure 2: Rev. Thomas Bayes

3.1.2 Interactive Elicitation of Opinion

Dipak Dey, Z. Q. John Lu, Kimball Kniskern
Statistical Engineering Division, ITL

A quote from A. O'Hagan (1998) describes aptly the importance of prior elicitation in statistical practice: "Prior elicitation is a key component of any serious Bayesian analysis in which the data are not so numerous as to swamp whatever prior information there might be. Yet to elicit a genuine prior distribution (and typically what is needed is a joint prior distribution in several dimensions) is a complex business demanding a substantial effort on the part of both the statistician and the person whose prior beliefs are to be elicited." In this project, a methodology will be developed with underlying interactive computer program for eliciting the hyperparameters of a subjective prior distribution. The software will be useful for many NIST problems such as the estimation of consensus mean or regression parameters in the presence of covariates.

Most NIST problems can be formulated in the form of a general linear model, such as consensus mean estimation (and key comparisons), multi-lab/multi-method studies, linear calibration, and polynomial regression, etc. The elicitation algorithm is based on an expert's opinion on the median and a few selected quantiles of the predictive distribution of the dependent variable. If the expert provides only the mean and the standard deviation, approximate quantiles can be obtained by a normal or a $t-$distribution. The following assumptions are needed in our framework:

1. The mean of the dependent variable is, within the relevant range, a linear function of the independent variables,
2. The errors in this relationship are independent normal with zero mean and equal variance,
3. The expert's postulated prior distribution is a member of the conjugate family of prior distributions,
4. The expert is coherent in the sense that her/his opinions can be modeled by subjective probability.

The main thrust of prior elicitation and modeling expert opinion is to facilitate a fully subjective Bayesian analysis of metrology problems. Consequently, this method provides a theoretically justifiable approach for incorporating scientific judgments from the expert in the form of a prior distribution. The resulting uncertainty (credible) intervals will be more realistic when the number of observations, methods, or labs is small.

3.1.3 Hierarchical Bayesian Analysis of SRM 1946: Lake Superior Fish Tissue for Fatty Acids and PCBs

Blaza Toman, Stefan Leigh
Statistical Engineering Division, ITL

Curtis Phinney, Michele Schantz
Analytical Chemistry, CSTL

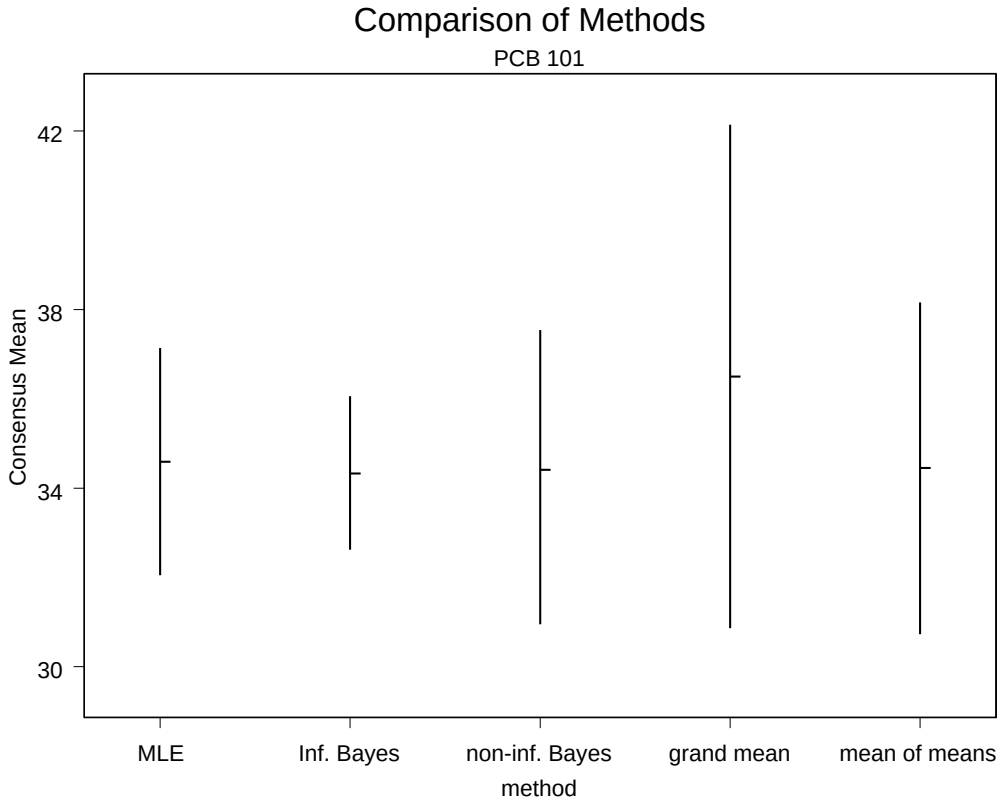


Figure 3: This figure shows a comparison of the 95 percent confidence intervals of the classical consensus mean estimators with the 95 percent HPD intervals for the non-informative (flat) and informative (in terms of between lab variance) Bayes estimators

The purpose of this work was to certify individual fatty acids and PCBs in SRM 1946, Lake Superior Fish Tissue. The concentrations of sixty-two PCBs were measured at NIST and at other environmental laboratories by multiple methods. Concentrations of forty-four fatty acids were measured at NIST as well as at several other laboratories experienced in nutritional measurements. The data from the different sources were then combined using classical (maximum likelihood) as well as Bayesian methods to arrive at the consensus values

The specific mathematical model used in the Bayesian analysis of the 1946 SRM data is very similar to the model used by the maximum likelihood method. As is the case for the MLE, the individual lab's data are considered to be random draws from a distribution with the specific lab mean and variance. Under the Bayesian model, the lab means are considered random draws from a distribution whose mean is the consensus mean and whose variance is the between lab variance. For SRM 1946, the prior distribution of the consensus mean is taken to be a flat (close to uniform) distribution, reflecting the fact that the prior knowledge is minimal. The prior distribution of the between and within lab variances were also taken to be noninformative. The Bayesian model automatically pools data from different labs to estimate the various parameters, a property called "borrowing strength". This property gives the Bayesian method an advantage over classical methods when data are relatively scarce. That is, when each lab has very few observations. When data are plentiful, in terms of number of labs as well as sample sizes within each lab, the Bayesian estimates are close to the MLEs. In order to evaluate the robustness of the method, sensitivity analysis was performed. It showed that the prior distributions of the consensus mean and of the within lab variances had little effect on the results. The prior distribution of the between lab variance did have some effect on the posterior variance of the consensus mean. The most conservative estimates were used for the actual analysis.

Even though the classical, and not the Bayesian values were used for the certification of SRM 1946, the project served as a successful test case for the use of hierarchical Bayesian methods in SRM certification.

3.1.4 Bayesian Approach to Combining Results from Multiple Methods

Hung-kung Liu, Nien Fan Zhang
Statistical Engineering Division, ITL

Method	Measurement Eq.	Uncertainty^2
Discrete	$\theta = \mu_I$	$\frac{u_1^2+u_2^2}{2} + \frac{(c_1-c_2)^2}{4}$
Standard	$\theta = \bar{\mu} + s_\mu^2 \times \nu$	$\frac{u_1^2+u_2^2}{2} + \frac{(c_1-c_2)^2}{4}$
Trapezoid	$\theta = \bar{\mu} + \lvert\mu_2 - \mu_1\rvert \times R(\frac{1}{4})$	$\frac{2(u_1^2+u_2^2)}{3} + \frac{5(c_1-c_2)^2}{12}$
BOB	$\theta = \bar{\mu} + \lvert\mu_2 - \mu_1\rvert \times R(0)$	$\frac{u_1^2+u_2^2}{3} + \frac{(c_1-c_2)^2}{12}$
Hierarchical	$\theta = \bar{\mu} + \sqrt{2}s_\mu^2 \times T$	∞
EB-MLE	$\theta = \bar{\mu} + \omega \times U$	$\frac{u_1^2+u_2^2}{4} + \frac{(c_1-c_2)^2}{12}$
EB-UE	$\theta = \bar{\mu} + \omega \times U$	$\frac{u_1^2+u_2^2}{4} + \frac{3(c_1-c_2)^2}{4}$

Table 3.1: In this table, the measurement equations and uncertainties squared are listed for different methods studied when $p = 2$, with c_i and u_i defining the certified value and uncertainty based on the ith method, respectively, and $\bar{\mu}$ and s_μ^2 are the sample mean and variance, respectively, of the method means $\mu_i,\ i = 1, \ldots, p$. Details of the methods and their associated ancillary variables I, ν, $R(\cdot)$, T, and U are in Liu and Zhang (2001).

Many solutions to the problem of estimating the consensus mean from the results of multiple methods or laboratories have been proposed. In the Bayesian analysis, the consensus mean is specified through probabilistic dependency as either a 'parent' or a 'child' of the method means. We proposed a unified approach to some of these Bayes solutions by expressing the consensus mean as a measurable function of the method means and some ancillary variable.

Let the measurand θ be the true value to be estimated from data using multiple methods (or different laboratories). Assume data obtained by p methods/labs with

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \ldots, p \ (\geq 2), \ \ j = 1, \ldots, n_i,$$

with μ_i defining the true value of θ for the ith method. Denote $Y_i = (y_{i1}, \ldots, y_{i,n_i})$, $\vec{Y} = (Y_1, \ldots, Y_p)$, and $\vec{\mu} = (\mu_1, \ldots, \mu_p)$. Our goal is to find a Bayes estimator (certified value) of θ whose posterior variance (uncertainty squared) can be decomposed according to the ISO Guide to the Expression of Uncertainty in Measurement (ISO GUM).

Our prior on θ is a description of our current knowledge about θ. In the Bayesian analysis, as a distinct component in the probability model for $(\theta, \vec{\mu}, \vec{Y})$, θ can be specified through probabilistic dependency in two possible ways:

(i) θ is a 'parent' of $\vec{\mu}$, as in hierarchical models, or

(ii) $\vec{\mu}$ is the 'parent' of θ, as in Bayesian BOB (Levenson et al. (2000)).

In either way, to make the distribution of $(\theta, \vec{\mu}, \vec{Y})$ identifiable, conditional independence of θ and $\vec{Y}$ given $\vec{\mu}$ is usually also assumed.

We proposed to unify these two approaches by specifying the measurand $\theta = \theta(\vec{\mu}, \nu)$ as a measurable function of $\vec{\mu}$, the true unknown state of nature of the methods, and ν, some ancillary variable, which is independent of $(\vec{Y}, \vec{\mu})$. In doing so, (i) and (ii) become special cases as illustrated in the examples in the table above. We also developed a "Bayes-type theorem" to simplify the computation of uncertainties.

This Measurement Equation Approach is the standard one used by ISO GUM. When the measurement equation is linear in the ancillary variable, the uncertainty of our Bayes estimator is shown to have a decomposition that is ISO GUM compliant.

3.1.5 Consensus Values - A Bayesian Approach

C. M. Wang, H. K. Iyer, D. F. Vecchia, J. D. Splett, and D. C. Boes
Statistical Engineering Division, ITL

Figure 4: BOB (Type B distribution On Bias) denotes an approach to calculating the uncertainty of a consensus value presented by Division authors Levenson et al. (2000). What about the sequel B-BOB, a Bayesian version of BOB?

Certification of reference materials is often based on data from two or more measurement methods or from two or more laboratories. One of the main reasons for the use of multiple measurement methods is the unavailability of a single method whose accuracy and precision have been satisfactorily characterized. When multiple methods are used, they are generally chosen in such a way that it is reasonable to expect the true value to be bracketed by the smallest and the largest results obtained by the different methods. The problem then is one of determining a consensus value, along with a valid uncertainty statement, that can be used to certify the reference material in question. The traditional approach for determining a consensus value is based on the one-way random effects model, with methods assumed to be a random sample from a population of methods. Levenson et al. (2000) have pointed out two problems with this assumption. First, it cannot be justified in some cases. Second, even if we make the assumption, when the number of methods is small (say, 2 or 3), the number of degrees of freedom associated with the estimate of uncertainty is also small, resulting in a confidence interval that is sometimes too wide to be useful. They suggested an alternative approach in which a type-B assumption is placed on the bias. They called the method BOB. They showed BOB can be partially justified by a Bayesian hierarchical model. In this work, we use a full Bayesian formulation to obtain a consensus value and an associated uncertainty interval. Our treatment is consistent with ISO uncertainty guidelines.

We assume noninformative priors on method means and standard deviations. We further assume an informative prior on the true value γ and derive the posterior of γ given the measurement data. The mean of the posterior is proposed as the consensus value, the standard deviation is used as a standard uncertainty, and a symmetric interval about the mean with a specified probability content is used to obtain the extended uncertainty, or equivalently, the degrees of freedom. Further, we show how the consensus value may be reported in a manner that is consistent with the ISO uncertainty guidelines. We call this Bayesian solution B-BOB for "Bayesian version of BOB." We give a computational algorithm and illustrative examples.

B-BOB provides a theoretically justifiable approach for incorporating scientific judgment in the form of a prior (type-B) distribution on the true value γ. The method is easy to implement. The resulting uncertainty intervals are expected to be much more realistic and not unduly conservative as is the case with frequentist approaches when the number of methods is small.

3.1.6 International Study of the Sublethal Effects of Fire Smoke on Survivability and Health

Blaza Toman
Statistical Engineering Division, ITL

Richard Gann,Julie Neviaser
BFRL

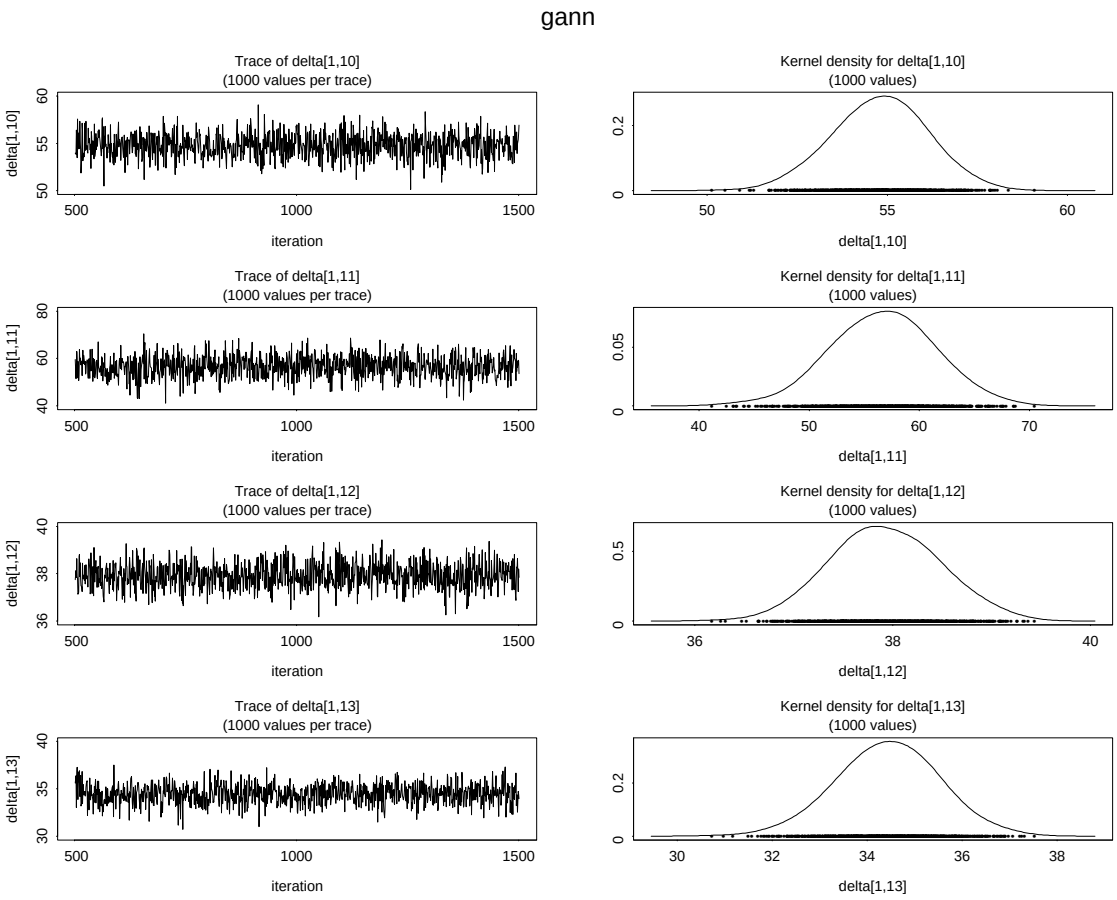


Figure 5: CODA diagnostics output for the SEFS project

In May 2000, the Fire Protection Research Foundation (FPRF) and the National Institute of Standards and Technology (NIST) began a major private/public fire research initiative to provide scientific information for public policy makers. Entitled the "International Study of the Sublethal Effects of Fire Smoke on Survival and Health" (SEFS), the project objectives are to: 1. Identify fire scenarios for which sublethal exposures to smoke lead to significant harm; 2. Compile the best available toxicological data on heat and smoke and their effects on escape and survival of people of differing age and physical condition, identifying when existing data are insufficient for use in fire hazard analysis; 3. Develop a validated method to generate product smoke data for fire hazard and risk analysis; and 4. Generate practical guidance for using these data correctly in fire safety decisions. The first phase of the research began in May 2000 and included the following tasks: a. provide decision-makers with the best available lethal and incapacitating toxic potency values for the smoke from commercial products for use in quantifying the effects of smoke on people's survival in fires. b. assess the potential for using available data sets to bound the magnitude of the U.S. population who are harmed by sublethal exposures to fire.

The statistical objectives of this study were to arrive at consensus values and uncertainty measures for lethal and incapacitating toxic potency values for a large number of building materials under three different combustion conditions. The consensus values were to be based on data obtained from many different studies found in the literature. The quality of the data varied from study to study in terms of the information provided and of accuracy. The sample sizes varied greatly. Some studies provided confidence intervals and some did not. Because of the lack of consistency of the data, it was decided that a Bayesian meta analysis would be the best approach to arrive at the consensus values. A Bayes hierarchical model was constructed using all vague priors at the lowest level of the hierarchy. The structure of this model allowed for the combination of data from studies that included uncertainty measures and those that did not. The posterior means provided consensus values and the posterior standard deviations provided the uncertainty measures. The computations were carried out using the Markov Chain Monte Carlo methods based on the BUGS software.

The impact of this work is potentially large because the study is being disseminated throughout the building industry, both nationally and internationally. The Bayesian methodology was developed at SED after attempts at classical analysis by the scientists did not produce satisfactory results, with the Bayesian approach subsequently receiving a lot of positive attention at NIST.

3.1.7 Using Bayesian Methods to Improve a Calibration Program: Calibration of a Pressure Transducer Using Dynamic Linear Modeling

Charles Hagwood
Statistical Engineering Division, ITL

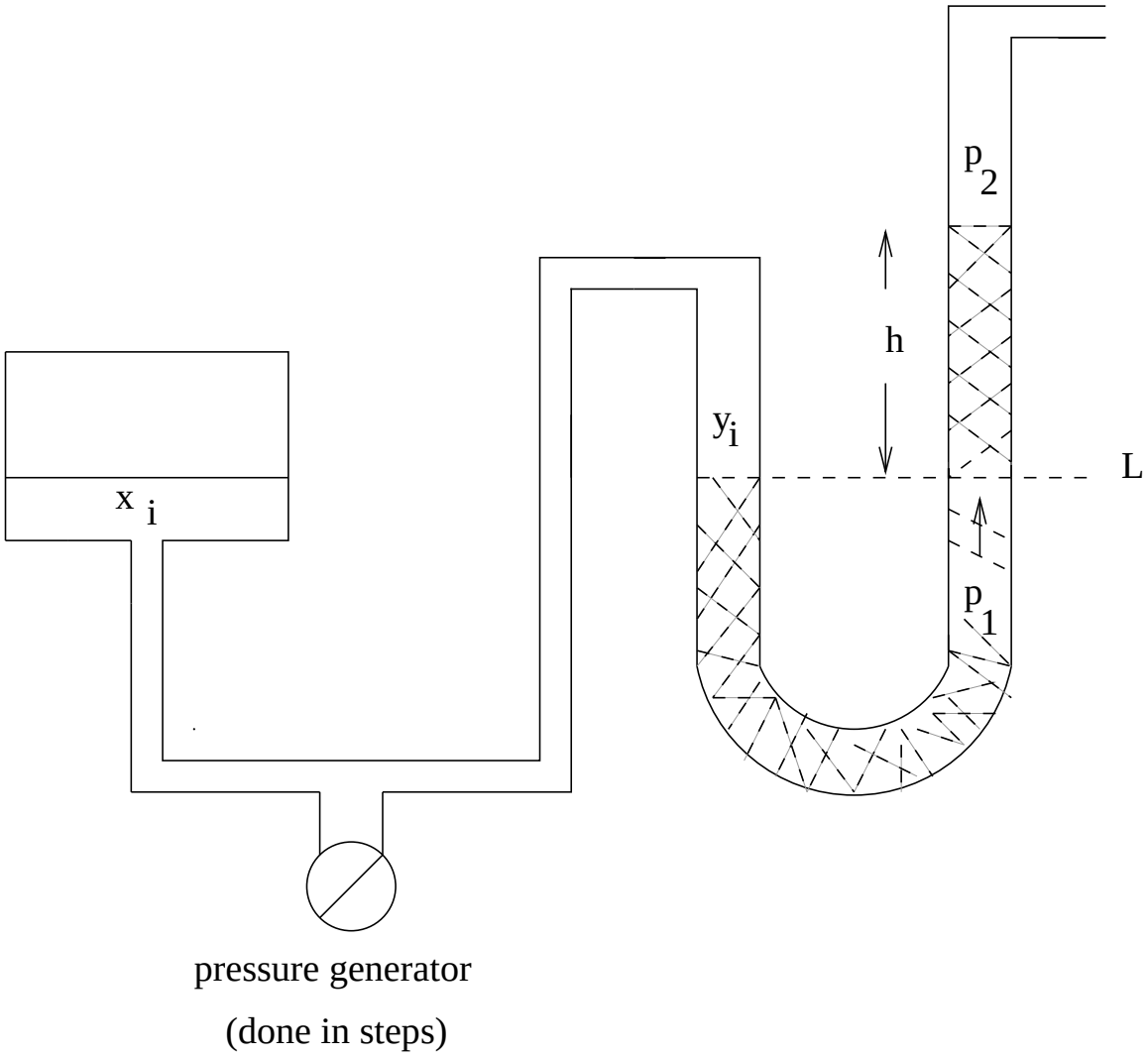


Figure 6: Calibration Set-up.

The Pressure Measurements Division often calibrates and recalibrates an instrument several times during the instrument's useable lifetime. Each recalibration is done without applying prior calibration data. An improvement of this program is being studied that makes use of the prior calibration data via Bayesian methods. By using the previous calibration data as prior information, the Bayesian procedure may provide a reduction in uncertainty of the calibration estimates.

One instrument among the many that the NIST Pressure Measurements Division calibrates is the pressure transducer gauge. The pressure transducer gauge is an instrument that converts pressure into an electrical signal. Because these are low pressure gauges, a manometer is used as the calibrating instrument. The manometer is essentially a liquid filled $U - tube$ for which the vertical separation of the liquid's surfaces gives a measure of the difference between the pressures at the ends. The liquid is usually mercury, water or oil, whose densities are well known. An illustration of the coupling used in the calibration is shown in the figure above. The gauges are connected by a manifold, which is connnected to a pressure generator. At the start of the calibration, pressure at a certain level is released into the manifold from the generator. The released pressure produces pressures at both the manometer and the transducer and they are recorded as a pair (x_1, y_1). This starting pressure is ramped up to create two new pressures at the two instruments, (x_2, y_2), and so on until say, n comparison values are produced, (x_n, y_n). At each pressure setting the pressure induced at the manometer can be very accurately determined. Based on the manometer reading, a correction is found for the transducer. That is, the correction is based on a calibration curve between the x's and y's.

The calibration curve is just a functional relationship between the measurements $y = f_\theta(x)$, with θ a parameter that is estimated using the calibration data, (x_i, y_i). When there is more variability in the transducer than the manometer, y is taken as the manometer reading. Then, an inversion provides the correction for x, i.e., $x = f_{\hat{\theta}}^{-1}(y)$. Of course, a monotonic portion of the calibration curve is taken as the relevant domain for which the relation holds. Often when there is no appreciable difference between the variances of x and y, no inverse is necessary, i.e., the relation $y = f_\theta(x)$ gives the correction, where now y denotes the manometer reading.

The transducer studied was calibrated and recalibrated 5 times over the last twelve years, 1990, 1992, 1994, 1996 and 1999. The data collected at each calibration period were recorded as $(x_i(t), y_i(t)),\ t = 1990, 1992, 1994, 1996, 1999$

$$\begin{aligned} x_i(t) &= ith\ transducer\ reading\ at\ time\ period\ t \\ y_i(t) &= ith\ manometer\ reading\ at\ time\ period\ t. \end{aligned}$$

The report to the customer is the calibration curve

$$y(t) = a_0 + a_1 x(t) + \cdots + a_j x(t)^j + b_1 Log_{10}(x(t)) + b_2 (Log_{10}(x(t)))^2 \qquad (3.1)$$

along with an uncertainty for the transducer when using the calibration equation. Note that inverse calibration is used, i.e., the primary standard is taken as the y-value.

Bayesian calibration of the transducer uses the predictive density to compute a calibration estimate, as well as an uncertainty interval. The previous calibrations are used as prior

information and a dynamic linear model is used to tie the calibrations together. The dynamic linear model is

$$Y(t) = X(t)\theta(t) + \phi\epsilon(t) \quad (3.2)$$

$$\theta(t) = \theta(t-1) + \phi\omega(t) \quad (3.3)$$

with $X(t), Y(t)$ as given in Eq.(1) and $\epsilon(t)$ is a $n(t) \times 1$ vector of multivariate normal errors with mean zero and variance-covariance matrix the identity matrix. The vector of regression parameters is denoted by $\theta(t)$. It is assumed to have a multivariate normal distribution. The error $\omega(t)$ is independent of $\theta(t-1)$ and $\epsilon(t)$. This is the standard dynamic linear model with parameters (θ, ϕ), see Pole, West and Harrison (1994). They give ways to determine $V(t)$, the variance-covariance matrix of $\omega(t)$.

At each time period, $t = t_i, i = 1, \ldots, 5$ ($t_i = 1990,\ 1992,\ 1994,\ 1996,\ 1999$), the priors are chosen recursively, starting with a vague Normal Gamma prior on $(\theta(t_0), \phi)$.

The dynamic linear model algorithm goes as follow:

• Start with a Vague Normal Gamma Prior on $(\theta(t_0), \phi)$ with parameters $\mu(t_0), \lambda(t_0), \alpha(t_0), \beta(t_0)$.

• Use the state equation $\theta(t_1) = \theta(t_0) + \phi\omega(t_1)$ to compute a derived prior for $(\theta(t_1), \phi)$. It will be normal gamma with parameters $\mu_d(t_1), \lambda_d(t_1), \alpha_d(t_1), \beta_d(t_1)$.

• Compute the posterior density of $(\theta(t_1), \phi)$ given the data $y(t_1), X(t_1)$. It will be normal gamma with parameters $\mu_{pos}(t_1), \lambda_{pos}(t_1), \alpha_{pos}(t_1), \beta_{pos}(t_1)$.

• Compute the marginal density $\theta(t_1)$. It will be Student's-t with known parameters. Use its mean to estimate $\theta(t_1)$. This can be used to check fit with the data, $\hat{y}(t_1) = X(t_1)E[\theta(t_1) \mid Data]$.

• Compute the predictive density of $y_{fut}(t_1)$; it will be Student's-t with known parameters.

• At time t_2, use the posterior of $(\theta(t_1), \phi)$ as the prior.

• Repeat.

Several comparisons are made between the Bayesian and the Classical calibration solutions, e.g., width of the uncertainty intervals. A comparison of predicted values is performed. Here one row is removed from each data set and a comparison is made to determine which method is better in predicting the value of the deleted row.

The NIST Pressure Measurements Division has shown interest in this methodology. Some of these results were presented at the National Conference of Standards Laboratories in Washington, D.C. in August 2001. Several calibrations laboratories expressed interest.

3.1.8 Measurement Protocol Development, Inhomogeneous Experimental Units, and Markov Chain Monte Carlo

Walter Liggett
Statistical Engineering Division, ITL

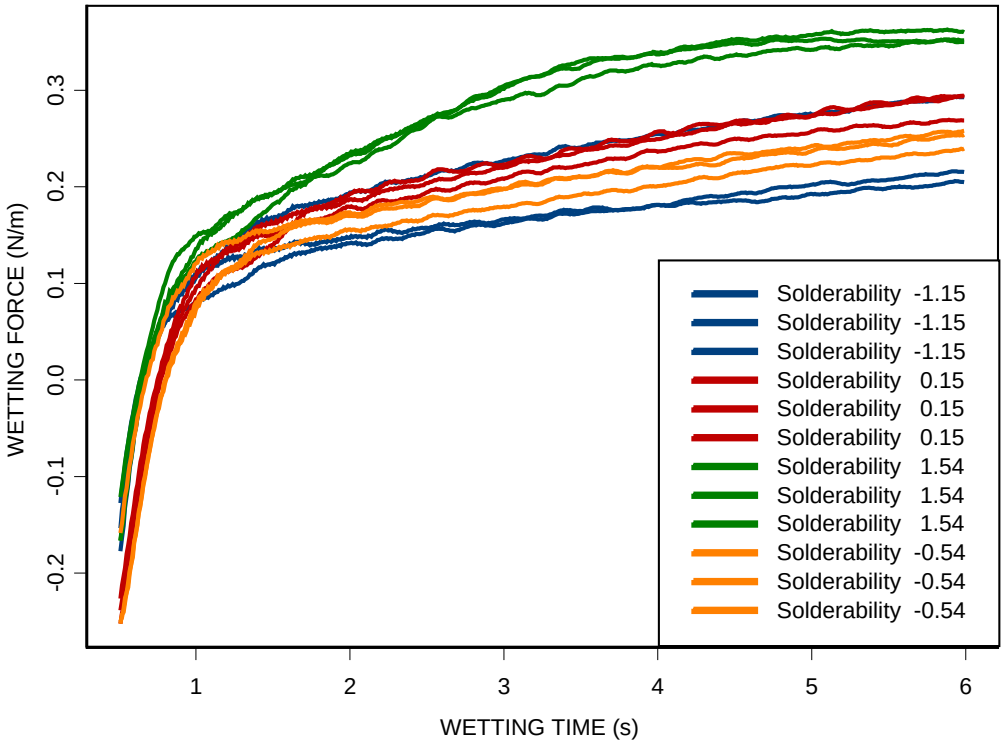


Figure 7: Wetting force curves for four test-lead sets (estimated solderabilities in legend). Within set variation due to variability in wetting balance and test-lead fabrication.

In the physical sciences, a measurement protocol specifies how a response, a raw measured value, is to be obtained from a unit, a physical object. In the simplest case, the response Y is given by

$$Y = \alpha + X\beta + E,$$

with X denoting the property of the unit that is to be measured, α and β are characteristics of the protocol, and E is a random variable with variance σ^2 that is also a characteristic of the protocol. A basic attribute of a protocol is its sensitivity, the ratio β^2/σ^2. In the simplest case, protocol development is maximization of the sensitivity. Subsequent steps in the development of a measurement system include calibration, which is the determination of α and β, and assessment of the uncertainty in measurements that the system produces. As has long been recognized, protocol development should precede interlaboratory testing of a measurement system.

Approaches to protocol development are distinguished largely by physical aspects. Although sometimes there is scientific theory to guide protocol development, of interest here is the case in which experiments provide primary guidance. Such experiments call for responses obtained when alternative protocols are applied to various experimental units. In many cases, the protocols are destructive; that is, each experimental unit yields only one response. These more difficult cases are the focus of this work.

Approaches to experimental protocol development are distinguished by the availability of experimental units. The experiments do not require experimental units with known values of X, the property of interest, but do require that the experimental units vary appreciably in the property of interest and that other variation in the experimental units match the intended application of the protocol. With an appropriate set of experimental units, one can compare protocols according to whether they respond selectively to the property of interest and according to the degree of their sensitivity.

Consider a protocol development experiment with a simple design. Assume that the values of X for the experimental units cluster tightly around each of three or more levels that are widely separated. Say that there are three or more alternative protocols, each with its own calibration curve and random error. The experiment consists of obtaining responses from all the protocols for some experimental units at each level. In the analysis of the experimental data, there is some lack of identifiability between the variation in execution of the protocol and variation in the experimental units. Nevertheless, if for each protocol and level, a sufficient number of experimental units are measured, then the sensitivity of the protocols can be ranked. The challenge is inference when only a modest number of experimental units is measured. This challenge can be met with Bayesian analysis based on Markov Chain Monte Carlo. A script for the computing can be created through modification of scripts for the case of longitudinal data. There are, of course, many questions that must be answered in performing such an analysis.

In reaching the ultimate goal of making Bayesian analysis available to engineers who do protocol development, several milestones must be passed. The first is doing a Bayesian analysis of wetting-balance data for which a frequentist analysis has already been done. The second is the comparison of alternatives in the statistical design and analysis. The

third is pursuit of a collaboration with someone who is using combinatorial methods for studies in material science. The fourth is teaching an SED course on protocol development.

As examples of the importance of protocol development, consider determination of the hardness of materials, interpretation of combinatorial experiments on materials, and improvement of manufacturing quality. Indentation protocols for hardness measurement involve choice of indenter shape, amount of force applied to the indenter, and holder for the experimental unit. Experimental units with a broad hardness range can be obtained through process variation such as variation in heat treatment in the case of steel. Combinatorial experiments are attractive because they utilize very small specimens for each experimental run. Success requires that protocols be developed that allow material properties easily measured on much larger specimens to be determined on small specimens. Improvement of manufacturing quality generally requires measurement of the manufactured units. By varying the manufacturing process, one can obtain a range of experimental units that can be used to improve the measurement protocol. With this new protocol, one can reduce the variation in the manufacturing process. This process can be iterated further. These examples show that protocol development is often a key step in reaching experimental goals.

3.2 Key Comparisons and Uncertainty Principles

3.2.1 Key Comparisons and Uncertainty Principles

Nien Fan Zhang, Will Guthrie, Blaza Toman, Nell Sedransk
Statistical Engineering Division, ITL

With the recent signing of the Mutual Recognition Arrangement (MRA) by the International Committee for Weights and Measures (CIPM), National Metrology Institutes (NMI's) and Regional Metrology Organizations (RMO's) around the world have committed themselves to establishing the equivalence of their measurement standards. Currently, however, no agreement exists as to the best statistical procedures for doing this. To assure accurate, efficient assessment of equivalence, the Statistical Engineering Division (SED) has proposed to provide a unified statistical framework and detailed guidance for this not-yet-defined process.

The MRA responds to a growing need for an open, transparent, and comprehensive scheme to give users reliable quantitative information on the comparability of national metrology services. It will also provide the technical basis for wider agreements negotiated for international commerce and regulatory affairs. A key to meeting the objectives of the MRA, however, is a sound and accepted set of procedures for establishing the equivalence of national standards.

Interlaboratory studies establish and ensure measurement capability for commerce since accurate measurements are necessary for assessing product specifications. SED statisticians have been responsible for the statistical design and analysis of interlaboratory studies for many years. Key Comparisons are international interlaboratory comparison studies chosen by the Consultative Committees under the CIPM to establish the degree of equivalence between national measurement standards. The Consultative Committees are responsible for identifying a set of key comparisons in each field, which covers a range of standards, so as to test the principal techniques in the fields. Recently, Key Comparisons have provided many new opportunities for SED to collaborate with scientists across NIST. SED has been involved in international Key Comparison projects in collaboration with eight out of 10 Consultative Committees under the CIPM.

A Key Comparison data base has been developed jointly by NIST and the International Bureau for Weights and Measures (BIPM). However, at the present time, there is no consensus among the various international labs and Consultative Committees as to the best choice of procedures performed at each step. Key Comparison testing is at its core a statistical process. Data are collected, statistically analyzed, and a reference value and degrees of equivalence among the participating laboratories are estimated. We see great benefits to the international community in developing a statistical roadmap that will clarify the choices and optimize the process. This year SED formed a project team to promote a unified approach to experimental design and analysis for Key Comparisons.

Specifically, the data collection phase needs statistically sound and efficient experimental

designs. This includes decisions as to the number of traveling standards and the pattern of the comparisons. It also includes determination of the sample size for each measurement at each lab, and possibly the layout of the experiment at each lab if more than a single comparison is being performed. We propose to study the issues of the experimental design phase, ultimately identifying a core set of conditions and physical constraints under which a design needs to be efficient.

The second phase of the Key Comparison process is the determination of the reference value and the assessment of NMI standard uncertainty. The question of whether and when a reference value is needed, and if it is needed how to estimate it, must be addressed. The uncertainty analysis consists of the decomposition of variance between the transfer standard and each NMI's standard. The basic principles for this are well known and internationally accepted. However, there are various valid alternate procedures and it will be valuable to study these and set out some guidelines for their use.

The final phase of the Key Comparison process is the determination and reporting of the level of equivalence among the participating labs and related uncertainties. Presently, there are several ways used to quantify the degree of equivalence. We believe that it would be beneficial to have a standard process for this task.

Meeting the research challenges described above will reduce the potentially duplicated effort expended in planning and executing key comparisons and improve the quality of results obtained. Participants and users of Key Comparisons will also gain a clearer understanding of Key Comparison data and results from the unified interpretations following from the statistical research.

This project will directly support NIST's new efforts to establish equivalence with other NMI's and RMO's under MRA. Recent collaborations in this area between SED and staff from other OU's have clearly identified the desire and need for guidance in carring out Key Comparisons. In a larger context, this project is in keeping with the recent trend toward open markets favored by a broad range of economists, industry leaders, and governmental and inter-governmental organizations.

3.2.2 Comparison of Temperature Realizations from 83.8058 K to 933.473 K

William F. Guthrie
Statistical Engineering Division, ITL

Greg Strouse, Billy Mangum
Process Measurements Division, CSTL

Erich Tegeler
Physikalisch-Technischen Bundesanstalt, Germany

John Connolly
National Measurement Laboratory, Australia

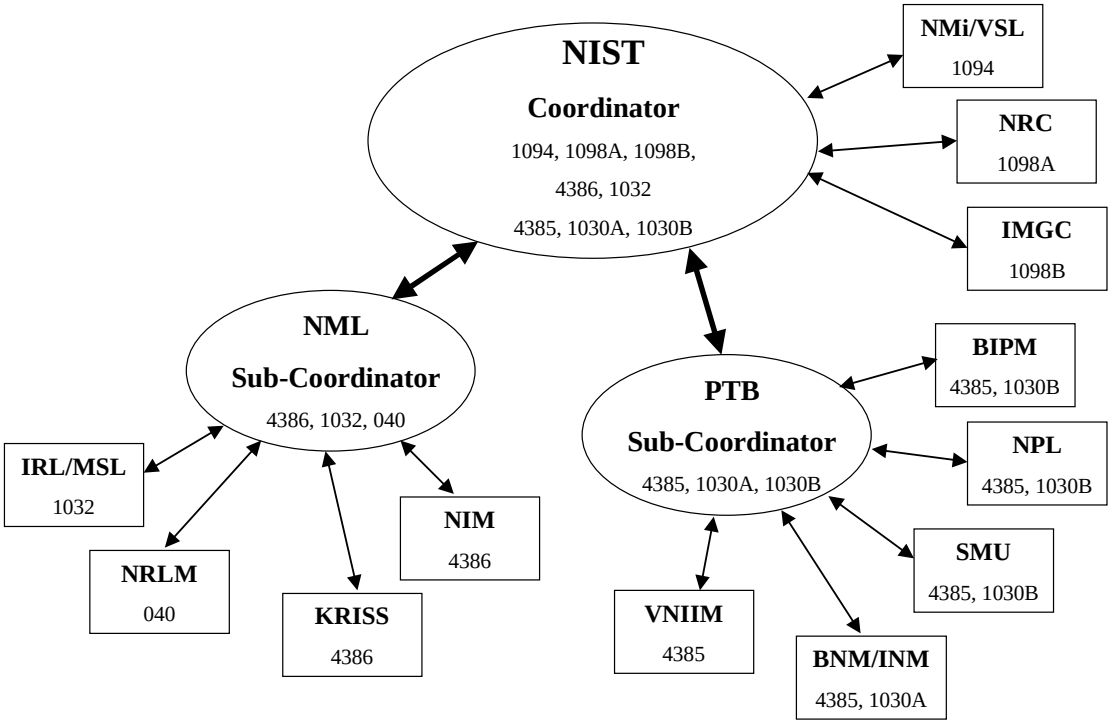


Figure 8: Schematic showing the experiment design of an interlaboratory experiment to compare realizations of the ITS-90 at different NMI's. Each oval or rectangle in the schematic represents an NMI and lists which transfer instrument (SPRT's) that laboratory calibrated in order to compare their realizations of the ITS-90 with the other laboratories. The NMI's that participated in the experiment are BIPM (International), BNM/INM (France), IMGC (Italy), IRL/MSL (New Zealand), KRISS (South Korea), NIM (China), NIST (United States), NMi/VSL (Netherlands), NML (Australia), NPL (United Kingdom), NRC (Canada), NRLM (Japan), PTB (Germany), SMU (Slovakia), and VNIIM (Russia). The arrows connecting the laboratories show how the transfer instruments (1030A, 1030B, 1032, 1094, 1098A, 1098B, 4385, 4386, and 040) were circulated between the laboratories.

