

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, 2003

**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 2003

Covering Period: January 2002 – December 2002

Publications: January 2001 – December 2002

Covers:
With homage to Leestemaker and Brueghel,

Latin Square : Life in SED

Contents

1	Division Overview	6
2	Staff	8
3	Project Summaries	9
3.1	Bayesian Metrology	10
3.1.1	ITL : Bayesian Metrology – Overview	10
3.1.2	PL : Bayesian Analysis of CCPR Key Comparison on Near-Infrared Spectral Responsivity	12
3.1.3	ITL : MCMC in StRD	15
3.1.4	CSTL : Dynamic Calibration	18
3.1.5	ITL : Bayesian Modeling in Inverse Problems	20
3.1.6	ITL : Bayesian Methods for Combining Data: Using Independence, Common Mean, Hierarcarchical and Partitions Models	23
3.1.7	ITL : Parameter Design for Measurement Protocols by Latent Variable Methods	28
3.2	Key Comparisons	34
3.2.1	ITL : International Key Comparisons and Uncertainty Principles	34
3.2.2	ITL : Uncertainty Analysis for Key Comparisons with Trends	36
3.2.3	ITL : A Robust Key Comparison Reference Value in Cases of Dominant Type B Error	40
3.2.4	EEEL : Two New Estimators of the Variance of the Graybill-Deal Estimator of a Common Mean	43
3.2.5	ITL : Some Statistical Methods Applicable to Key Comparisons Studies	48

3.2.6	ITL : Models and Confidence Intervals for True Values in Interlaboratory Trials	52
3.2.7	ITL : Simulation Study of Estimation Procedures for Key Comparisons	56
3.3	IT Performance	59
3.3.1	ITL : Statistical Analysis and Prediction of Extreme Network Performance	59
3.3.2	EEEL : Modeling the Recovery Process for Notification and Polling .	67
3.3.3	ITL : Fusion of Biometric Algorithms	70
3.4	Process Characterization	75
3.4.1	ITL : Process Characterization - Overview	75
3.4.2	EEEL : Characterization of High-Speed Optoelectronic Devices . . .	76
3.4.3	EEEL : Properties of Dielectric Materials	83
3.4.4	EEEL : Residual Resistivity Ratio Metrology for Superconductors . .	90
3.4.5	PL : Stochastic Approximation using Twin Processes	94
3.4.6	PL : Lifetime of Magnetically Trapped Neutrons	96
3.4.7	ITL : Cryogenic Detection of Weakly Interacting Particles	100
3.4.8	PL : Neutron Detector Calibration	104
3.4.9	CSTL : Consistency of Nuclear Methods for Thin Film Analysis . . .	108
3.5	Measurement Services	112
3.5.1	BFRL : Range Imaging and Registration Metrology	112
3.5.2	BFRL : Sulfate Performance Prediction for Infrastructure Abatement	115
3.5.3	CSTL : Half-life of Arsenic-76	118
3.5.4	MEL : Effect of PAC Tube Cooling on Machine Tool Thermal Deformation	122
3.5.5	EEEL : Standard Reference Materials	125
3.5.6	CSTL : Army CCG Project 474 Gas Mask Verification	128
3.5.7	CSTL : SRM 2396–Oxidatively-Modified DNA Biomarkers	131
3.5.8	CSTL : IAEA Isotope Reference Materials Intercomparison: Carbon and Oxygen	135

3.5.9	MSEL : Charpy V-notch Reference Value Uncertainty	137
3.5.10	CSTL : Standard Reference Materials for the Food Industry	139
3.6	New Methods for Metrology	141
3.6.1	CSTL : Errors in Variables for Gas Standard Calibration	141
3.6.2	ITL : Generalized Tolerance Intervals	145
3.6.3	BFRL : Consensus Curve Estimation	147
3.6.4	ITL : A Study on the Variance Estimation for a Stationary Process in SPC	149
3.6.5	EEEL : SVD-based Structural Approach For Locally Weighted Re- gression	154
3.6.6	ITL : A Tutorial Argument for Orthogonal Experiment Designs . . .	157
3.7	Web Products	162
3.7.1	ITL : NIST/SEMATECH e-Handbook of Statistical Methods	162
3.7.2	EEEL : HELP for Missing Data	166
3.7.3	ITL : Development of a Web Application for Statistical Analysis . . .	168
3.7.4	ITL : Development of a Bayesian Software Library	170
3.7.5	ITL : Web Database Project	172
4	Special Programs	175
4.1	International Activities	175
4.1.1	International Organization for Standardization (ISO)	175
4.1.2	Hands-On Workshop for SIM Members	176
4.1.3	SED Visits to AIST of Japan	176
4.1.4	Clinical Biochemistry	177
4.1.5	American Society of Mechanical Engineers	177
4.1.6	CIPM/CCM Working Group for Fluid Flow	177
4.2	Education	179
4.2.1	Education and Training	179
4.2.2	Conference on Designs for Generalized Linear Models	185