As part of the ongoing effort by National Metrology Institutes (NMI's) around the world to assess the equivalence of their measurement standards and procedures, an interlaboratory comparison of different countries' realizations of the International Temperature Scale of 1990 (ITS-90) was lead by scientists at NIST. Fifteen laboratories participated in this experiment, a Key Comparison, which compared the temperature realizations at seven defining fixed points on the ITS-90 and one supplementary fixed point. Each fixed point on the temperature scale is the melting, freezing, or triple point of a pure metal. The fixed points for which comparisons were made are argon (Ar), mercury (Hg), gallium (Ga), indium (In), tin (Sn), cadmium (Cd), zinc (Zn), and aluminum (Al), in order of increasing temperature. Standard platinum resistance thermometers (SPRT's) were used to compare the temperature realizations obtained with each fixed point. For this experiment, each laboratory calibrated one or more SPRT's circulated between countries by NIST. Seven different SPRT's were used in the experiment to parallelize the running of the experiment. The design of the experiment is shown in the figure on the preceding page. The results of the experiment, relative to the NIST results, are shown on the figure on the next page.

As part of the team leading this Key Comparison, SED staff helped with the analysis of the data, which had been collected over several years. Because the protocol for the experiment, agreed to by the participating laboratories prior to the signing of the Mutual Recognition Arrangement, did not specify exactly how many measurements each laboratory would make or what format would be used to submit their results, one of the first tasks in the analysis of the data was to confirm whether the submitted components of uncertainty were for an individual measurement or for the mean of replicate measurements. Information on the number of measurements made by each laboratory and the degrees of freedom for each uncertainty estimate were also obtained. The next step in the analysis was to determine how each uncertainty component contributed to the measurement error of the process so that covariances between measurements within each laboratory could be accounted for. Other issues that affected the analysis of the data included determination of uncertainties associated with the SPRT's used as transfer instruments, appropriate computation of coverage factors to obtain expanded uncertainties with correct confidence levels, explanation of effects arising from different paths for computing temperature differences across subsets of laboratories, interpretation of Key Comparison Reference Values, and demonstration of methods for linking these results with future supplementary comparisons between Regional Metrology Organizations.

The results of this study will be included in the Key Comparison Database maintained by the BIPM according to the Mutual Recognition Arrangement. There they will be available for use by anyone interested in determining how differences in international temperature standards might impact the sales of temperature-related equipment or services between specific countries. A summary of this Key Comparison will also be published in Metrologia. *With these results widely available, temperature differences between countries can be further studied to determine their causes and to find ways to reduce their magnitude. In addition, the questions raised and experience gained in this study can be used to improve the quality and reduce the costs of future Key Comparisons.*

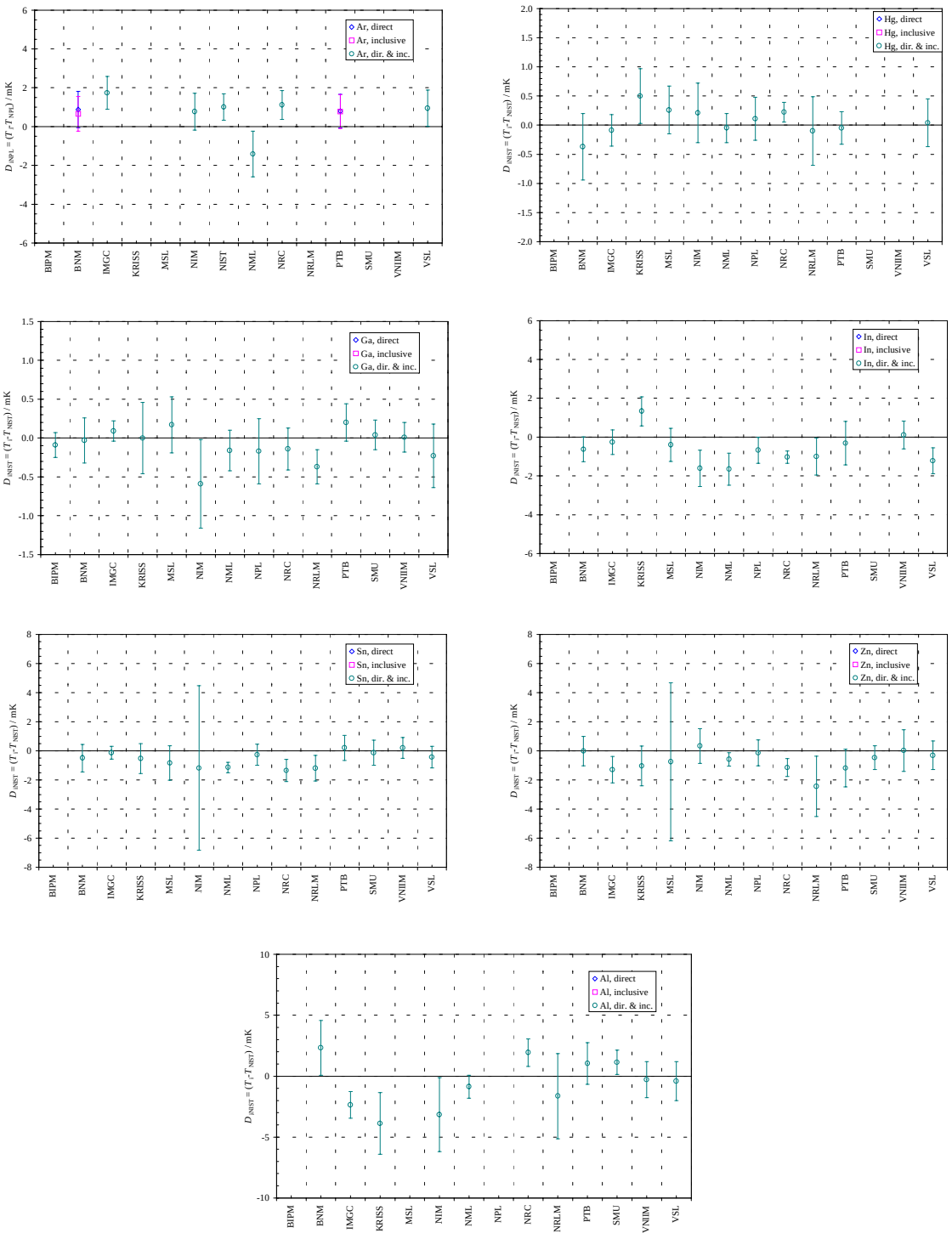


Figure 9: The plots above show the results of this Key Comparison for the seven defining fixed points of the ITS-90. The results are shown here as temperature differences from the NIST results. Similar results, relative to each participating laboratory, will be included in the Key Comparison Database. The uncertainties shown for each result are approximate 95% confidence intervals for the difference in temperatures realized by each laboratory. Intervals that include $D_{i,NIST} = 0$ indicate that there is no statistically significant difference between the temperature realizations of the specified laboratory and NIST.

3.2.3 Expression of Uncertainty in Functional Measurement

Walter Liggett
Statistical Engineering Division, ITL

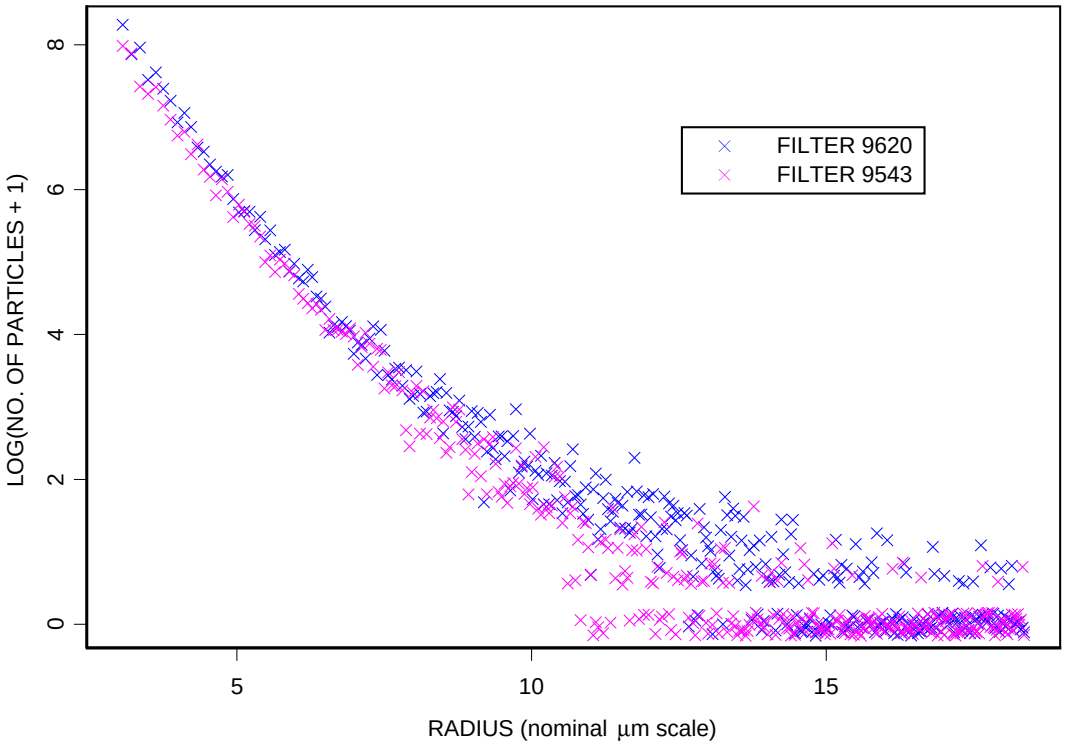


Figure 10: From two different filters, particle count as a function of particle size (plot positions jittered).

Commonplace at NIST are data of interest that represent functions of a continuous independent variable. Examples include chemical spectra (spectroscopy, chromatography), material test methods (hardness measurements, indicators based on a wetting balance, instrumented indentation), measurements on combinatorial libraries constructed with material gradients (polymer films, laser deposition of dielectric films), particle-size distributions, and traces of physical motion (handwriting). Generally, such functional measurements are few in number, and each has high dimension. Thus, such measurements are properly considered functional data rather than longitudinal data.

A primary activity in physical sciences metrology is comparison of measurements for the purpose of understanding underlying sources of variation. This contrasts with the summarization of measurements for the purpose of understanding the population from which the measurements were obtained. In metrology, one often seeks to determine whether the difference between two measurements can reasonably be attributed to known sources of variation. So that this determination can be made for the difference between an existing measurement and a future measurement, the existing measurement must be accompanied by its uncertainty. This simple scenario provides a basis for developing methods for expressing uncertainties.

Development of a way to express the uncertainty in a functional measurement can begin with methods for comparison of functional measurements. Comparisons based on naive adoption of elementary statistical methods have obvious weaknesses. In fact, even for comparison of two functional measurements, methods of inference are a topic of current research.

Comparison of two functional measurements is an issue only if the underlying functions are observed with noise. The figure shown above contains two functional measurements of a particle-size distribution, each measurement being a series of counts subject to Poisson variation. Because of the noise, the comparison method, that is, the method for testing the null hypothesis that the known sources of variation account for the observed difference, must include smoothing. The need for smoothing stems from the fact that the dimension of the functional measurement is much larger than the number of parameters needed to specify the underlying function. However, when the smoothing needed to increase the power of the comparison must be data-driven, which it usually must, the process of statistical inference becomes complicated.

There are many smoothing techniques, each one with its own advantages and each extensible, with some difficulty, to include statistical inference. The three major classes of techniques are local polynomial smoothing, penalized smoothing splines, and adaptive regression splines. Local polynomial smoothing gives results at a particular value of the independent variable that can be interpreted without concern for the effects of the whole domain of the observation. Methods based on penalized smoothing splines can be automated and can be extended to analyses involving diverse groups of functional measurements. Adaptive regression splines can accommodate parsimoniously a wider range of variation in the underlying function. In choosing a smoothing technique for a particular NIST functional measurement, these major differences and many others must be considered.

In a search for unexpected differences between two functions, a minimum of assumptions about the form of the difference under the alternative hypothesis is clearly preferable. In the case of two functions, one can think in terms of representations of the sum and the difference rather than representations of each function separately. Under the null hypothesis, the difference is zero or at least has a specific form, and the sum can have any form. Under the alternative hypothesis, the difference can be assumed to be smooth but otherwise to be arbitrary. There are difficulties in developing a test appropriate for these assumptions. If one assumes cubic spline representations with given knots, then a test can be specified. If one applies an adaptive regression spline technique to find the knots, then one has to be concerned about the bias introduced in the test by the knot selection. If one applies a penalized smoothing spline technique, there may be less reason to be concerned with selection bias because only two smoothing parameters, one for the sum and one for the difference, have to be selected. On the other hand, penalized smoothing splines may not fit the underlying functions well. Clearly, choice of a test for comparing two functional measurements involves several statistical trade-offs.

It is possible that knowledge of the science underlying the functional measurements can be used in the choice of a test. Even though a difference between two underlying functions is unexpected, the likely causes might be such that a form for the difference under the alternative hypothesis can be specified. For example, in the case of the data shown in the figure above, one might expect the difference to have no more than one or two peaks. In such a case, the testing problem can be thought of as semiparametric with the difference specified parametrically and the sum specified nonparametrically. When a semiparametric approach is justified, the power of the resulting test may be much greater than the power of an applicable nonparametric test.

As practiced by scientists, functional data analysis is almost exclusively graphical. When two functions are to be compared, they are compared graphically. If there is a difference discernable amidst the noise and if the reason for this difference is apparent, then the scientist will pursue the problem. Otherwise, the scientist is not likely to take any action. In using statistical methods, scientists will have an objective criterion for pursuing apparent differences. With this criterion, the chances of wasting time on appearances caused by noise will be controlled.

There is a vast and growing literature on functional data analysis, some of which is of value to NIST. Sorting through this literature on the basis of the character of NIST functional measurements and the statistical terminology in this literature will have the benefit of uncovering those techniques that have clear value in NIST's functional measurements.

3.2.4 Bayesian Analysis of the CCPR Key Comparison on Near-Infrared Spectral Responsivity

Blaza Toman, Dipak Dey
Statistical Engineering Division, ITL

Steven Brown, Thomas Larason, Keith Lykke
Optical Technology Division

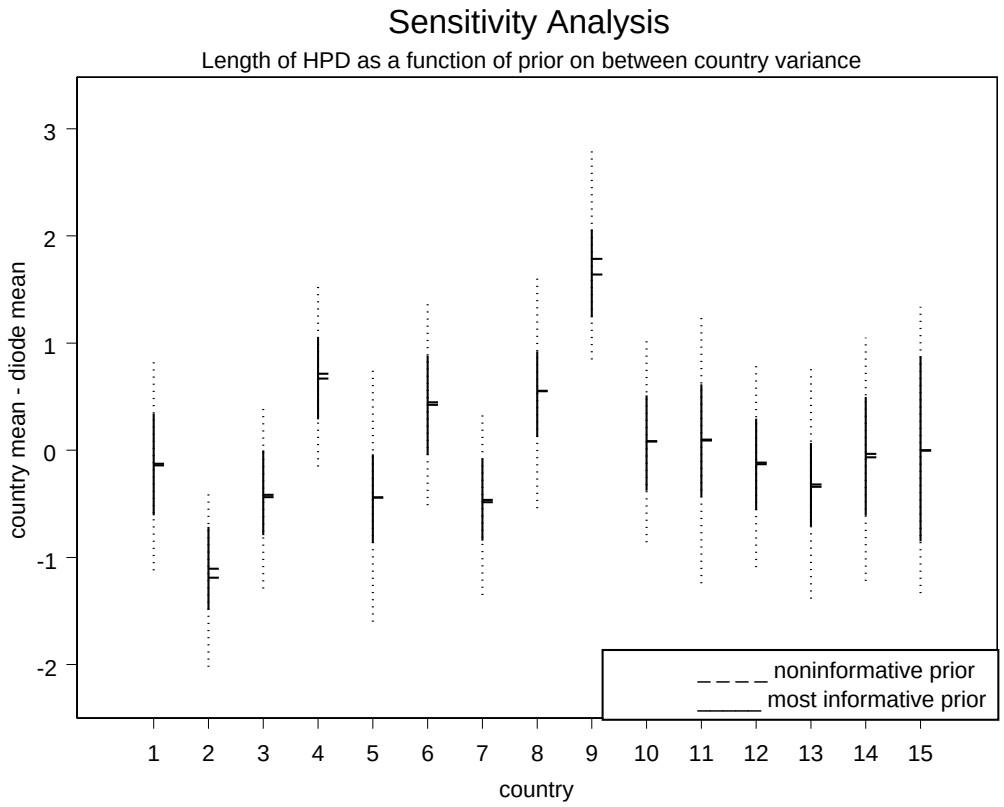


Figure 11: This figure shows the differences in the lengths of the HPD intervals as the prior of the between-laboratory variance becomes more informative

An international comparison of near-infrared spectral responsivity was executed under the direction of the Consultative Committee on Photometry and Radiometry (CCPR). The comparison was structured in a star pattern, with the National Institute of Standards and Technology as the host laboratory. A total of fifteen national laboratories participated in one of 4 rounds. Three indium gallium arsenide (InGaAs) detectors were sent to each laboratory, one of a particular type and two randomized detectors from 5 different vendors. NIST measured the photodetectors between rounds to establish their radiometric stability.

The goals of the statistical analysis were to assess the agreement between the laboratories and to arrive at a common reference value for each wavelength. Achieving these goals was complicated by the fact that in addition to the usual laboratory effect, there are very strong detector and manufacturer effects, and the star pattern of the experimental design allowed for only an incomplete overlap of individual detectors with individual laboratories. That is, each detector was sent to at most four laboratories (plus NIST) and some were sent to only one laboratory (plus NIST). The complexity of the problem made it a particularly good candidate for a Bayesian hierarchical model analysis. The model was constructed so that the detector effects were related by a common prior distribution, and so were the manufacturer effects, thus allowing for "borrowing of strength". Type B uncertainty was incorporated into the model and vague priors were used at the last level of the prior hierarchy. Posterior means provided the common reference values for each wavelength, with posterior standard deviations used for the uncertainty measure. Predictive distributions were used to assess the agreement between the laboratories. Sensitivity analysis was used to evaluate the degree to which the prior distributions affected the results. It showed that the functional form of the prior distributions had little effect on the results. It further showed that for most of the parameters, the prior variances also had little effect. Only the size of the prior variance of the between-laboratory variance had a significant effect on the sizes of the HPD intervals.

The Bayesian analysis used in this Key Comparison should be applicable to other Key Comparisons with similar structure and objectives.

3.2.5 A Generalized Confidence Interval for the Consensus Mean

C. M. Wang and H. K. Iyer
Statistical Engineering Division, ITL

Thomas Mathew
University of Maryland, Baltimore County

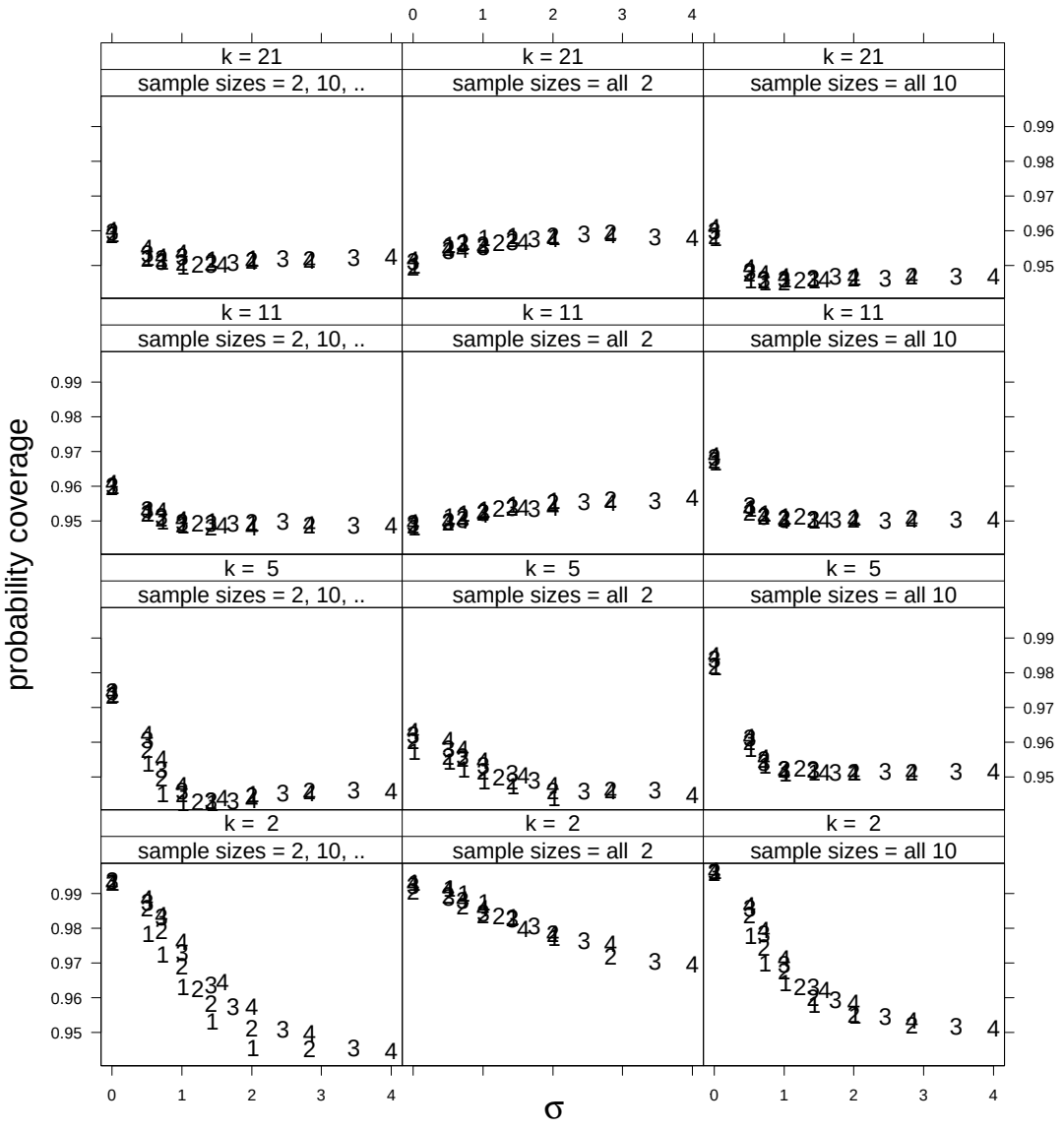


Figure 12: Each panel plots the empirical simulated coverage probability of the interval vs. the value of σ for a specific combination (or pattern) of k and n_i. The plotting symbols are designated for cases with different values of the within-laboratory error variance. The nominal confidence coefficient is 0.95.

We consider a one-way random effects model with unequal sample sizes and unequal variances. Using the method of Generalized Confidence Intervals, we develop a new confidence interval procedure for the mean. Statistical simulation is used to demonstrate that the procedure maintains coverage probabilities close to the nominal value.

Sometimes it is necessary to combine results from multiple, independent methods such as in the case of interlaboratory comparisons. In an interlaboratory comparison, measurements of an artifact are made by each of k laboratories. A one-way analysis of variance model, given below, is usually used to describe the measurements

$$Y_{ij} = \mu + b_i + \epsilon_{ij},$$

with Y_{ij}, $i = 1, 2, \cdots, k$, $j = 1, 2, \cdots, n_i$, denoting the jth measurement of the ith laboratory, μ is the unknown common mean, b_i is the amount by which the true response of laboratory i deviates from μ; that is, the effect or bias of laboratory i, and ϵ_{ij}, $j = 1, 2, \cdots, n_i$, are measurement errors and assumed independently distributed as normal with mean 0 and variance σ_i^2. The purpose of interlaboratory comparisons is to estimate the common mean μ and to provide a standard error of this estimate for constructing a confidence interval for μ.

The particular statistical approach that is appropriate for the estimation of μ depends on what assumptions are made about the b_i. We consider a one-way random effects model, so the b_i, $i = 1, 2, \cdots, k$, are normally distributed with mean 0 and variance σ^2. Using the method of generalized confidence intervals, we develop a new confidence interval procedure for μ. The main advantage of the generalized confidence interval in this problem is that we can avoid computing point estimates of nuisance parameters, especially that of σ^2. The point estimate of σ^2 can be "unstable". Unlike the maximum likelihood solutions, the generalized confidence intervals perform well even for small k. The accompanying figure summarizes the coverage probabilities of the generalized confidence intervals for μ, obtained using simulation. The resulting generalized confidence interval procedure appears to hold the coverage probability sufficiently close to the nominal value in all the cases considered in our simulation study to make it a useful procedure for practical applications.

There appears to be no other satisfactory solution for setting confidence intervals on the mean of an interlaboratory comparison. The proposed method provides an adequate solution for a class of problems that are important to NIST activities.

3.2.6 Statistical Uncertainty Analysis of CCEM-K2 Comparisons of Resistance Standards

Nien Fan Zhang, Nell Sedransk
Statistical Engineering Division, ITL

Dean Jarrett
Electricity Division, EEEL

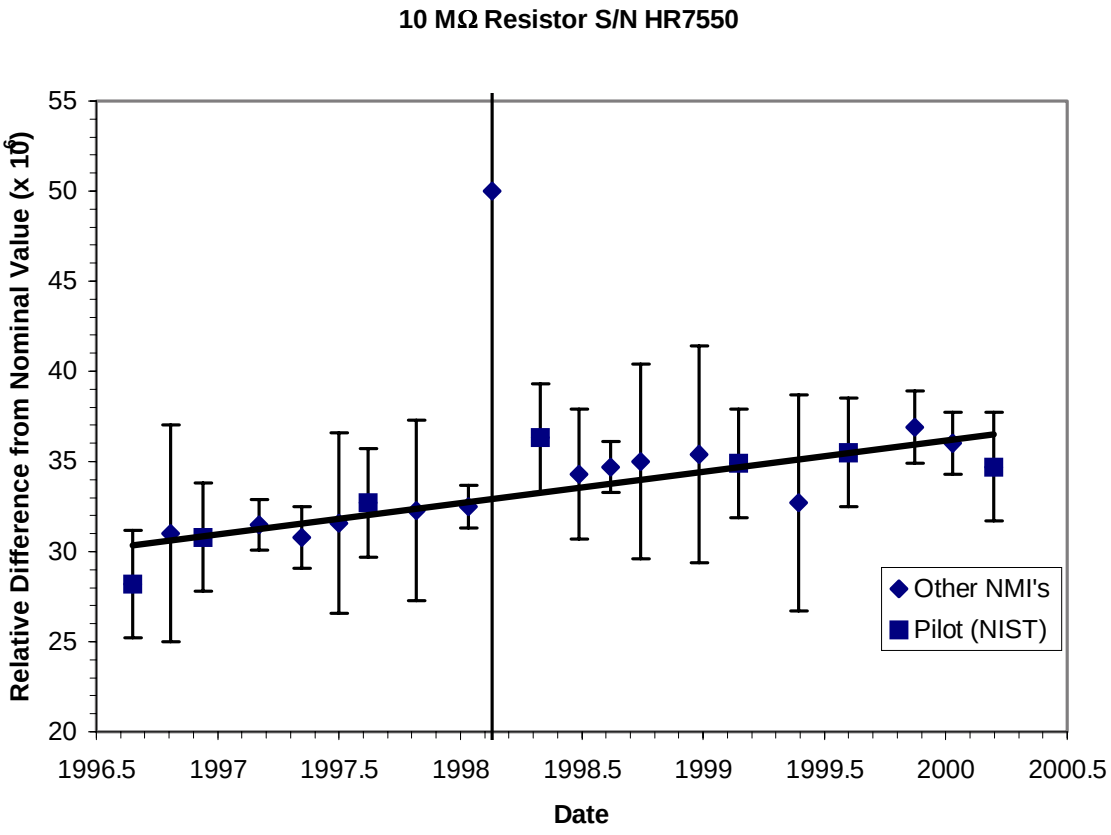


Figure 13: Measurement of 10M Ω standard S/N HR7550 by all participants. Error bars denote individual NMI's expanded relative uncertainty using $k = 2$. Linear regression line based only on the pilot lab measurements.

An international comparison of dc resistance at 10 MΩ and 1 GΩ was organized under the auspices of Consultative Committee for Electricity and Magnetism (CCEM) and piloted by NIST with 14 other national metrology institutes (NMIs) participating.

In this Key Comparison, three 10 MΩ wirebound resistors (S/Ns HR7550, HR7551, and HR7552) and three 1 GΩ film-type resistors (S/Ns HR9101, HR9102, and HR9106) were used as the traveling standards. Each participating NMI was requested to complete their measurements of the traveling standards within a two-month period. During the comparison, the traveling standards were measured at the pilot lab, NIST, for seven separate periods by two different measurement systems that are used on a regular basis to calibrate customer high-resistance standards.

In each case, for each of the three traveling resistors a simple linear regression line was fit to the NIST measurements of resistors using the mean date of each period of measurements as the predictor variable in the regression. See the attached figure. For a fixed traveling standard (the jth, $j = 1, 2, 3$) and each of the non-NIST NMIs, the difference between its measurement and the corresponding prediction or the reference value is $D_i(j) = x_i(j) - x_{i,p}(j)$, with $x_i(j)$ denoting the measurement for the jth traveling standard made by the ith ($i = 1, 2, ..., 15$) NMI and $x_{i,p}(j)$ is the prediction of the measurement of the ith NMI based on the jth regression ($j = 1, 2, 3$). Then, the differences for each of the three traveling standards are combined as a weighted average:

$$D_{i,comb} = \sum_{j=1}^{3} w(j) D_i(j),$$

with the weight $w(j)$ based on the reciprocals of the residual variances of the regressions. $D_{i,comb}$ represents the difference between the measurements made by the ith NMI and the prediction for this NMI based on measurements made by the pilot lab NIST. For a non-pilot lab NMI, the variance of $D_{i,comb}$ is obtained as

$$\sigma^2_{D_{i,comb}} = \sigma^2_{x,B,i} + \sigma^2_{x,A,i} \frac{\sum_{j=1}^{3} \frac{1}{\sigma_r^4(j)}}{(\sum_{j=1}^{3} \frac{1}{\sigma_r^2(j)})^2} + \frac{1 + \frac{1}{n} + \frac{(t_i - \overline{t}_{NIST})^2}{\sum_{k=1}^{n} (t_{NIST,k} - \overline{t}_{NIST})^2}}{\sum_{j=1}^{3} \frac{1}{\sigma_r^2(j)}},$$

with $\sigma_{x,A,i}$ and $\sigma_{x,B,i}$ denoting the root-sum-squares of the Type A and Type B uncertainties, respectively, of the ith NMI based on the NMI's uncertainty budget. $\sigma_r(j)$ is the standard deviation of the residual from the jth linear regression corresponding to the jth resistor ($j = 1, 2, 3$) and n (in this case $n = 7$) is the number of the periods that the resistors were measured by the pilot lab. t_i is the date when the measurements were made by the ith NMI and $t_{NIST,k}$ is the date for the kth period when the measurements were made by the pilot lab. $\overline{t}_{NIST}$ is the average of $t_{NIST,k}$ for $k = 1, 2, ..., 7$. For the pilot lab, NIST, the term corresponding to $D_{i,comb}$ is the mean of $D_{NIST,k,comb}$, denoting by $\overline{D}_{NIST,comb}$ and the variance of $\overline{D}_{NIST,comb}$ is

$$\sigma^2_{\overline{D}_{NIST,comb}} = \frac{\sigma^2_{x,A,NIST}}{n} \frac{\sum_{j=1}^{3} \frac{1}{\sigma_r^4(j)}}{(\sum_{j=1}^{3} \frac{1}{\sigma_r^2(j)})^2} + \sigma^2_{x,B,NIST},$$

with $\sigma_{x,A,NIST}$ and $\sigma_{x,B,NIST}$ denoting the root-sum-squares of the Type A and Type B uncertainty, respectively, of NIST. The reference value of the Key Comparison, X_{KCRV}, is defined as the weighted mean of the $D_{i,comb}$, including the mean of $D_{NIST,k,comb}$, i.e., $\overline{D}_{NIST,comb}$. Namely,

$$X_{KCRV} = \sigma^2_{KCRV} \sum_{i=1}^{15} \frac{D_{i,comb}}{\sigma^2_{D_{i,comb}}}.$$

The variance of X_{KCRV} is also obtained.

The degree of equivalence between two (ith and jth) NMIs is defined as

$$D_{i,j} = D_{i,comb} - D_{j,comb}.$$

The uncertainty of $D_{i,j}$ depends on whether any one NMI is the pilot lab or not. When neither NMI is the pilot lab, the variance of $D_{i,j}$ is

$$\sigma^2_{D_{i,j}} = \sigma^2_{x,B,i} + \sigma^2_{x,B,j} + (\sigma^2_{x,A,i} + \sigma^2_{x,A,j}) \frac{\sum_{j=1}^{3} \frac{1}{\sigma_r^4(j)}}{(\sum_{j=1}^{3} \frac{1}{\sigma_r^2(j)})^2} + \frac{2 - \frac{(t_i - t_j)^2}{\sum_{k=1}^{n} (t_{NIST,k} - \overline{t}_{NIST})^2}}{\sum_{j=1}^{3} \frac{1}{\sigma_r^2(j)}}.$$

If one NMI is the pilot lab, the corresponding uncertainty of the degree of equivalence between the pilot lab and the kth non-pilot lab is

$$\sigma^2_{D_{NIST,j}} = \sigma^2_{x,B,NIST} + \sigma^2_{x,B,j} + (\sigma^2_{x,A,j} + \frac{\sigma^2_{x,A,NIST}}{n}) \frac{\sum_{j=1}^{3} \frac{1}{\sigma_r^4(j)}}{(\sum_{j=1}^{3} \frac{1}{\sigma_r^2(j)})^2} + \frac{1 + \frac{1}{n} + \frac{(t_{NIST,j} - \overline{t}_{NIST})^2}{\sum_{k=1}^{n} (t_{NIST,k} - \overline{t}_{NIST})^2}}{\sum_{j=1}^{3} \frac{1}{\sigma_r^2(j)}}.$$

The described approach has been used to determine the uncertainty of the combined difference between the measurements of an NMI and the corresponding predictions based on the pilot NMI for each participating NMI in CCEM-K2. These results can be applied to other Key Comparisons with traveling standards.

3.3 Process Characterization

3.3.1 Flow Measurements for Multi-Meter Transfer Standards

James J. Filliben, Will Guthrie, Ivelisse Aviles
Statistical Engineering Division, ITL

Mark Vangel
Dana-Farber Cancer Institute

George Mattingly, Pedro Espina
Process Measurements Division, CSTL

Dual Meter Experiment Design for Flow Transfer Standard Testing

Run	Day	Flow Rate	Upstream Meter	Downstream Meter
1	1	high	1	2
2	1	low	1	2
3	1	low	1	2
4	1	high	1	2
5	1	low	2	1
6	1	high	2	1
7	1	high	2	1
8	1	low	2	1
9	2	high	2	1
10	2	low	2	1
11	2	low	2	1
12	2	high	2	1
13	2	low	1	2
14	2	high	1	2
15	2	high	1	2
16	2	low	1	2

Figure 14: This 16-run replicated 2**3 full factorial design takes full advantage of the duality of meters, meter positions, flow rates, and days. The data that result from this design will be well-suited to specialized 2-level design analyses plus classical Youden analyses, and will readily yield information about the relative importance of main effects and interactions.

The elimination of measurement-based trade barriers is essential for international commerce. The buyer/seller of a commodity in the worldwide market place must be assured of an accurate measurement of the quantity under purchase. For fluid and gaseous products, the quantity bought/sold is most commonly determined by flow measurements: the volume of material flowing through a pipe. Since the volume of such pipe-conducted materials is frequently enormous (e.g., oil), small errors in flow measurements may have enormous engineering and financial consequences.

To assure the accuracy of flow meters, flow standards are established and maintained by National Measurement Institutes (NMIs) in various countries worldwide.

From these NMI's, there exist 2 traceability components:

- Within-country: whereby buyers/sellers within a country have traceability to their country's NMI via local flow laboratories;
- Between-country: whereby a country's NMI has traceability to NMIs in other countries. (In this international arena, such traceability is typically referred to as "equivalence" or "comparability".)

For both national traceability and international comparability, 2 components are essential:

- transfer standards (= portable ultra-accurate flow meters), and
- proficiency testing protocols (how the meters will be used).

Proficiency testing tells us how good a laboratory is; that is, how good a laboratory's meters (and procedures) are. Proficiency testing is an umbrella term that includes such items as

- Under what conditions the meters will be used;
- How (and how often) the meters will be calibrated;
- How the data will be collected;
- How the data will be analyzed.

(From a statistics point of view, the first 3 issues are experimental design-related in nature, and the last issue is data analytic in nature.)

Not all NMIs agree on their flow measurements; small differences frequently exist (as confirmed by differing measured values for common (international) transfer standards). The formal mechanism whereby such NMI differences are estimated, agreed on, disseminated, and implemented is the (International) Mutual Recognition Agreement (MRA), of which (for flow measurements) the U.S. (and NIST) is a participating member. Under the MRA, the formal term that describes such NMI-to-NMI differences is the "Key Comparison" (KC). For a given pair of NMIs, the Key Comparison will consist of 2 values (and the specification and execution of all proficiency tests that lead to these 2 values):

- Difference: How much do the 2 labs differ "on the average";
- Uncertainty: How confident are we of the above estimated difference value?

For a given measurement unit of interest, the international coordination effort for the MRAs is carried out by the Committee Internationale des Poids et Mesures (CIPM) . For flow measurements, The CIPM created a Working Group for Fluid Flow (WGFF). The WGFF focuses on the following 6 flow measurement areas:

- water flow;
- hydrocarbon liquid flow (including oil);
- low pressure air flow;
- high pressure gas flow;
- air speed;
- volume.

NIST's CSTL Process Measurements Division is the United States' NMI for water flow. Over the last 2 years, George Mattingly of the Process Measurements Division has played a leadership role in serving as the chair of the WGFF's subgroup focusing on the first (and most active) of the 6 categories (water flow). It is the task of this subgroup to identify and address all of the relevant issues dealing with both water flow transfer standards and water flow proficiency testing protocol.

Specifically, in regard to the WGFF water flow proficiency testing protocol, the SED/CSTL collaboration has resulted in 2 major improvements:

- Experimental Design: the commonly-employed solo meter test set-up has been replaced by a drift-eliminating 2**3 full factorial dual-meter set-up. This 2-meter design serves to be much more informative about meter and laboratory issues such as meter effects, meter position effects, flow rate effects, lab effects, repeatability, reproducibility, and drift.

- Data Analysis: The Youden Plot will become a formal part of the test protocol and will greatly assist in the determination of meter effects and lab effects. In addition, an SED-developed standardized analysis of 2**3 full factorial designs will also become part of the formal protocol to ascertain position effect, flow rate effects, and (most importantly) interactions.

The proficiency testing protocol that results from this WGFF water flow subgroup is serving as the prototype for the remaining 5 areas, and so this particular effort has high leverage and will serve as the basis for all international flow comparisons.

The application of the proficiency testing protocol to a pilot in-house NIST system has already resulted in a significant improvement in insight regarding the relative importance of meter, meter position, flow rate, and interactions, and thus serves as the foundation for the NIST component in subsequent Key Comparisons.

3.3.2 Nonlinear Network Measurement System

Dominic Vecchia, Kevin Coakley, Jolene Splett
Statistical Engineering Division, ITL

Don Degroot, Jim Booth
RF Technology Division, Electromagnetic Technology Division, EEEL

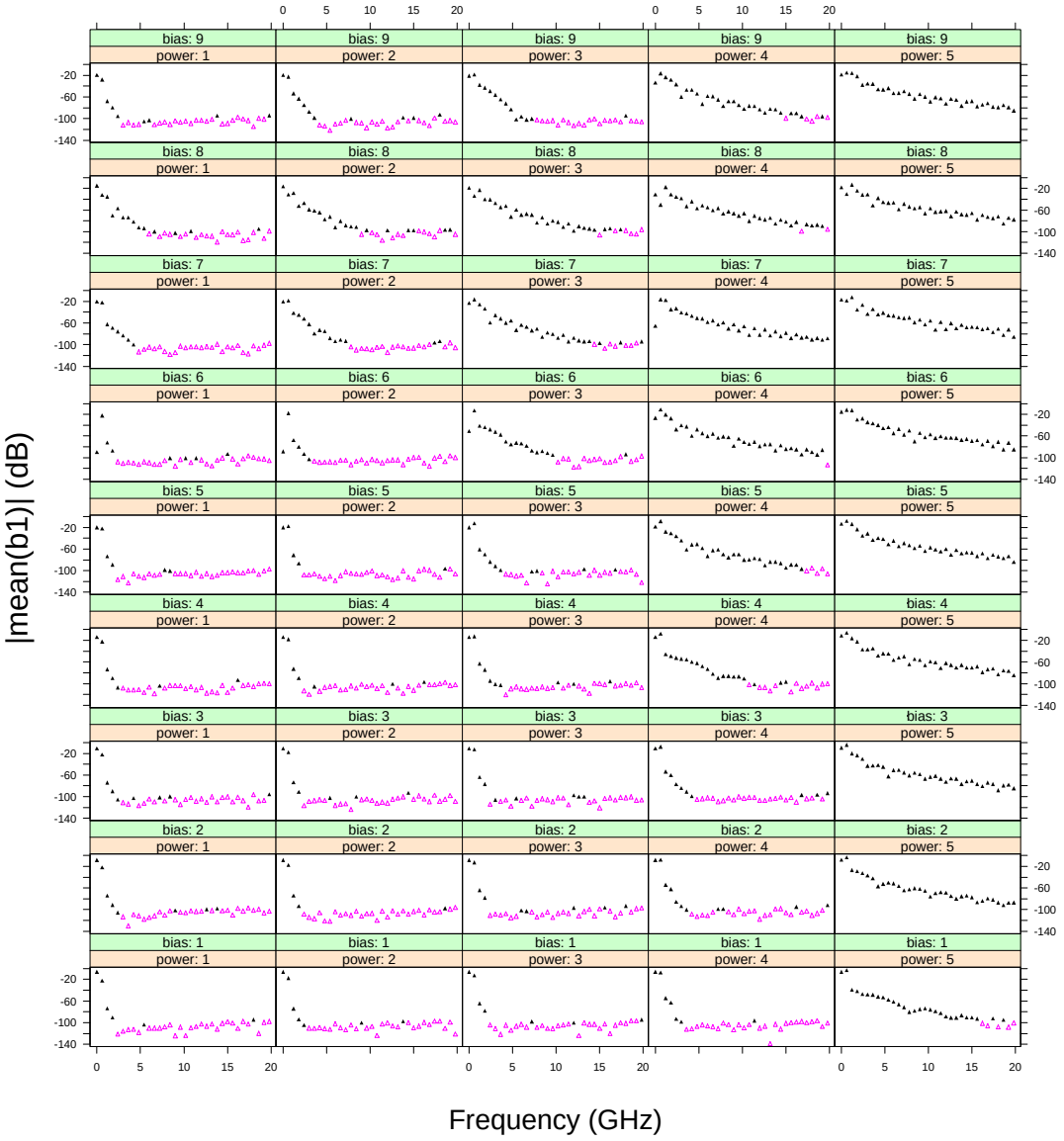


Figure 15: Frequency-domain measurement of a reflected wave from a diode verification circuit for 160 power-bias sweeps over 5 equally spaced powers from -10 to 10 dBm and 9 equally spaced biases from -1.0 to 0.6 V. Solid black triangles indicate harmonics deemed to be statistically significant.

As wireless networks are pushed beyond the limits of network analysis, large signal descriptions are required to characterize devices. Our long-term goal is to dramatically improve state-of-the-art measurements and modeling of high-frequency wireless systems and components under large signal conditions. NIST has assembled a multidisciplinary team and established a new measurement facility for large signal device-level characterizations. Existing techniques are limited by measurement uncertainty in phase; we require additional external information to obtain a high-frequency characterization. Moreover, general system-level modeling techniques are inadequate and fail to exploit the wealth of data that is now available at the device level. We are developing a standard nonlinear device in order to reduce measurement uncertainty, and are developing and evaluating new models for nonlinear devices.

SED contributions to the Nonlinear Network Measurement System (NNMS) project during the initial year include:

1. We designed measurement repeatability studies and identified significant warm-up effects and more troubling trends in the experimental data collection system. We are investigating whether these systematic effects are connected to the device-under-test (DUT) or to the measurement system itself.

2. We developed a statistical procedure for detecting harmonics in the measured signals. Based on this procedure, we selected favorable experimental operating conditions for model development. At these conditions, there is no significant harmonic content beyond the capabilities of the measurement system.

3. We developed a general class of metrics for quantifying the relative performance of candidate prediction models. The metrics we are currently considering allow for prediction errors to be weighted differently by frequency. In general, the choice of the weights depends on the particular application and the particular device-under-test (DUT). We have demonstrated that the equal-weight metric has a clear interpretation in the time domain. Metrics will be used to compare competing prediction models now being developed for various nonlinear DUTs, especially superconducting devices and example diode circuits. In addition, we plan to develop metric-based diagnostics for quantifying the stability of the measurement system.

We laid the technical groundwork for achieving long-term program goals including, (1) validation/verification of NNMS-based models of standard devices; (2) completion of a phase uncertainty evaluation of NNMS data; and (3) development of system models and simulation capabilities. These programs will enable more efficient wireless system design and have potential impact on every segment of the wireless community in the U.S. Electronics and Communications industries.

3.3.3 Statistical Analysis of High-Speed Optoelectronic Measurements

K.J. Coakley and C.M. Wang
Statistical Engineering Division, ITL

P. Hale and T. Clement
Optoelectronics Division, EEEL.

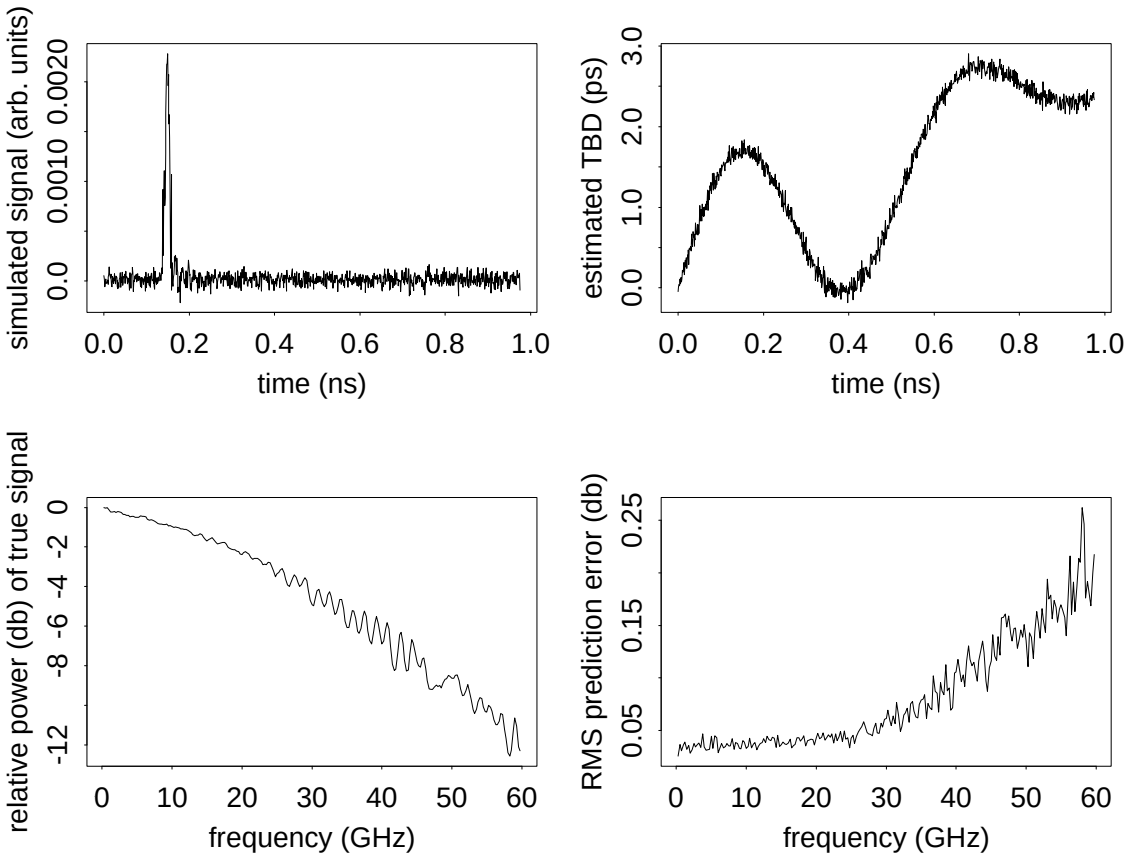


Figure 16: In a simulation experiment, the true signal is similar to the measured impulse response function of a photodiode. Simulated signals are contaminated by additive noise, timing jitter noise, time shift errors, and time base distortion (TBD) errors. The standard deviations of the jitter noise and the time shift errors are 1.95 ps and 0.59 ps, respectively. To obtain our TBD estimate, we contaminate the true TBD with additive Gaussian noise. During a 4 ns interval, the signal is sampled 4096 times. We align 1000 noisy realizations of the observed signal based on estimates of the relative time shift errors. Based on the TBD estimate, we interpolate the average of the aligned signals onto an equally spaced time grid using a regression spline approach. We estimate the power spectrum of the true signal by computing the periodogram of our interpolated signal. The estimated power spectrum is then adjusted based on our estimate of the variance of the jitter noise.

High bandwidth measurements are needed to support high-performance systems that take advantage of the potential bandwidth of optical fiber. Systems presently being installed operate at 5 to 10 gigabits per second using pure optical time division multiplexing (OTDM), and research is being done on the next generation of OTDM systems which operate at 20 to 40 gigabits per second. Methods are needed to characterize the impulse and frequency response of high-speed sources and detectors to at least three to five times the system modulation rate. NIST is developing the key underlying statistical technology towards the realization of actual measurement capability needed by industry.

In principle, one would like to sample a high speed optoelectronic signal at equally spaced time intervals. The random component of the timing error is called jitter. The systematic component is called time base distortion (TBD). To reduce additive noise effects, we average many observed signals. Before averaging, the signals are aligned based on estimates of the relative time shift errors. Using a regression spline model, we interpolate the average of the aligned signals onto an equally spaced time grid based on our estimate of the TBD. We estimate the power and phase spectrum from this interpolated signal. To correct for jitter effects, the power spectrum is multiplied by $\exp(\omega^2\hat{\sigma}_\tau^2)$, where $\hat{\sigma}_\tau$ is our estimate of the standard deviation of the jitter noise.

Over the last year, we accomplished the following.

- "Uncertainty of time base distortion estimates" accepted for publication in *IEEE Transactions on Instrumentation and Measurements*.

- Improved our procedure for estimating the variance of the timing jitter noise. Wrote a draft paper, "Adaptive estimation of Root-Mean-Square (RMS) timing jitter noise."

- Developed a fast implementation of our signal alignment code.

- Developed a Monte Carlo simulation code to quantify random and systematic errors in phase and power spectral estimates due to the joint effects of additive noise, timing jitter noise, errors in TBD estimates, errors in RMS jitter noise estimates, and errors in signal alignment.

Industries that will benefit from the development of the new measurement capability are involved in the following technologies: Gigabit Ethernet networks, Fibre Channel, CATV, satellite TV, tethered microwave antennas, optical telecommunications, optical components and test equipment, SONET/SDH (synchronous optical network/synchronous digital hierarchy industry) and Wireless.

3.3.4 A Statistical Model for Cladding Diameter of Optical Fibers

C. M. Wang
Statistical Engineering Division, ITL

T. J. Drapela
Optoelectronics Division, EEEL

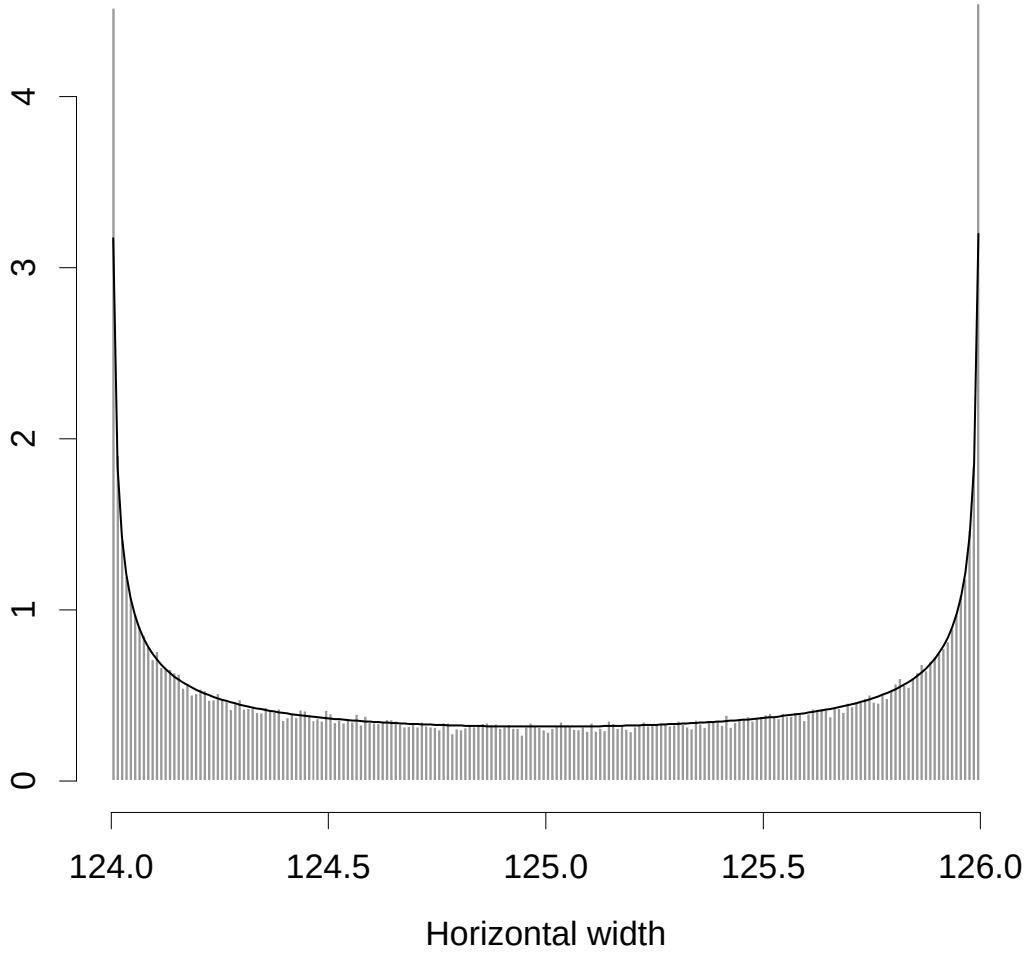


Figure 17: Histogram and probability density function of horizontal widths of an ellipse with $M = 63$ and $m = 62$.

NIST has developed a contact micrometer for accurate measurement of optical-fiber outer diameter. The contact micrometer is used to measure reference fibers that are artifacts used by the telecommunications industry for calibrating their own measurement systems. We present a model for diameters measured by the contact micrometer. Based on this model, the probability distribution of the diameter is derived and two diameter estimates are presented.

A typical single-mode telecommunications optical fiber is nearly circular and has a glass *core* of about 10 μm in diameter surrounded by a glass *cladding* with an outer diameter of about 125 μm. Gray-scale systems are used by the industry to determine the cladding diameter that must be measured and controlled within 0.1 μm to enable the manufacture of efficient fiber connectors that do not require manual adjustment. Measurements made with gray-scale systems may suffer from a systematic error of a few tenths μm. NIST developed a contact micrometer that can make cladding diameter measurements accurate to 0.04 μm. The contact micrometer is used to certify Standard Reference Material (SRM) fibers so that the industry has artifacts for calibrating their gray-scale systems.

A contact micrometer consists of a stationary post called an anvil, and a moveable part called a spindle. Measurements are performed by first pressing the fiber between the spindle and the anvil, and the position of the spindle is monitored interferometrically. Then, the fiber is removed and the spindle is brought into contact with the anvil. The difference between the two positions is the diameter of the fiber. If we model the cross section of the fiber by an ellipse, the contact micrometer measures the horizontal width of the ellipse as the diameter of the fiber. The horizontal width of the ellipse is shown to be

$$W = 2\sqrt{m^2 \sin^2\theta + M^2 \cos^2\theta}$$

with θ denoting the angle between the major axis and the positive x-axis, M is the semi-major axis length, and m is the semiminor axis length. The pdf of W can be derived by assuming that θ is distributed uniformly in $[0, 2\pi]$. The accompanying figure displays a histogram of W obtained from 100,000 simulated values of θ with $M = 63$ and $m = 62$. The superimposed line is the pdf of W.

Based on the proposed model for diameters measured by the contact micrometer, two methods for estimating the cladding diameter and its uncertainty for the SRM fiber are presented. In preparing SRM fibers, we take measurements at equally spaced angular orientation and use both methods to assure an accurate estimate of the cladding diameter of fibers.

The proposed methods have been used to estimate the cladding diameter and its uncertainty for a reference fiber (SRM 2520). The methods can also be used to obtain a diameter estimate for a measurement system having the fiber oriented at any angle.

3.3.5 Lifetime of Magnetically Trapped Neutrons

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

P.R. Huffman, M.S. Dewey, D.M. Gilliam
Ionizing Radiation Division, PL

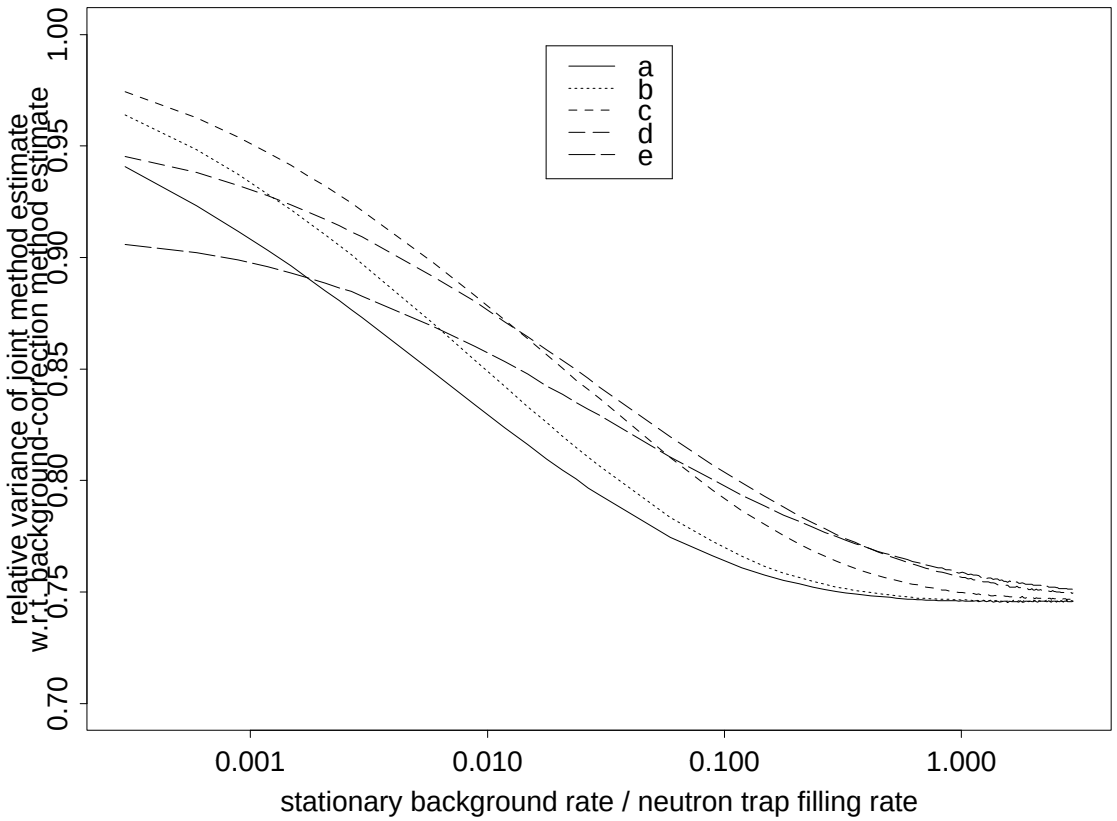


Figure 18: We estimate the neutron lifetime from joint realizations of the primary "neutron decay plus background" data and "background-only" data. The asymptotic variance of this "joint" method estimate is uniformly less than that of an estimate computed from "background-corrected" data for all cases studied. The overall background signal is modeled as the sum of an exponential activated aluminum signal plus a stationary background signal. In (a-e), the expected number of activated aluminum nuclei at time 0, divided by $\lambda\tau$ (the rate at which neutrons enter the trap times the neutron lifetime) equals: 0, 0.0071, 0.071, 0.356 and 0.712.

Stochastic modeling, planning, and analysis for neutron lifetime experiments.

Recently, for the first time, a team of researchers from Harvard University, Los Alamos National Laboratory, University of Berlin, and NIST produced and confined polarized Ultra Cold Neutrons (UCN) in a magnetic trap. Data from this breakthrough experiment yielded a neutron lifetime estimate of 660 s. The 68 percent confidence interval for this estimate is (490 s, 950 s). Although this proof-of-principle result is not as precise as the currently accepted value of 885.7 ± 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.

Over the last year, while participating in the planning of a second generation experiment, we studied the performance of a new approach for estimating the mean lifetime of the neutron. Our goal was to better account for background noise in order to reduce the variability of the estimate of the neutron lifetime. In the new approach, we assume that the analytical form of the background signal is known. In the fill stage of each run of the multi-run experiment, we load the magnetic trap with neutrons. During the subsequent observation stage of each run, we measure a realization of the primary "neutron decay plus background signal." In a separate experiment, we refill the trap and measure independent realizations of the "background-only" signal. We estimate the neutron lifetime by fitting a model to the joint realizations of the "primary" and the "background-only" data. The asymptotic standard error of the new estimator is consistently lower than those of the alternative estimators. In one alternative estimation method, the lifetime is estimated by fitting an exponential model to background-corrected data. In the other alternative estimation method, a full model is fit to the uncorrected primary data.

- A manuscript "Neutron Lifetime Experiments Using Magnetically Trapped Neutrons: Optimal Background Correction Schemes" was published in *Nuclear Instruments and Methods in Physics Research A*.

- A manuscript A manuscript "Estimation of the Neutron Lifetime: Comparison of Methods with Account for Background" was submitted to *Physical Review C*.

- Determined optimal data collection strategies and computed asymptotic standard error of neutron lifetime estimate for new background models related to ongoing experiment.

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

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 planning is critical for maximizing the information extracted from the data, and for selecting the optimal redesign of the apparatus.

3.3.6 Stochastic Modeling and Estimation of Intensity of a Spray Process

Grace L. Yang
Statistical Engineering Division, ITL

John F. Widmann
Fire Research Division, BFRL

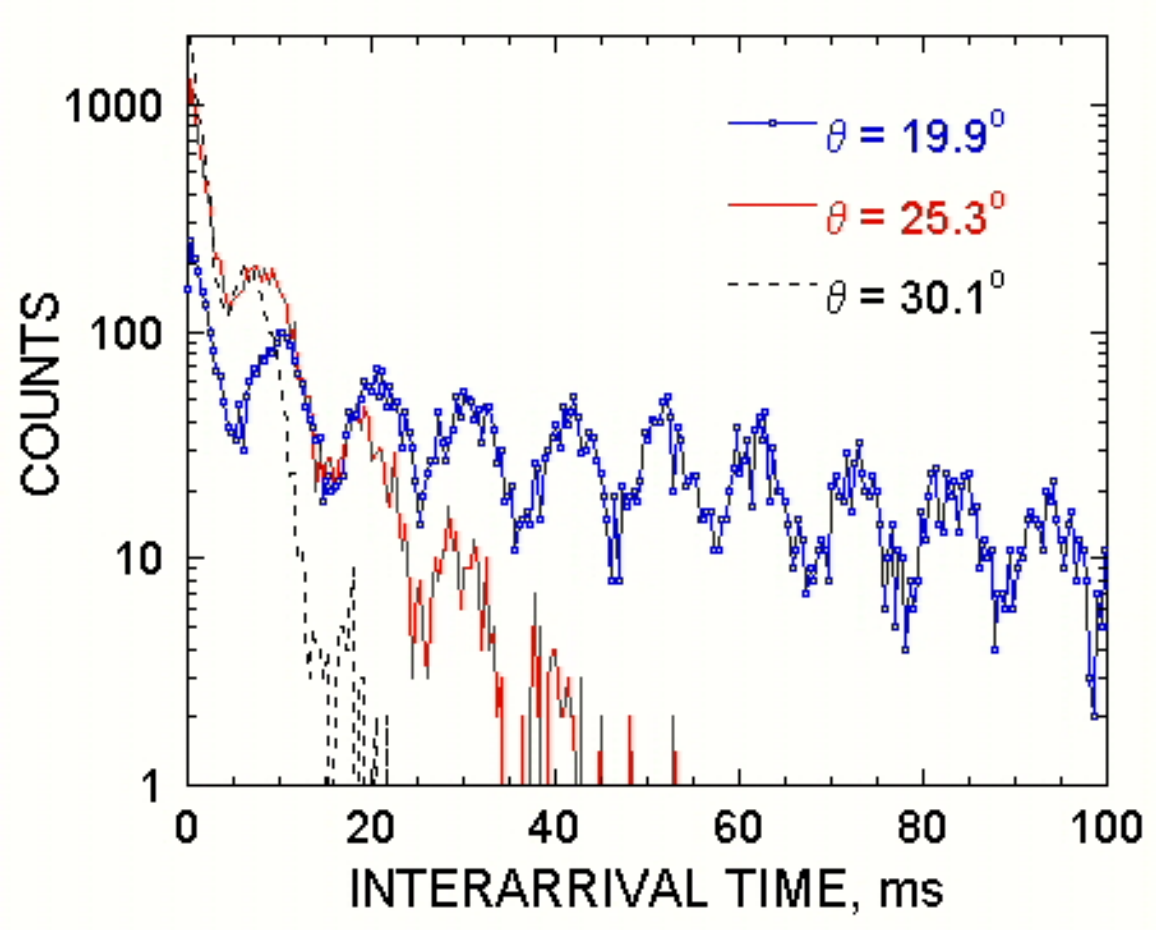


Figure 19: Experimental interarrival time distributions (not normalized) obtained 35 mm downstream of a pressure swirl atomizer. The spray angle, θ, is measured from the center axis of the nozzle.

A phase Doppler interferometry (PDI) is a non-intrusive technique frequently used for obtaining information about spray characteristics. It has been used widely in areas including liquid fuel spray combustion, spray coatings, agriculture pesticides, fire suppression and others. Due to the design of the instrument, the recordings of the PDI contain gaps, called dead times. We construct consistent estimates of the spray intensity in the presence of dead times under very general conditions.

As the droplets in the spray diffuse, the PDI, installed at a fixed location, records the arrival time and size of each droplet that reaches its probe volume. The recordings of the PDI are not continuous. Rather, there are gaps that can be characterized by an alternating sequence of random variables,

$$W_1, Y_1, W_2, Y_2, W_3, Y_3, \ldots,$$

with the $W's$ denoting the time intervals during which the PDI can record the arrival times and the $Y's$ being the dead time periods during which the PDI is inactive.

We model the spray process as a Poisson process and validate its appropriateness. Once the model is established, we construct consistent estimates of the Poisson intensity in the presence of dead times. The estimation is complicated by the fact that when dead times are present, the inter-arrival times are no longer exponentially distributed, but have a multi-modal distribution. Profiles of the three inter-arrival time distributions are given in the figure. They were obtained at three different locations, as indicated by the values of θ.

• The paper "A correction method for spray intensity measurements obtained via phase Doppler interferometry" has appeared in *Aerosol Science and Technology* **32:6, June 2000.**

• A manuscript "Consistent estimation of Poisson intensity in the presence of dead time" is to be submitted to a statistics journal.

The problem of gaps in the PDI recordings has been reported in the aerosol literature at least over the last decade. Our work is a contribution to a better understanding of the incomplete data problem (due to the presence of the unavoidable dead time). Our estimation procedure provides a new method of utilizing the data that eliminates the bias in large samples.

3.3.7 Cryogenic Low Energy Assay of Neutrinos (CLEAN)

Kevin J. Coakley
Statistical Engineering Division, ITL

D. McKinsey
Harvard University

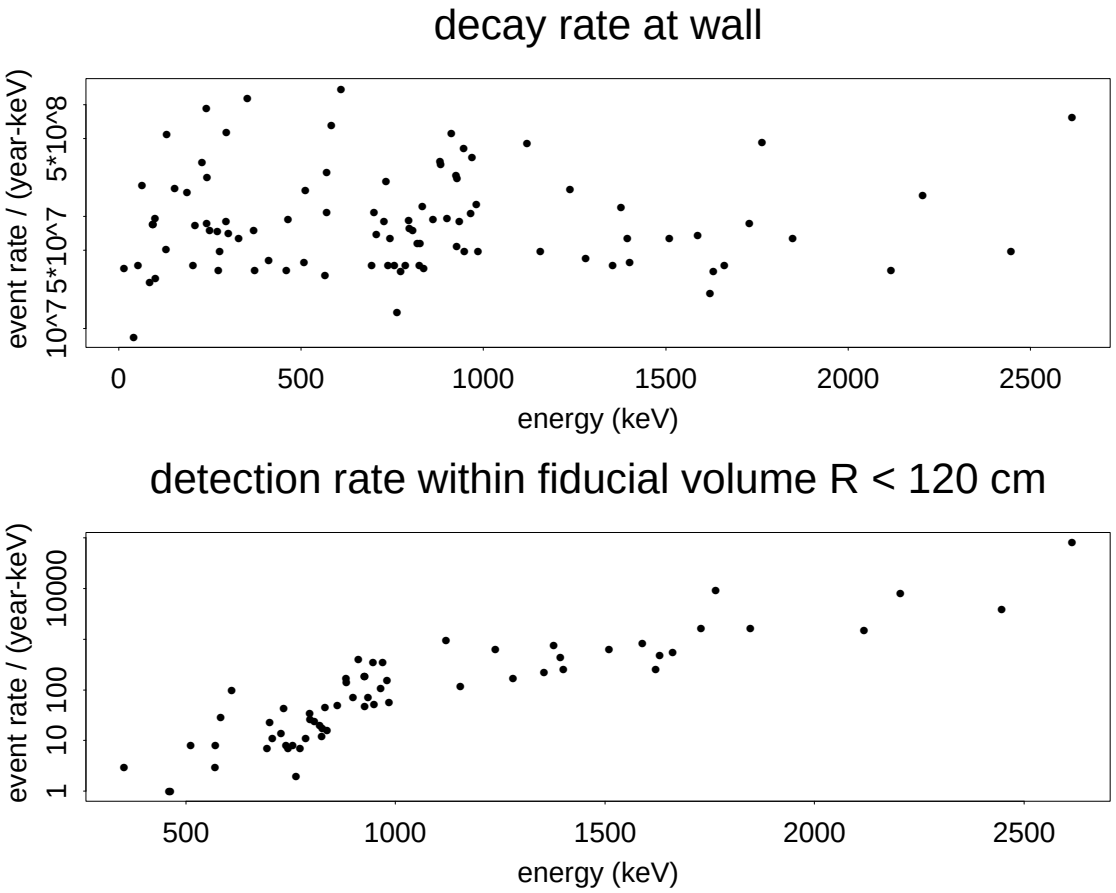


Figure 20: In the proposed CLEAN experiment, due to natural radioactivity of isotopes in materials, background photons are emitted from the outer wall of the detector. As a background photon propagates inward, it can deposit energy in the neon by Compton scattering or by photoelectric absorption. Above, we predict the intensity of background events when the total photon energy is deposited within the inner detector region known as the fiducial volume.

Statistical modeling, planning and analysis for Cryogenic Low Energy Assay of Neutrinos (CLEAN).

In the proposed CLEAN experiment, we consider a spherical detector region filled with neon. Neutrino-electron scattering events produce scintillation light in the neon. The expected number of photons per event depends on the energy of the neutrino. As the photons propagate toward detectors at the outer walls of the spherical region, they can Rayleigh scatter. This diffuse scattering affects photon transit times as well as photon spatial intensities at the walls of the spherical region. In addition to neutrino scattering events, background photons (due to the natural radioactivity of detector materials) will also produce scintillation light. Background photons deposit energy in the neon by Compton scattering or by photoelectric absorption.

To demonstrate the feasibility of CLEAN, we must demonstrate the ability to discriminate background events from neutrino scattering events. In some cases, an energy criterion is sufficient since highly energetic background photons will produce a large number of scintillation photons, whereas low energy (< 700 keV) neutrino scattering events will produce fewer photons. Since not all background photons are highly energetic, we need an additional method for identifying background events. By knowing the spatial location of the event, we can also discriminate between background events and neutrino scattering events. This is so because most of the background photons are attenuated before they reach an inner fiducial volume of the spherical detector. Neutrino flux rates are estimated from neutrino scattering events detected within this fiducial region. Over the past year, we accomplished the following.

- Developed Monte Carlo simulation code to model scattering and attenuation of background photons emitted from walls of detector.

- Developed Monte Carlo simulation code to model Rayleigh scattering of scintillation photons.

- Developed a first generation maximum likelihood procedure to estimate the spatial location of an event based on spatial intensity of detected photons at the walls of the detector.

- Gave a talk at collaboration meeting at Princeton University, September 25, 2001.

Better measurements of the low energy spectrum of solar neutrinos are of interest for two reasons. First, they would help clarify understanding of the neutrino oscillation phenomenon. Second, they would improve understanding of how nuclear reactions occur in the sun. Cold Dark Matter should be detectable using CLEAN technology.

3.3.8 Statistical Analysis of Decay Data: Determining Uncertainty in Half-life Estimation of Radionuclides

Z. Q. John Lu, Grace L. Yang
Statistical Engineering Division, ITL

Larry L. Lucas
Ionizing Radiation Division, PL

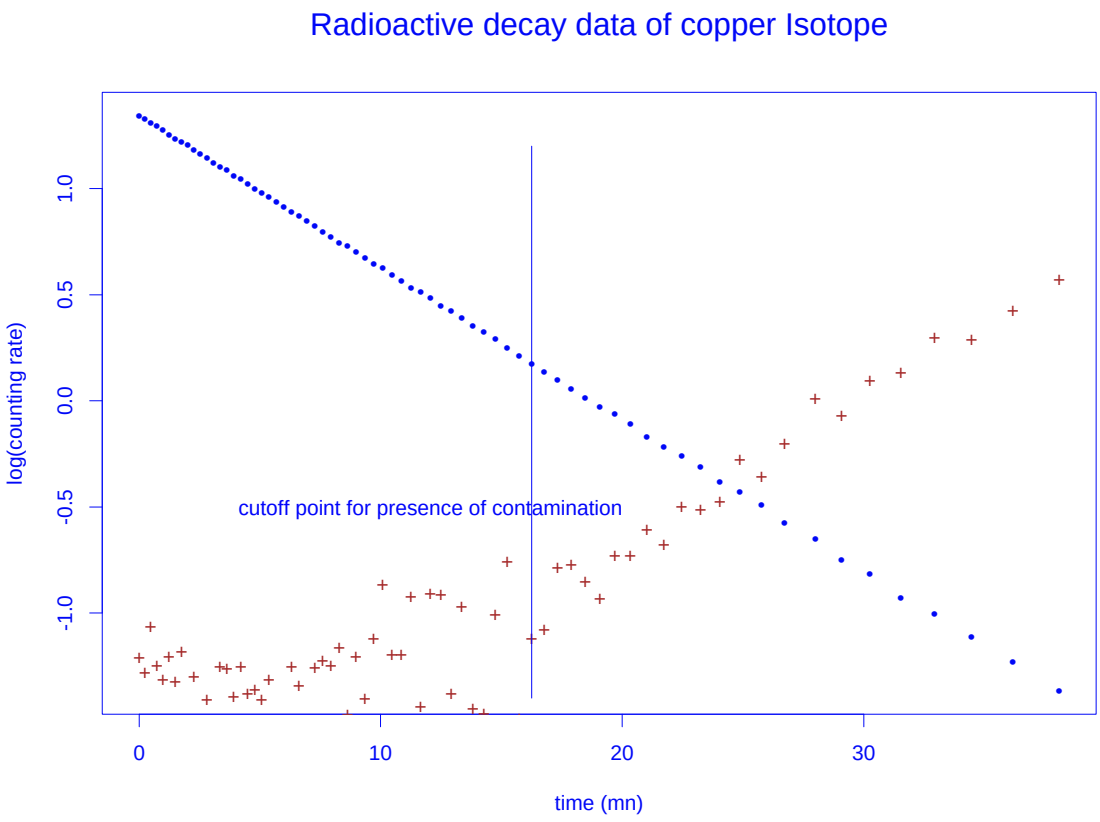


Figure 21: ^{62}Cu is a new short-lived positron-emitting radionuclide used in positron emission tomography (PET). The measurements are made from NIST's ionization chamber facility. The time index is the clock times of midpoint values of waiting time intervals at which a fixed electrical charge was recorded. (Legend: "." denotes log mean counting rate and "+" denotes the log percentage standard deviation (CV).) The fact that log CV has an upward trend may indicate potential presence of contamination from another radioactive source.

The discovery of radioactivity has led to widespread applications of radioactive materials in medicine and industry such as medical imaging and food preservation. In many applications, accurate determination of half-life values of radionuclides is very important. The Radioactivity Group at NIST has maintained a website (http://physics.nist.gov/Halflife), providing the most updated values on half-lives of many radionuclides. Recently, high precision demand for half-life determination of new short-lived radiopharmaceuticals brings to our attention the need for more efficient statistical analysis of radioactive decay data and more realistic uncertainty statements. This project represents our first investigation into this important long-standing statistical/physics problem.

Radioactive decay is itself a random process. We cannot predict exactly when an unstable nucleus will decay. However, for a given sample of radioactive materials, that contain up to a *very large* number (e.g., 10^{12}) of unstable nuclides, we can correctly predict and characterize the decay process through the statistical process. An important quantity is the *half-life* $t_{1/2}$, the time it takes for half of the unstable nuclides (particles) contained in a given sample to decay. To use the more familiar statistical language, we can assume that random lifetime for the population of a given species of atom has a lifetime probability distribution. Assume that this lifetime distribution is exponential $F(t) = P(X > t) = \exp(-\lambda(t - t_0))$, then $t_{1/2} = \log(2)/\lambda$, and $t_{1/2}$ is also the median value of $F(\cdot)$. Strictly speaking, the decay process is a Bernoulli process: the number of decays and the number of surviving atoms has a binomial distribution. At any given time instant, the decays in a very small time interval are approximated by an inhomogeneous Poisson process with intensity function $N_0\lambda \exp(-\lambda(t - t_0))$.

This model, however, does not address the measuring mechanism of experimental data. The observed process is often a mixture of several processes, leading to the phenomenon of overdispersion. Furthermore, two types of data collection schemes have often been used: one involves observations of decay counts (or some proportional quantities) at fixed time intervals and the other involves records of the random time intervals at which a certain *fixed* amount of decay counts have registered. These different data collection techniques call for different models and transformations for computing half-life values and subsequent uncertainty theory. We have developed some new distribution theory for the count rate statistics, when the measurements are the observed waiting times (order statistics) from an exponential population.

Though half-life as the decay constant of a given species does not vary much with environmental influence, it is, however, never known and can only be estimated by observations of the decay process of selected testing materials and possibly by combining knowledge from past experiments. So the estimate of half-life is invariably subject to statistical errors inherent in data and impurity of testing materials. Any inference based on data or past knowledge has to take into account the intrinsic uncertainty from impurity of radioactive testing materials, noise environment, instrument errors, and testing/lab effects. The approach by treating the half-life value as a random variable is the most direct way of incorporating uncertainty from multiple sources and will be pursued further in the

context of Bayesian metrology.