4.2.3	Summer Students Program	188
4.2.4	Minority Internship Announcement	193
4.3	New Staff	196
4.3.1	Juan Soto	196
4.3.2	Dennis Leber	197
4.3.3	Bill Strawderman	198
5	Staff Publications and Professional Activities	199
5.1	Publications	199
5.1.1	Publications in Print	199
5.1.2	NIST Technical Reports	201
5.1.3	Publications in Process	202
5.1.4	Working Papers	205
5.1.5	Acknowledgements in Publications	206
5.2	Talks	206
5.2.1	Technical Talks	206
5.2.2	General Interest Talks	208
5.2.3	Lecture Series	208
5.3	Professional Activities	209
5.3.1	NIST Committee Activities	209
5.3.2	Standards Committee Memberships	209
5.3.3	Other Professional Society Activities	210
5.4	Professional Journals	210
5.4.1	Editorships	210
5.4.2	Refereeing	210
5.5	Review Panels	211
5.6	Honors	211
5.7	Trips Sponsored by Others and Site Visits	211

5.8 Training & Educational Self-Development	211
5.9 Special Assignments	212

1. Division Overview

Nell Sedransk, Chief
Statistical Engineering Division, ITL

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

The role of SED extends across NIST; SED staff actively collaborate with more than 90% of the scientific Divisions at NIST, both on the Gaithersburg and Boulder campuses, and provide statistical support for some of the administrative offices as well. Basic collaborations include core support to provide a statistical basis for certification for Standard Reference Materials produced at NIST, statistical methodology and documentation for NIST calibration services, and education and training of NIST scientists and engineers in the implementation of appropriate statistical methodology.

As members of multidisciplinary teams, SED staff collaborate more fundamentally to scientific research at NIST to define research objectives, to formulate statistical strategies and develop statistical methods for process characterization and to analyze experimental data. The statistical expertise central to a particular multidisciplinary research project may lie in any of many sub disciplines of statistics (including experimental design, generalized linear models, stochastic models, Bayesian inference, time series analysis, reliability analysis, statistical signal processing, image analysis, spatial statistics, quality control, exploratory data analysis, 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 scientist colleagues.

SED also contributes internationally to metrology through the development of statistical methodology and statistical tools for metrologists to use worldwide. Increasing attention to international intercomparisons and international acceptance of standards from national metrology laboratories has increased the prominence of statistics in metrology. Problems unique to metrology as well as problems unique to NIST that require extraordinarily high precision in their formulation often necessitate innovations in statistical methodology. This fundamental statistical research at SED, whether in probabilistic modeling, in design of experiments, in theory and methodology of inference, in computationally intensive statistical tools or in Bayesian inference and modeling, expands the statistical methodology available to NIST scientists, to US industry and to metrolo-

gists worldwide. This research also contributes in a fundamental way to the discipline of statistics.

The educational role of SED within NIST extends to development and presentation of short courses and workshops on topics in statistical methodology. These 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. Increasing attention is being given to making statistical tools available via both internal NIST and external NIST web pages. A highly successful program of statistical research opportunities gives undergraduate and graduate students the chance to explore the sub discipline of statistical metrology while contributing to SED activities.

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

This report provides technical summaries of some of the significant projects undertaken during the year 2002. The projects presented here have been selected to provide a sampling 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, Ph.D. *Division Chief*

Metrology Statistics and Computation

Nien Fan Zhang, Ph.D. *Group Manager*

Will Guthrie, M.S.

Charles Hagwood, Ph.D.

Alan Heckert, M.S.

Walter Liggett, Ph.D.

John Lu, Ph.D.

Juan Soto, M.S.

Statistical Modeling and Analysis

James Filliben, Ph.D. *Group Manager*

Ivelisse Aviles, Ph.D.

Dennis Leber, M.S.

Stefan Leigh, M.S.

Hung-kung Liu, Ph.D.

Andrew Rukhin, Ph.D.

Blaza Toman, Ph.D.

James Yen, Ph.D.

Boulder Statistics

Jack Wang, Ph.D. *Group Manager*

Kevin Coakley, Ph.D.

Jolene Splett, M.S.