Inspired by the need to improve half-life estimation of new short-lived radiopharmaceutical materials, statistical theory for analysis of radioactive decay data will be developed from a modern statistical viewpoint using maximum likelihood and Bayesian statistics. The applications of this work will include half-life estimation and uncertainty quantification of many radionuclide materials and should facilitate standardization efforts on radionuclide measurements and radiopharmaceutical devices at NIST. The industries that will likely benefit include pharmaceuticals, medicine, and environment.

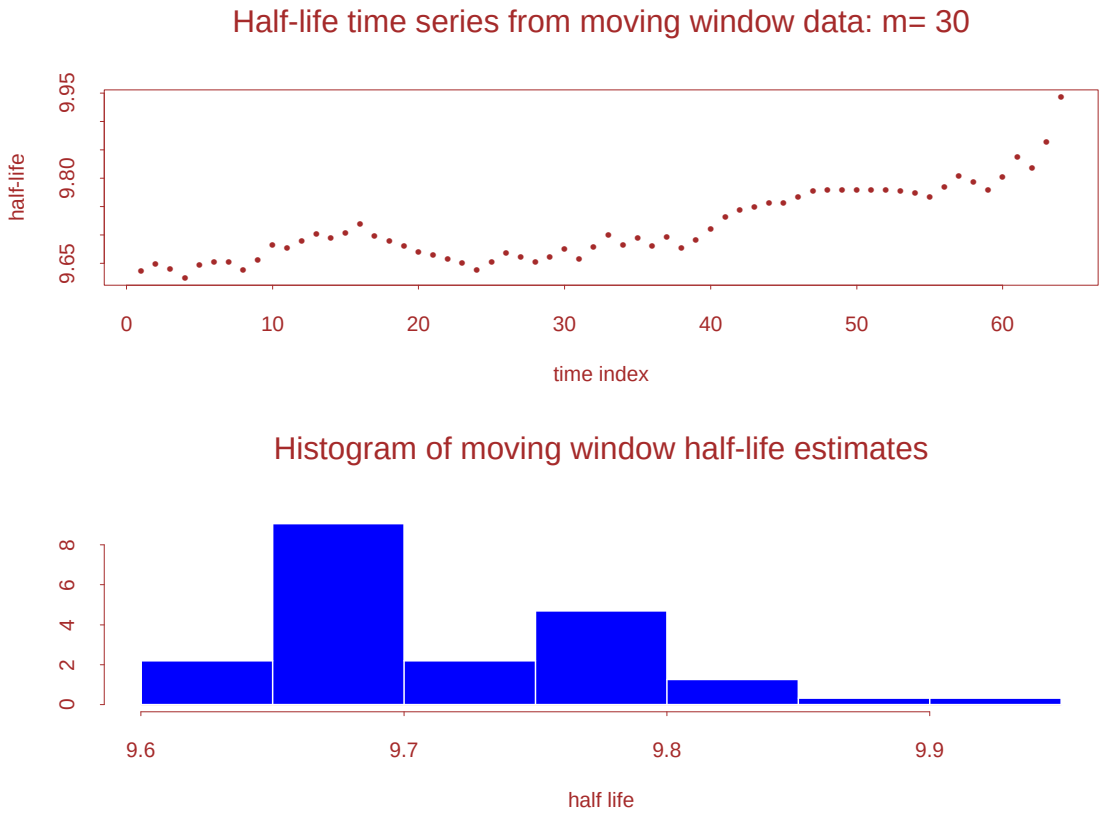


Figure 22: The variability of half-life can be seen by computing half-life values from each 30-point moving window subset of data. The top panel plot exhibits a drift (upward) as time increases and at the low-radioactive count period, which may be caused by the presence of some contamination of longer half-life particles. The lower panel shows the histogram of half-life values, which essentially covers a range $[9.6, 9.85]$, significantly larger than the uncertainty reported in the earlier published results.

3.3.9 Combining Process Capability Indices from a Sequence of Independent Samples

Nien Fan Zhang
Statistical Engineering Division, ITL

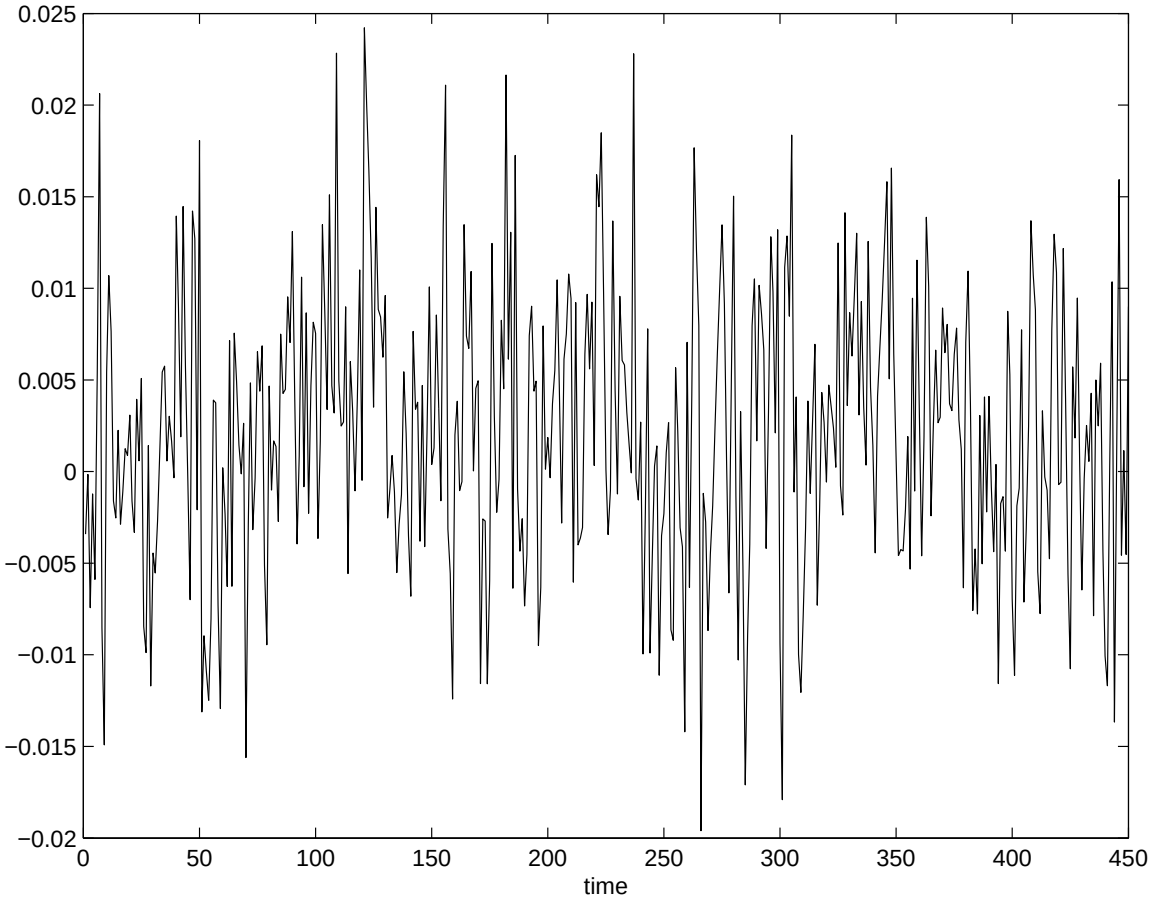


Figure 23: This figure shows the mass imbalance data from a chemical process. A transfer tank with one inflow valve and one outflow valve is used.

Process capability indices such as C_p, C_{pk}, and C_{pm} have been used in manufacturing for process assessment and for evaluation of purchasing decisions. They provide management with a single-number summary describing the extent to which a process has conformed to its specifications. To track process capability in order to prioritize process improvement projects, quality engineers and process operators are often asked to summarize the quality of a process by combining a sequence of C_{pk} or C_{pm} results for the same production process but from different batches.

The C_{pk} index is defined as

$$C_{pk} = \frac{min[USL - \mu, \mu - LSL]}{3\sigma}$$

with USL and LSL denoting the upper and lower specification limits, respectively, and μ and σ^2 are the mean and variance, respectively of the process $X_i, i = 1, 2, ...$ An estimator of C_{pk}, $\hat{c}_{pk}$ is constructed by replacing μ and σ by the sample mean $\overline{X}$ and the sample standard deviation S, respectively based on the assumption of a normal distribution.

Since S is a biased estimator of σ, $S/c_4(n)$ was proposed as an unbiased estimator of σ with $c_4(n)$ a ratio of two gamma functions. $S/c_4(n)$ can thus be used to form another estimator of C_{pk},

$$\hat{c}'_{pk} = \frac{min[USL - \overline{X}, \overline{X} - LSL]}{3\frac{S}{c_4(n)}}$$

Note that when $n > 300$, $c_4(n) \approx 1$. Although $S/c_4(n)$ is an unbiased estimator of σ, $\hat{c}'_{pk}$ is not necessarily an unbiased estimator of C_{pk} because $\hat{c}'_{pk}$ is not a linear function of S. The expectation of $\hat{c}'_{pk}$ can be calculated analytically based on Zhang et al. (1990) and it can be shown that $\hat{c}'_{pk}$ is a biased estimator of C_{pk}.

When we combine estimators of C_{pk} from a sequence of independent samples, we assume that the process X_i is in statistical control. That is, the process mean and variance are the same when each sample is obtained. Thus, the process has a constant true C_{pk}. Then, given a sequence of independent samples, a sequence of C_{pk} estimators are obtained.

Three common approaches of combining the estimators of C_{pk} from a sequence of independent samples are discussed. The first approach is to use the average value of the estimators of C_{pk} based on the samples weighted by the subsample size. The second approach is to use the overall sample mean and the pooled sample variance. The third approach is similar to the second one, but instead of the pooled sample standard deviation, the average of the subsample standard deviations is used.

Assume that we have k random samples from a process. For the ith sample of size $n(i)(i = 1, ..., k)$, the sample mean and the sample variance are $\overline{X}_i$ and $S^2(i)$, respectively. Let $N = \sum_{i=1}^{k} n(i)$. The overall sample mean and the pooled sample variance are $\overline{X}$ and S^2_{pool}, respectively. The criterion of minimum MSE is applied to make comparisons among six estimators based on the three approaches mentioned above. MSE's are either analytically calculated or obtained by simulations. It is concluded that the estimator based on the

overall sample mean and the square root of the pooled variance divided by the c_4 factor performs better than the other estimators. That is,

$$\tilde{c}'_{pk} = \frac{min[USL - \overline{X}, \overline{X} - LSL]}{3S'_{pool}}$$

wth $S'_{pool} = S_{pool}/c_4(N - k + 1)$.

The data from a chemical process, for which mass imbalance must be monitored as the process runs, are used to demonstrate the estimation of process capability using a sequence of samples obtained at different times. A transfer tank with one inflow valve and one outflow valve is used in this process. The mass imbalance is the difference between the input and output streams. The sampling frequency for data collection was two minutes. The upper and lower specification limits for mass imbalance are 0.025 and -0.025, respectively. Figure 1 displays six sets of mass imbalance data of size 75 each. C_{pk} was used to monitor the process capability. For each subset of data, $\tilde{c}'_{pk}(i)(i = 1, ..., 6)$ was calculated: 1.0955, 1.0314, 0.9337, 0.9636, 1.0325, and 1.1854.

The $\hat{c}'_{pk}$ value, which is based on the whole data set of N=450, is 1.0175. Note $c_4(450) \approx 1$. The combined estimate, based on the overall sample mean and the pooled sample variance is $\tilde{c}'_{pk} = 1.0400$, while the simple average of the six C_{pk} subsample estimates $\tilde{c}'_{pk}(i)(i = 1, ..., 6)$ equals 1.0404. Although the difference between these two combined estimates is very small, $\tilde{c}'_{pk} = 1.0400$ may be the best combined estimate of C_{pk} based on our results.

Similar results are obtained for other capability indices such as C_{pm} and C_{pmk}.

For stable processes, estimates of a process capability index can be combined to give better assesssment of the process. We have shown that in general the sample process capability indices based on the overall sample mean and the pooled sample variance will have smaller MSE than those using weighted averages of subsample estimates of process capability. The results can be applied to manufacturing industry for process assessment and for evaluation of purchasing decisions.

This paper is published in International Journal of Production Research (2001).

3.3.10 A Weekly Cycle in Atmospheric Carbon Dioxide

K.J. Coakley
Statistical Engineering Division, ITL

R.S. Cerveny
Arizona State University

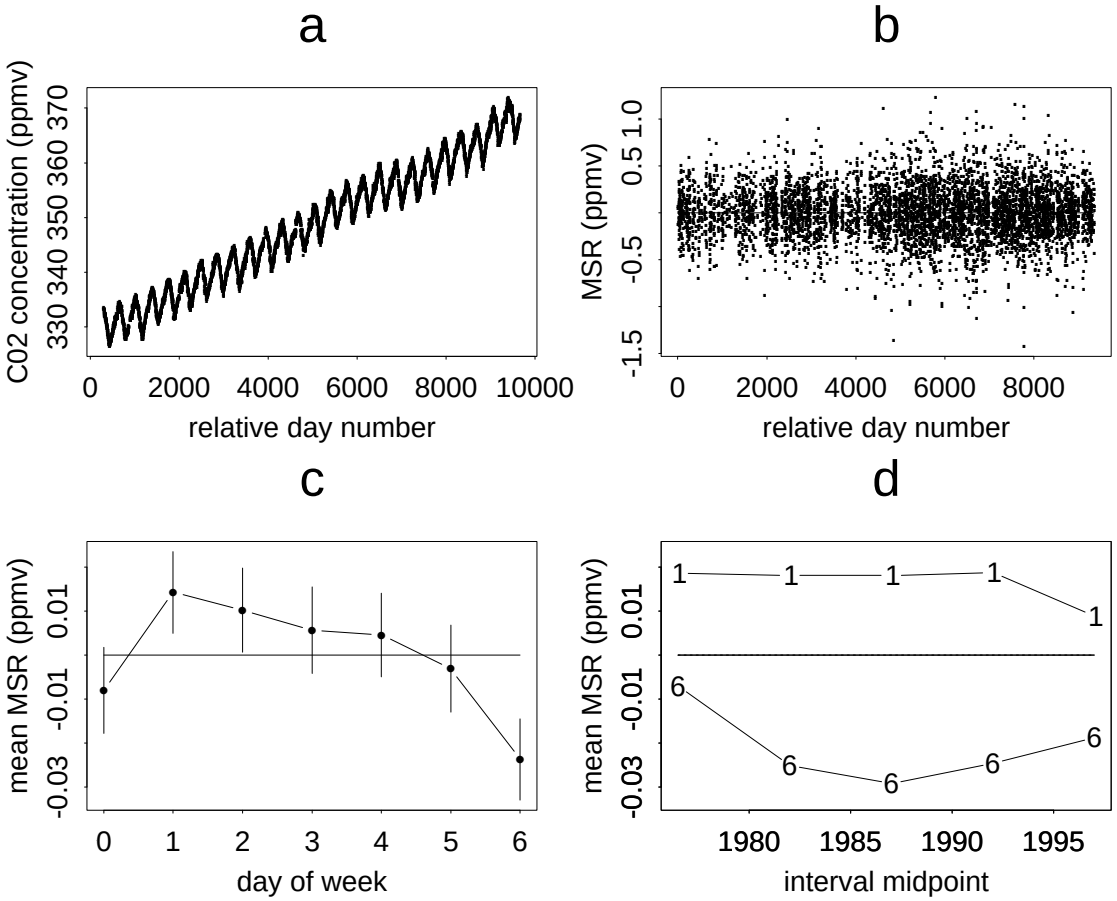


Figure 24: In (a), carbon dioxide measurements at the Mauna Loa observatory from 1974-1999 are plotted as a function of relative day number. There are gaps in the time series. In order to detect a possible weekly cycle in this time series, we filter out the longer-term variation. In this digital filtering approach, we select the filter objectively. In (b), we plot the output of the filter (called the mean symmetrized residual (MSR)). In (c), for each day of the week (Sunday-Saturday), we plot the mean MSR value and an approximate 68 percent confidence interval (mean ± 1-sigma). Sunday is the zeroth day of the week, and Saturday is sixth day of the week. The difference between weekday and weekend mean MSR values is statistically significant at the 0.02 level. We partition the 26-year record into five intervals. In (d), we plot the mean MSR values for Monday (1) and Saturday (6) for each interval.

Detection of a high frequency quasi-periodic signal in a noisy time series can be difficult when the data are contaminated by lower frequency signals. Here, we present a method for detecting a weekly cycle using a digital filtering scheme for which we select the form of the filter objectively. We focus on detection of a weekly cycle in atmospheric carbon dioxide.

We filter out lower frequency trends that complicate detection of a higher frequency quasi-periodic signal by computing a statistic called the mean symmetrized residual (MSR). For each of all possible seven-day time windows that includes an observation, we computed the difference between the observed value and the mean of all observations in the window of interest. The MSR for the ith day is the average of all seven residuals. The MSR can be viewed as the output of a hi-pass digital filter:

$$MSR(i) = \sum_{m=-6}^{6} h(m)x(i+m)$$

with

$$h(m) = \begin{cases} -\frac{7-|m|}{49} & \text{if } m \neq 0 \\ \frac{42}{49} & \text{if } m=0 \end{cases}$$

We sort the MSR values according to weekday. For each of the seven days of the week, we compute the mean of the MSR values and an associated standard error of the mean MSR value.

At the Mauna Loa Observatory in Hawaii, we conclude that CO_2 concentrations, on average, are significantly lower (0.022 parts per million by volume, ppmv) on weekends (Saturday-Sunday) than during the rest of the week. To test the significance of this difference, we compute a test statistic equal to the the standardized difference between weekday and weekend mean MSR values. Using a parametric bootstrap procedure, we estimate the null distribution of the test statistic. The weekend-weekday MSR difference is significant at the 0.02 level.

We speculate that the observed weekend-weekday difference in CO_2 at Mauna Loa is the result of anthropogenic emissions on Hawaii and nearby sources. We did not detect a weekly cycle in atmospheric carbon dioxide measurements at the South Pole.

A paper based on this study was published *Geophysical Review Letters*.

With high confidence, we detected a weekly cycle in atmospheric carbon dioxide. We presented an objective criterion for designing a digital filter for detecting a weekly cycle in a noisy time series. The methodology can be extended to detect cycles of different length.

3.3.11 SRM 4356: Low-Level Radionuclide Ashed Bone

James J. Filliben
Statistical Engineering Division, ITL

Zhichao Lin, Ken Inn
Ionizing Radiation Division, PL

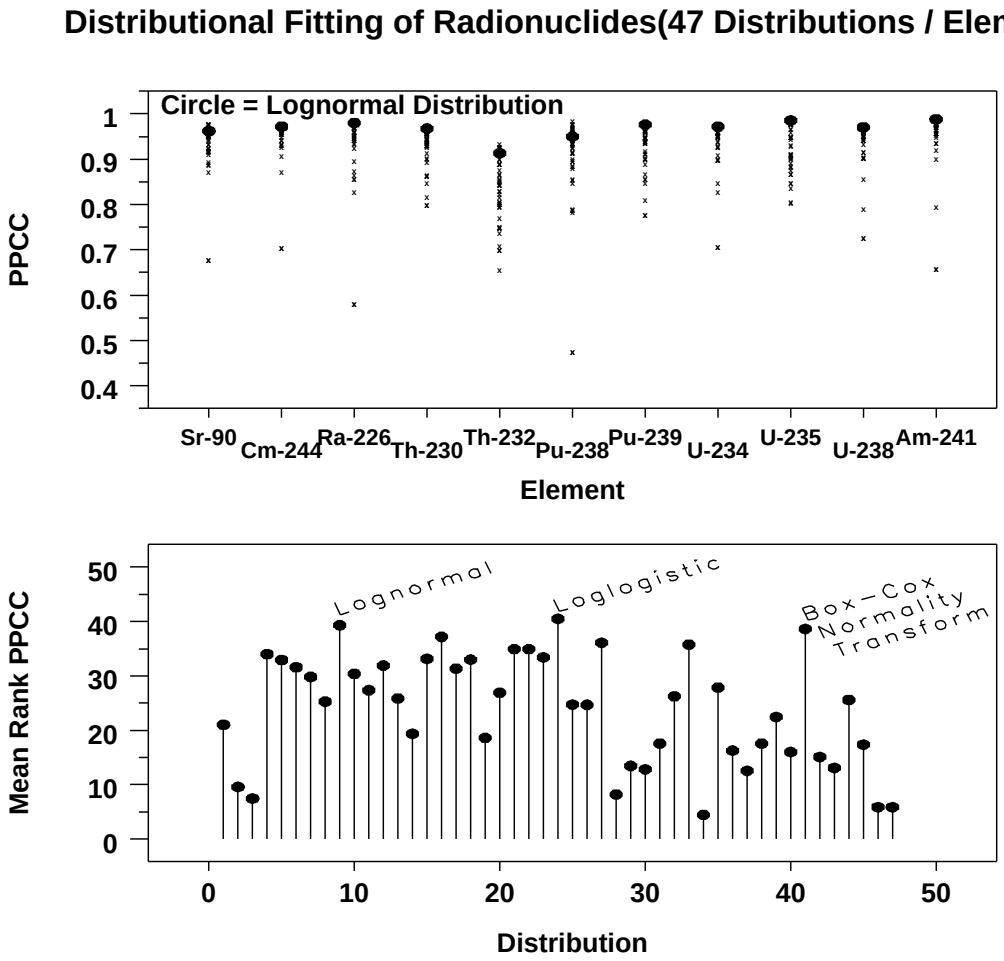


Figure 25: On the top plot, it is seen how the lognormal distribution provides a good fit across all 11 radionuclides. On the bottom plot, the PPCC for each distribution is then ranked (within each radionuclide), and then the average (across the 11 radionulides) ranked PPCC is computed and displayed. The net result is that the lognormal distribution (along with the loglogistic distribution and a Box-Cox normality-transformed distribution) provides the best fit among the 47 distributions.

The human body has 3 biological sinks for long-lived radionuclides: the lungs, the liver, and bones. The lowest of these sinks is bone, and so the study of bone is especially important for determining–even posthumously–the exposure of a person to such radionuclides. Accurate determination of the radionuclides in bone is essential for improving biokinetics modeling and for assessing occupational and public internal radiation dosage.

Two methods of measuring such radiation are alpha spectrometry and mass spectrometry. A problem exists, however, with the use of such devices for low-level radiation because it is difficult to have the specimen physically close enough to the spectrometer to detect a signal. Hence a preliminary step is needed (radiochemical separation) to isolate the alpha atoms from the specimen. These "purified" atoms are then presented to the spectrometer for measurement.

Spectrometric devices undergo continuous improvement, and calibration methods for the devices do exist, but what has been previously lacking is a standard by which the joint accuracy of both steps (separation + measurement) may be assessed.

The lack of such a standard for method validation and quality control in analytical measurements limits the reliability of the current analytical results and the comparability of the data shared among the national and international laboratories.

In collaboration with the International Committee on Radionuclide Metrology (ICRM), NIST's Radioactivity Group in the Ionizing Radiation Division, along with the Statistical Engineering Division, led an international group of experienced laboratories to develop a unique ashed bone standard reference material (SRM-4356) for low-level radionuclides in bone specimens. The SRM is a composite material containing 4% occupationally contaminated human bone and 96% bovine bone. The massic activities of 90Sr, 226Ra, 230Th, 232Th, 234U, 238U, 238Pu, (239+240)Pu, and (243+244)Cm which were certified using a variety of radiochemical procedures and detection methods.

The data analysis addressed a number of problems including homogeneity assessment, distributional fitting, and generalized tolerance limit calculations. Some standard techniques (ANOVA) were applied, but more advanced techniques including Maximum PPCC (Maximum Probability Plot Correlation Coefficient) estimation, in concert with bootstrapping, were employed to determine required tolerance limits. The data analysis indicated that the heterogeneities of the certified radionuclides are undetectable down to a sample size of 5 grams. In order to compute tolerance limits, a common robust data distribution across all certified radionuclides was sought. A set of 47 possible distributions and distributional families was analyzed via the Maximal PPCC (Maximum Probability Plot Correlation Coefficient) criterion. The lognormal distribution provided the most robust distributional fit.

Because the combined uncertainty resulted from inseparable contribution of factors that included interlaboratory variance, material heterogeneity, counting statistics, and analytical methodologies, and because tolerance limits for non-normal distributions are not readily available (or are too wide to be useful), a bootstrap approach was subsequently utilized for the actual calculation of the lognormal tolerance limits. Such limits were com-

puted for each element, and they were reasonable and consistent from both a physics point of view and from a data point of view.

It was found that a unique characteristic of this SRM material is the disequilibrium of U and Th decay chains. The disequilibria were the results of mixing occupationally contaminated human bone with natural bovine bone and the fractionation during internal biological processes. Radionuclide disequilibria prevented the certification of the U and Th daughters, 210Pb and 228Th.

The release of SRM 4356 serves to complete the suite of reference materials for the long-term radionuclide sinks in the human body:

- *the lungs (SRM 4351)*
- *the liver (SRM 4352)*
- *bones (SRM 4356)*

Further, SRM 4356 joins the family of low-level natural-matrix SRMs:

- *SRM 4350B (river sediment)*
- *SRM 4354 (freshwater lake sediment)*
- *SRM 4355 (Peruvian soil)*
- *SRM 4357 (ocean sediment)*

The above comprehensive collection of SRMs serves as a national and international standard for accurate monitoring of low-level radionuclides for the next decade.

This omnibus joint approach (PPCC + bootstrap) has proven to be a significant improvement in both distributional "estimation" and in tolerance limit estimation. This methodology has recently been the subject of a paper "An alternative statistical approach for interlaboratory comparison data evaluation" (Lin, Inn, and Filliben) in the Journal of Radioanalytical and Nuclear Chemistry (Vol. 248, No. 1 (2001), pages 163-173).

3.4 Collaboration

3.4.1 Statistical Visualization of Network Performance in Terms of Upper Quantiles

Hung-kung Liu, Nell Sedransk, Z. Q. John Lu
Statistical Engineering Division, ITL

David Su, Doug Montgomery
Advanced Network Technologies Division, ITL

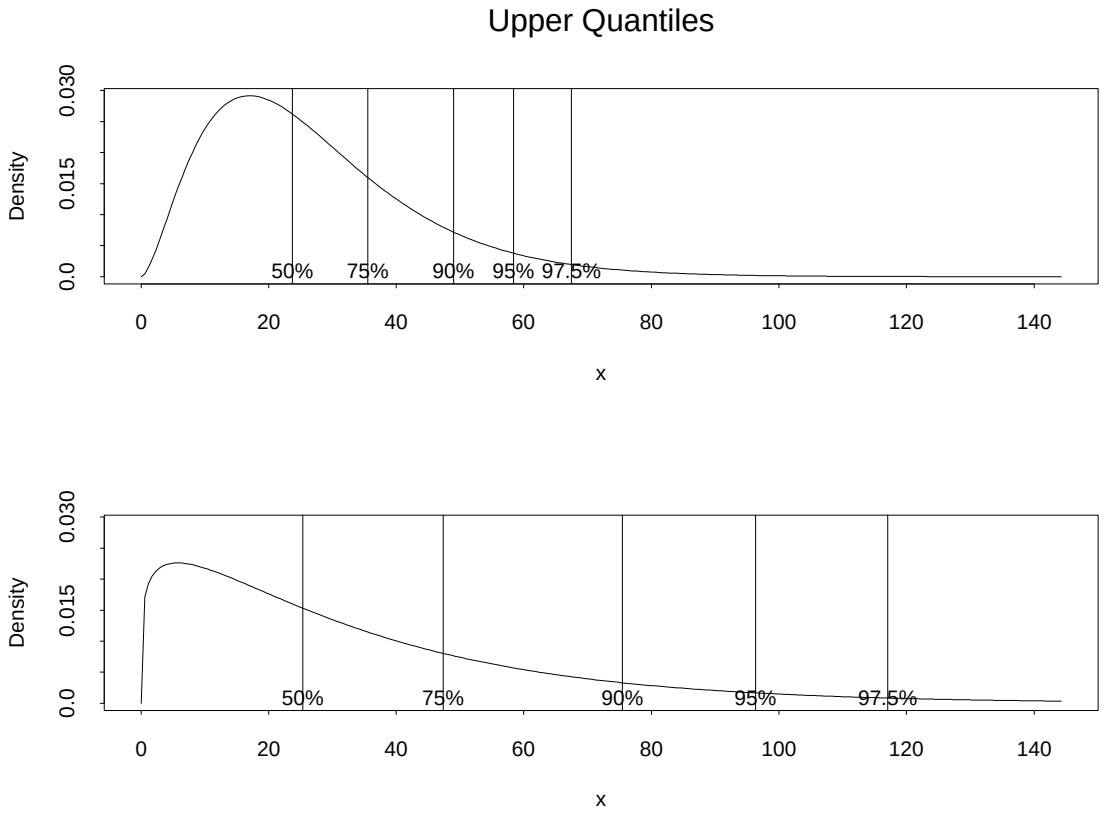
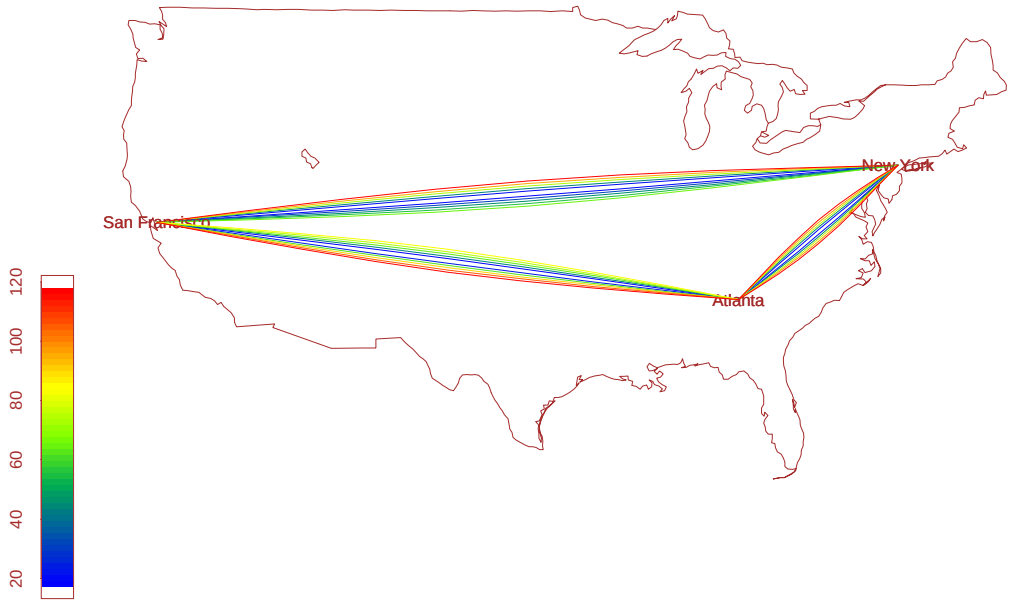


Figure 26: Sample heavy-tailed distributions.

The Statistical Engineering Division and the Advanced Network Technologies Division are collaborating with industrial consortia that are developing metrics and tools to measure the performance of Internet services.

The critical issues for statistical diagnostics and summaries of network traffic behavior are characterizing the tail behavior for the "heavy-tailed" empirical distributions of network measurements and detecting changes in distribution patterns. We have begun development of statistical graphical tools for visualizing Internet data that do not depend on a particular Internet model by focusing on the upper quantiles rather than means or medians. A spatial representation of the two different scenarios of heavy-tailed distributions in the previous plot is shown below.



Current plans include augmenting these general-purpose statistics and graphics to include trend summaries and diagnostics for changes in the mixture proportions.

This work is funded in part by DARPA.

Our efforts in statistical modeling and visualization are intended to create statistical models sophisticated enough to cover a broad range of real network behavior, and yet simple and intuitive enough to be easily employed by researchers developing network simulators, emulators, and control models.

3.4.2 Ranking Algorithms for Face Recognition

Andrew Rukhin, Stefan Leigh, Alan Heckert,
Mariama Moody, Kimball Kniskern, Susan Heath
Statistical Engineering Division, ITL

P. Jonathon Phillips, Patrick J. Grother
Information Access Division, ITL

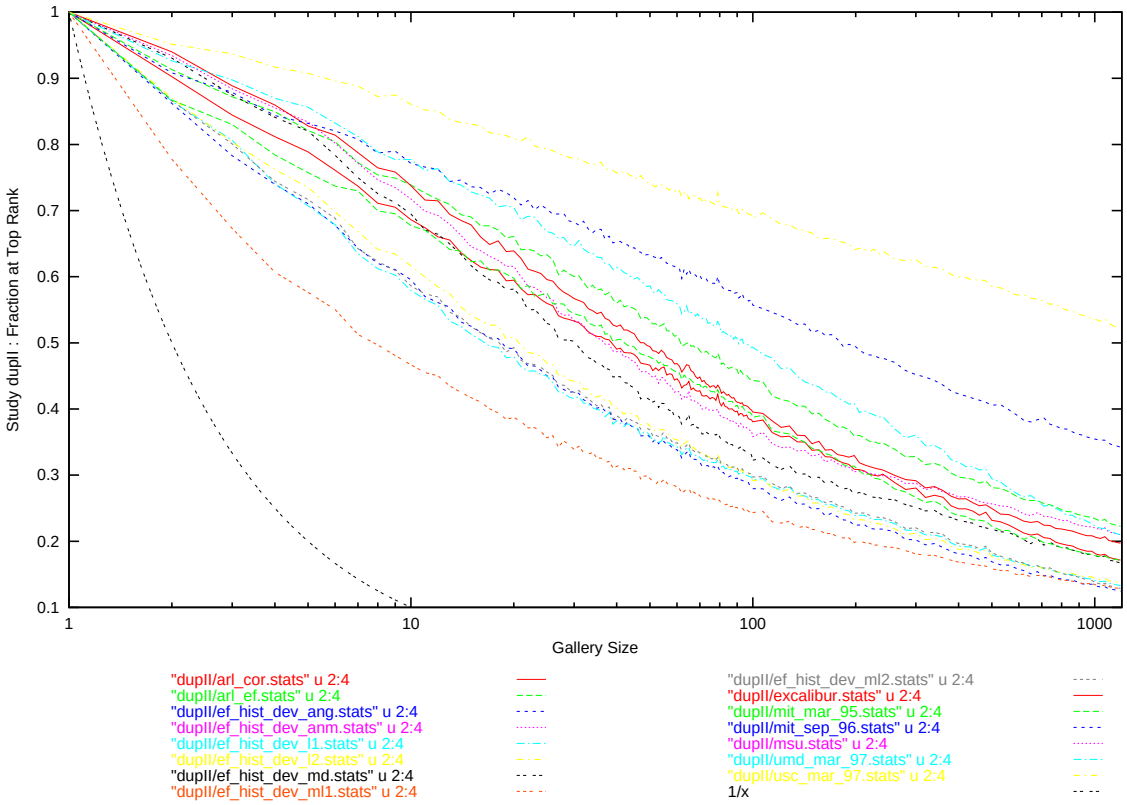


Figure 27: Probability that an algorithm will find the correct match in a Gallery of faces as a function of the number of subjects represented in the Gallery. Random guessing leads to the straight line. All reasonable real-world algorithms outperform this. The best roll off logarithmically. The images used here represent matching recent Probe images against Gallery images 18 months old.

Identification and verification of a person's identity are two generic application areas of face recognition systems. In identification applications, an algorithm identifies an unknown face in an image by searching through an electronic mugbook. In verification applications, an algorithm confirms the claimed identity of a particular face. Proposed applications have the potential to impact all aspects of everyday life by controlling access to physical and information facilities, confirming identities for legal and commercial transactions, and controlling the flow of citizens at borders.

For face recognition systems to be successfully fielded, one has to be able to evaluate their performance. To evaluate an algorithm, its behavior is scored on a test set of matchable images in a mugbook known as the Gallery. One computes a similarity matrix that quantifies the proximities of images of a subset of the Gallery (called the Probe set) to each image in the Gallery.

The task of SED is to develop methods for comparing algorithmic performance based solely on comparison of similarity matrices generated by the algorithms under test running against prefixed Probe/Gallery sets. Since the algorithms are proprietary, comparison methods may not presuppose detailed knowledge of any particular algorithm.

Large collections of test images are already in existence (FERET/Army Research Lab/ George Mason Univ./93-96) or currently undergoing development (Human ID/DARPA/99-04). These databases (which include IR, still, video, and hyperspectral images of the face, gait, and iris of thousands of human subjects) provide the Human ID research community with de facto database standards for algorithm development and comparison.

A first, simple approach (to appear: Proc. paper accepted for the Third Workshop on Automatic Identification Advanced Technologies, Tarrytown, March 2002) is to limit the comparisons to replicated same-face match scores, transform the scores from the multiple algorithm outputs to a common scale, and examine ranking's and clustering's produced by application of standard Multiple Comparisons procedures, e.g., Student-Newman-Keuls. A useful common scaling is achieved by Probability Integral Transform-ing (PIT) each algorithm's scores using knowledge of its characteristic score EDF based on larger heterogeneous (FERET) experiments, then applying the inverse Gaussian cumulative distribution. Application of this procedure to a sizable extract of the FERET database yields a credible ranking of 15 algorithms dated 1996-1997.

An extension to a mixture of same-subject and different-subject match scores can be achieved by use of ordinary 2-dimensional MultiDimensional Scaling (MDS). MDS translates similarity matrices into pictorial maps with matrix row/column headers converted into mapped locations with appropriate inter-location distances. A good algorithm should cluster same-subject images and cleanly discriminate among different-subject images. The ability to discriminate, and tightness of clusters as quantified, e.g., by circumscribed Voronoi ellipse aspect ratios, can be used to rank algorithm performance. Demonstration tests against small-scale FERET extracts show this clearly.

While Φ^{-1}-PIT and use of Multiple Comparisons and MDS have the advantage of retaining the ratio scale of the original similarity scores, much of the work already published

and currently being done in this area makes use of rank statistics. We are exploring (submission, pending acceptance: International Conference on Pattern Recognition, Quebec, August 2002) multiple properties and statistics derived from the use of partial rank correlations (PRC). This involves extending the known distributional theory for PRC's based on Kendall and Spearman statistics and applying them to the study of interesting dependency patterns among different algorithms. Loosely, the ID community recognizes that most current algorithms perform most reasonably in scoring true (close) matches and (far) dramatically disparate non-matches: i.e., algorithms perform best at the far ends of the performance scale. It is commonly presumed that enhanced understanding of algorithmic performance (and the dual issue of image difficulty) will come from "pushing in" at either end of the match/nonmatch performance scale. The application of nonparametric dependence via copula theory to partial rank co-occurrences seems to hold promise for enhanced understanding here.

In addition, the team has performed rough draft work on All Possible Subsets and Alternating Conditional Expectation (ACE) modeling of covariates' explanatory power for FERET similarity scores, as well as the application of simple Stochastic Matrix Ordering techniques for ranking similarity matrices. This may lead to better designs of image recognition studies.

Of all the security technologies thrust into the spotlight by the events of September 11th and the aftermath, systems that try to identify people by analyzing computerized images of their faces are among the most prominent, controversial, and potentially promising. Demands for workable technologies have exploded, as have demands for objective means to evaluate existing technologies and provide guidance in constructively improving existing technologies. The DARPA Human ID Project, as a major funder of ongoing U.S. research in this area, serves as a magnet, mixing bowl, and ultimately test bed provider for the rapidly evolving knowledge base and set of systems. SED is tasked, by DARPA and the NIST-resident Human ID Group, to help with the algorithm comparison test collection.