Visiting Faculty and Guest Researchers

Duane Boes, Ph.D. *Colorado State University*

M. Carroll Croarkin, M.S. *NIST Guest Researcher*

Dipak Dey, Ph.D. *University of Connecticut*

Hari Iyer, Ph.D. *Colorado State University*

Don Malec, Ph.D. *Bureau of Census*

Joan Rosenblatt, Ph.D. *NIST Guest Researcher*

Tom Ryan, Ph.D. *University of Michigan*

Bill Strawderman, Ph.D. *Rutgers University*

Dominic Vecchia, Ph.D. *NIST Guest Researcher*

Grace Yang, Ph.D. *University of Maryland*

Administrative Staff

Stephany Bailey

Mary Clark

Lorna Buhse

3. Project Summaries

3.1 Bayesian Metrology

3.1.1 Bayesian Metrology – Overview

Blaza Toman

Statistical Engineering Division, ITL

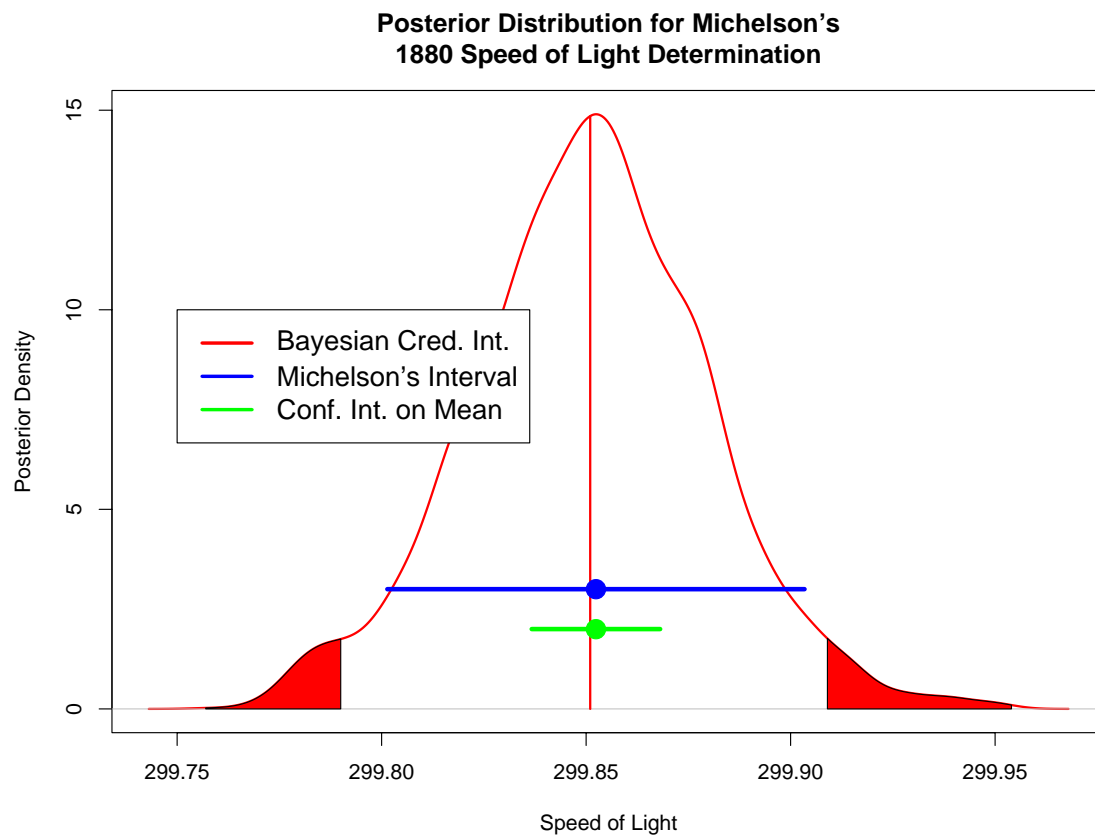


Figure 3.1: Bayesian analysis of Mickelson speed of light data.

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 four years, Bayesian methods have been applied in collaborations with virtually every NIST Laboratory.

The first objective of the project is to develop within SED a cadre of Bayesian statisticians. This has been accomplished so that now most members of SED are successfully applying Bayesian methods to their projects. 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), statistical modeling of Type B error, multivariate analysis, 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 many professional presentations and numerous publications in professional statistical journals. The goal of integrating Bayesian methodology into NIST practice is also being met; thus far, a number of 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 and Bayesian Computation Using BUGS was very well attended. 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.