3.4.3 Archetypal Topics in Evaluation of Information Retrieval Systems

Walter Liggett
Statistical Engineering Division, ITL

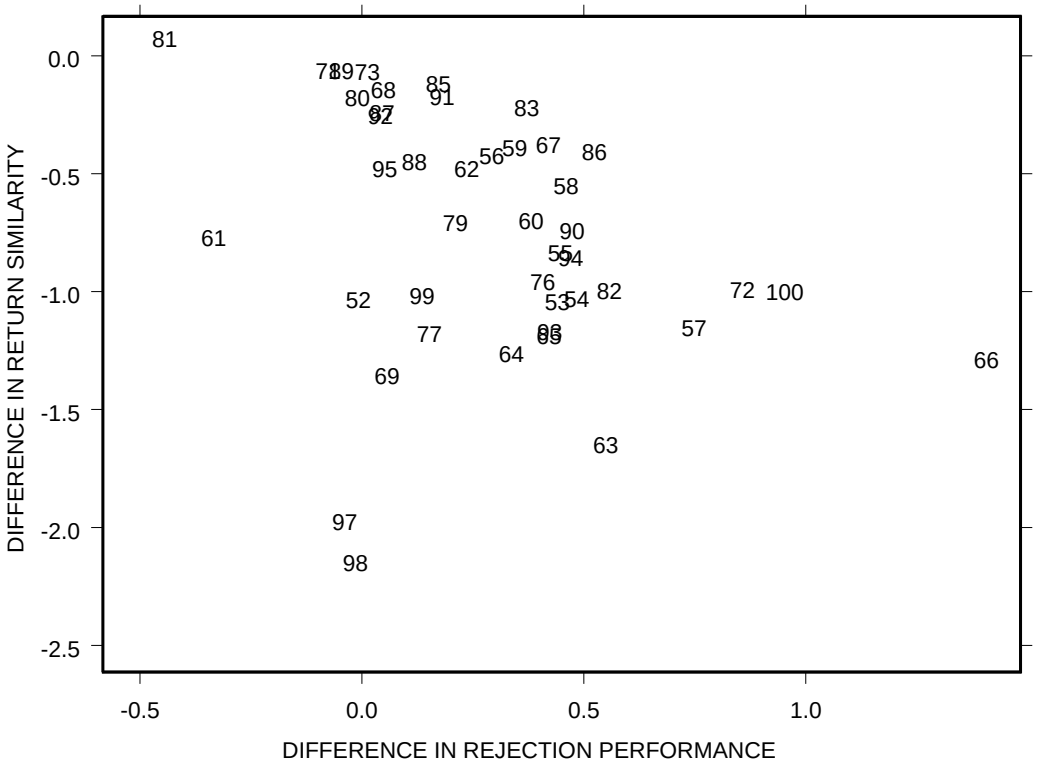


Figure 28: Topic-by-topic comparison of two systems in terms of two performance indicators.

A prominent information retrieval (IR) task is the search of a known, stable collection of documents for those relevant to a novel, unanticipated topic. Comparative evaluation of IR systems in their performance of this task requires a group of topics and associated expert judgements of the retrieved documents as relevant or not relevant. The major challenge in this evaluation is inference from the group of topics to the topic population for which the systems are intended to be useful. This is a major challenge because the effect of the topic is larger overall than the effects of other factors in the evaluation.

Consider a large document collection, a group of topics (information needs), a designation of each document in the collection as relevant or not relevant to each topic, and information retrieval systems that respond to a query (a natural language statement of a topic) with an ordered list of 1000 documents from the collection. The goal is to draw conclusions about systems from the relation between the system responses and the relevance designations. Alternative queries for the same topic make possible a new approach to this goal because system responses to a battery of queries provide insight into the topic properties that influence system behavior. These topic properties, which can be regarded as latent variables, arise from the possibilities for natural language expression that the queries exhibit. An example of such a property is the extent to which a single key word or phrase distinguishes relevant documents from the others. Summarizing such insights over a group of topics using a form of archetypal analysis leads to general conclusions about system behavior.

Our approach to summarization consists of two steps. First, we summarize each of the fifty topics with indicators that seem to provide meaningful comparisons among topics. Second, inspired by archetypal analysis (Cutler and Breiman 1994, Technometrics), we use these summaries to choose a few topics that are as different as possible. These topics are as close to being archetypes as the set of topics allow. Because of the way these topics are chosen, we assume that the influence of language structure on system behavior shown by these topics applies to all fifty topics and beyond these to all similar topics.

For each topic, the values of the indicators are computed from the 1000-document lists, of which there is one for each alternative query, and from the designations of documents as relevant or not relevant to the topic. One indicator is a measure of rejection performance; that is, performance in rejecting irrelevant documents. The steps in computing this consist of (1) reducing each 1000-document list to the positions of the relevant documents, (2) finding for each list, the number of irrelevant documents that occur before one quarter of relevant documents are returned, (3) applying a log transform to each of these values, and (4) averaging over the queries for the topic. Almost without exception, IR system evaluation has been predicated on indicators such as those that are based only on the positions of the relevant documents among the irrelevant ones, not on the document identifiers.

Another indicator is a measure of return similarity: similarity in the return order of the relevant documents. The return order of relevant documents can be used to compare system responses for the same topic through computation of dissimilarities. The computing begins with reducing each 1000-document list to the sublist consisting of just the relevant documents, and obtaining for each relevant document its position (rank) in this

sublist. An obvious choice for checking dissimilarity is Spearman's coefficient of rank correlation. The indicator is computed by summing over the (squared) dissimilarities for all pairs of queries. This indicator is a measure of success with query expansion. One problem that affects IR system behavior is a mismatch between phrases used in a query and different phrases with the same meaning used in some of the documents relevant to the topic. Query expansion is intended to address this problem. A system that was completely successful with this problem would have the same relevant document return order for all queries.

In the figure, we compare two systems by plotting the difference in return similarity versus the difference in rejection performance. It is apparent that topic 66 is unusual. Other evidence, which is not interesting enough to be discussed further, shows that topic 66 is an outlier. The next topic in from the right is topic 100. This topic and the others on the periphery of the central cluster seem worthy of special attention because they have the potential for providing clear insights. These are the archetypal topics.

If one could measure the properties of a topic, one could model relative system performance as a function of topic properties. One could then specify the class of topics for which the chosen system is to be used and choose the system on the basis of the model. This approach cannot yet be implemented because the topic properties are latent variables, not ones that we yet know how to measure directly. Nevertheless, this work offers as an alternative to probability sampling, a procedure consisting of defining the types of topics on which a system is to be used in terms of these latent variables and then deciding for each type of topic the system that will perform best. This is not a foolproof procedure because our knowledge of these natural language factors is still murky. Nonetheless, this procedure is an alternative to the assumption that the topics in the evaluation are a random sample from the appropriate population.

The value of the activities associated with NIST's Text REtrieval Conference (TREC) is threefold: the challenges in completing the tasks put forward, comparison of ideas at the TREC conference, and discovery of algorithmic innovations that lead to better systems. The last one of these is the one to which this work on system comparison contributes.

3.4.4 Range Imaging and Registration Metrology

Stefan Leigh, Andrew Rukhin
Statistical Engineering Division, ITL

Christopher Witzgall, David Gilsinn
Mathematical and Computational Sciences Division, ITL

Geraldine Cheok
Structures Division, BFRL

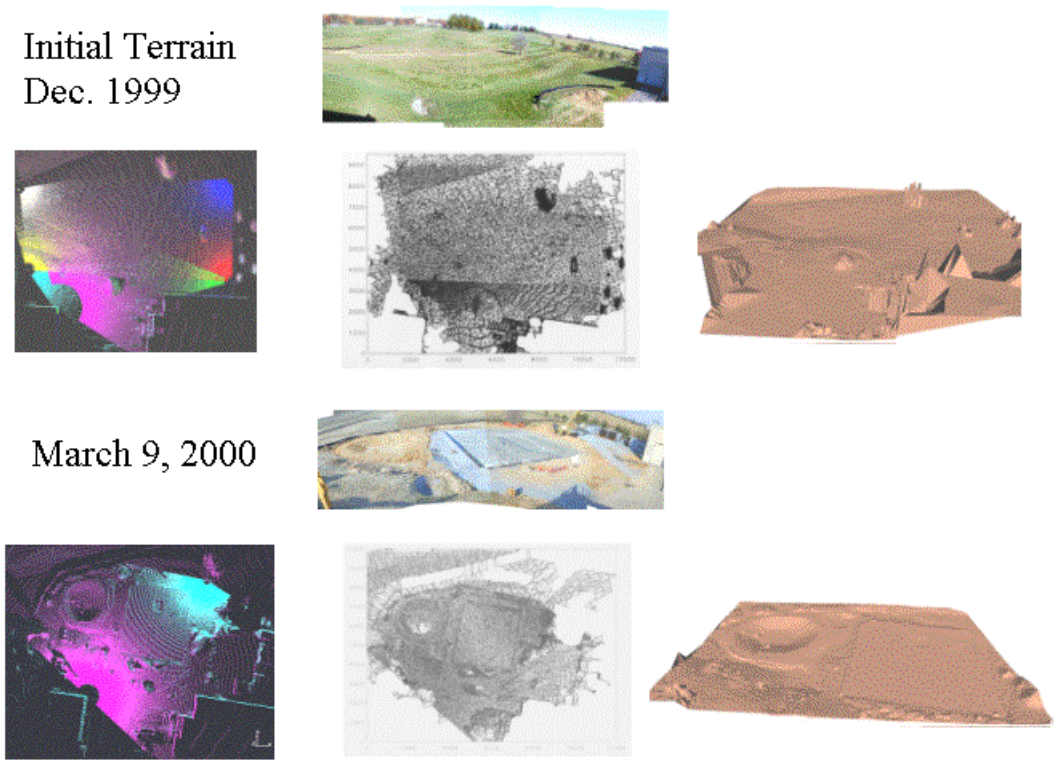


Figure 29: Digital photographs of the NIST campus construction site with LADAR scans, composite 3D mesh numerically generated from the LADAR images, and the resulting reconstructed 3D surface model of the site.

Implementations of range imaging sensing, such as LADAR (laser distance and ranging), are anticipated to have manifold applications in both military and commercial settings soon. In civil engineering, LADAR could be used to rapidly track terrain changes due to excavation at a construction site, with procedures and methods developed to display ongoing results in real time. Such capabilities would enable visualization and feedback-based corrective measures by on-site or off-site contractors, engineers, and designers.

The Construction Metrology and Automation Group of BFRL continues to work on the use of interactive LADAR for rapid assessment of status and quantitative changes in amorphous objects on construction sites. A Non-intrusive Scanning for Construction Status Assessment project identified 3 key areas for research: registration of data from different scan locations, determination of the accuracy of surfaces reconstructed from LADAR data, and object recognition. The first two areas interrelate in obvious ways: poor registration results in the generation of an incorrect reconstruction, and both registration and surface generation methods involve calibration issues that require development of a set of protocols and statistics-based evaluation criteria to measure actual performance.

In order to objectively evaluate surface reconstruction algorithms, first the accuracy/precision characteristics of the sensor used must be determined. Calibration experiments, varying distance/size/color/reflectivity of targets and variations in size and disposition of laser beam, are ongoing. The analysis of such data represents SED's current major contribution to this project. In a second phase, characteristics of the device and reconstruction algorithm with respect to handling of missing points, outliers, discontinuities, vertical surfaces etc. are being determined. In a third phase, all such knowledge is to be integrated into a credible calculation of statistical uncertainty for the reconstructed scene or volume.

In the first and second phases, a set of metrics has to be established to assess accuracy. For sensor evaluation, such metrics depend largely on the sensor characteristics and are relatively straightforward for range calibrations. However, determination of the angular accuracy of the scanner is turning out to be more complicated due to divergence of the laser beam(s) and because some scanners use lasers outside the visible range. For the evaluation of the surface generation algorithms, these metrics are harder to establish. One approach is comparison with simple reference surfaces ("ground truth"), such as simple geometric objects of preestablished shape and volume, progressing to more complex shapes. Algorithm accuracy can be evaluated based on how well known volumes are reproduced.

The Construction Industry Institute FIATECH Consortium has identified 3D laser scanning as one of their highest priority technical development programs in the coming years. NIST is providing the technical leadership for this project. A CRDA with Reality Capture Technologies, Inc. was initiated in FY01 to study various methods for processing scan data from the NIST 205 construction tests. Riegl, Cyra and Metric Vision are working with NIST to help define the test and calibration needs of industry. It is anticipated that other companies will be invited to join this collaboration in FY02.

3.4.5 Bioactivity of Ultra-High Molecular Weight Polyethylene (UHMWPE) Wear Microparticles

James Yen, Stefan Leigh, James Filliben
Statistical Engineering Division, ITL

Hsu-Wei Fang
Ceramics Division, MSEL

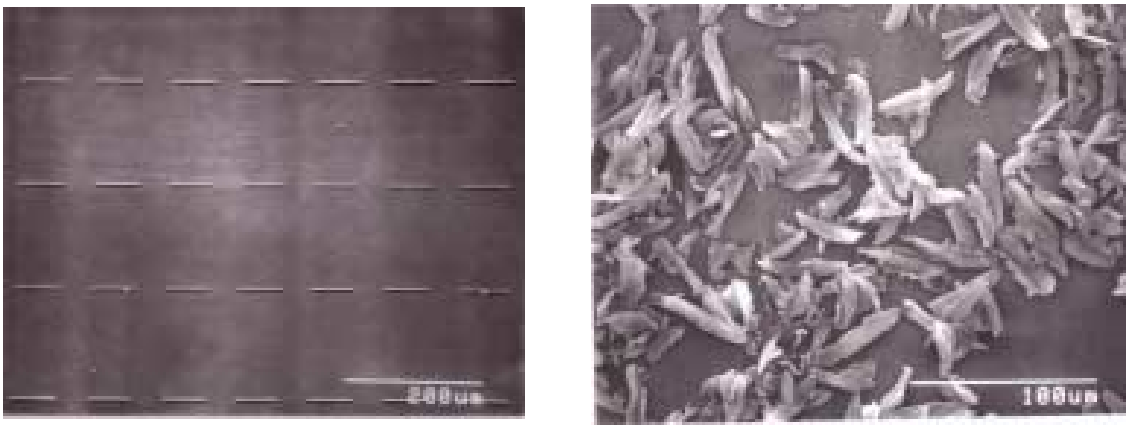


Figure 30: Hsu-Wei Fang of the Ceramics Division has developed the technique of rubbing solid polyethylene pins against the finely patterned surface of silicon wafers (shown at left), causing tiny wear particles of relatively controlled size and shape to rub off (shown at right). Scientists hope to use such particles to investigate how the size and shape of particles sloughed from artificial joints affects the inflammatory response from the surrounding tissue.

SED statisticians are helping investigate how the shape of wear particles sloughed by artificial joints during normal use affects the degree of damage to the surrounding tissue.

Artificial joints typically contain parts made of ultra-high molecular weight polyethylene (UHMWPE). Normal use causes microscopic UHMWPE wear particles to slough off the artificial joint into the surrounding tissue. These wear particles have been linked to inflammation, bone resorption, and eventual loosening of the joints. Investigators believe that the biological reaction to wear particles depends on the shape of the particles; in particular, they hypothesize that longer, thinner particles will cause more harmful bioactive responses than rounder particles. If that hypothesis proves true, then future artificial joints may be constructed so that the generation of needle-like (longer and thinner) wear particles will be minimized, thus increasing the life and functionality of artificial joints.

Hsu-Wei Fang of the Ceramics Division is perfecting methods of generating UHMWPE particles of uniform size and shape. Using micro-lithographic techniques borrowed from the semiconductor industry, he etches extremely fine patterns onto polished silicon wafers; the picture on the left shows the surface of part of such a wafer. A special apparatus then rubs pins made of solid UHMWPE against the patterned surface, causing tiny UHMWPE particles to slough off and be caught in a water bath. The size and shape of the wear particles can be controlled to some extent by the pattern of ridges etched on the silicon wafer. The picture on the right shows part of a batch of needle-like particles.

SED statisticians have been in the process of helping Mr. Fang analyze his data and plan future experiments. Initial bioactivity studies from test tube experiments indicate that there is a shape effect to the biological reactions that the wear particles provoke. Future studies involving further lab experiments and possible animal studies will try to capture and disentangle the effects on bioactivity of the shape, size, and dosage of the particles.

If the shape of wear particles does have a significant impact on their bioactivity, then future artificial joints may be constructed so that the generation of the more harmful wear particles will be minimized, thus increasing the life and functionality of artificial joints.

3.4.6 The Effects of Hydrolytic Degradation on the Biocompatibility of a Polylactic Acid Used for Bone Repair

William F. Guthrie
Statistical Engineering Division, ITL

Sumie Yoneda and Francis Wang
Polymers Division, MSEL

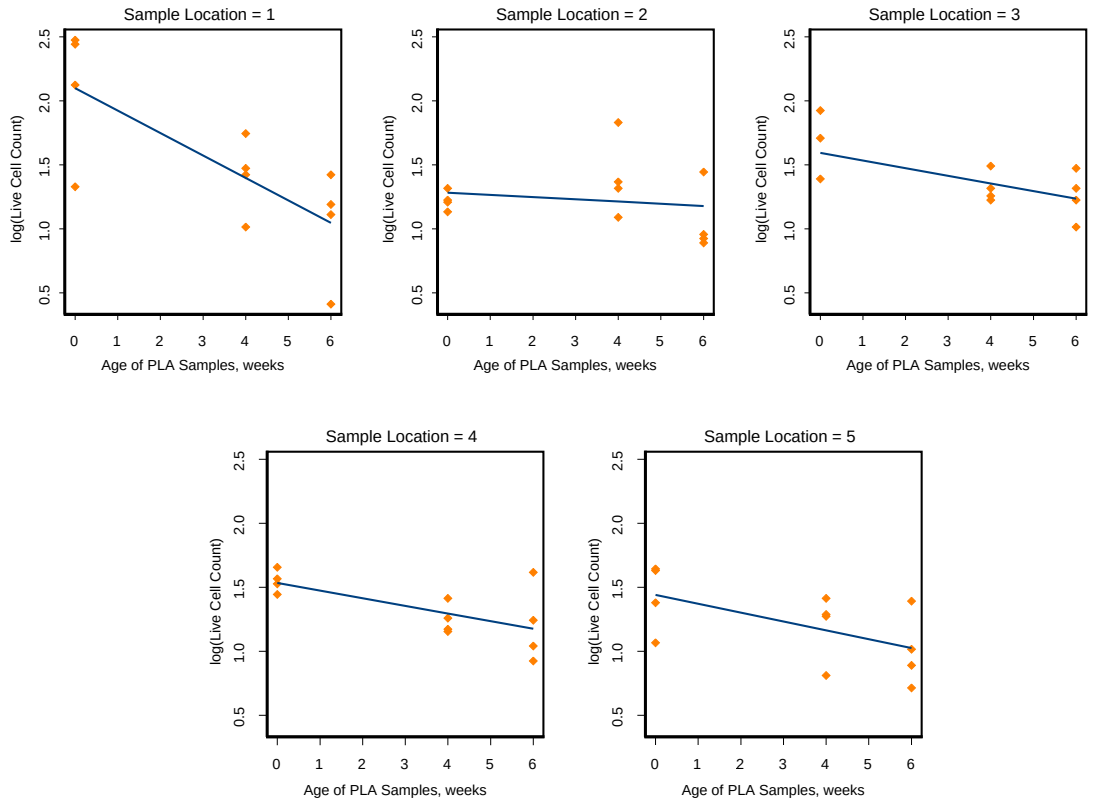


Figure 31: Plots comparing the effects of hydrolytic degradation on the number of live cells at different locations on samples of a polylactic acid (PLA) used for bone reconstructions. The samples were allowed to degrade by aging them for different amounts of time in an aqueous solution. The samples were then seeded with cells and the number of live cells was assayed after 16 hours of incubation. The negative slopes observed for all locations on the sample indicates that fewer cells survived on the degraded samples. Of the five locations on each sample, cells were less likely to survive in the center of the sample (Location 1), than near the edges (Locations 2-5). The cells survived at equal rates on fresh PLA samples and on tissue-culture polystyrene, indicating that the non-degraded PLA samples were not toxic to the cells.

Recently biodegradable polymers have begun to be considered for use in a variety of advanced medical applications, such as temporally vascular grafts, artificial skin and orthopedic implants like bone nails, screws, or staples. Amorphous polylactic acid (PLA), a polyester, is a promising candidate for inclusion in composite scaffolds used to repair bone defects because of its high biodegradability, which provides space for the formation of new bone. However, knowledge of the adhesive interactions between bone cells and biodegradable polymers is needed for the design of novel biomaterials and the development of new strategies for bone repair. To help provide this information, in vitro experiments were done to assess the biocompatibility of PLA under degradation conditions similar to those that would be experienced in vivo. Among these experiments was one that compared the numbers of cells surviving at five different locations of PLA samples that had been allowed to degrade for different amounts of time. It was hypothesized that if the biocompatibility of the PLA decreased as it degraded, live cells would also be less likely to be found near the centers of the PLA samples than at their edges, where they might be less exposed to degradation byproducts.

The design of this experiment involved real trade-offs due to constraints caused by the nature of the factors being studied. If the scientists had only a general interest in potential location effects on the samples, a completely randomized, nested design could have easily been used to estimate both age and location effects on the biocompatibility of the PLA. With the nested design, the scientists could control how many levels of degradation they would like to study, the total number of samples to be prepared, and the number of locations on each sample for which cells would be counted. However, their desire to estimate the effects of degradation at specific locations on the sample required either that each sample be measured at only one location or the assumption that no significant sample-to-sample variation would be observed. Given the levels of degradation and location to be studied, measuring each sample at only one location would have required 60 samples to be prepared, a significant amount of work that could only be done in sequential batches. On the other hand, using only 12 samples, each measured at all 5 locations, would yield correlated data if sample-to-sample variation were present. Since it was believed that the samples could be fabricated nearly identically, the 12 sample design was chosen. During the analysis of the data, which did show significant sample-to-sample variation, SED staff were able to confirm that the correlation of the samples only made the typical standard errors for the estimated model parameters more conservative and, based on variance component estimates, probably had little impact on the results, justifying the difficult design choices that were made.

From the results of this experiment, shown in the figure on the preceding page, it is clear that degradation of PLA reduces its biocompatibility. It is notable, however, that a reduced number of cells were able to survive degradation. This suggests that PLA used in composite bone scaffolds may support bone infiltration by providing room for growth of new bone in the scaffold if the remaining scaffolding materials are biocompatible. Knowledge of the biocompatibility characteristics of PLA fills in another piece in a complex puzzle that will lead to improved methods for bone repair when solved.

3.4.7 Thermal Properties of Pyroceram 9606

James Yen, James Filliben
Statistical Engineering Division, ITL

Dan Flynn, Bob Zarr, Erik Hohlfeld
Building Environment Division, BFRL

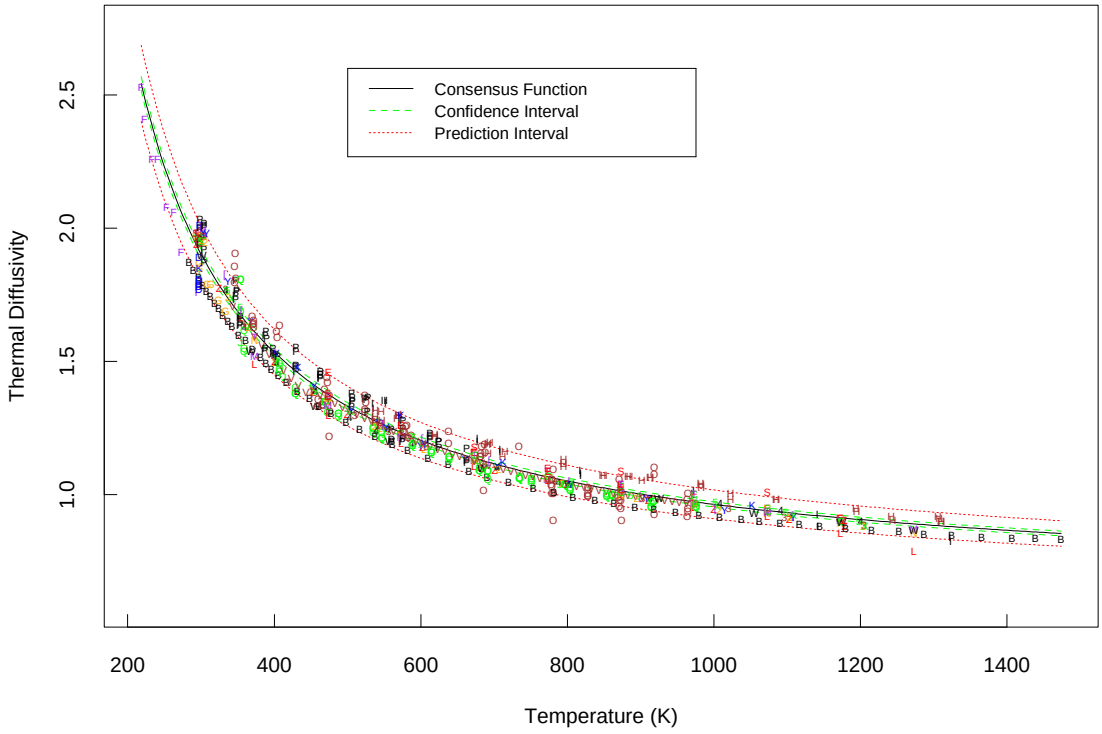


Figure 32: The picture shows the data from 31 thermal diffusivity data sets; the points from each data set are depicted by different symbols. The solid black line shows the estimated consensus diffusivity function. The red dotted lines depict estimated pointwise 95% confidence intervals, while the green broken lines bracket estimated pointwise 95% prediction intervals.

NIST will soon market Pyroceram 9606 as a reference material for certain thermal properties. SED statisticians helped pool the data from various experiments to estimate consensus thermal functions for Pyroceram.

Corning developed a glass ceramic material, Pyroceram 9606, especially suited for high temperature applications. NIST intends to use Pyroceram as a reference material for use in calibration and performance evaluation of instruments measuring thermal properties such as thermal conductivity, thermal diffusivity, and specific heat (heat capacity). All of these quantities are temperature–dependent. Therefore, the reference values would ideally take the form of a function of temperature. Usually, however, the reference values are given only for a subset of temperatures (e.g., 100K, 200K, 300K, etc.).

Dan Flynn of the Building Environment Division (BFRL) has gathered data sets from numerous laboratories around the world where thermal measurements were made for Pyroceram. There were 18 data sets for thermal conductivity, 31 for thermal diffusivity, and 5 for heat capacity. There were other data sets that were excluded for scientific and/or statistical reasons.

The figure depicts the data for thermal conductivity, with each symbol denoting a different data set. Some experiments only contained measurements over a small temperature range; also, the original data from some of the experiments are not available—only the estimated or smoothed values from those experiments are left, leading to possible excessive smoothness or arbitrariness of sample size.

In theory, the thermal properties satisfy the constraint Diffusivity=Conductivity/ (Density * Heat Capacity). The thermal quantities were first estimated independently, and then small adjustments were made at some temperatures to help satisfy the constraint. Thermal theory was used for estimating the heat capacity as a function of temperature, using a function involving the Debye function. The conductivity and diffusivity were fit using nonlinear regression to an empirical model.

The uncertainties in the reference values take the form of pointwise 95 percent prediction intervals and include the uncertainties due to regression error of the fitted values and the variability among the different data sets. These uncertainties were computed using an analysis of residuals from the smoothed values. The figure also shows the much narrower 95 percent pointwise confidence intervals.

Pyroceram 9606 will provide a needed reference material for use in calibration and performance evaluation of instruments measuring thermal properties in high temperature applications.

3.4.8 Errors in Variables for Gas Metrology Calibrations

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

Frank Guenther
Analytical Chemistry Division, CSTL

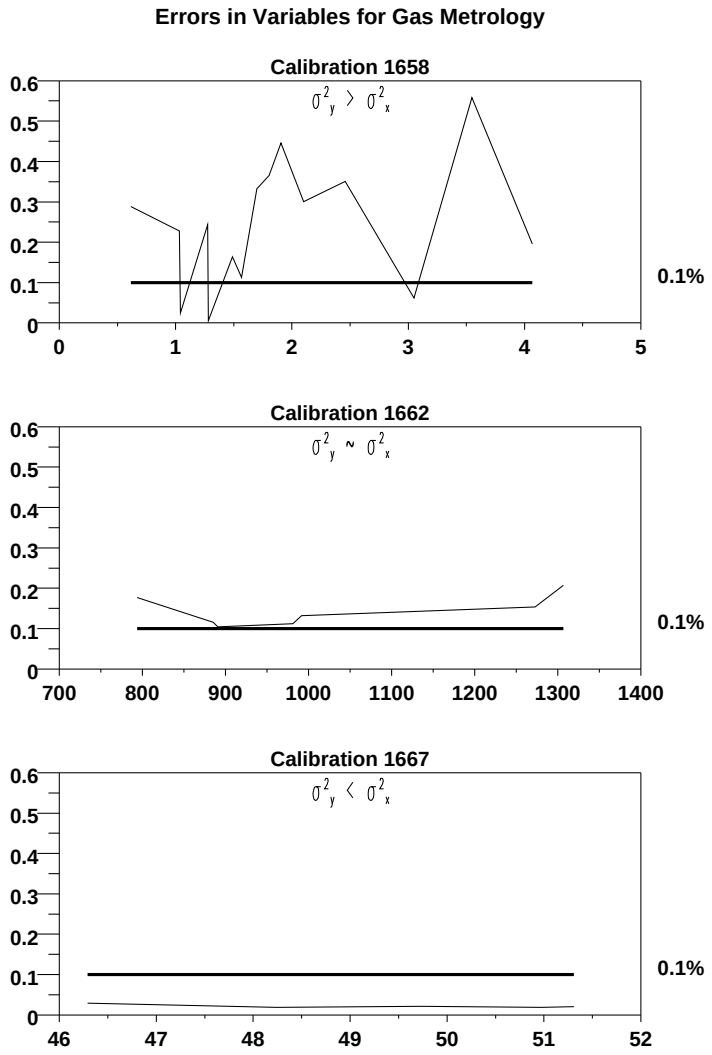


Figure 33: Relative standard deviation plots derived from actual calibration lines showing the occurrence of each of the three modes for regression analysis: $\sigma_Y > \sigma_X$, $\sigma_Y \approx \sigma_X$, $\sigma_Y < \sigma_X$. The solid horizontal 0.1% line represents the assumed relative standard deviation in the X variate, 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 regression methodology: 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: specifically, the crucial $\sigma_Y^2 \gg \sigma_X^2$. It was the latter that led the PTB (German) delegation to the ISO Gas Analysis Working Group TC 158 to develop and sponsor through to adoption a new Gas Analysis Standard (ISO 6143.2), which is based on the use of Errors in Variables methodology for the analysis of gas mixture composition calibrations.

A working draft standard was circulated over a period of 3 years prior to adoption and commented on by all concerned national standards organization participants, except for the U.S. which had no official representation at the meetings during this time period. NIST noted several troubling features of the new standard. Incorrect language is employed: Errors in Variables is termed Generalized Least Squares. The adaptation of Errors in Variables methodology, to supplant OLS, is recommended across the board, irrespective of the variance patterns in the data. Also, the Errors in Variables technique in the new standard is implemented in the form of an executable code. It can be surmised that the computer program is an adaptation of a "Generalized Least Squares" routine from Press *Numerical Recipes in C*.

We are in the process of carefully reviewing part of the large corpus of NIST's Gas Metrology Group's archival calibration data, with a view toward developing a statistically justifiable set of procedures for the analysis of such data. While we agree with the push for the use of EiV technology in this area, we believe the new 6143.2 to be untenable: specifically in its substitution of a black-box computer code for careful statistical understanding and analysis of datasets on a case-by-case basis.

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 practioners with clearcut guidelines and procedures. We seek to lay down a clear logic for the linear calibration problem with specific reference to NIST's Gas Metrology calibration experience. We expect those guidelines to be documentable, with clear reference to broadly accepted inferential principles, such as maximum likelihood or Bayesian estimation. Whether an existing calibration uncertainty prescription will be used or a new one developed, 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 ultra fast adaptation of such standards by national industries translate to direct commercial leverage in OECD and third-world markets. NIST, and the other world standards organizations, bear direct responsibility for ensuring the technical integrity of such standards.

3.5 SRMs and General Consulting

3.5.1 Standard Reference Materials

James J. Filliben, Alan Heckert
Statistical Engineering Division, ITL

Standard Reference Materials (SRMs) are artifacts or chemical compositions that are manufactured according to strict specifications and certified by NIST for one or more chemical or physical properties. NIST SRMs are developed on a continuing basis to meet the measurement and calibration needs of public health and safety, environmental monitoring, U.S. industry, and science and technology.

The Statistical Engineering Division provides technical support to the SRM Program by collaborating directly with laboratory chemists and other scientists engaged in the development and certification of these SRMs. Development of a new SRM typically takes two to five years and encompasses:

- Validation of the measurement method;
- Design of the prototype specimen;
- Verification of statistical control;
- Testing for homogeneity;
- Characterization of the measurement error;
- Design of the production specimen;
- Estimation of the certified value;
- Estimation of the uncertainty for the certified value.

SED statisticians advise on the design and analysis of experiments at all phases; and combine all information to produce a final value and uncertainty.

A summary of SED SRM accomplishments includes the following:

Lab Customers
ITL excepted, all NIST laboratories have scientists working on SRMs who receive SED help. As in years past, CSTL continues to be the heaviest consumer of SED SRM services, with CSTL's Analytical Chemistry Division being the division within NIST using the most SED SRM technical assistance.

Number of SRMs Certified
The trend for SRMs actually certified during the calendar year is again on the increase. In 1999, the number of such certifications was 30; in 2000 it dropped to 20; and in 2001 it increased to well above 30.

SED Staff Participation
Doing SRM certification is a task in which most SED staff members participate. The number of SRMs per staff varies widely from a few per year to many tens per year. Stefan Leigh is our most active SED staff member with SRMs typically in excess of 50 per year.

Gamut of SRMs
The scope of SRMs that benefit from SED assistance cuts broadly across all of the aforementioned 4 general categories (health/safety, environment, industry, and science/technology);

examples of SRMs would include: silicon resistivity, dielectrics, polarized dispersion, surface roughness, SEM magnification, toluene, asbestos, natural gas, urban dust, tita alloys, nickel-chromium film depth, lead in bovine blood, diesel oil, mercury cement, lead paint, Europium spectra, uric acid, optical fibers, human serum, D-glucose, waterway sediment, fish tissue, drugs in hair, mussel tissue, water PCB, ashed bone radionuclides, peanut butter, chocolate, octanol, araclor, trans-fatty acids, anion solutions, spinach, volatile organics, conductivity, and titanium dioxide.

Internal Web Page
Alan Heckert continues to maintain SED's NIST-accessible internal web page displaying the status of division SRMs.

SRM Complexity
The demands of science and technology continue to grow, as do the complexity of SRMs. An increasing number of multi-component SRMs are appearing, some requiring separate certified values (and uncertainties) for as many as 50 to 100 constituents.

Software
Through the joint efforts of Stefan Leigh, Alan Heckert, and Jim Filliben, a Consensus Mean command was inserted into Dataplot. This code embodies a variety of solutions developed over the years (primarily by NIST statisticians: Mandel-Paule, Schiller-Eberhardt, Vangel-Rukhin, BOB (Levenson), etc.) to address the NIST-typical multi-lab/multi-method problems that are so commonplace for certain classes of SRMs. The code was developed to save time and computational effort, and has been successful in that regard.

Bayesian Solutions
Bayesian methods have historically been applied to SRM problems only occasionally and only for those difficult SRMs with an especially complicated modeling structure. In conjunction with the Bayesian Metrology Project, however, a more focused effort has been initiated to systematically apply Bayesian methods across a broader range of SRM problems, and to do so in a routine fashion. Blaza Toman has spearheaded the efforts in this regard.

SRM Accounting
The accounting of SRMs has improved over the last year. The matching and melding of 4 different (and differing) SRM accounting systems was accomplished with the net effect that tracking (across NIST organizational units) the status of individual SRMs is now markedly easier and more accurate. This reconciliation was complicated by NIST issues of SRM funding, by SED divisional issues of budget balancing, and by within-laboratory issues of SRM prioritization. The year has seen an unprecedented openness and agreement among labs, divisions, SRMP, Budget, and SED in determining how to streamline and account the SRM process.

3.5.2 Characterizing Dielectric Materials

Jolene Splett, Kevin Coakley
Statistical Engineering Division, ITL

Mike Janezic, Raian Kaiser
Radio Frequency Technology Division, EEEL

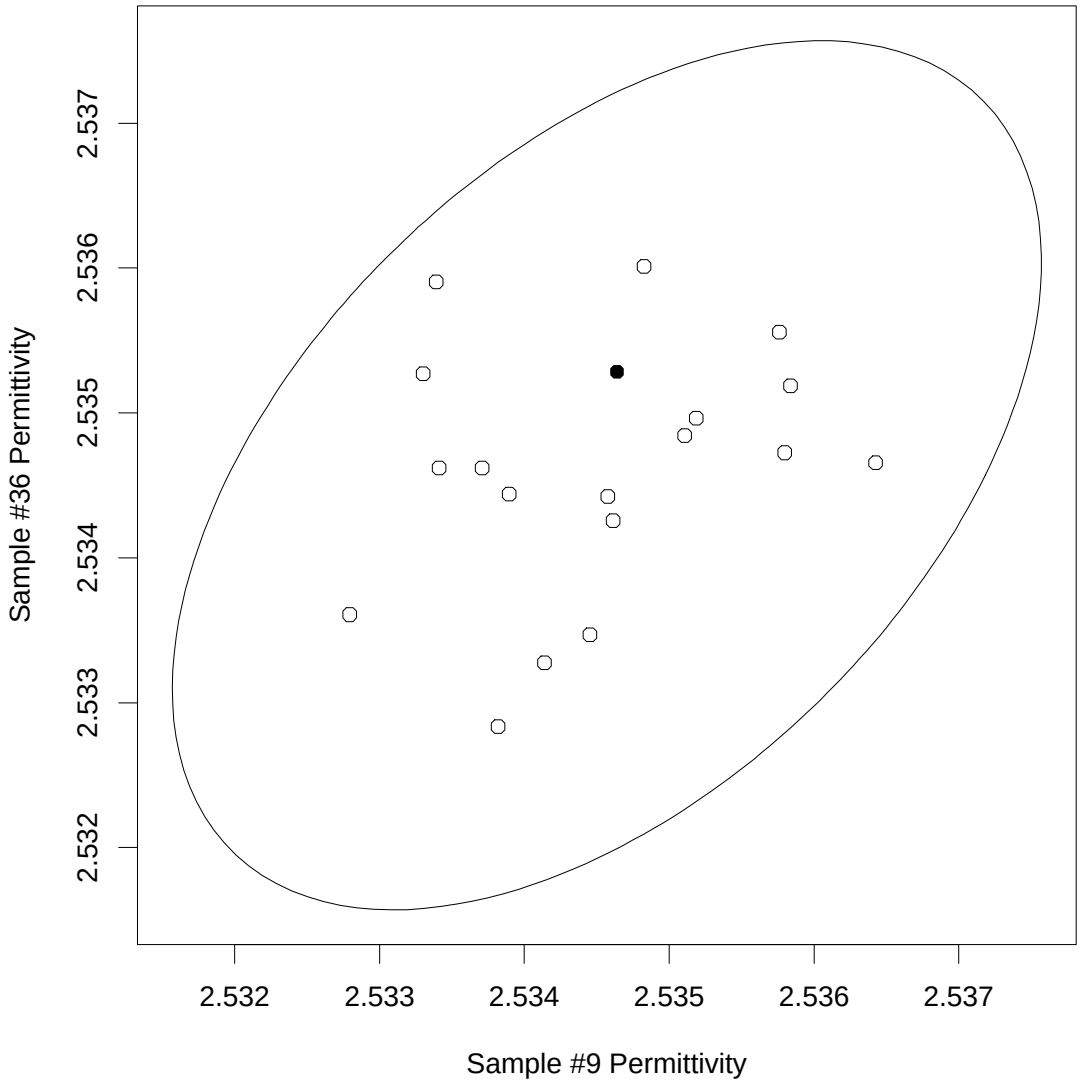


Figure 34: An example of a confidence ellipse to monitor two correlated responses. The circles represent historical data while the dot represents a new observation.

NIST is developing new measurement methods for characterizing dielectric materials with respect to permittivity and loss tangent for the purpose of developing standard reference materials. SED has been collaborating with EEEL staff for several years on this project, and the first cross-linked polystyrene SRMs should be available for sale some time during FY2002.

The main SED contributions to this project in the past year are listed below.

1. Derived analytically the variance function for the additive noise associated with the resonance curve for the special case of equal variances and no covariance between the real and imaginary components of the scattering parameter. The derivation provides justification for the variance function developed earlier.

2. Quantified systematic uncertainties associated with estimates of the cavity length, cavity diameter, and sample thickness using actual data in conjunction with a Monte Carlo program.

3. Designed experiments to demonstrate the stability of the measurement process and analyzed the resulting data. Used the repeatability study data and Monte Carlo study results to develop an uncertainty statement.

4. Developed a measurement assurance program to monitor the behavior of the measurement system over time. The permittivity and loss tangent of two cross-linked polystyrene samples will be measured on the same day. These measurements will be correlated largely due to environmental factors. To monitor both measurements simultaneously, we use a procedure for generating a confidence ellipse. Traditional control charts will also be used to monitor individual samples over time.

5. Drafted a paper describing procedures for estimating the quality factor and resonant frequency that will be submitted to *IEEE Transactions on Microwave Theory and Techniques*.