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

Blaza Toman

Statistical Engineering Division, ITL

Steven Brown, Thomas Larasen

Optical Technology Division, PL

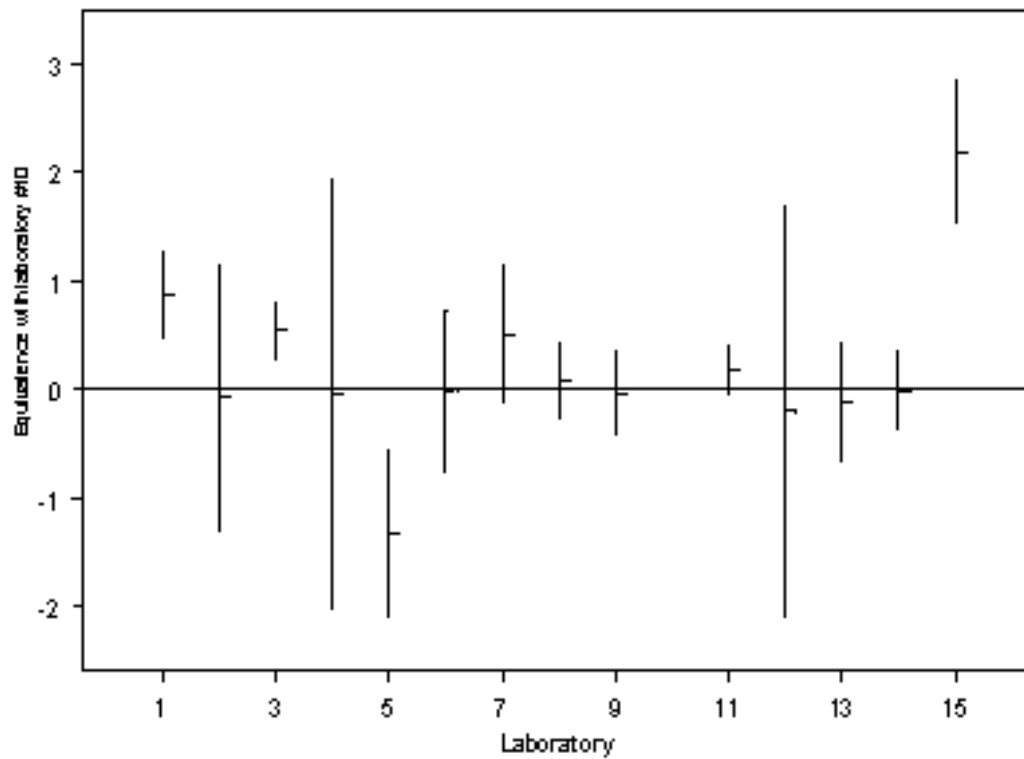


Figure 3.2: As an example of the analysis, the following graph gives the 95% HPD for $\alpha_i - \alpha_{10}$ for the fifteen laboratories.

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 main goal of the statistical analysis was to assess the agreement between laboratories.

The information provided in the key comparison consisted of an average measurement \bar{y}_{ij} (average of n_{ij} measurements), the type A uncertainty represented by the standard deviation $s_{ij}/\sqrt{n_{ij}}$, and the type B uncertainty, call it τ_{ij} , for each laboratory i and photodetector j . The physical situation can be approximated by the following model:

$$\bar{Y}_{ij} | \delta_{ij}, b_{ij}, \sigma_{ij}^2 \sim N(\delta_{ij} + b_{ij}, \sigma_{ij}^2 / n_{ij}) \quad i = 1, \dots, k$$

and

$$\delta_{ij} | m_{ij}, \omega_{ij}^2 \sim N(m_{ij}, \omega_{ij}^2)$$

$$b_{ij} | \tau_{ij}^2 \sim N(0, \tau_{ij}^2).$$

In this model,

$$\delta_{ij} = \mu + \alpha_i + \beta_j$$

where μ represents the overall mean responsivity value, the α_i represents a possible (but not known) bias or systematic error of the i th laboratory, and the β_j represents a possible bias or systematic error due to the j th photodetector. The b_{ij} represents the Type B error (the known systematic error) of the laboratory i for photodetector j , and σ_{ij}^2 represents the usual measurement error variance (Type A). The model can be written in matrix form as

$$\underline{Y} = 1\mu + F\underline{\alpha} + G\underline{\beta} + \underline{b} + \underline{\varepsilon}.$$