6. Completed draft documentation for the SRM that will be submitted to the *NIST Journal of Research*.

The electronic, microwave, communication, and aerospace industries have many applications for dielectric materials including: printed circuit boards, substrates, electronic and microwave components, sensor windows, antenna radomes and lenses, and microwave absorbers.

3.5.3 Charpy V-notch Reference Value Uncertainty

Jack Wang, Jolene Splett
Statistical Engineering Division, ITL

Chris McCowan, Tom Siewert, Dan Vigliotti
Materials Reliability Division, MSEL

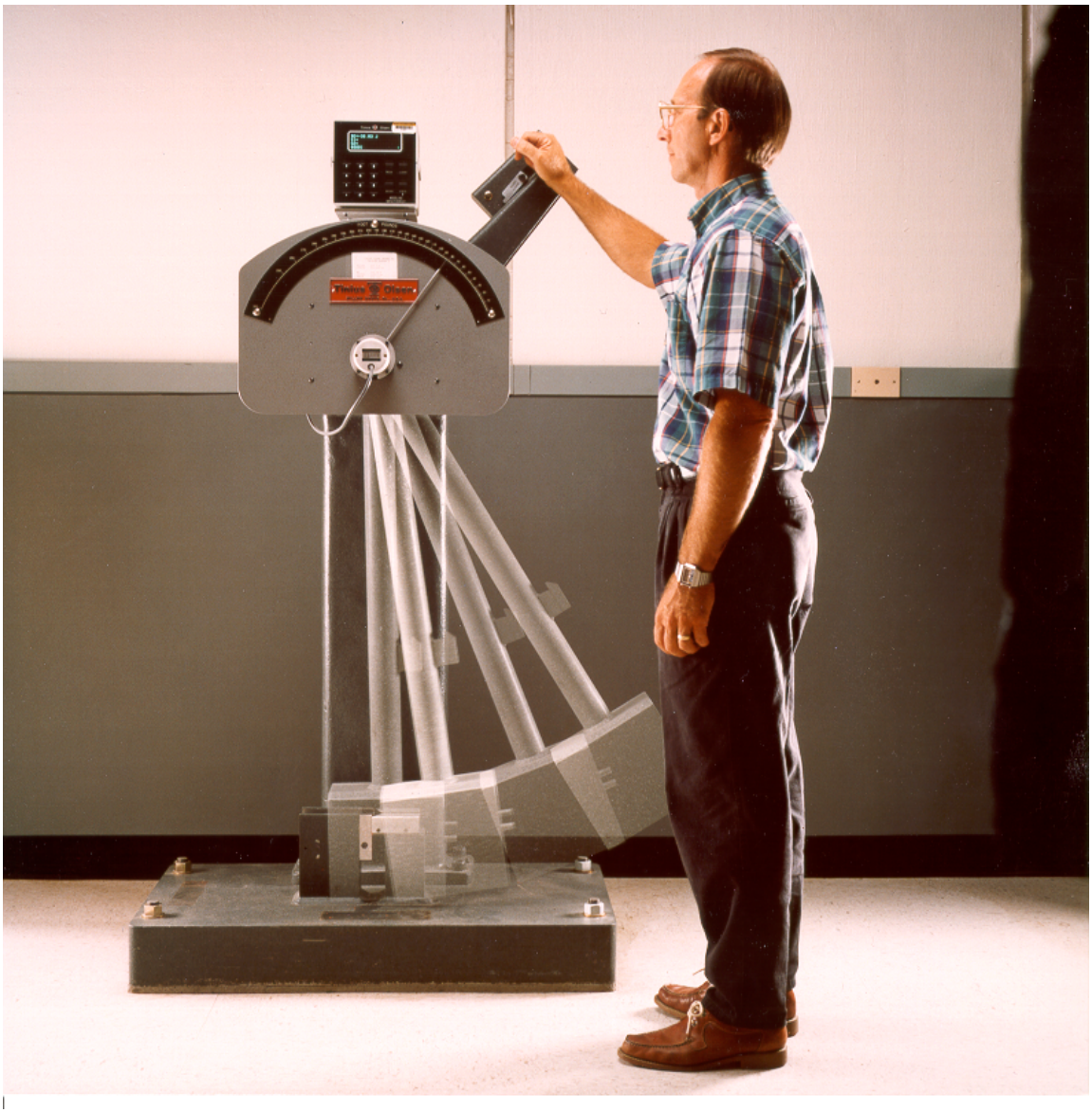


Figure 35: A master Charpy impact machine in action.

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 reference 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. Sets of five specimens are sold to companies that want to certify their Charpy machines.

Basically, the Charpy verification program is conducted in accordance with ASTM Standard E23. However, the standard does not provide guidelines for computing the uncertainty associated with individual test specimens or with the certified value of the reference specimens. SED has established a statistically valid uncertainty statement for the certified value, with degrees of freedom computed using the Satterthwaite approximation.

Future work involves comparing the NIST verification program and uncertainty statements to those of other countries, including Belgium and Japan, and working with MSEL staff to harmonize Charpy programs around the world.

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.

3.5.4 Certification of Moisture in Crude Oil: Reference Standard, RMs 2721,2722. Not a routine Analysis

Charles Hagwood
Statistical Engineering Division, ITL

Samuel Margolis
Analytical Chemistry Division, Div. 839

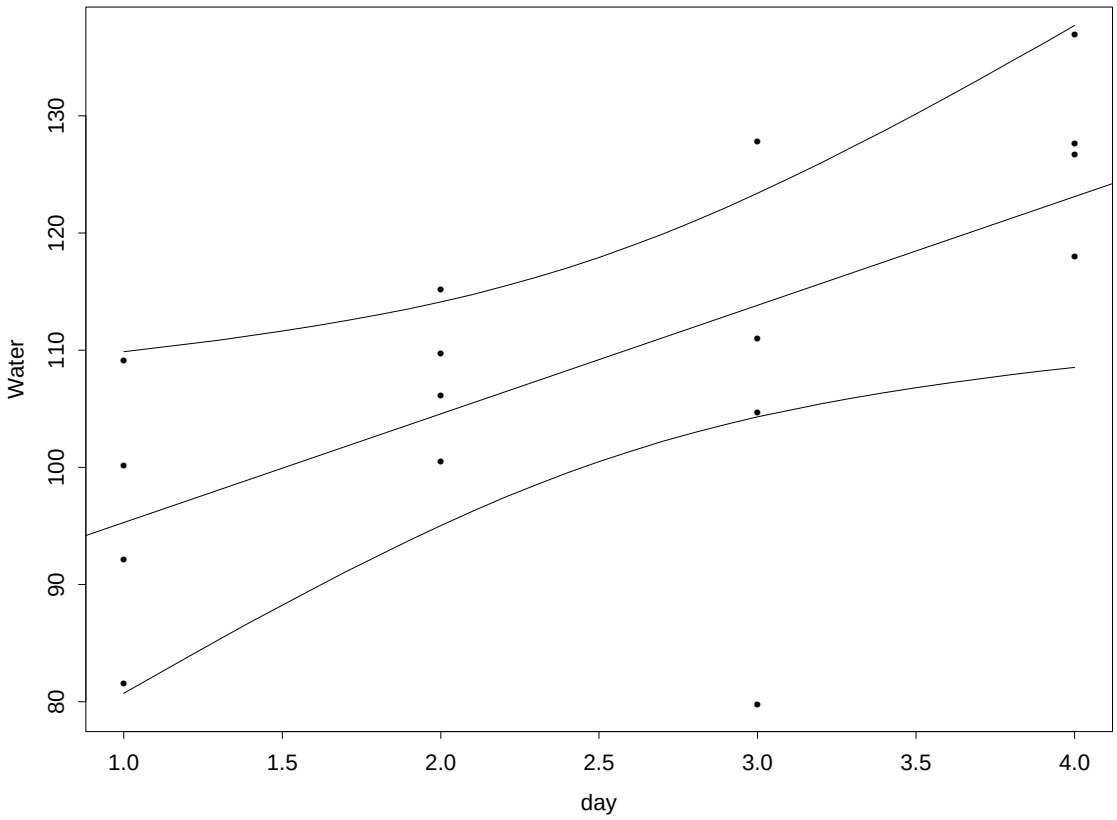


Figure 36: SRM 2722: Simultaneous Confidence Bands for Volumetric Water

Oil pumped from wells contains a certain amount of moisture. Buyers want to certify that the oil they buy contains very little moisture. RMs 2721 and 2722 were produced to help in this certification. The statistical analysis of that certification is discussed. This note illustrates that statistical certification can oftentimes be neither novel nor routine.

A vat of crude oil and water is well-mixed for a length of time and the contents are allowed to stand and separate. The top layer of water-saturated crude oil is decanted into a second container and put in ampoules. The ampoules are sequentially filled and capped according to a determined design. The evaluation of moisture content is based on two titration techniques: volumetric and coulometric. Also, a round-robin study is conducted. Titration, a common analytical chemistry method, is used to determine the unknown concentration of a substance in solution by allowing it to react with a known reagent until a neutralization process occurs. For example, to determine the concentration of a base in a solution, an acid is added until the solution is neutralized. Then the amount of acid added determines the concentration of the base.

The Karl Fisher reagent is the common titrant used in moisture analyses, because it contains reagents that react with water. Because the oil in RMs 2721 and 2722 is crude, there are interfering agents other than water that react with the Karl Fisher reagent, making the laboratory analyses difficult. Thus, the data are not very clean, making the certification non-routine. This is not due to the scientist's carelessness, but to the nature of the material with which one is working. For example, one has to back out the water concentration from the water plus interference measurements minus the interference measurements. Furthermore, some of the nice things one expects from a certification process failed. The data for RM 2721 showed nonhomogeneity and the coulometric measurements showed a day effect. Even with these difficulties the reference material was certified.

The measurement of water in crude oil is an economically important measurement because the oil industry wishes to eliminate charges for water when purchasing tanker lots of crude oil.

3.5.5 Atoms on Demand

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

Jabez McClelland, Shannon Hill
Electron and Optical Physics Division, PL

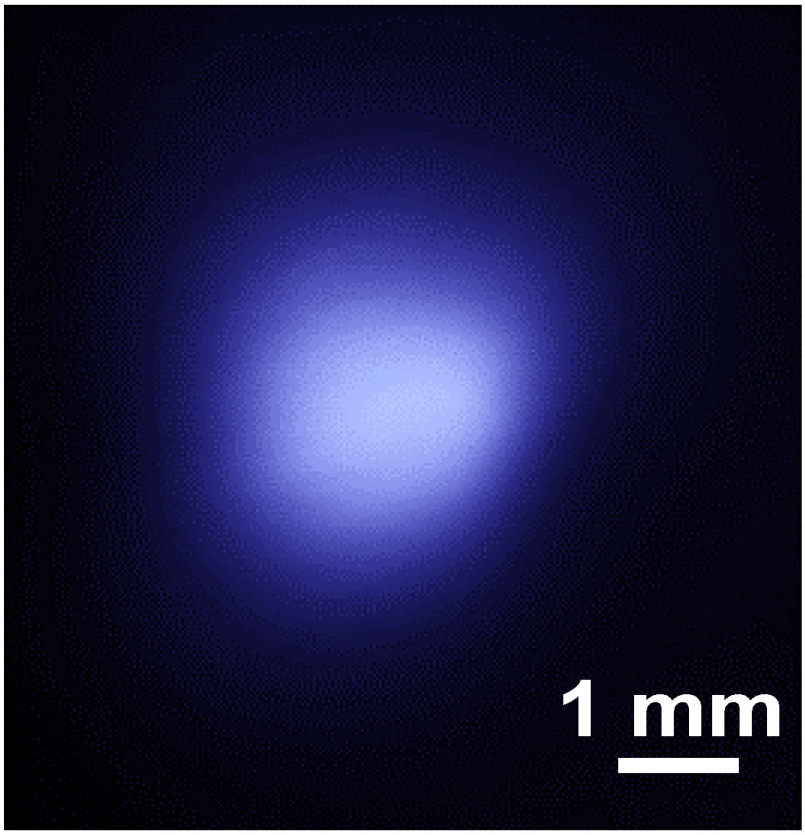


Figure 37: Image of a cloud of ultracold Cr atoms trapped in a magneto-optic trap. This type of trap is used as a source of single atoms in an atom-on-demand implementation.

The Electron and Optical Physics Division's Atom on Demand project seeks to develop tools and actual hardware for working with single atoms in a controlled deterministic way. A first milestone is delivery of a device that can produce and sustain one atom of a given species, upon demand, for a useful time interval. The physicists designing this device constantly update, as their understanding of the underlying problems deepens, simulation codes to describe what is happening in their developmental device. The leader of the team contacted SED to (1) troubleshoot the overall mathematical logic underlying the device for possible problems, and (2) determine whether any of the issues being dealt with via simulation could be handled more simply, in a closed form.

For the problem of isolation of single neutral atoms stored in a magneto-optical trap (MOT), a steady-state distribution was determined. The recursive model for the number of atoms present in the MOT at any given time under the feedback regime was elucidated. The optimal regime for the Poisson loading was found. The values of the loading and loss parameters that maximize the probability of exactly one atom in the trap were found through extensive Monte Carlo simulations. The desirable relationship between loading probabilities and the detection time needed to balance the advantage of longer detection time and the possibility of the MOT changing its content was also found.

If there were no attempt to control the atom population of the MOT using feedback, i.e., opening or closing the door to the trap, then the number of atoms in the trap could be modeled as a simple birth-death process. Suppose the door to the MOT could be opened the instant it became empty or closed at the instant it became occupied. Then, the number of atoms in the trap again could be modeled as a birth-death process. However, the number of atoms in the trap cannot be monitored continuously, and because of physical limitations, the trap can only be checked every T seconds. In that interval T, more than one atom can sneak inside the trap before the door is closed. One can use the conditional probabilities of a Poisson process to modify the birth and death parameters so as to approximate the effect of a non-zero T. That creates a "fraternal twin" birth-death process that has a stationary distribution quite close to that of the real process. Formulas for the stationary distribution of the twin process can then be used to adjust the loading rate of atoms so as to maximize the proportion of time that a single atom is in the trap.

The Division demonstrated fast-turnaround responsiveness in support of a potentially very important project for NIST's Physics Laboratory. Atom on Demand technology will enable novel quantum computation schemes, unprecedented control over dopants in materials, and eventual realization of the ultimate in nanotechnology: atom-by-atom construction.

3.5.6 Function Registration in Materials Research

Walter Liggett
Statistical Engineering Division, ITL

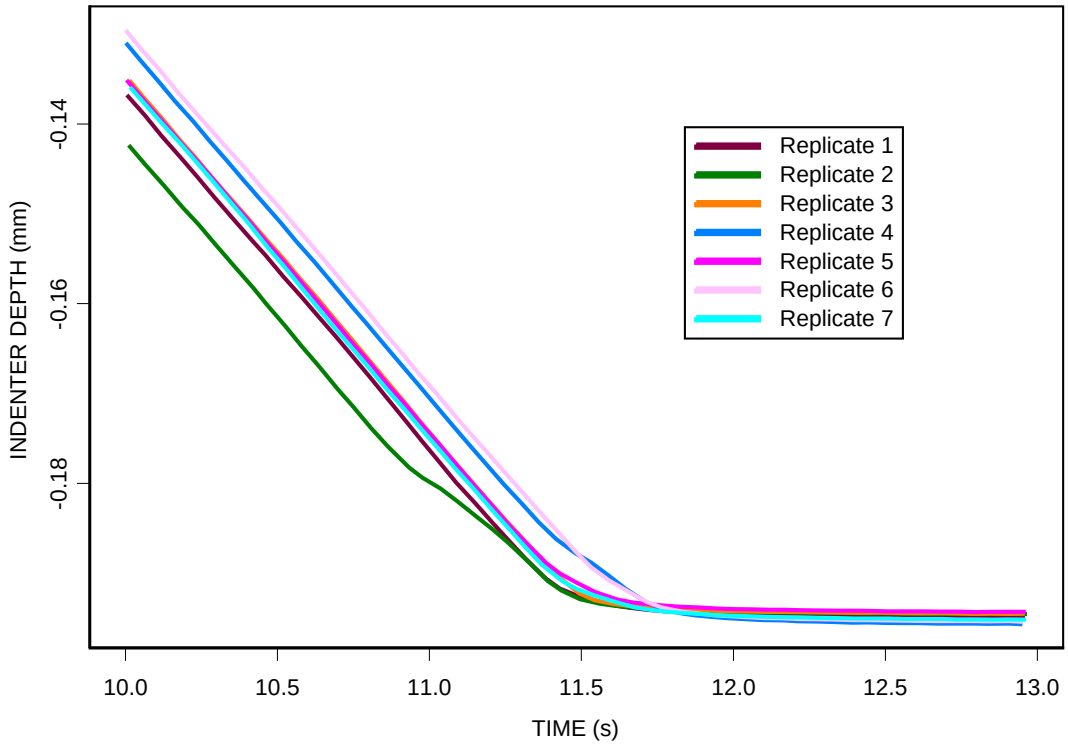


Figure 38: Progress of the indenter into the test block with application of the major load.

Functional measurements that are part of materials research at NIST and elsewhere impose some special requirements on function registration and thus inspire statistics research. Function registration is feature alignment through transformation of the independent variable. The proper approach to this is application dependent.

Functional measurements arise in test methods for materials. Some of these methods consist of probing the material and observing the response. An example is forcing an indenter into the material and observing the resulting depth. The force applied to the indenter as a function of time is specified. Thus, this example of functional data differs from examples in the statistics literature in that the cause of the variation observed is not latent but specified.

Functional measurements also arise in combinatorial methods for material science. In one type of combinatorial method, variation across a specimen is achieved by incorporating gradients in the material processing factors. The resulting variation across the specimen is the functional measurement. If the gradients vary from specimen to specimen, function registration is needed to compare specimens. When the variation is two-dimensional, as it often is, the function registration must be two-dimensional instead of one-dimensional, as in most cases discussed in the literature.

As an example of a functional measurement arising from a test method, consider the time variation in indenter depth caused by force applied as specified for a Rockwell hardness measurement. The specification includes applying and holding the minor load, applying and holding the major load, and releasing the major load, thereby leaving only the minor load. The first figure (above) shows the depth as a function of time during the transition from applying the major load to holding it. Comparison of the seven replicates shown seems a reasonable step in exploratory analysis of these data. Clearly, such comparison requires registering the replicates and centering them.

Because the transition in applied force has a discontinuous first derivative and because registration requires smoothing, selecting a registration method is difficult. One possibility is registration by aligning the force-versus-time observations with each other. Another possibility is registration by aligning each force observation with the nominal load curve consisting of applying the force at a constant rate followed by holding the force at a constant value. The second figure (below) is the result of the first possibility.

Graphing the aligned replicates as deviations from a central value is also important in analyzing these data. Similar to the problem with registration, there is a problem with centering at the average because this requires smoothing, and smoothing the corner in the depth function that is caused by the corner in the load function can create what appears to be variation common to all replicates. Figure 2 avoids this by centering on the nominal depth curve. This figure shows that the depth curves are rounded at the corner and that there is other variation common to the application of the load.

The second figure has no major implications for Rockwell hardness measurement, which uses only two points from the depth curve: the depth before application and the depth

after release of the major load. There is, however, interest in the entire depth curve for different load specifications. Instrumented indentation, which is what this area is called, and combinatorial methods are both just developing. Clearly, each of these areas in material science can benefit from further statistical development.

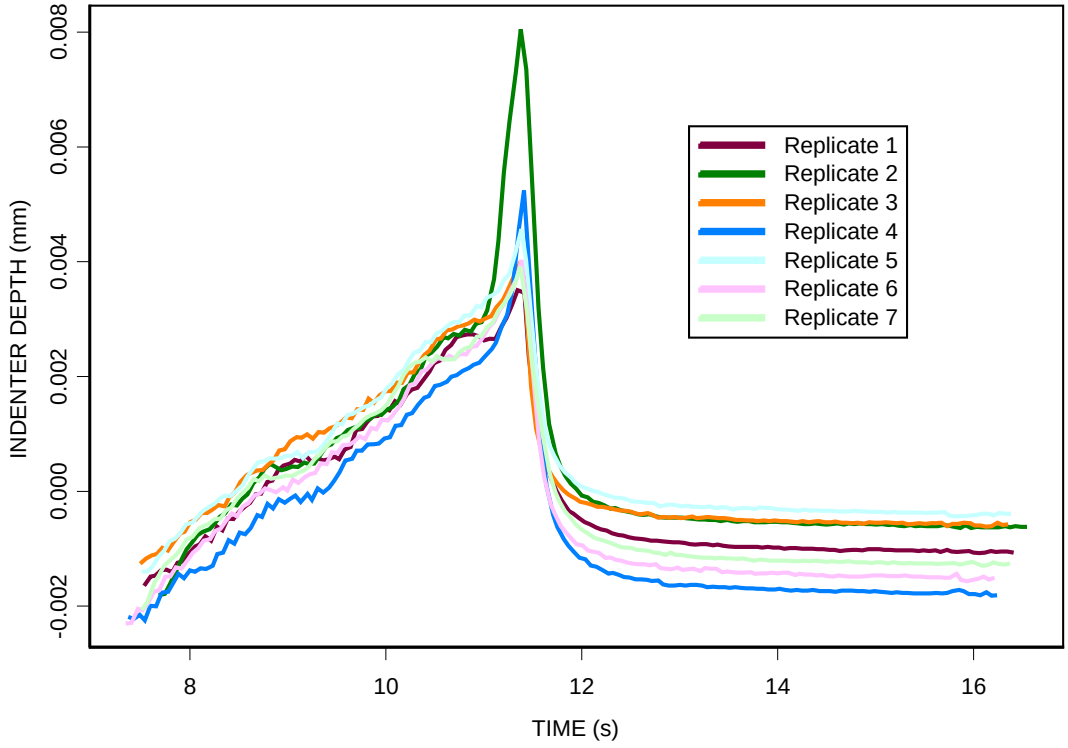


Figure 39: Indenter curves registered by matching replicates and centered on the nominal curve.

3.5.7 Improving Building Codes via PPCC Estimation

James J. Filliben, Alan Heckert
Statistical Engineering Division, ITL

Emil Simiu
Structural Systems and Design Division, BFRL

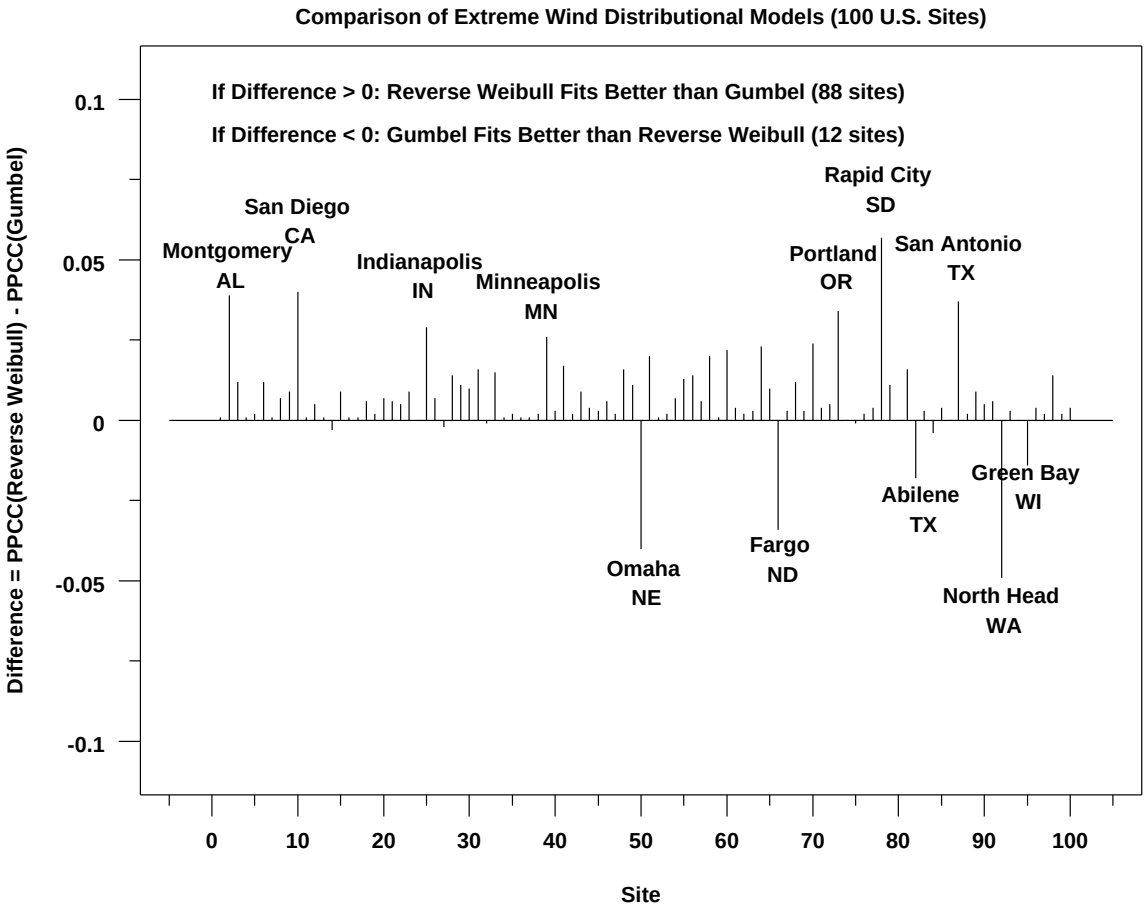


Figure 40: For a given distribution, the PPCC (Probability Plot Correlation Coefficient) is a measure of the linearity of the probability plot, and hence a measure of the distributional goodness of fit. This plot illustrates how a Reverse Weibull fit to the extreme wind speeds is better than a Gumbel fit to the extreme wind pressures– at 88 out of the 100 U.S. sites.

There is a controversy in the building structural safety community as to how to best estimate certain critical percent points of the distribution of wind forces on a building. These percent points correspond to certain accepted industry max-wind recurrence times. The importance of these percent point estimates is that they ultimately convert into updated building code safety values.

It is known that the wind force on a building is proportional to the square of the wind speed. Given that, there have been 2 general approaches in the structural safety literature for estimating building wind forces corresponding to various mean recurrence intervals:

- **Method 1 uses the time series of extreme winds, fits an extreme value distribution to the data, estimates percent points of the extreme winds (corresponding to certain recurrence times), and then squares such percent points to yield corresponding percent points for the building wind forces.**
- **Method 2 starts with a time series of squared wind speeds (= wind pressures), determines a best-fit extreme value distribution to these squares, and then computes percent points of this distribution directly.**

Method 1 has been in use for many years. Method 2 has been proposed and promulgated more recently by certain authors (Cook, Naess). A particular twist that Cook/Naess advocate for method 2 is the use of a specific distribution (namely, the well-known Gumbel distribution) for the fitting of these extreme wind pressures; they claim that this method and distribution yield superior results.

The concern is that these 2 methods yield significantly different percent points and hence would yield significantly different building code values.

NIST BFRL's Emil Simiu (of the Structural Systems and Design Division) has drawn on his 20+ years of experience in this building wind loads arena (in conjunction with long-standing collaborations with several members of SED) to refute the Cook/Naess method. Making use of distributional modeling techniques developed and implemented in SED, Simiu has shown that a method 1 approach is superior to the proposed Cook/Naess method 2 approach.

The crux of the refutation was the application of both methods to 100 representative data sets drawn from around the country, and showing that the Cook/Naess method was consistently inferior.

Simiu demonstrated that he could find and produce a particular distribution that yielded a consistently better fit with the extreme winds data than Cook/Naess's Gumbel method yielded with the corresponding extreme pressures data. In order to achieve this, 3 problems had to be addressed:

- How to measure and compare the quality of 2 distributional fits (especially when the 2 distributions were fit over different units)?
- How to find an optimal-fitting distribution (better than the Gumbel)?
- How to estimate best-fit parameters for this optimal distribution?

The answers to these questions were provided via Simiu/SED collaboration:

- To measure and compare distributions, the probability plot correlation coefficient was used (this unit-independent metric relies on the quantification of probability plot linearity).
- To determine the optimal distribution, liberal use was made of the large collection (70+) of possible distributions in Dataplot. For a variety of reasons, Simiu settled on a (3-parameter) reverse Weibull distribution.
- To determine best-fit distributional parameters, Simiu make use of the Maximum PPCC Criterion (Maximum Probability Plot Correlation Coefficient Criterion) for computing parameter estimates that maximized the linearity of probability plots within the reverse Weibull distributional family–first by shape parameter, and then conditionally for location and scale parameters. This methodology circumvents the computational pitfalls inherent in Maximum Likelihood Estimation when the distributions have a location parameter that serves as a truncation point.

To demonstrate the effectiveness of the approach, it was shown that for the 100 sites analyzed, the reverse Weibull distributional fit to the extreme wind speeds data was overwhelmingly better (88 out of 100 sites) than the Gumbel distributional fit to the extreme pressures data (see plot).

This study comes at a critical time in the structural safety community. Standards committees are considering how to update engineering safety estimates; such estimates will find their way into building codes. This study provides convincing evidence that the proposed codifications based on the Cook/Naess Gumbel method would be inferior, and that further investigations for the most appropriate probabilistic model are in order.

3.5.8 COST: A Web Tool for State Highway Concrete Mixture Optimization

James J. Filliben
Statistical Engineering Division, ITL

Dale Bentz
Building Environment Division, BFRL

Marcia Simon
Special Projects and Engineering Division, Federal Highway Administration

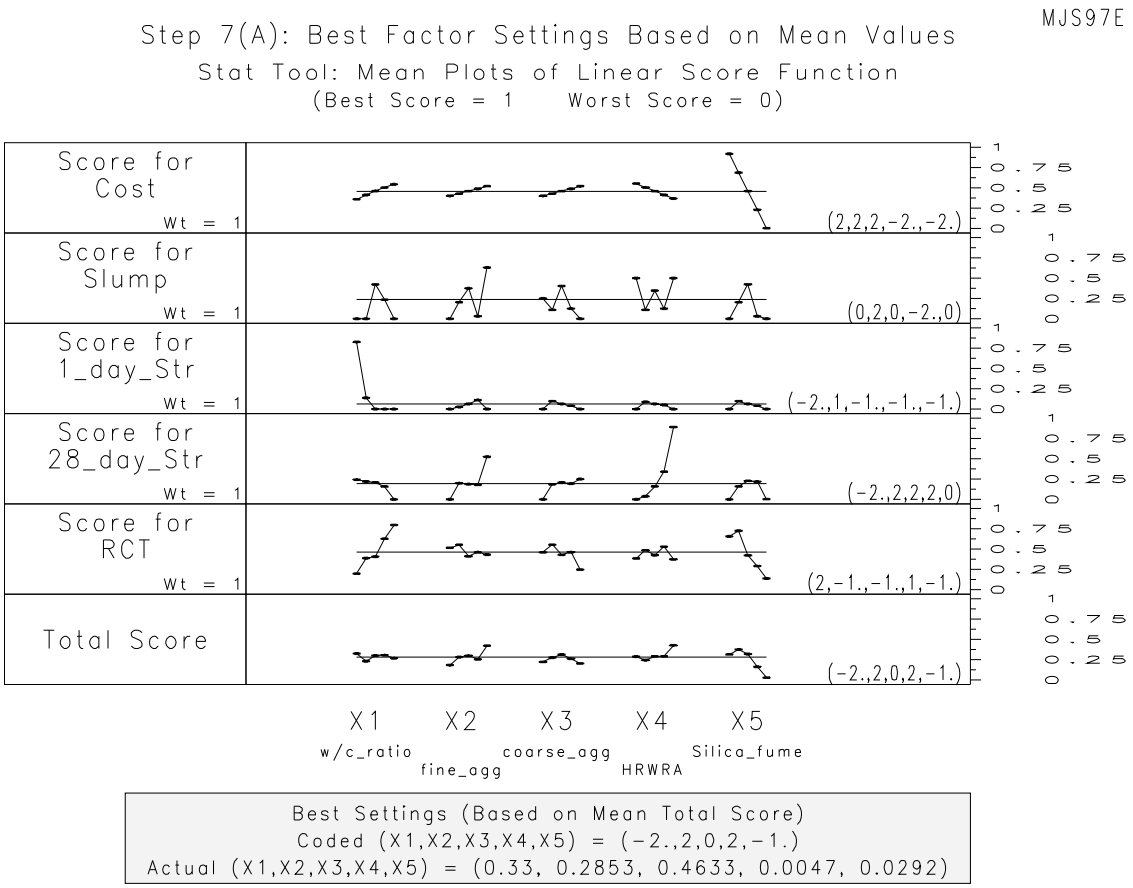


Figure 41: This figure shows the effect of each of the 5 factors (along the horizontal axis) on the scores for each of the 5 responses (cost, slump, 1 day strength, 28 day strength, and RCT) as well as on a total (= summary) score. For each of the 5 raw responses, we see which factor has the dominant effect on the response (for example, for response 1 (cost), the most important factor is X5 (silica fume) while for response 3 (1 day strength), the most important factor is X1 (water/concrete ratio)). We also see the best (average) coded settings for each response (for example, for cost, the best setting for the 5 factors is (X2 = 2, X1 = 2, X3 = 2, X4 = -2, X5 = -2). Finally, for the total score, we see that the most important factor across all responses is X5 (silica fume), and the best (average) coded setting for the 5 factors is (X2 = -2, X1 = 2, X3 = 0, X4 = 2, X5 = -1), while the best setting (in units of the original factors) is (X1 = 0.33, X2 = 0.2853, X3 = 0.4633, X4 = 0.0047, X5 = 0.0292).

On a state-by-state basis, a serious problem exists regarding highway concrete formulation; practical difficulties exist in "designing" the concretes so as to meet all concrete specifications at a minimum cost, or to provide maximum durability within a specified cost range. In the context of highway concrete proportioning, since the raw materials (aggregates, etc.) for highway concrete are usually local (in state), and since such materials vary considerably from one state to another, there exists no universal proportion formula for optimal concrete that will hold across all states. Hence, the fine tuning of optimal concrete must be done on a state-by-state basis.

In its simplest form, portland cement concrete is a mixture of water, portland cement, fine aggregate, and coarse aggregate. Additional components, such as chemical admixtures (air entraining agents, superplasticizers, etc.) and mineral admixtures (coal fly ash, silica fume, blast furnace slag, etc.) may be added to enhance certain properties of the fresh or hardened cement. So-called "high-performance" concrete (HPC) may be required to simultaneously meet multiple performance criteria (e.g., compressive strength, elastic modulus, rapid chloride permeability, etc.). Such concretes typically contain at least 6 components.

As the concrete-community authors Rougeron and Aitcin point out: "... the optimization of the composition of HPC is at present more of an art than a science...." The ACI (American Concrete Institute) 211 publication serves as a useful guideline for proportioning normal concrete mixtures for compressive strengths less than 6000 psi. The ACI 211 falls short, however, for higher-performance / multiple-criteria concretes. As a result, trial and error, "one-factor-at-a- time" experiments are often employed, which are inefficient, costly, and decidedly non-optimal.

To address this problem, a joint project was initiated between the Federal Highway Administration, the Building Materials Division of NIST, and the Statistical Engineering Division of NIST. The primary customer for this project is the research and production engineer at state highway administrations across the United States. The formal goal of the project is to provide an online design/analysis tool to assist concrete producers, engineers, and researchers in determining optimal concrete mixture proportions.

At the core of any data-based optimization project must be the application of appropriate experimental design principles and techniques. Design of experiments has served well in the food, detergent, gasoline, and metal alloys industries, but has seen little application in the concrete industry. Using experimental design as the foundation, the project had 3 operational components:

1. Experimental Design: Introduce the use of efficient, optimal designs for the study of concretes;

2. Data Analysis: Perform an effective battery of data analysis tests on the resulting data.

3. Internet: Provide both of the above to the general public via an internet-accessible link.

The joint FHWA/NIST project was several years in duration. The product of the collaboration is the online design/analysis system entitled COST (Concrete Optimization Software System). COST is accessible from both the FHWA and NIST web sites. The NIST website is http://ciks.cbt.nist.gov/cost/. COST is free of charge. The battery of data analysis procedures available to be applied to the data are extensive. The analysis software engine that underlies COST is NIST/SED's Dataplot.

Using COST, the concrete producer, engineer, or researcher determines

1. important mixing factors affecting individual or multiple criteria; and
2. optimal concrete proportions across individual or multiple criteria.

Regarding operational details, COST prompts the user for information about preferred performance variables, mixture components, optimization criteria, and other engineering specs. COST then automatically generates a set of trial batches (an experimental plan) to be prepared and evaluated. The engineer then carries out the specified experiment, recording the performance characteristics of interest. After re-entry of the performance data back into the system, COST applies an extensive (10-step) battery of data analysis procedures (both graphical and quantitative). Based on the user-specified performance criteria, COST then produces summary information specifying critical factors and optimum mixture proportions.

The web site was activated in May, 2001. To date (January, 2002) there have been approximately 31,000 hits (about 1000 hits per week). More modestly, the number of experiments generated has been 130, with 11 complete analyses. The site has been mentioned been mentioned in several of the trade magazines.

It is suspected that the net impact of the site will be cumulative and the maximal impact will take a few years. Even with the discrepancy between the 31,000 hits and the 130 experiments, there is still a significant value added in that the site visitor sees the structured design approach taken, and this plays a seminal role in introducing to the concrete community the fact that a fundamentally better method for doing concrete optimization exists beyond the historical hit-and-miss, one-factor-at-a-time approach.

3.5.9 Recommended Default Values for Cyclic Degradation in the DOE Air Conditioners and Heat Pumps Test Procedures

Ivelisse Aviles, James J. Filliben
Statistical Engineering Division, ITL

Brian Dougherty
Building Environment Division, BFRL

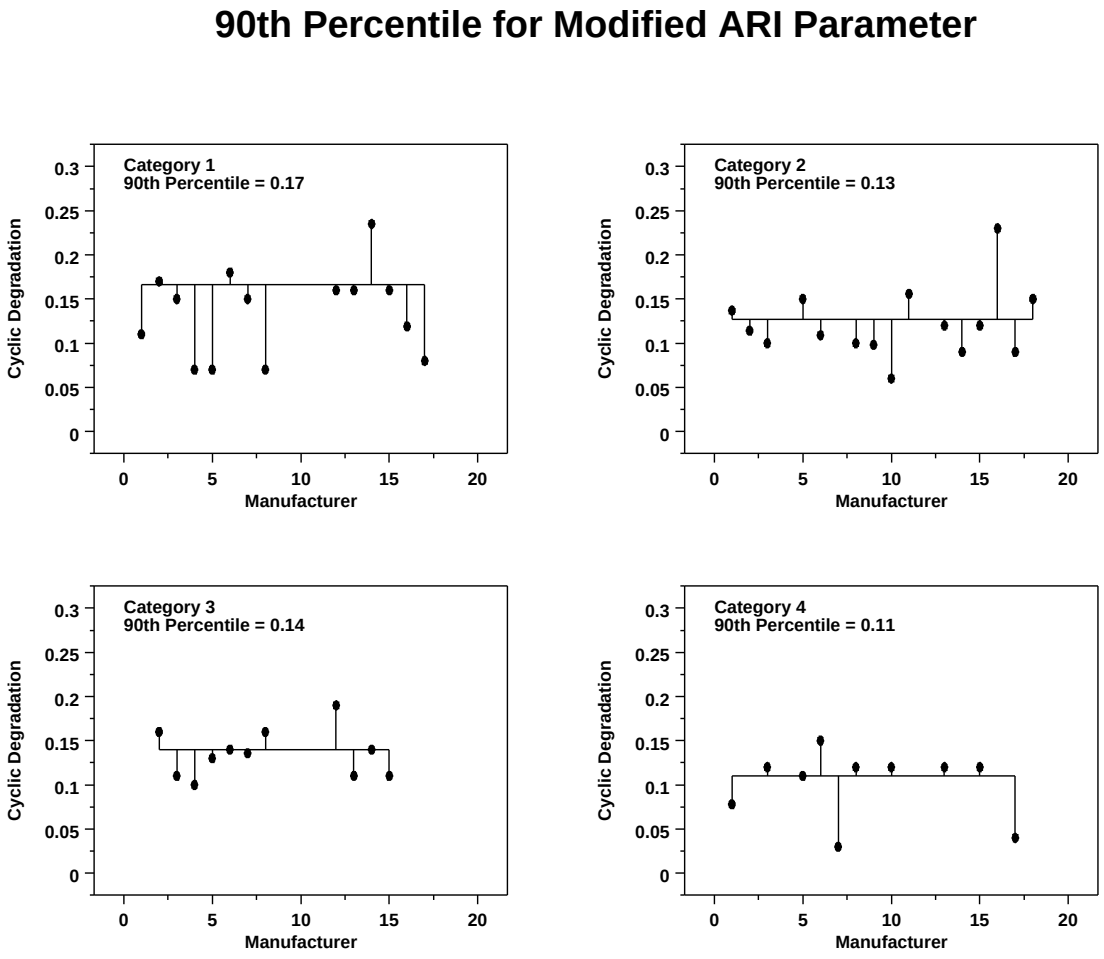


Figure 42: Based on certain hardware features of air conditioners and heat pumps, the data were divided into four categories. For each category, the 90th percentile for the cyclic degradation data is the recommended default value for C_d. Most manufacturers will benefit with the recommended defaults. Note that Manufacturers 11 and 18 can eliminate testing, Manufacturers 4, 7, 9, and 17 might choose to test because their equipment can easily obtain a lower C_d.