The F and G are design matrices which identify, for each element of Y , the laboratory (in F) and the photodetector (in G). Using noninformative prior distributions for the δ_{ij} , after integration over the nuisance parameters β_j , the laboratory – to – laboratory comparisons $\alpha_i - \alpha_{i'}$ can be estimated by the posterior means

$$c_{ii'} \left(F' V_0^{-1} F \right)^{-1} F' V_0^{-1} Y.$$

In this expression $c_{ii'} = (0, 0, \dots, 1, 0, \dots, 0, -1, 0, \dots, 0)$ with the 1 in the i th position and the -1 in the i' th position, and E is a diagonal matrix with entries $\tau_{ij}^2 + s_{ij}^2/n_{ij}$. Further,

$$V_0^{-1} = E^{-1} - E^{-1}G \left(G'E^{-1}G \right)^{-1} G'E^{-1}.$$

The posterior standard deviations of the $\alpha_i - \alpha_{i'}$ are

$$\sqrt{c_{ii'} \left(F'V_0^{-1}F \right)^{-1} c_{ii'}}.$$

There are several aspects of this analysis which are innovative and should prove useful in other key comparisons. Most importantly, the type B uncertainty is carefully differentiated from the type A uncertainty. Further, the experimental design contains blocking by the photodetectors and this is taken into account in the estimation of the laboratory – to – laboratory effects. Finally, the analysis is completed without the use of a KCRV.

3.1.3 MCMC in StRD

Hung-kung Liu, Will Guthrie, Don Malec, Grace Yang
Statistical Engineering Division, ITL

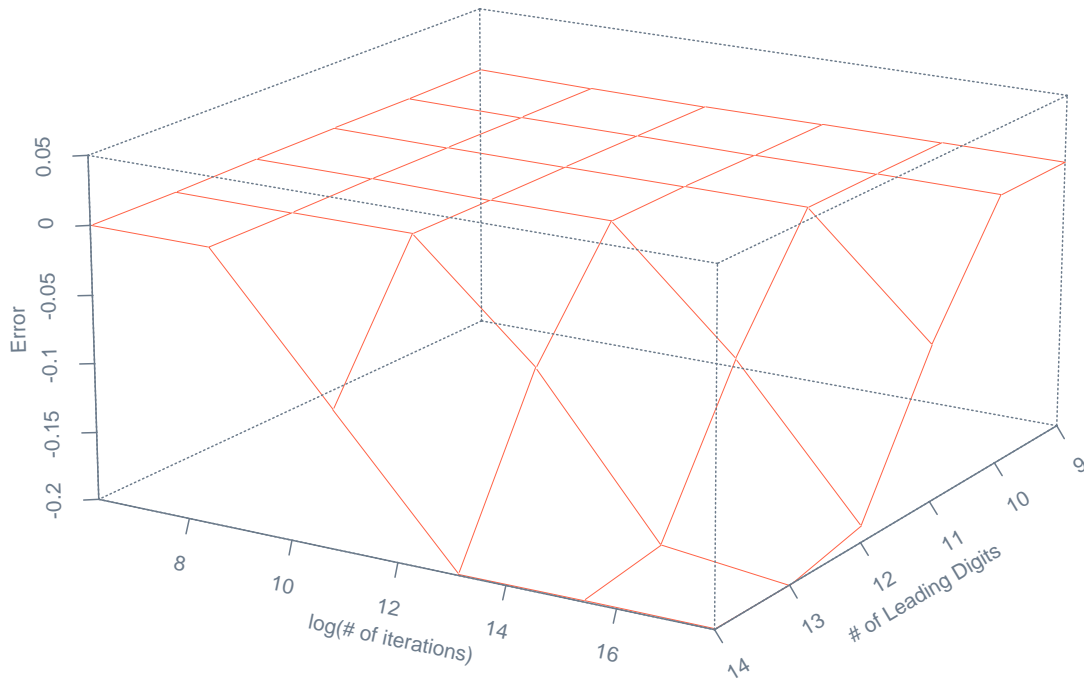


Figure 3.3: The difference between the MCMC estimate of the posterior mean and its theoretical value is plotted versus the number of leading digits common to each data point and the number of MCMC iterations.

The numerical accuracy of statistical software has been of concern to statisticians since computers started to become widely available in the 1960's. Numerical inaccuracies caused by floating point arithmetic, although often not important, can change the conclusions of an analysis. Computational accuracy is of even more concern today because the number of software packages has exploded as computers have evolved and statistical software is increasingly written and used by non-statisticians who may not be aware of potential computational problems.

To address this problem, SED developed the Statistical Reference Datasets (StRD) web site (<http://www.itl.nist.gov/div898/strd/index.html>) which provides datasets with certified values for assessing the numerical accuracy of software. Four areas of statistical computation were originally addressed, univariate statistics, linear regression, nonlinear regression, and analysis of variance. Recently Bayesian analysis using Markov chain Monte Carlo (MCMC) has become popular and is a new area in which intensive statistical computations are used. Despite its importance, however, the numerical accuracy of the software for MCMC is largely unknown.

To help users and software developers assess the numerical accuracy of MCMC software, we have constructed new StRD datasets to assess its numerical accuracy using a simple model. Let $\{Y_1, Y_2, \dots, Y_n\}$ denote the values in a dataset. We assume that the Y_i are independent and identically distributed normal random variables with unknown mean μ and unknown variance σ^2 . We assume further that μ and σ^2 are independent random variables such that μ has an improper uniform prior density with respect to the Lebesgue measure and σ^2 has a density function proportional to $1/\sigma$. Certified values are then obtained for the dataset using theoretical derivations. The numerical accuracy of MCMC software can then be assessed by comparing the MCMC results with the certified values. Using these datasets computational inaccuracies from at least five potential sources may be identified. These include:

- truncation error;
- cancellation error;
- accumulation error;
- simulation error;
- random number generation error.

Using the benchmark of Simon and Lesage (1989), our generated datasets are designed to have from seven to fourteen constant leading digits to allow computational accuracy to be examined at different stiffness levels. Currently all our datasets have eleven observations, although larger datasets, which may identify a different source of accumulation error, are also planned.

To illustrate the results that may arise from testing typical MCMC software, we compute the posterior mean and variance of μ and σ^2 using these datasets and a univariate Gibbs

sampler MCMC. The number of MCMC iterations used ranged from 500 to 50M and increased geometrically. In the plot above, the difference between the MCMC estimate of the posterior mean and its theoretical value is plotted versus the number of leading digits common to each data point and the number of MCMC iterations. The plot clearly shows where the computational accuracy starts to break down. The majority of the errors appear to come from the accumulation of individual errors in representing the MCMC samples/draws from the posterior distribution. Interestingly, this error can be minimized by not letting the simulation run too long.

Using datasets like these new additions to the StRD web site, users and developers of MCMC software will be better able to assess the numerical accuracy of their computations, helping ensure the accuracy of scientific and engineering conclusions drawn from statistical analyses. Software testing will also offer additional insight into the properties of MCMC algorithms and their implementations, furthering the use of Bayesian methods for statistical modeling.

3.1.4 Dynamic Calibration

Charles Hagwood

Statistical Engineering Division, ITL

Industrial quality is governed by the ability to control errors in mechanical equipment. To achieve error control of any quality in a measurement system, one must first determine the relationship between the system's output and the actual value of the quantity being measured. Calibration is this process. To calibrate an instrument is to compare its output with a known input. Most calibrations done at NIST are based on comparisons between a NIST primary standard and an industry secondary standard. Primary standards are instruments whose measurements are traceable to the System International (SI) unit, mass, length, time, electric current, thermodynamic temperature, amount of substance, luminous intensity. Primary standards are held and maintained by designated labs, usually national labs. NIST is the main primary standards laboratory in the United States. Secondary standards are one tier lower instruments. They must be periodically recalibrated directly against a primary standard and are used commercially to calibrate other instruments. So, an instrument calibrated at NIST is often calibrated and recalibrated several times during its useable lifetime.

The calibration services of the National Institute of Standards and Technology (NIST) are designed to help the makers and users of precision instruments achieve the highest possible levels of measurement quality and productivity. Since all U.S. calibration labs trace their uncertainty back to NIST, all effort at NIST is made to ensure accurate uncertainty statements. One effort being considered jointly by the Statistical Engineering Division and the Pressure Measurements Division is to incorporate all of an instrument's prior calibration data in determining a final calibration statement. As of now, prior calibrations are not used. Harrison and West (1994) developed a tool to deal with dynamically evolving data collection called the dynamic linear model. Like the Kalman filter, the dynamic linear model is a recursive Bayesian algorithm that allows the parameters of the model to evolve over time, in the end producing an estimate of the predictive density. The mean of the predictive density is the calibration estimate and a Bayesian credible interval is used to compute uncertainty. The previous calibration data serve as prior information. The dynamic calibration model developed is based on Harrison and West's dynamic linear model. The results are illustrated with the calibration of a transducer gage by direct comparison to the NIST UIM (Ultrasonic Interferometer Manometer). This transducer was calibrated and recalibrated in 1990, 1992, 1994, 1996, 1999.