An energy test procedure is the technical foundation for all energy efficiency standards, energy labels, and other related programs. It provides manufacturers, regulatory authorities, and consumers a way of consistently evaluating energy use and savings across different appliance models.

Air conditioners and heat pumps provide different services but operate on the same principle; accordingly, the test procedures are very similar. The fundamental measure of efficiency for these devices is the ratio of heat delivered (or extracted) divided by the electrical input energy. For heating performance, the ratio is called a *Coefficient of Performance* (COP), and for cooling performance the ratio is called an *Energy Efficiency Ratio* (EER).

Seasonal performance of air conditioners and heat pumps does not depend only on the unit's steady-state COP or EER. The steady-state efficiency test does not necessarily reflect true annual energy consumption because there is significant energy loss and/or inefficiencies during the start-up and shut-down periods. The unit's performance during the transient periods at the beginning and end of an on-cycle and the parasitic electrical energy used during the off-cycle also affect seasonal performance.

North American standards have made estimating seasonal energy consumption the primary goal of its test procedures. In the United States, estimating seasonal energy consumption began mainly because of energy legislation passed in the mid-1970s that required test procedures to provide an estimate of annual operating cost and/or seasonal efficiency. As a result, the United States Department of Energy (DOE) developed a procedure that includes both steady-state and cycling tests coupled with a calculation procedure that accounts for changing weather conditions throughout the seasons. The DOE test procedure for air conditioners and heat pumps has included a cooling mode cycle test and a heating mode cycle test. A measure of the degradation associated with cyclic operation is gained by using the results of these cyclic tests plus the results from the steady-state tests conducted under the same test conditions. The *degradation coefficient* (C_d) is the measure of the efficiency loss due to the cycling of the units. The goal is make C_d as low as possible.

The C_d can be determined by conducting two lab tests. In lieu of testing, the manufacturer may choose to use a default value. Presently, a single default value ($C_d = 0.25$) is provided in the DOE's AC/HP test procedures. Most manufacturers choose to test because their equipment can easily obtain a lower C_d. It is in the best interests of the industry, DOE, and the nation to lower the default C_d specified in the DOE test procedures. The current default was set over two decades ago and the technology has advanced so drastically that the default value is now virtually useless. NIST was tasked with recommending new default values for C_d.

Air-Conditioning and Refrigeration Institute (ARI), the manufacturer's trade association, provided NIST with data from 339 units that were tested in 2000 as part of the AC/HP industry's certification program. All the testing was performed by one independent laboratory (ITS). The data were analyzed considering unit type, seasonal performance, capacity, compressor type, expansion hardware, and cyclic enhancement devices. Analysis of the data demonstrated that unit type, compressor type and AC/HP type all had minor differences in C_d values. The major effects on C_d were expansion and enhancement devices.

Based on certain hardware features of AC/HP (i.e., type of expansion device and whether it has an indoor fan delay), the data were divided into four categories. Category 1 has equalizing expansion and no indoor fan delay; Category 2 has equalizing expansion and indoor fan delay; Category 3 has non-equalizing expansion and no indoor fan delay, and; Category 4 has non-equalizing expansion and indoor fan delay. For each category, the 90th percentile for the cyclic degradation data is the recommended default value for C_d. These defaults are 0.17, 0.13, 0.14, and 0.11 for Categories 1, 2, 3, and 4, respectively. Thus, four new default values were been recommended to replace the current single value for cyclic degradation ($C_d = 0.25$) in DOE's AC/HP test procedures.

Test procedures provide a means for manufacturers, regulatory authorities, and consumers to compare the energy consumption of different models of appliances in a consistent manner. Given this diverse range of users, it is no surprise that test procedures are compromises between representing realistic usage patterns and performing measurements that are reliable and cost-effective. Manufacturers seek to make the testing requirements simple and inexpensive and want to ensure that their products appear as energy efficient as possible, while consumers want the test procedures to reflect results that are as realistic as possible.

With the recommended default values, it is possible to eliminate about 10% of the DOE testing. This reduction in cycling tests represents a decrease in test burden for the industry. In addition, the time and the money saved with fewer tests will help industry bring new products to the market at a faster rate. Since there is a need to promote less testing but discourage excessive complaints at the same time, more information on data collection needs to be disclosed in order to recommend smaller defaults. Also, continuing technological innovations will undermine the ability of existing procedures to reasonably represent an appliance's energy efficiency and thus default values should be recalculated at least every decade.

4. Special Programs

4.1 Web Products

4.1.1 NIST/SEMATECH Engineering Statistics Internet Handbook

Carroll Croarkin, Will Guthrie, Alan Heckert, James Filliben
Statistical Engineering Division, ITL

Paul Tobias, Jack Prins, Chelli Zey
SEMATECH

Barry Hembree
AMD

Ledi Trutna
Consultant to Motorola University

The Statistical Engineering Division is completing a joint (5-year) project with the Statistical Methods Group of SEMATECH (the semiconductor consortium based in Austin, Texas) in the development and publication of a hyperlinked web-based handbook on statistical methods for scientists and engineers. This electronic handbook was inspired and patterned after NBS handbook 91, Experimental Statistics (1963) by SED's Mary Natrella. This updated handbook is intended to provide modern statistical and graphical techniques that are appropriate for the problems confronting U.S. industry, particularly the semiconductor industry, and the NIST laboratories. Among the many unique aspects of the handbook is its ability to access statistical software from within its pages to analyze built-in case studies or to apply the case study analyses to other data sets.

The handbook is extensive–about 3000 pages in length. It has 8 chapters:

1. Explore (exploratory data analysis)
2. Measure (Measurement Process Characterization)
3. Characterize (Manufacturing Process Characterization)
4. Model (Regression)
5. Improve (Experimental Design)
6. Monitor (Process Monitoring)
7. Compare (Process/Product Comparison)
8. Reliability (Product and System Reliability)

Many of the chapters follow a common structure: 1) introduction; 2) assumptions; 3) experimental design; 4: analysis; and 5) case studies. In addition, a preliminary Tools and Aids section provides many features that allow the inquiring analyst to considerably short-circuit the process of finding a specific answer (to a specific question) among the 3000 pages.

Engineering Questions Without/With Flowcharts: A collection of common engineering questions provide links directly to the relevant sections within the handbook. Topics here include model adequacy, distributional fitting, outlier detection, propagation of error, assessing important factors, etc. To assist those who learn "top-down", an additional collection of questions has been assembled with answers given in the form of flow charts. Topics here include reliability, regression, statistical control, etc.

Gallery of Graphical Techniques: A gallery of 40+ graphical procedures has been included, which cover a broad range of exploratory data analysis tools. This collection is useful if the analyst has a particular graphical tool in mind and wants to link immediately to details of how the tool is defined, created, and interpreted.

Gallery of Quantitative Techniques: In a similar fashion, this collection of 40+ quantitative procedures allows for the quick linking to details of such numerical techniques.

Gallery of Case Studies: A collection of 18 case studies is provided to assist those users who prefer to "learn by doing". This collection links to the detailed case studies at the end of each of the 8 chapters.

Gallery of Probability Distributions: A collection of 18 distributions/distributional-families have been provided. Equations and plots for standard form probability density functions, cumulative distribution functions, percent point functions, hazard functions, and survival functions are given, along with formulas for common summary statistics.

Glossary: An extensive glossary of statistical terms has been integrated into the handbook to provide immediate links/definitions to terms as they arise.

Statistical Software: A unique aspect of the handbook is that it allows the reader to run analyses of case study data– or his or her own data–directly from within the handbook. This is accomplished via the integration of the handbook with NIST's own public domain software, Dataplot, which was developed and has been maintained by SED's Jim Filliben and Alan Heckert. In addition, the handbook has built-in accessability "hooks" which encourage case study linking with commercial software. Several software houses have demonstrated interest in this (Mathsoft, Mathpoint, Minitab, JMP, etc.), and with the widespread positive acceptance of the handbook, this software extensibility will no doubt continue to grow.

Course Builder: To assist the educator and trainer in both academia and industry, a Java-based "course-builder" utility allows for the custom building of an individualized course from user-selected handbook pages.

Editing: Over the last year, the handbook has had the benefit of an outstanding "cover to cover" editing by Tom Ryan. Tom has contributed invaluably to both the technical content and the linquistic integrity of the entire document.

Conference Talks: Handbook-dedicated sessions have occurred at several conferences in recent years. The latest such session was at the Quality and Productivity Conference (May 2001) with a well-received session entitled "A Brief Tour of the NIST/SEMATECH Engineering Statistics Handbook".

Public Release: A beta version of the handbook has been available to the public since September, 1999; the site URL is http://www.itl.nist.gov/div898/handbook/index.html. Public release of the web pages (and CD ROM) is scheduled for the Spring.

The handbook has already received thousands of hits–and has been extremely well received (nationally and internationally), and across all sectors: industry, academia, government, and R/D. The feedback has been virtually all positive–some very positive (and simultaneously very pointed):

> *"I wanted to congratulate you on your SUPERB site ... your site is an excellent tutorial (all chapters are very good and the examples are well presented). I hope many visit your site and stop wasting hundreds of thousands of dollars on these Six Sigma seminars" [...]*
>
> *"This is great! I'm a biologist (with an engineering physics BA), and have been looking for a way to go from data to model rather than the other way 'round ..." [academia]*
>
> *"... this is an excellent resourse/site! dataplot is great too! many will benefit from the efforts of all those contributing to the above." [IT industry]*
>
> *"The handbook is a great tool for looking up reliability information. I believe the work is very complete and well written." [telecomm industry]*
>
> *"I have found the level of technical information outstanding at your website ... The depth and scope of the content is very valuable for my product development and QA needs" [U.S. industry]*
>
> *"I just discovered your online handbook and so far it looks like a fantastic thing. Thank you for building it." [academia]*
>
> *"Thank you for generating the Engineering Statistics Handbook on Line (sic). My dad gave me a set of statistics books developed by the US Army years ago that had really clear concise explanations of the theory, practice, and interpretation ... Your handbook is even better than what I had." [chemical industry]*
>
> *"... thank you for an invaluable on-line book! A few weeks ago I did not know much about reliability testing and I now feel so much more confident." [semiconductor industry]*
>
> *"Thank you very much for the Handbook! It became immediately my most favorite link." [chemical industry]*
>
> *"I am a Chemical Engineering professor and have just recently discovered the handbook. I must congratulate the authors on assembling a really useful collection of knowledge and structuring it to take maximum advantage of hyperlinking and WWW presentation." [academia]*
>
> *"Your "ENGINEERING STATISTICS HANDBOOK" is an excellent resource enriched with numinous (sic) statistical tools that are very practical in engineering. It helps us a lot lot (sic)!!!" (U.S. industry)*
>
> *"I am an engineer/researcher working at ... on a robotics research project for which I wanted to do some reliability calculations. Section 8 of the Handbook had just what I needed" [IT industry]*
>
> *"Superb handbook–One of the best I've seen." [U.S. industry]*

"... I really think you have a great site, its very useful for students like me." [academia]

"I just discovered your online stats handbook. I'm quite impressed! In only a few minutes, I was able to find a wealth of information on testing for non-randomness. The prose is engineer-friendly (not too much math) and the hyperlinks make it easy to figure something out when a term or two is holding back my understanding" [academia]

"Your Handbook is the most complete document I have ever seen on the Web. Additionally, it is written in an engineering language that makes it very easy to understand and assimilate ... Congratulations to all the people who worked in this marvelous project ... I would like to use the translation of some themes of your handbook in the software we are developing." [software company]

"Love your site! This is a great reference tool. Good content and layout." [oil industry]

"The handbook is fantastic. It is the first place to which I turn when I have statistics questions, particularly in EDA" [IT industry]

"... Great Job!!! I support several hundred scientists, engineers, and managers here. I've been looking for a good reference on basic statistics that emphasizes graphical, exploratory methods. I believe I've finally found it!" [U.S. industry]

4.1.2 A 10-Step EDA Procedure for the Analysis of 2-Level Factorial Designs

James J. Filliben
Statistical Engineering Division, ITL

Step 4

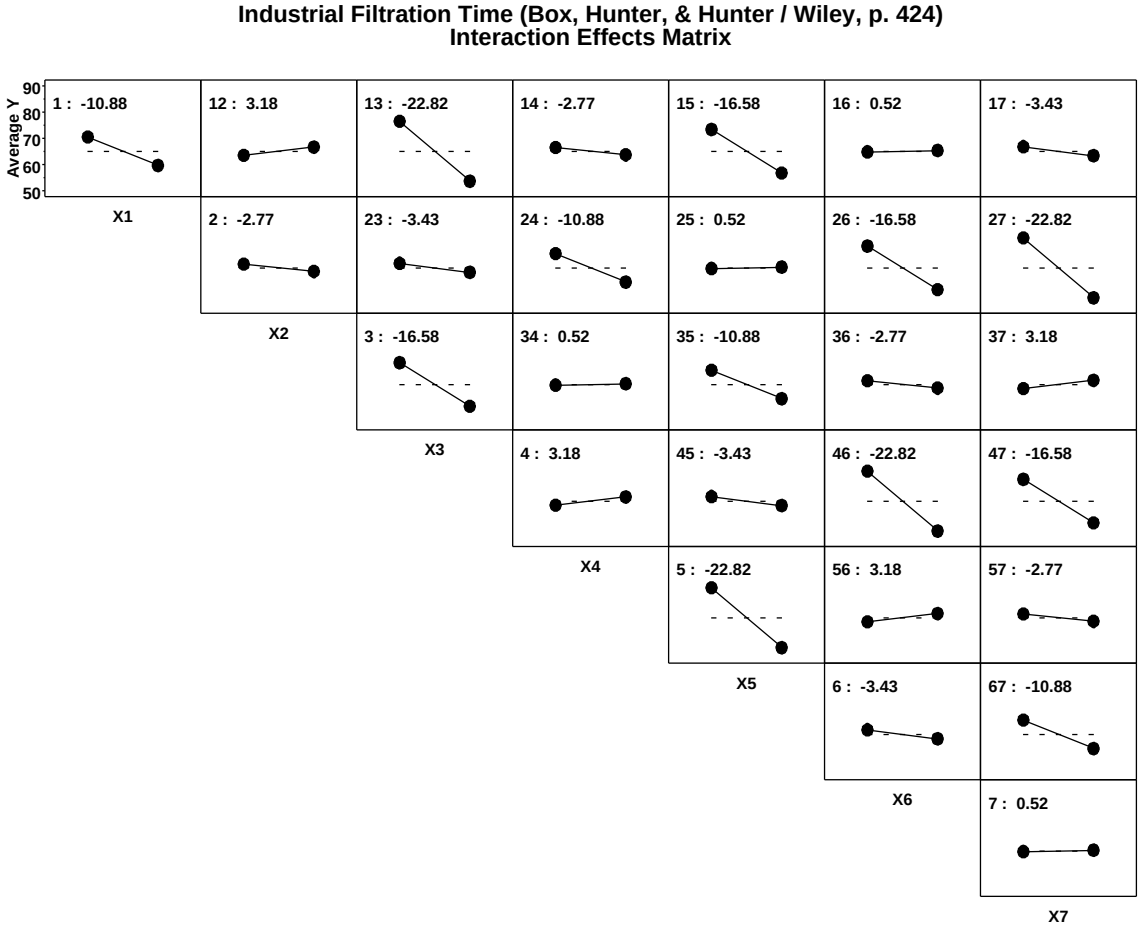


Figure 1: This is step 4 (interaction effects matrix plot) of the 10-step analysis sequence. From the plot we note that the most important factors are factor 5 (with an effect of -22.82 units), factor 3 (effect = -16.58 units), a factor 1 (-10.88 units). Just as importantly, we note that factor 5 is confounded with 3 interactions: X1*X3, X2*X7, and X4*X6; factor 3 is also confounded with 3 interactions: X1*X5, X2*X6, and X4*X7; and factor 1 is confounded with the 3 interactions X2*X4, X3*X5, and X6*X7. Finally, a first pass at best settings (to minimize filtration time) for the 7 factors are (+,.,+,.,+,.,.) where "." indicates the factor setting is unimportant.

One-factor-at-a-time experimentation still has firm entrenchments in the daily world of research experimentation in industry, science, and engineering. The attraction of such designs is due to the undeniable fact that they are conceptually, logistically, and analytically simple. Unfortunately, in the omnipresent world of interactions (which are existent in almost all physical and chemical environments), such designs yield estimates that are grossly incorrect and conclusions that are grossly invalid.

A growing positive trend over the last decade has been the increasing acceptance of the design-side and analysis-side virtues of orthogonal experimental designs in general, and 2-level orthogonal designs in particular. The marked superiority of these highly efficient full and fractional factorial designs is due to the fact that they yield unparalleled insight into the phenomenon under study, while providing estimates and conclusions which are valid, defensible, and reproducible.

The advantages of these 2-level designs are multifold:

- unparalleled insight
- efficient in number of runs/time/$
- able to estimate interactions
- balanced in main effects and 2-factor interactions
- yield "perfect fit" models
- yield optimal test statistics
- are easily augmentable
- yield best settings
- yield ranked list of factors
- serves as a basis for global optimization
- screens out / identifies important factors

With the increased usage of these designs, there comes the challenge of how to efficiently analyze them. The standard least squares estimates for factor effects in these designs are (due to orthogonality) computationally trivial, but there is much more insight that can be extracted from these designs, above and beyond a list of estimates with uncertainties along with a series of F-test tail probabilities. Such insight invariably flows from "knowing" what is in the data rather than "believing" what is in the data, and this quest for knowledge places us squarely in the domain of the proven Exploratory Data Analysis (EDA) foundation of graphics.

In this regard, a 10-step EDA battery of graphical tests has been developed over the last 12 years to facilitate the analysis of these increasingly-popular 2-level designs. These 10 graphical techniques have been gleaned from a larger collection of tools. The techniques deliver to the research analyst clear and convincing answers to the following questions:

- What are the most important factors?
- What are the most important interactions?
- What are estimated effects for the factors and interactions?
- What is the optimal setting for the system?
- What is a good, parsimonious model?
- Where should we run our next experiment?
- What confounding structure exists for the factors?

The 10 graphical techniques are

1. Ordered Data Plot
2. DEX Scatter Plot;
3. Main Effects Plot;
4. Interaction Effects Matrix Plot;
5. Block (Robustness) Plot;
6. DEX Youden Plot;
7. Absolute Effects Plot;
8. Halfnormal Probability Plot;
9. Cumulative Residual SD Plot;
10. Contour Plot;

Several of these techniques are new to the statistics community; three, in particular, are of note:

- Interaction Effects Matrix Plot;
- Block (Robustness) Plot;
- DEX Youden Plot;

To facilitate the use of this 10-step EDA methodology, the following two dissemination actions have been put into place:

- Web: A 13-page write-up of the approach has been included in Chapter 5 (DEX) of the recently completed web-based NIST/SEMATECH Engineering Statistics Internet Handbook. This 3000-page web site is at:

  http://www.itl.nist.gov/div898/handbook/

- Software: An 800-line macro (DEXPLOT.DP) has been written to completely automate the 10-step procedure. This macro is part of NIST/SED's Dataplot software–the free, public-domain graphics and analysis NIST/SED package, freely available and downloadable from

  http://www.itl.nist.gov/div898/software/dataplot.html

Operationally, the analyst enters the data (in Yates order) and provides a general project title, and the software automatically produces the 10 pages of plots–with automatic scaling, annotation, estimation, factor ranking, and derived confounding structure. The software has been extensively tested.

Given the data from a 2-level, full or fractional factorial design, the analysis literally reduces to a 5-minute operation. Since Dataplot receives about 100 hits per day (about 30,000 hits per year), the inclusion of the DEXPLOT.DP macro has the potential for wide usage.

The development of the 10-step procedure and its implementation/dissemination in the NIST/SEMATECH Handbook and Dataplot will serve to routinely and efficiently enhance the insight from the analysis of 2-level full and fractional factorial designs, and also significantly improve the ability of the analyst to determine important factors, important interactions, best settings, and a good prediction equation.

For the experimentalists who take advantage of the orthogonal 2-level designs, the net result will be improved scientific/engineering/industrial experimental research conclusions both within and outside of NIST.

4.1.3 Redesign of the SED Web Pages

Alan Heckert, Kiamkia Moore, Nell Sedransk
Statistical Engineering Division, ITL

YC Chang, James Graham
Information Services and Computing Division, ITL

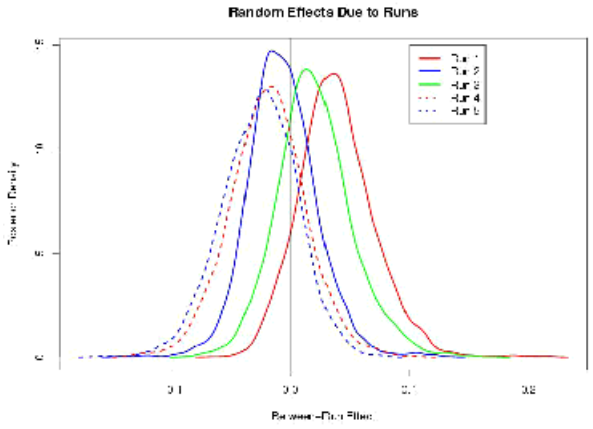
Statistical Engineering Division
Bayesian Metrology
NIST National Institute of Standards and Technology

Home | SED Staff | Programs | Products & Publications | Search

Bayesian Metrology Project

The NIST Statistical Engineering Division is developing the use of Bayesian statistical methods to solve NIST metrology problems as a NIST Competence Project.

- Overview
- Impetus
- Customers
- Goals
- Milestones
- Achievements
- Future Events/Activities
- Data Gallery
- Contacts/Personnel
- Additional information



The above graph contains a link to a Bayesian analysis of the Michelson light data (in PDF format).

If you cannot view PDF format files from your browser, the Adobe Acrobat Reader can be freely downloaded. You may also contact the SED webmaster to request a printed copy of the document.

Date created: 8/28/2001
Last updated: 9/12/2001
Please email comments on this WWW page to sedwww@cam.nist.gov.

[SED Home | Next]

Figure 2: Initial page for Bayesian web pages.

The SED Web pages were originally designed in 1995. Although SED has implemented several successful Web products (e.g., the StRD project and the NIST/SEMATECH Handbook), the SED Web pages needed to be redesigned. In addition to modernizing the appearance, the long-term goal is to more effectively utilize the Web in serving the SED mission. In particular,

1. Promote the work of SED to NIST scientists and management. One example is the project Web pages discussed below.
2. Promote the work of SED outside NIST. For example, the Web pages could help in recruiting potential job applicants.
3. Assist the educational/training role of SED. The Handbook is one obvious example of training via the Web.

The redesign of the SED Web pages can be divided into the following subtasks:

1. Conformance to NIST/ITL and accessibility standards
2. Project Web pages
3. Database-driven content

Conformance to NIST/ITL and accessibilty standards

In the past, NIST divisions had great freedom in designing Web pages. Although many excellent Web pages were implemented, there was no consistency between the Web pages of different divisions. To address this issue, NIST assigned a team to develop standard templates for NIST Web pages, which were initially implemented on the main NIST and ITL Web pages. SED delayed the Web page redesign until it became clear what the NIST/ITL standards were going to be for division Web pages. Alan Heckert and Kaimkia Moore implemented the new NIST/ITL design standards for all current SED Web pages (with the exception of the StRD and Handbook pages), using a combination of Perl scripts and manual editing. This redesign gives the SED Web pages a more modern appearance.

Federal law now requires that government Web pages conform to accessibility standards. For SED, this primarily affected the StRD and Handbook pages. Kaimkia Moore, a summer student, implemented most of these changes for the StRD pages. The edits for the Handbook pages were performed by various members of the Handbook team.

Project Web pages

SED previously provided summary Web pages for major projects. An effort is under way to develop more extensive Web pages for the major SED projects.

Extensive Web pages were developed for the Bayesian metrology project in preparation for the Director's review of the Bayesian competence project. These Web pages support the long-term goals listed above. For example, staff of the NIST Program Office reviewed

these pages, which helped them provide guidance for the Director's review. NIST scientists can view these Web pages to help determine if a Bayesian approach might be useful for solving their problems. Job candidates with an interest in Bayesian statistics can learn what SED is doing in this area.

The Bayesian Web pages will be used as a template for developing useful Web pages for other major SED projects.

Database-driven content

Building Web pages on top of database technology can make Web pages more flexible, dynamic, and easier to maintain (e.g., SED staff could update information by filling out a Web form). SED is contracting with Division 896 (YC Chang is the primary technical person for Division 896) to store the content of SED pages, where appropriate, in databases.

Databases are currently under development for the SED Yellow Book and for the SED publications list. Initial pages are available for both inputting and viewing the relevant information. Additional databases are planned for an SED calendar and for monitoring SRM's.

In summary, the SED redesign, in addition to modernizing the appearance of the Web pages, is laying the foundation for making the web a more effective tool for furthering the SED mission.

4.2 International Activities

4.2.1 SED Activities with International Organization for Standardization(ISO)

Nien Fan Zhang, Will Guthrie, Nell Sedransk
Statistical Engineering Division, ITL

The Statistical Engineering Division (SED) supports the development of international standards, particularly those that impact measurement science. SED participates at ISO Technical Committee (TC) 69 on Applications of Statistical Methods and its Subcommittee 6 on Measurement Methods. SED is also involved in the activities of the ISO/REMCO Committee on Reference Materials.

ISO/TC69 is an international standards group that develops generic statistical standards. The TC has five active subcommittees that develop documents in the following subject areas: SC1-Vocabulary, SC3-Bulk Sampling, SC4-Process Control, SC5-Acceptance Sampling, and SC6-Measurement Methods.

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. Carroll Croarkin of SED took an active part in the activities of ISO/TC69. She served as Chair of the US TAG (2000-2002) and was the convenor of working group ISO/TC69/SC6/WG7 on statistical approaches to uncertainty analysis.

At the 2001 ISO/TC69 meeting in Copenhagen, Nien Fan Zhang became the project leader of PDTS 21749. The ISO/PDTS 21749 "Measurement Uncertainty for Metrological Applications from Experimental Data Analysis" 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. Carroll Croarkin previously served as leader of this project. This document will be brought to the committee revision stage in the spring of 2002.

At the 2001 ISO/TC69 meeting in Copenhagen, ISO/TC69 recognized that the revision of ISO Guide 35-Certification of Reference Materials has relevance to the interests of TC69 by appointing Nien Fan Zhang as a liaison between ISO/TC69/SC6 and ISO/REMCO. SED staff have been involved in the revision of ISO Guide 35. The statistical and uncertainty portions of the original ISO Guide 35 were written by Keith Eberhardt of SED as a joint effort of ISO/REMCO and ISO/TC69. The revision of Guide 35 is focused on rewriting sections on the quantification of uncertainties associated with certified values of reference materials. Nien Fan Zhang and Will Guthrie of SED have reviewed the draft revision of Guide 35 and provided comments and suggestions to ISO/REMCO/WG1.

4.2.2 American Society of Mechanical Engineers

Hari Iyer
Statistical Engineering Division, ITL

The American Society of Mechanical Engineers (ASME) has developed, over the years, performance test codes that "provide test procedures which yield results of the highest level of accuracy consistent with the best engineering knowledge and practice currently available." These codes specify procedures, instrumentation, equipment operating requirements, calculation methods, and uncertainty analysis. ASME performance test code committee number 19.1 (ASME-PTC 19.1) has developed a supplement to ASME performance test codes titled "Test Uncertainty." The work on the supplement was begun in 1991 and was completed in 1998. It was approved as an American National Standard by the ANSI Board of Standards review on March 20, 1998. This supplement is now undergoing a revision/updating process with attention given to harmonization with the ISOGUM (Guide to the Expression for the Uncertainty in Measurement).

ASME PTC 19.1 committee consists of about a dozen individuals representing various industries along with some university scientists. Hari Iyer now serves as a member on this committee, representing NIST, and provides guidance on the theoretical and foundational aspects of measurement uncertainty. He is also involved in writing selected sections of the revised test codes scheduled for completion in 2003. He began his term in February 2001. The committee meets twice a year at which time the test uncertainty document is reviewed, and revisions, additions and deletions are made.

4.2.3 Workshop on Uncertainty Assessment for Chemical Measurements

William F. Guthrie
Statistical Engineering Division, ITL

Robert L. Watters, Jr.
Director's Office, TS

In October 2001, NIST staff were invited to participate in an Inter-American Metrology System (SIM) Chemical Working Group Meeting organized by Gabriela Massif, Director of the Center of Chemical Metrology at INTEC, in Santiago, Chile. At the meeting, staff from SED and NIST Technology Services gave a one-day workshop on Uncertainty Assessment for Chemical Measurements. During the workshop NIST staff compared different methods and issues affecting uncertainty computations and shared NIST's views on the correct assessment of uncertainties with approximately 100 students from National Measurement Institutes (NMI's) across the Americas and from the South American chemical industry. At the meetings there was also an opportunity to learn how some of our counterparts from other countries compute uncertainties in chemical measurements, especially in Key Comparisons. Opportunities like this to work with our colleagues from the Americas will help us harmonize the values we use to compare between NMI's and to improve the uncertainty estimates we all make in different kinds of experiments.

4.3 Education

4.3.1 Advanced Mass Measurements

Hung-kung Liu, James Filliben, Carroll Croarkin
Statistical Engineering Division, ITL

Georgia Harris
Office of Weights and Measures, Technology Services

A 5-day workshop on mass measurements was given at NIST in November 2001 for industrial and state metrologists, weight manufacturers and balance manufacturers. Participants take two lower-level courses before registering for this workshop which, in large part, focuses on statistical methods for controlling mass measurement processes and assessing uncertainties of mass determinations.

Attendees were introduced to concepts of: theory and solution of weighing designs, error analysis, components of variance, and procedures for controlling measurement processes. Specific applications to mass measurements included:

- weighing designs for common weight sets;
- propagation of uncertainties through several series of designs;
- computations of uncertainty from check standard measurements;
- control of balance precision; and
- computation of total uncertainties.

Concepts relating to weighing designs are also explored in computer sessions in which specific weighing designs are critiqued and modified by each attendee.

4.3.2 Bayesian Tutorial

Blaza Toman
Statistical Engineering Division, ITL

A three-day tutorial in Bayesian methods for NIST scientists will be given starting on February 14, 2002. This class is an expanded version of a seminar that was given to the CSTL scientists in early November 2001, and is intended to provide an understanding and some practical knowledge of this field of statistics. The anticipation is that the scientists will then be able to take advantage of some of the benefits of the Bayesian paradigm such as the relative ease of incorporating expert opinion and type B uncertainty into the statistical analysis of experimental data.

The course will provide an introduction to Bayesian methods applied to some of the specific problems often encountered by NIST scientists. These include analysis of interlaboratory experiments including Key Comparisons, linear calibration and some aspects

of experimental design. Real examples of these will be used to motivate and explain the methodology. At the end of the course, the student will be able to perform these analyses using freely available software called BUGS. The course will be taught from notes with references to a textbook.

4.3.3 Minority Internship Announcement

Charles Hagwood
Statistical Engineering Division, ITL

The **Statistical Engineering Division** of the National Institute of Standards and Technology (NIST) announces its **2002 Student Internship** of supervised practical work experience in applied statistics for minority undergraduates. Students may participate in the internship during summer vacations and/or during semester breaks. A continuation of the program may also be possible for a student who elects to go on to graduate study in statistics. The purpose of the program is to interest minority students in a statistics career by providing hands-on experience.

Statisticians are in great demand, particularly at the M.S. and Ph.D. level, throughout government, industry and business. This career path allows individuals to focus on their mathematical, computing and statistical skills in many different areas of application, depending upon the individuals' interests and aptitudes. In most areas, the prospects for advancement are excellent. Graduates with experience in applied statistics are particularly sought.

NIST, an agency of the Department of Commerce, was established to assist industry in the development of technology needed to improve product quality, to modernize manufacturing processes, to ensure product reliability and to facilitate rapid commercialization of products based on new scientific discoveries. The technical part of NIST consists of engineers and physical scientists doing basic science and research to accomplish these goals. The Statistical Engineering Division provides collaborative statistical consulting for these scientists. As an additional point of pride about NIST, in 1997 and 2001 two of our physicists won Nobel Prizes in Physics. More information about NIST and the Statistical Engineering Division can be found at the Web sites http://www.nist.gov and http://www.nist.gov/itl/div898.

We envision this program as a joint effort of university faculty and the Statistical Engineering Division staff at NIST with two phases. In the preliminary phase, if a suitable student can be identified early on, the college program can be organized to meet the needs for work in the Statistical Engineering Division at NIST. During this preliminary phase, the student may visit NIST for orientation and work of a less technical nature. To participate in the internship phase, a student needs:

- a general knowledge of statistical methodology
- elementary computing skill
  - experience using a PC or a Unix-based operating system

- use of statistical software, SAS, S+,SPSS, Minitab, Matlab is desirable

- coursework requirements
  - complete calculus sequence (usually 3 semesters), preferably with one semester of linear algebra
  - one or more semesters of statistics and probability
  - one semester of statistical methods or regression or data analysis

As an intern, the student will work under the supervision of a member of the Statistical Engineering Division staff on the design of experiments and/or analysis of experimental data. Salaries for undergraduate student employees depend upon qualifications, but generally range from 475 to 600 dollars per week. Summer internships last either 10 or 12 weeks.

To Professors:

If you have a student interested in this internship, who already meets these requirements, immediate entry into the internship phase of this program would be possible. If you have a freshman or sophomore who is interested in the preliminary phase of the internship, a commitment to pursue the necessary coursework in statistics would be necessary.

To Apply

Please contact:

Dr. Charles Hagwood (301) 975-2846 hagwood@nist.gov

Highest priority will be given to US citizens. For students who are not US citizens, an FBI background check is required before employment; this can take 6 months or longer.

4.4 New Staff

4.4.1 Ana Ivelisse Aviles

BIOGRAPHICAL SKETCH

Ana Ivelisse Avils has been a Mathematical Statistician in the Statistics Engineering Division at NIST since July 2001. She received her Ph.D. degree in Industrial Engineering and Management Sciences from Northwestern University in June 2001. She is the recipient of a NSF Graduate Fellowship, Ford Foundation Fellowship, Mary Natrella Scholarship, Ellis R. Ott Scholarship, and Richard A. Freund International Scholarship. She is also committed to developing techniques to facilitate the study of Sciences, Mathematics, and Engineering to visually impaired people. Her research interests include statistical design and analysis of industrial experiments, linear and nonlinear mixed-effects models, quality improvement and quality control.

ON-GOING PROJECTS

*Develop an appropriate uncertainty analysis for Special Test 65100-Impulse Spectrum Amplitude. (Together with N.G. Paulter and D.R. Larson, Div 811.) *Flow Measurements for Multi-Meter Transfer Standards. (Together with W.F. Guthrie and J.J. Filliben, Div 898 and P.I. Espina and G.E. Mattingly, Div 836.) *Standard Reference Materials: SRM 1003c-Spherical Glass Beads. *Co-Instructor (with J.J. Filliben, Div 898), Course on Design of Experiments (02/02). *Co-Chair (with A.I. Khuri, University of Florida), Conference on Designs for Generalized Linear Models (04/02). *Basic Research: "Assembled Designs for Estimating Dispersion Effects".

STATISTICAL RESEARCH

As a project supervisor in the pharmaceutical industry, she found that most engineers do not have access to a wide variety of statistical methods. Her research addresses this problem through the development of simple to use, yet statistically powerful tools for the design and analysis of experiments. Below is the abstract from her paper with Bruce Ankenman (Northwestern University) and Jos Pinheiro (Lucent Technologies) that is under revision for *Statistica Sinica*.

Optimal Designs for Mixed-Effects Models with Two Random Nested Factors

The main objective of this paper is to provide experimental designs for the estimation of fixed effects and two variance components, in the presence of nested random effects. Random nested factors arise from quantity designations such as lot or batch and from sampling and measurement procedures. We introduce a general class of designs for mixed-effects models with random nested factors, called assembled designs. We provide parameters and notation for describing and enumerating assembled designs. Theorems establishing conditions for existence and uniqueness of D-optimal assembled designs for the case of two variance components are presented. Specifically, we show that, in most practical situations, designs that are most balanced result in D-optimal designs for maximum likelihood estimation.

4.4.2 John Lu

BIOGRAPHICAL SKETCH

John Lu joined the Statistical Engineering Division in July 2001. He received M.S. and Ph.D. degrees in statistics from the University of North Carolina, Chapel Hill, and a BS degree from Peking University in Beijing, China. John came from Insightful Corporation (formerly MathSoft Inc) and was a member of the Geophysical Statistics Project at the National Center of Atmospheric Research (NCAR).

ONGOING PROJECTS

John Lu has involved with several projects at NIST. One is the Bayesian metrology project, in which his focus is on developing computational tools and Bayesian methodology for many NIST problems for which the Bayesian approach may offer a more suitable framework for uncertainty analysis: examples include combining information from multi-lab and multi-method studies, use of type B information, and interactive elicitation of expert opinion. Another project is the Network Measurements and Modeling Project funded by DOD, collaborating with the Advanced Network Technology Division and other statisticians at SED, in which the statistical trust is to develop realistic statistical and time series models for heavy-tailed traffic data and to demonstrate applications in real-time emulation/simulation/prediction of network traffic.

STATISTICAL RESEARCH

John Lu has broad research interests in statistics, computation, and collaborative research. Specific statistical areas in which he has worked include time series analysis, local polynomial regression, data assimilation and design for computer-based models, and Bayesian statistics. Here is the abstract of a recent paper he co-authored with Mark Berliner and Chris Snyder, and was published in the Journal of Atmospheric Sciences.

Statistical Design for Adaptive Weather Observations

Suppose that one has the freedom to adapt the observational network by choosing the times and locations of observations. Which choices would yield the best analysis of the atmospheric state or the best subsequent forecast? Here, this problem of "adaptive observations" is formulated as a problem in statistical design. The statistical framework provides a rigorous mathematical statement of the adaptive observations problem and indicates where the uncertainty of the current analysis, the dynamics of error evolution, the form and errors of observations, and data assimilation each enter into the calculations. The statistical formulation of the problem also makes clear the importance of the optimality criteria (for instance, one might choose to minimize the total error variance in a given forecast) and identifies approximations that make calculation of optimal solutions feasible, in principle. Optimal solutions are discussed and interpreted for a variety of cases. Selected approaches to the adaptive observations problem found in the literature are reviewed and interpreted from the optimal statistical design viewpoint. In addition, a numerical example, using the 40-variable model of Lorenz and Emanuel, suggests that some other proposed approaches may often be close to the optimal solution, at least in this highly idealized model.

4.4.3 Blaza Toman

BIOGRAPHICAL SKETCH

Blaza Toman has been a Mathematical Statistician in the Statistical Engineering Division at NIST since December 2000. She received her Ph.D. in Statistics from Ohio State University in 1987. Prior to joining NIST she was an Associate Professor of Statistics at George Washington University, specializing in Bayesian Experimental Design. She taught both undergraduate and graduate classes and supervised several Ph.D. students. She also worked as an independent statistical consultant for several medical device companies, designing clinical trials using Bayesian methodology.

ON GOING PROJECTS

*Develop and implement a Bayesian analysis of the CIPM Key Comparison: CCPR-K2.a Spectral Responsivity part 1 (900 nm to 1600 nm). *Design an experiment for a study of dust wipes used to obtain samples for lead testing from household paint. *Develop (and disseminate to NIST scientists) Bayesian methods for the Consensus Mean problem. *Develop an inventory of Key Comparison experimental designs, study their properties and disseminate results of this study to NIST scientists and the international metrological community (together with W. Guthrie and N.F. Zhang). *Study and evaluate the current methodology for the calculation and use of the Key Comparison Reference Value, disseminate results to NIST scientists and the international metrological community (together with W. Guthrie and N.F. Zhang). *Teach a course Introduction to Bayesian Analysis for NIST scientists (February 2002).