The dynamic linear calibration model is

$$\begin{aligned} Y(t) &= X(t)\theta(t) + \lambda\epsilon(t) & \epsilon &\sim N(0, I_{n_t}) \\ \theta(t) &= \theta(t-1) + \lambda\omega(t) & \omega(t) &\sim N(0, W_t) \end{aligned}$$

Initial Conditions

$$\theta(t_0), \lambda \sim Ng(m_0, C_0, \alpha_0, \beta_0)$$

where Ng denote the normal gamma density, $\lambda\omega(t)$ is a known random vector, independent of $\epsilon(t), \theta(t-1)$, representing the instruments degradation or increase in uncertainty

about its values between times $t - 1$ and t . $Y(t)$ denotes the manometer values at calibration time t and $X(t)$ denotes the transducer's values.

Dynamic Calibration versus Static Calibration (torr)

<i>manometer</i>	<i>transducer</i>	halfwidth (static)	halfwidth (dynamic)
0.00322	0.00335	0.0159207	0.009266
0.02525	0.02587	0.0157891	0.009245
0.10391	0.10504	0.0157397	0.009243
0.64223	0.64572	0.0157865	0.009253
1.01895	1.02425	0.0158106	0.009262
6.33155	6.36687	0.0161406	0.0093235
7.77857	7.82111	0.0166988	0.009361
10.1962	10.2503	0.0167308	0.009499

Here a pressure transducer has been calibrated and recalibrated against a NIST manometer in 1990, 1992, 1994, 1996, 1999. The static calibrations done at the NIST Pressure Measurements Division only take into account data taken at the calibration session. Dynamic calibration takes into account not only the comparison data taken at the calibration session, but all prior calibration data. The table above is taken from a cross validation study that compares the static calibration procedure with the dynamic calibration procedure. One row is deleted from year 1999 and the remaining 1999 data are treated as data from the calibration session. The table shows how well the static model performs versus dynamic model in terms of predicting the value of the manometer for the deleted row. The dynamic calibration model always produces a shorter confidence interval for the manometer value.

The goals of this project are: (1) To convince Dr. Albert Lee and Dr. Archie Miller of the Pressure Measurements Division that dynamic calibration can be superior to static calibration. (2) To develop software so that calibration statements can be automatically updated when instruments are brought in for recalibration.

3.1.5 Bayesian Modeling in Inverse Problems

Z.Q. John Lu, Kevin J. Coakley
Statistical Engineering Division, ITL

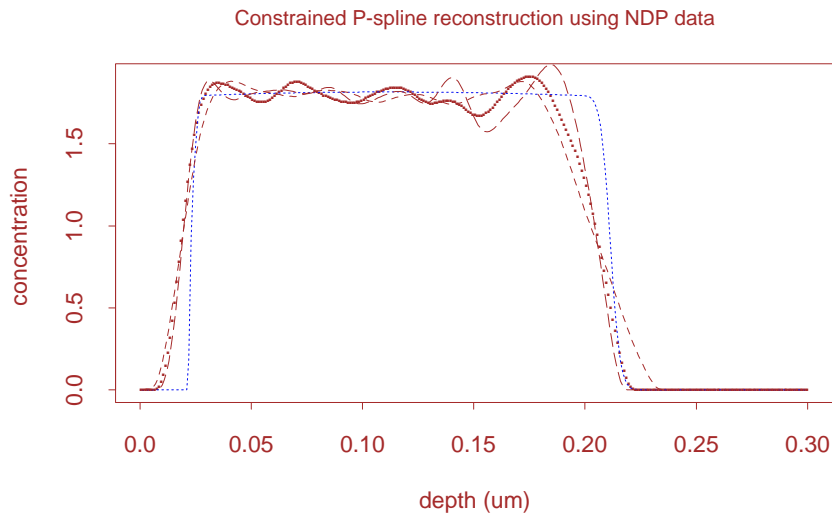


Figure 3.4: Constrained P-spline reconstructed concentration profiles using different amount of smoothing parameter (dashed, curly lines, in red color), and comparison to profile from a direct measurement (dotted line, in blue color).