STATISTICAL RESEARCH

Blaza Toman's main research interests are in experimental design and Bayesian statistics. Currently she is particularly interested in the use of hierarchical Bayesian models and various MCMC algorithms in the analysis of Key Comparison data.

The following abstract is from a paper published in the *Journal of Statistical Planning and Inference*, co-authored with Athan Katsis from the University of Cyprus.

Bayesian sample size calculations for binomial experiments

ABSTRACT

This paper proposes new methodology for calculating the optimal sample size when a hypothesis test between two binomial proportions is conducted. The problem is addressed from the Bayesian point of view. Following the formulation by DasGupta and Vidakovic (1997, *J. Statist. Plann. Inference* 65, 335-347), the posterior risk is determined and set not to exceed a prespecified bound. A second constraint deals with the likelihood of data not satisfying the bound on the risk. The cases when the two proportions are equal to a fixed or to a random value are examined.

4.4.4 Dipak K. Dey

Dipak K. Dey is a Professor and Head of the department of Statistics at the University of Connecticut. He is a fellow of the American Statistical Association and the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. After receiving his Ph.D. from Purdue, he was a visiting scholar at Stanford University and then taught at the University of Kentucky and Texas tech University. He held visiting position at Macquire University, Indian Statistical Institute and the University of British Columbia. His research interests include Bayesianm Modeling, Statistical Decision Theory, Statistical Shape Analysis, Computational Statistics, Reliability and Survival Analysis. He has published three books and 120 refereed papers.

Currently he is a Visiting Faculty at the NIST. He spent his sabbatical during Fall 2001 in the Statistical Engineering Division(SED)at NIST. During his sabbatical he was involved in various projects in the SED. In particluar he worked on problems on combining information from interlaboratory studies and modeling expert opinion and prior elicitation. He also partly supervised a student from summer internship program at NIST.

4.4.5 Tom Ryan

Tom Ryan received his Ph.D. in statistics from the University of Georgia in 1977, and he also has two degrees in business administration from Georgia State University. He is a Fellow of the American Statistical Association and the Royal Statistical Society, and is a Senior Member of the American Society for Quality. He was most recently Visiting Professor in the Department of Statistics at the University of Michigan during 2000-2001, and previously served as Director of Statistical Consulting at Case Western Reserve University during 1996-2000. While at Case Western Reserve his work included studying the school funding issue in Ohio, working on a grant from The Abbington Foundation, in addition to teaching courses and other duties which included directing graduate students. He is an applied statistician with approximately 25 refereed publications in the areas of statistical quality improvement, regression analysis, and experimental design. Additionally, he is the author of Statistical Methods for Quality Improvement (Wiley; first edition, 1989; second edition, 2000) and Modern Regression Methods (Wiley, 1997). He is presently under contract to Wiley to write books on engineering statistics and experimental design and is currently working on both books. He is also an active researcher with papers in several journals within the past few years (including one in JASA in 2000), with papers on control charts and logistic regression to appear in the near future. His present research interests are, in order, exact logistic regression, design of experiments, and control charts. Dr. Ryan also serves on the Editorial Review Board of the Journal of Quality Technology, a position that he has held since 1990.

Since February, 2000, Dr. Ryan has been working on editing and rewriting the entire NIST/SEMATECH Engineering Statistics Internet Handbook. He has also edited the Yellow Book for 2002

4.5 Students Program

4.5.1 Summer Students 2001

Charles Hagwood
Statistical Engineering Division, ITL

Figure 3: 2001 Summer Students (from left to right): Susan Heath (Rutgers University), Anand Kesavan (University of Michigan), Mariama Moody (Hampton University), Kanika Chadda (Boston University), Kia Moore (Norfolk State University), Kimball Kniskern (University of California, Berkeley, Co-Op student).

The Statistical Engineering Division's Summer Students Program was restarted afresh in 2001. It provides statistical work experience for university undergraduate and graduate students. Each student is assigned a project and works under the supervision of an attentive staff member. Undergraduate students gain a greater appreciation for statistics and the successful ones are encouraged to pursue graduate work in statistics. Graduate students use the program to supplement understanding of their course work. Thus, the Statistical Engineering Division plays a direct role in increasing the pool of well trained professional statisticians.

In 2001, students applied from all over the United States. The application process requires the students to provide: 1. A resume which includes an official transcript. 2. A paragraph describing their career goals. 3. The names of two or three references.

A student's salary depends on their educational level. Salaries are based on an official pay scale and range from $10.88/hr to $22.42/hr (freshman - second year Ph.D. student). Students are expected to spend at least two summer months at NIST.

The program is coordinated with other student programs at NIST, such as the NIST SURF program and the U.S. Department of Commerce Post-Secondary Internship Program (Oak Ridge Associated Universities Program, American Indian Science and Engineering Society Program, Hispanic Association of Colleges and Universities National Internship Program, Minority Access, Inc and Lee College Programs Partnership). A special part of the SED Summer Students Program is an outreach to minority students in historically black colleges and Hispanic-serving institutions. The minority aspect of the program is explained in the announcement above.

Student Projects

Andrew Rukhin and Stefan Leigh worked with Mariama Moody, Kimball Kniskern and Susan Heath on aspects of the ongoing DARPA/SED/Div 894 (Info Access Div) collaboration in the area of Human ID. Each student collaborated on different aspects of the same broad project, attended project meetings, and attended the DARPA Human ID biannual meeting held in Rosslyn, July, 2001.

Mariama specifically worked on organizing and summarizing extracts from the FERET database, experimenting with transformations of similarity matrices extracted therefrom, and doing 2-way analyses of variance with associated graphics. Her name will appear as coauthor of "Transformation, Ranking and Clustering for Face Recognition Algorithm Performance", which will appear in the Proceedings of the IEEE Third Workshop on Automatic Identification Advanced Technologies (AutoID 02), as well as in expanded NISTIR form.

Kimball worked on the same paper (and will appear as coauthor) doing data reduction and BMDP ANOVA's and Multiple Comparisons. He also worked extensively in 2 other areas of the same project: (1) Multidimensional Scaling - reading text, running S-plus and SAS MDS codes, reducing data: a conference proceeding-type publication is anticipated and he will appear as coauthor; (2) a Covariates for Similarity Score Prediction study - ongoing, reading text, reducing and reformatting data, running BMDP All Possible Sub-

sets Regression: again, a conference proceeding-type publication is anticipated and he will appear as coauthor.

Susan, at our request, studied the potential applicability of Canonical Correspondence Analysis for ranking algorithm performance, and reported to the group on that (advising, credibly, against). She also contributed innovative and constructive ideas to the overall project, the MDS portion of the project, and the Covariates portion of the project. Her name will appear as coauthor on each of the publications.

Kimball also worked with Dipak Dey and John Lu on the Prior Elicitation project. He wrote an S-plus program for computing the prior distribution of hyperparameters for a linear model. His work will be part of an ongoing research effort in developing an interactive system for elicitation of an expert's opinion. The goal is to integrate his work into a formal Bayesian analysis of many NIST problems.

Kia helped Alan Heckert redesign the SED Web pages and the SED STRD Web pages. They converted everything to the new NIST standard format.

Kanika helped Joan Rosenblatt organize the papers of Churchill Eisenhardt. Also, she helped with the reorganization of the SED library.

Anand worked with Hung-Kung Liu on inference for local regression fitting.

Figure 4: Future SED Superstars.

5. Staff Publications and Professional Activities

5.1 Publications

5.1.1 Publications in Print

1. A.I. Aviles, (with B.E. Ankenman, J.C. Pinheiro) Assembled Designs for Dispersion Effects, *ASA Proceedings of the Section on Quality and Productivity*, 2000, pp. 82–87.

2. A.I. Aviles, *Optimal Designs for Mixed-Effects Models with Random Nested Factors*, Ph.D. Thesis, Northwestern University, Evanston, IL, 2001.

3. K.J. Coakley (with P.R. Huffman, C.R. Brome, J.S. Butterworth, M.S. Dewey, S.N. Dzhouyuk, R. Golub, G.L. Greene, K. Habicht, S.K. Lamoreaux, C.E. Mattoni, D.N. McKinsey, F.E. Wietfeld, J.M. Doyle), Magnetic Trapping of Neutrons, *Nature*, 403, 62, (2000).

4. K.J. Coakley (with P.R. Huffman, C.R. Brome, J.S. Butterworth, M.S. Dewey, S.N. Dzhouyuk, R. Golub, G.L. Greene, K. Habicht, S.K. Lamoreaux, C.E. Mattoni, D.N. McKinsey, F.E. Wietfeld, J.M. Doyle). Progress towards magnetic trapping of ultracold neutrons, *Nuclear Methods in Physics Research A*, 440,3, (2000).

5. K.J. Coakley and G.L. Yang (with P.R. Huffman, A.K. Thompson, C.R. Brome, J.S. Butterworth, M.S. Dewey, S.N. Dzhouyuk, R. Golub, G.L. Greene, K. Habicht, S.K. Lamoreaux, C.E. Mattoni, D.N. McKinsey, F.E. Wietfeld, J.M. Doyle, L. Yang, K.J. Alvine). Magnetic Trapping of Ultracold Neutrons: Prospects for an Improved Measurement of the Neutron Lifetime, in *Proceedings of the Conference on Fundamental Physics with Pulsed Neutron Beams*, Durham NC, June 1-2, 2000.

6. K.J. Coakley, with (D.K. Walker, R.F. Kaiser, D.F. Williams), Lumped-Element Models for On-Wafer Calibration, Proceedings of 56th ARFTG Microwave Measurements Conference, Denver, C0, Nov. 30 - Dec. 1, 2000.

7. K.J. Coakley and P.D. Hale, Alignment of noisy signals, *IEEE Transactions on Instrumentation and Measurement*, 50, 1, 144 (2001)

8. K.J. Coakley (with C.R. Brome, J.S. Butterworth, 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).

9. K.J. Coakley, Neutron lifetime experiments using magnetically trapped neutrons: optimal background correction strategies, *Nuclear Instruments and Methods in Physics Research A,* 469, 3, 354, (2001)

10. K.J. Coakley and M.S. Levenson, Guest editorial: Advances in quantitative image analysis, *International Journal of Imaging Systems and Technology*, 11, 4, 209 (2001)

11. J. J. Filliben and N. A. Heckert (with E. Simiu and S. K. Johnson), Extreme Wind Load Estimates Based on the Gumble Distribution of Dynamic Pressures: An Assessment, *Structural Safety*, 23, 2001, pp. 221-229.

12. J. J. Filliben (with Z. Lin and K. Inn), An Alternative Statistical Approach for Interlaboratory Comparison Data Evaluation, *Journal of Radioanalytical and Nuclear Chemistry*, Vol. 248, No. 1, 2001, pp. 163-173.

13. J. J. Filliben (with M. Simon (FHWA) and D. Bentz), Concrete Optimization Software Tool: User's Guide, *Federal Highway Administration Report*, March 2001.

14. J. J. Filliben (with E. Simiu, R. Wilcox, and F. Sadek), Wind Speeds in the ASCE 7 Standard Peak-Gust Map: An Assessment, *NIST Building Science Series* 178, September 2001.

15. W.F. Guthrie (with M.R. Winchester, W.R. Kelly, J.L. Mann, B.S. MacDonald, and G.C. Turk, An Alternative Method for the Certification of S Mass Fraction in Coal Standard Reference Materials, *Fresenius Journal of Analytical Chemistry*, Vol. 370, 2001, pp. 234-240.

16. W.F. Guthrie (with G.C. Turk, L.L. Yu, and M.L. Salit) Using High-Performance Spectrometric Methods for Calibration Transfer Between Environmental CRMs, *Fresenius Journal of Analytical Chemistry*, Vol. 370, 2001, pp. 259-263.

17. W.F. Guthrie (with M.R. VanLandingham, J.S. Villarrubia, and G.F. Meyers) Nanoindentation of Polymers: An Overview, *Macromolecular Symposia: Recent Advances in Scanning Probe Microscopy of Polymers*, Vol. 167, 2001, pp. 15-44.

18. W.F. Guthrie (with D.L. Banks, K.R. Eberhardt, L.M. Gill, M.S. Levenson, H.K. Liu, M.G. Vangel, J.H. Yen, and N.F. Zhang, An Approach to Combining Results From Multiple Methods Motivated by the ISO GUM", *Journal of Research of the National Institute of Standards and Technology*, Vol. 105, No. 4, 2000, pp. 571-579.

19. W.F. Guthrie, (with M.W. Cresswell, N.M.P. Guillaume, W.E. Lee, R.A. Allen, R.N. Ghoshtagore, Z. Osborne, N. Sullivan and L.W. Linholm) Extraction of Sheet Resistance from Four-Terminal Sheet Resistors Replicated in Monocrystalline Films with Non-Planar Geometries, *IEEE Transactions on Semiconductor Manufacturing*, Vol. 12, No. 2, 1999, pp. 154-165.

20. C. Hagwood (with Lynne Rosenthal) Reliability of Conformance Tests, *IEEE Transactions on Reliability*, 50 (2), 204-208, 2001.

21. C. Hagwood, (with D. Shepherd, R. Fields) Evaluation of the Elevated Temperature Creep Strength of Three Lead-Free Solder Alloys in Soldered Joints, em Journal of Testing and Evaluation, July 2001, 380-386.

22. S.D. Leigh, (with P. Stutzman) Compositional Analysis of NIST Reference Material Clinker 8486, *Proc. International Cement Microscopy Assoc. Meeting*, April 30, 2000, Montreal.

23. S.D. Leigh, (with R.D. Shull, R.D. McMichael, L.J. Swartzendruber) Absolute Magnetic Moment Measurements of Nickel Spheres. *Journal of Applied Physics*, 87 (9), May 2000, p. 5992–5994.

24. S.D. Leigh, (with G.L. Yang, J.F. Widmann, S.R. Charagundla, C. Presser) A Correction for Spray Intensity Measurements Obtained via Phase Doppler Interferometry. *Aerosol Science and Technology*, 32 (6), June 2000, p. 2–19.

25. S.D. Leigh, (with J.F. Widmann, C. Presser) Improving phase Doppler volume flux measurements in low data rate applications. *Measurement Science and Technology*, 12, June 2001, p. 1180–1190.

26. S.D. Leigh, (with A. Rukhin, Y. Carlinet, V. Galtier, K. Mills) Calibrating an Active Network Node. *Proc. Second Workshop on Active Middleware Services*, August 2000.

27. S.D. Leigh, (with J. Sieber) Certification of SRM 1848 PCMO Additive Package. *Proc. 49th Annual Denver X-Ray Conference*, August 2000, Denver, Colorado.

28. S.D. Leigh, (with J. Verkouteren) New Low-Index Liquid Refractive Index Standard: SRM 1922. *Fresenius' Journal of Analytical Chemistry*, 367 (3), June 2000, p. 226–231.

29. S.D. Leigh, (with A. Rukhin, V. Galtier, Y. Carlinet, K. Mills) Expressing Meaningful Processing Requirements among Heterogeneous Nodes in an Active Network. *Proc. 2nd WOSP - Workshop on Software Performance*, September, 2000, Ottawa.

30. S.D. Leigh, (with A. Rukhin, J. Soto, J. Nechvatal, et al) A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. NIST Special Publication 800-22. October, 2000.

31. S.D. Leigh, (with J.F. Widmann, C. Presser) Effect of Burst-Splitting Events on Phase Doppler Interferometry Measurements. *Proc. 39th AIAA Aerospace Sciences Meeting*, paper 2001-1130, 8-12 January, 2001, Reno, NV.

32. S.D. Leigh, (with A. Rukhin, A. Heckert, P. Grother, J. Phillips, M. Moody, K. Kniskern, S. Heath) Transformation, Ranking, and Clustering for Face Recognition Algorithm Performance. *Proc. Third Workshop on Automatic Identification Advanced Technologies (AutoID02/IEEE)*, March 2002, Tarrytown, NY.

33. W.S. Liggett, Nonparametric and Semiparametric Models in Comparison of Observations of a Particle Size Distribution, *Proceedings of the International Conference on New Trends in Computational Statistics with Biomedical Applications*, Japanese Society of Computational Statistics, 2001, pp. 131–148.

34. W.S. Liggett, (with P. Over) Understanding TREC Results–the Role of Statistics, *Bulletin of the International Statistical Institute, 53rd Session Proceedings*, International Statistical Institute, 2001, pp. 45–48.

35. W.S. Liggett, (with C. Buckley) Query Expansion Seen Through Return Order of Relevant Documents, *The Ninth Text REtrieval Conference (TREC-9)*, eds. E.M. Voorhees and D.K. Harman, NIST Special Publication 500-249, 2001, pp. 51–70.

36. W.S. Liggett , (with S.R. Low, D.J. Pitchure, J. Song) Capability in Rockwell C Scale Hardness, *Journal of Research of the National Institute of Standards and Technology*, 105, 2000, pp. 511–533.

37. H.K. Liu, (with J.T. Hwang, and G. Stenbakken) HELP for Missing Data, *Proceedings of the 15th International Workshop on Statistical Modeling : New Trend in Statistical Modeling*, 2000, pp. 461–464.

38. H.K. Liu, (with N.F. Zhang) Bayesian Approach to Combining Results from Multiple Methods, *Proceedings of The 2001 Joint Stat. Meetings*

39. Z.-Q. Lu, Local Polynomial Prediction and Volatility Estimation in Financial Time Series. Chapter 5 of *Modeling and Forecasting Financial Data: Techniques of Nonlinear Dynamics* (Eds. A.Soofi and Ly Cao), expected publication March 2002, Kluwer, pp. 115-136 (tentative).

40. Z.-Q. Lu, (with L.M.Berliner, C.Snyder) Optimal Design for Spatial and Adaptive Observations. Chapter 5 of *Studies in Atmospheric Sciences*, Lecture Notes in Statistics, Vol.144, Springer-verlag, 2000, pp. 65-78.

41. Z.-Q. Lu, (with R.M. Errico, L.Fillion, D.Nychka), Some Statistical Considerations Associated With the Data Assimilation of Precipitation Observations. *Quarterly Journal of the Royal Meteorological Society*, Vol. 126, no. 562, part A , 2000, pp. 339-360.

42. J.D. Splett, (with G.E. Obarski) Transfer Standard for the Spectral Density of Relative Intensity Noise of Optical Fiber Sources Near 1550 nm, *Journal of the Optical Society of America B*, 18 (6), 2001, pp. 750–761.

43. C.M. Wang, (with S.M. Etzel, A.H. Rose) Dispersion of the temperature dependence of the retardance in SiO_2 and MgF_2, *Applied Optics*, 39 (31), 2000, pp. 5796–5800.

44. C.M. Wang, (with P.D. Hale) Heterodyne system at 850 nm for measuring photoreceiver frequency response, *Technical Digest of Symposium on Optical Fiber Measurements*, 2000, pp. 117–120.

45. C.M. Wang, (with A.H. Rose, S.D. Dyer) Fiber Bragg grating metrology round robin: Telecom group, *Technical Digest of Symposium on Optical Fiber Measurements*, 2000, pp. 161–164.

46. C.M. Wang, (with A.H. Rose, S.D. Dyer) Optical fiber Bragg grating metrology round robin, *NIST Journal of Research*, 105 (6), 2000, pp. 839–866.

47. C.M. Wang, (with C.N. McCowan, T.A. Siewert, D.P. Vigliotti) Reference materials for weld metal ferrite content: gauge calibration and material characterization, *Welding Journal*, 80 (4), 2001, pp. 106–114.

48. G. L. Yang, (with K. Coakley) Likelihood models for two stage neutron lif etime experiments, *Physical Review C. vol. 63,* 2000, pp. 014602-1–16.

49. G. L. Yang, (with J. Widmann, R. Gharagundla, C. Presser and S. Leigh) A correction method for spray intensity measurements obtained via Phase Doppler Interferometry, *Aerosol Science and Technology, v. 32*. 2000, pp. 18.

50. G. L. Yang, (with S.Y. He) The strong convergence of the integrals with respect to the product-limit estimate under random truncation and censoring, *Statistis and Probability Letters*. 49, 2000, pp. 235–244.

51. G. L. Yang, (with L. Le Cam) *Asymptotics in Statistics: Second Edition*, Springer-Verlag, 2000.

52. J.H. Yen, (with J. Brown Thomas, M.C. Kline, L.M. Gill, D.L. Duewer, L.T. Sniegoski, K.E. Sharpless) Preparation and value assignment of Standard Reference Material 968c fat-soluble vitamins, carotenoids, and cholesterol in human serum, *Clinica Chimica Acta*, 305, 2001, pp. 141–155.

53. N.F. Zhang, Combining Process Capability Indices from a Sequence of Independent Samples, *International Journal of Production Research*, 39(13), 2001, pp. 2769–2781.

54. N.F. Zhang, (with A.E. Vladar, M.T. Postek, R.D. Larrabee, S.N. Jones) Reference Material 8091: New Scanning Electron Microscope Sharpness Standard, *Proceedings of SPIE*, 4344, 2001,pp. 827–834.

55. N.F. Zhang, Statistical Control Charts for Monitoring the Mean of a Stationary Process, *Journal of Statistical Computation and Simulation*, 66, 2000, pp. 249–258.

56. N.F. Zhang, (with M.T. Postek, R.D. Larrabee, A.E. Vladar) Multivariate Kurtosis for Measuring Image Sharpness, *Proceedings of the 15th International Workshop on Statistical Modeling: New Trend in Statistical Modelling*, 2000, pp. 529–532.

57. N.F. Zhang, (with M.T. Postek, A.E. Vladar, R.D. Larrabee) Potentials of On-line Scanning Electron Microscope Performance Analysis Using NIST Reference Material 8091, *Proceedings of SPIE*, 3998, 2000, pp. 28–37.

5.1.2 NIST Technical Reports

1. W.F. Guthrie, M.S. Levenson, (with R.G. Gann, K.D. Steckler and S. Ruitberg), Relative Ignition Propensity of Test Market Cigarettes, NIST Technical Note 1436, 2000, 36 p.

2. S.D. Leigh, (with C.R. Schultheisz), Certification of the Rheological Behavior of SRM 2490, Polyisobutylene Dissolved in 2,6,10,14-Tetramethylpentadecane. NIST Special Publication 260-143, 2001, 75p.

3. S.D. Leigh, (with P.E. Stutzman), Phase Composition Analysis of the NIST Reference Clinkers by Optical Microscopy and X-ray Powder Difraction. NISTIR 1441, 2001, 44p.

4. J.D. Splett, (with G.E. Obarski), Measurement Assurance Program for the Spectral Density of Relative Intensity Noise of Optical Fiber Sources Near 1550 nm, NIST SP 250-57, 2000, 90p.

5. C.M. Wang, (with R.M. Craig) Measurement assurance program for wavelength dependence of polarization dependent loss of fiber optic devices in the 1535 – 1560 nm wavelength range. NIST SP 250-60, 2002.

6. C.M. Wang, (with P.A. Williams, S.M. Etzel, J.D. Kofler) Standard reference material 2538 for polarization-mode dispersion (non-mode-coupled). NIST SP 260-145, 2002.

5.1.3 Book Reviews

1. A.I. Aviles, Book Review-Linear Mixed Models for Longitudinal Data, G. Verbeke and G. Molenberghs, *Technometrics*, 43(3), 2001, pp. 375.

5.1.4 Publications in Process

1. A.I. Aviles, (with B.E. Ankenman, J.C. Pinheiro) Optimal Designs for Mixed-Effects Models with Two Random Nested Factors, *Statistica Sinica*, submitted.

2. A.I. Aviles, Robustness Experiments with Two Variance Components, *ASA Proceedings of the Section on Physical and Engineering Sciences*, to appear.

3. K.J. Coakley, H. Chen-Mayer, G. Lamaze, D. Simons, Calibration of a stopping power model for silicon based on Neutron Depth Profiling and Secondary Ion Mass Spectrometry Measurements, *Nuclear Instruments and Methods in Physics Research B*, to appear

4. K.J. Coakley and R.S. Cerveny, A Weekly Cycle in Atmospheric Carbon Dioxide, *Geophysical Review Letters*, to appear

5. K.J. Coakley, J.M. Richardson, Z. Chowdhuri, W.M. Snow, M.S. Dewey. Estimation of Neutron Mean Wavelength from Rocking Curve Data, *Journal of Applied Crystallography*, submitted

6. K.J. Coakley, G.L. Yang, Estimation of the Neutron Lifetime: Comparison of Methods which Account for Background, *Physical Review C*, submitted

7. W.F. Guthrie (with C.G. Simon and F.W. Wang), Cell Seeding into Calcium Phosphate Cement, *Journal of Biomedical Materials Research*, submitted.

8. W.F. Guthrie, Should (T1- T2) Have Larger Uncertainty Than T1?, *Proceedings of the 8th International Conference on Temperature: Its Measurement and Control*, to appear.

9. W.F. Guthrie (with B.W. Mangum and G.F. Strouse), Summary of Comparison of Realizations of the ITS-90 over the Range 83.8058 K to 933.473 K: CCT Key Comparison 3, *Metrologia*, to appear.

10. W.F. Guthrie (with B.W. Mangum and G.F. Strouse), CCT-K3: Key Comparison of Realizations of the ITS-90 over the Range 83.8058 K to 933.473 K, *Report of the Comité Consultatif de Thermométrie*, to appear.

11. D.C. Boes, (with A.M. Mood, A.P. Jayasumana) A Nonpreemtive Priority Delay Model with Modified-vacation Intervals for Homogeneous FDDI Networks, *IEEE Infocom*, to appear.

12. K.J. Coakley, A Bootstrap Medthod for Nonlinear Filtering of Maximum Likelihood Reconstructions, *International Journal of Imaging Science and Technology*, submitted.

13. S.D. Leigh, J. Yen, (with J. Sieber, D. Broton, C. Fales, B. MacDonald, A. Marlow, S. Nettles) Standard Reference Materials For Cements, *Cement and Concrete Research*, submitted.

14. W.S. Liggett, Nonparametric and Semiparametric Models in Comparison of Observations of a Particle Size Distribution, *Journal of the Japanese Society of Computational Statistics*, submitted.

15. W.S. Liggett, Empirical Evaluation of Information Retrieval Systems, *Information Retrieval*, to be submitted.

16. H.K. Liu, (with G. Stenbakken) Empirical Modeling Methods using Partial Data, under submission.

17. C.M. Wang, K.J. Coakley, (with P.D. Hale, T.S. Clement) Uncertainty of oscilloscope timebase distortion estimate, *IEEE Transactions on Instrumentation and Measurement*, in press.

18. C.M. Wang, (with T.J. Drapela) A statistical model for cladding diameter of optical fibers, *Technometrics*, submitted.

19. G. L. Yang, (with S.Y. He), Estimation of Regression Parameters with Left Truncated Data, submitted for publication.

20. J.H. Yen, An Estimator of Effect Size based on the Mann-Whitney Statistic, *Psychological Methods*, in revision.

21. N.F. Zhang, What the Generalized Moving Averages Can Do for the Process Monitoring, *2001 Proceedings of Section of Physical and Engineering Sciences of American Statistical Society*, to appear.

22. N.F. Zhang, (with N. Sedransk, D. G. Jarrett) Statistical Uncertainty Analysis of CCEM-K2 Comparisons of Resistance Standards, to be submitted.

5.1.5 Working Papers

1. K.J. Coakley, C.-M. Wang, P.D. Hale and T.S. Clement. Adaptive estimation of RMS jitter noise

2. S.D. Leigh, B. Toman, (with C.S. Phinney, M.M. Schantz, M.J. Welch), Certification of Standard Reference Material 1946 – Lake Superior Fish Tissue for Fatty Acids, *Journal of the American Oil Chemistry Society*, to be submitted.

3. H.K. Liu, (with J.T. Hwang) Does EM Algorithm Work for Identifying Principal components When Massive Data are Missing.

4. J.D. Splett, K.J. Coakley, (with M.D. Janezic, R.F. Kaiser), Complex Permittivity Measurement Using the NIST 60mm Cylindrical Cavity, to be submitted to *NIST Journal of Research*.

5. K.J. Coakley, J.D. Splett, (with M.D. Janezic, R.F. Kaiser), Estimation of Q–factors and Resonant Frequencies, to be submitted to *IEEE Transactions on Microwave Theory and Techniques*.

6. N.F. Zhang, Estimation of Process Variance in Using SPC Charts for a Stationary Process.

5.1.6 Acknowledgements in Publications

1. A.I. Aviles in: E. Simiu, R. Wilcox, F. Sadek,.J.J. Filliben, Wind Speeds in the ASCE 7 Standard Peak-Gust Map: An Assessment, NIST Building Science Series 178, 73 pages, September 2001.

2. A.I. Aviles in: B.E. Ankenman, H. Lui, A.F. Karr, J.D. Picka, A Class of Experimental Designs for Estimating a Response Surface and Variance Components, *Technometrics*, to appear.

3. K.J. Coakley in: D.C. Hurley, V.K. Tewary, A.J. Richards, Surface acoustic wave methods to determine the anisotropic elastic properties of thin films, Measurement Science and Technology, 12 1486-1494 (2001)

4. K.J. Coakley, in: Chaotic Transitions in Deterministic and Emil Simiu, Stochastic Dynamical Systems: Applications of Melnikov Processes in Engineering, Physics, and Neuroscience. Princeton University Press, 2002.

5. S.D. Leigh in: C. Elster and A. Link, Analysis of key comparison data: assessment of current methods for determining a reference value, *MeasurementScience and Technology*, 12 (2001), p. 1431–1438.

6. John Lu in: W. Constantine, D. B. Percival, and P. G. Reinhall, Inertial range determination for aerothermal turbulence using fractionally differenced processes and wavelets. *Physical Review* E, 64(3), 036301 (2001).

7. C.M. Wang, J.D. Splett in: C.N. McCowan, T.A. Siewert, D.P. Vigliotti, The NIST Charpy V–Notch Verification Program: Overview and Operating Procedures, *NIST IR–6618*, (2002).

5.2 Talks

5.2.1 Technical Talks

1. A.I. Aviles, Robustness Experiments with Two Variance Components, 45th Fall Technical Conference, Ontario, Canada, October 18, 2001.

2. K.J. Coakley, Statistical Problems in Optoelectronics, Institute of Statistics and Decision Sciences, Duke University, November 10, 2000.

3. K.J. Coakley, Estimation of Neutrino Event Location, Collaboration meeting at Princeton University, September 25, 2001.

4. K.J. Coakley, Correction of Optoelectronic Signals for Drift, Time Base Distortion, and Jitter, Symposium on Reference Receiver Calibration, Telecommunications Industry Association/ International Electrotechnical Commission, Working Group 4, TC-86, Kauai, January 23, 2002

5. J. J. Filliben, EDA on DEX, 2001 Quality and Productivity Conference, Austin, TX, May 23, 2001.

6. J. J. Filliben, Tukey and EDA: Reflections, Principles, and Applications, ASA CO-WY Fall Chapter Symposium Exploratory Data Analysis in the 21st Century: Research in Modern EDA: A Symposium in Honor of John W. Tukey, Colorado University/Denver, Denver, CO, October 27, 2001.

7. J. J. Filliben, An Exploratory Data Analysis Retrospective: The John Tukey Legacy, U.S. Census Bureau, Suitland, MD, November 27, 2001.

8. W.F. Guthrie, Should (T1- T2) Have Larger Uncertainty Than T1?, 8th International Conference on Temperature: Its Measurement and Control, Berlin, Germany, June 21, 2001.

9. W.F. Guthrie, An Overview of the Handbook, Quality and Productivity Research Conference, Austin, TX, May 24, 2001.

10. W.F. Guthrie, NIST/SEMATECH Engineering Statistics Handbook, Workshop on Monte Carlo Methods and NIST Web Handbooks, DLMF Seminar Series, Gaithersburg, MD, April 4, 2001.

11. W.F. Guthrie, Do Cigarettes Have to Cause Fires, NRC Panel Meetings, NIST, Gaithersburg, MD, February 14 and February 27, 2001.

12. W.F. Guthrie, Analysis of Data from Key Comparison 3: Comparison of the Realization of the ITS-90 over the Range 83.8058 K to 933.473 K, Working Group Meeting of the Comité Consultatif de Thermométrie, Gaithersburg, MD, January 18, 2000.

13. C. Hagwood, An Application of Stochastic Differential Equations to Particle Sizing, 11th INFORMS Applied Probability Society Conference, New York City, July 25-27, 2001.

14. C. Hagwood, Bayesian Calibration, 2001 NCSL Workshop and Symposium, July 28-August 4, 2001.

15. W.S. Liggett, Nonparametric and Semiparametric Models in Comparison of Observations of a Particle Size Distribution, International Conference on New Trends in Computational Statistics with Biomedical Applications, Osaka, Japan, August 30, 2001.

16. W.S. Liggett, Understanding TREC Results–the Role of Statistics, 53rd Session of the International Statistical Institute, Seoul, Korea, August 24, 2001.

17. W.S. Liggett, The Bayesian Paradigm for Expressing Knowledge of a Physical Quantity, National Measurement Institute of Japan, Tsukuba, Japan, September 7, 2001.

18. H.K. Liu, N.F. Zhang, Bayesian Approaches to Combining Results from Multiple Methods, Joint Statistical Meetings, Atlanta, August, 2001.

19. N. Sedransk, Choosing Problems - Choosing Solutions (Invited Plenary Speaker), Spring Research Conference on Statistics in Industry and Engineering, Roanoke, Virginia, June 18 - 20 2001.

20. N. Sedransk, Statistical Modeling and Analysis: Statistics for Measurement in a Virtual World (Invited), AUTOTESTCON Valley Forge, Pennsylvania, July 20 - 23 2001.

21. N. Sedransk, H.K. Liu, Statistical Modeling and Visualization of Network Data (Invited), DARPA NMS PI Meetings, Atlanta, Georgia, October, 2001.

22. J.H. Yen, Combining Information in the Physical Sciences and Engineering (poster presentation), Gordon Research Conference on Statistics in Chemistry and Chemical Engineering, Williamstown, MA, July 2001

23. N.F. Zhang, What the Generalized Moving Averages Can Do for the Process Monitoring, Spring Research Conference on Statistics in Industry and Technology, Roanoke, VA, June 2001.

24. N.F. Zhang, How to Combine Process Capability Indices from a Sequence of Independent Samples, Mathematical and Information Sciences Division, CSIRO, Sydney, Australia, December 14, 2001.

25. N.F. Zhang, Statistical Process Monitoring for Autocorrelated Data, Mathematical and Information Sciences Division, CSIRO, Melbourne, Australia, December 17, 2001.

26. N.F. Zhang, Statistical Process Monitoring for Autocorrelated Data, International Conference on Statistics, Combinatorics and Related Areas, Wollongong, Australia, December 20, 2001.

5.2.2 General Interest Talks

1. J. J. Filliben (and SED Staff Members), Pi Experiment, NIST Centennial Open House, Gaithersbug, MD, May 10 and May 11, 2001.

2. J.J. Filliben, Basketballs, Funnels, and Designed Experiments,

3. N.F. Zhang, Member of the Panel "Careers in Statistics" at University of Maryland sponsored by Department of Mathematics and Career Center of University of Maryland, April 4, 2001.

5.2.3 Workshops for Industry

1. W.F. Guthrie (with R. Watters): Workshop on Uncertainty Estimation in Chemical Measurements, Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy, New Orleans, LA, March 3, 2001.

2. W.F. Guthrie (with R. Watters): Workshop on Uncertainty Estimation in Chemical Measurements, SIM Chemical Working Group Meeting, Santiago, Chile October 16–17, 2001.

3. H.K. Liu, G. Harris (TS), Workshop on Advanced Mass Measurements, given at NIST, Gaithersburg, MD, October, 2001.

5.2.4 Lecture Series

1. J. J. Filliben, Statistics Tutorial and e-Handbook, Combined-Regional Measurement Assurance Programs (RMAP) Metrology Meeting, NIST, Gaithersburg, MD, 3 half-day Lectures: March 21 and March 22, 2001.

2. J. J. Filliben, Experiment Design for Hazard Identification & Reduction, Consumer Product Safety Commission, Bethesda, MD, 4-Day Workshop: March 15, March 27, April 12, April 26, 2001.

3. J. J. Filliben, Design of Experiments, 2 Lectures to SED Summer Interns, Gaithersburg, MD, August, 2001.

4. J.H. Yen, Meta-analysis and Combining Information, NIST, Gaithersburg, MD, May 1-2, 2001.

5.3 Professional Activities

5.3.1 NIST Committee Activities

1. A.I. Aviles, Member, NIST Employees Concerned with Disabilities (ECD).

2. K.J. Coakley, Member, ITL Awards Committee

3. K.J. Coakley, Member, Boulder Editorial Review Board

4. N. Sedransk, Member of NIST Task Force on Traceability.

5. N. Sedransk, Member of Measurement Services Group.

6. J.D. Splett, Member, ITL Diversity Committee.

7. C.M. Wang, Member, EEEL MCOM subcommittee on polarization-mode dispersion SRM.

8. C.M. Wang, Member, EEEL MCOM subcommittee on MAP for polarization dependent loss.

9. J.H. Yen, Member, NIST 2010 Information/Knowledge Management Strategic Focus Area.

10. N.F. Zhang, Member, EEEL, MCOM subcommittee on AC-DC Difference of Voltage.

5.3.2 Standards Committee Memberships

1. N.F. Zhang, Project Leader of PDTS 21749 of ISO/TC/69 on Application of Statistical Methods.

2. N.F. Zhang, Liaison between ISO/TC69/SC6 and ISO/REMCO on Reference Materials.

5.3.3 Other Professional Society Activities

1. W.F. Guthrie, Secretary, ASA Section on Quality & Productivity, 2001–2004.

2. W.F. Guthrie, Publications Chair, ASA Section on Physical & Engineering Sciences, 2001–2002.

3. C. Hagwood, Chair, 11th INFORMS Applied Probability Society Conference, New York City, July 25-27, 2001. 1994–2000.

4. G.L. Yang, Committee on Fellows, Institute of Mathematical Statistics.

5. G.L. Yang, COPSS Committee (Committee of Presidents of Statistical Societies) to nominate the F. N. David award.

5.4 Professional Journals

5.4.1 Editorships

1. K.J. Coakley, Editor of Special Issue, International Journal of Imaging Science and Technology.

5.4.2 Refereeing

1. A.I. Aviles, Technometrics.

2. K.J. Coakley, International Journal of Imaging Systems and Technology, Weather and Forecasting, Biometrics.

3. W.F. Guthrie, *Metrologia*

4. H.K. Liu, *Journal of Multivariate Analysis*.

5. N. Sedransk, *Metrologia*.

6. C.M. Wang, Journal of Statistical Computation and Simulation, Journal of Quality Technology, Psychometrika, Technometrics.

7. J.H. Yen, Statistics and Probability Letters.

8. N.F. Zhang, Journal of Quality Technology, Statistics.

5.5 Proposal Reviewing

1. A.I. Aviles, National Science Foundation, Course Curriculum and Laboratory Improvement Division.

2. S.D. Leigh, NIST Advanced Technology Program.

3. N. Sedransk, National Science Foundation.

5.6 Honors

1. H.K. Liu, Outstanding Contribution Award for Valued Contributions to ITL as a Member of the ITL Diversity Committee, January, 2001.

2. N.F. Zhang, Cash Award and Certificate from NIST Civil Rights Office.

5.7 Trips Sponsored by Others and Site Visits

1. N.F. Zhang, visit to CSIRO in Sydney and Melbourne, Australia, December 13-18, 2001.

5.8 Training & Educational Self-Development

1. A.I. Aviles, Assistive Technology for Blind/Low Vision Employee (include Voice Note QT, CCTV, and ZoomText), Gaithersburg, MD, December 21, 2001.

   Managing Diversity, NIST, Gaithersburg, MD, January 5–6, 1995.

2. H.K. Liu, English as a Second Language, NIST, Gaitherburg, December, 2000 – January, 2001.

3. H.K. Liu, Microsoft Project 98 Introduction, Gaitherburg, September, 2001.

4. C.M. Wang, Bayes and empirical Bayes methods for data analysis, Atlanta, GA, August 7, 2001.

5. N.F. Zhang, Attended 14 management training and other training courses.

5.9 Special Assignments

1. A.I. Aviles, Co-Chair (with A.I. Khuri), Conference on Designs for Generalized Linear Models (GLMs), (Gaithersburg, MD, April 2002).

2. A.I. Aviles, Co-Principal Investigator (with A.I. Khuri), NSF Proposal: Grants for Students and Junior Scientists to Attend the Conference on Designs for Generalized Linear Models (DMS-0207059).

3. C. Hagwood, Search Committe Member, NICHD.

4. S.D. Leigh, SED liaison for NIST/NRC postdoctoral associateship program.