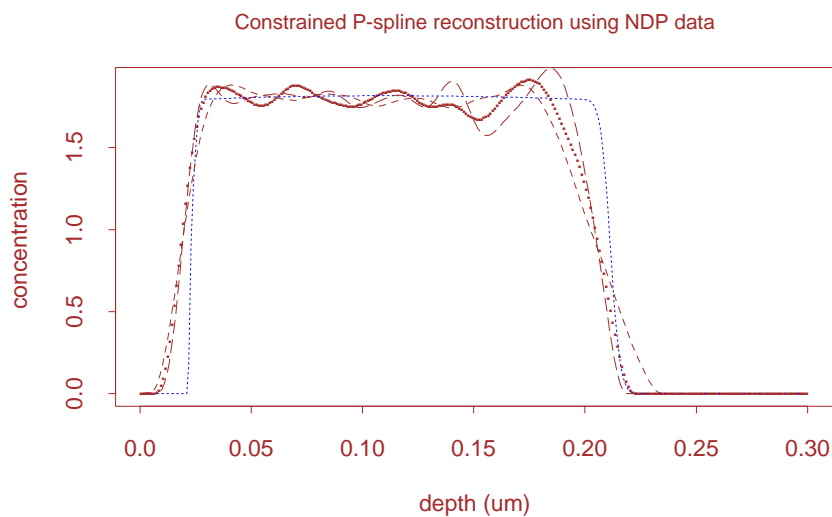


Figure 3.5: Predictions from reconstructed concentration profiles (lines in red color) and original data (dots in blue color).

Modern measurement technologies have led to the statistical problem of reconstruction of curves (profiles) or images based on indirect high-throughput measurements. Often such problems are ill-posed in the sense that there are many solutions that may fit the data perfectly well, even though most of the solutions may be very "unphysical", or too noisy. The need to incorporate physical information such as smoothness or boundary conditions and other prior information has made the Bayesian approach appealing. However, there is often not enough prior information to impose a proper probability model, consequently development of suitable probabilistic models for the underlying functional curves or image based on partial physical or prior knowledge is a challenging issue in implementation of the Bayesian approach. The benefits from accomplishing such a demanding task are also great, however, as uncertainty or precision of the reconstructed solution or calibration can be derived naturally by the posterior distribution on the underlying curve, which may be evaluated through numerical or multivariate simulational methods. In this work we focus on the development of Bayesian regression modeling methods, especially the constrained P-spline based regression techniques for concentration profiles. Because of the flexibility of regression modeling, various prior information may be easily modeled. We demonstrate this approach with neutron depth profiling (NDP) technique, a non-destructive method for measuring the spatial concentration profile of an element in a material.

A characteristic of many measurement methods in use today is indirect measurement, measuring what is observable and relating them to the true quantities (measurand) of interest based on a physical model. S. Twomey (1977) described many such examples in his now classic book *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements*. One advantage of such high-throughput measurements is that many measurements can be made cheaply. However, due to the ill-posed nature of the relationship between measurements and measurand, there is an intrinsic limit in the information content of indirect measurements. An example is the blurring process in optical measurement or remote sensing, in which measurements are related to the profile function $f(x)$ through an integral

$$y(\lambda_i) = \int K(\lambda_i, x) f(x) dx \quad (3.1)$$

for $i = 1, 2, \dots, n$, where $K(\lambda, x)$ denotes a kernel function. An example is $K(\lambda, x) = k(\lambda - x)$, where k is a integrable symmetric density function that attains its maximum at 0 and tends to zero rapidly in the tails. The integral operation is a "convoluted" version of the true signals and has the smoothing effect of reducing the high-frequency components of $f(x)$, thus making recovering certain high-frequency components of $f(x)$ difficult. Furthermore, in practice y is, of course, measured with error. Indeed, in the NDP experiment (Coakley et al 2002 in this volume), $y_i = y(E_i)$ at a given frequency in the energy spectrum is modeled as Poisson distributed $P(\mu(E_i))$, where the mean $\mu(E_i)$ is modeled through a probability transition matrix $p(E_i, x_k)$ for $i = 1, \dots, n = 391; k = 1, \dots, m = 300$ such that

$$\mu(E_i) = \sum_{k=1}^{300} p(E_i, x_k) f(x_k) \quad (3.2)$$

for $i = 1, \dots, n$. The goal is to estimate $f(x)$ as a function on $[0, 0.3]$ based on y_1, \dots, y_{391} .

The probability transition matrix $p(E_i, x_k)$ plays the role of the kernel function K in (3.1). To see the ill-conditioned nature of the problem, one can compute the singular value decomposition of the 391 by 300 matrix

$$[p(E_i, x_k)]_{i=1,\dots,n;k=1,\dots,m}$$

It turns out that the singular values $\delta_1 \geq \dots \delta_{300}$ decay to zero very rapidly. Actually, the ratio $\delta_i/\delta_1 < 10^{-3}$ for $i \geq 28$. Thus, the components in the eigen-directions corresponding to $\delta_i, i \geq 28$, will be difficult to recover from data alone. Additional information is necessary and this has to come from elsewhere such as physical constraints. To generate a large class of flexible models that allow easy incorporation of physical and partial prior information, we consider nonparametric regression models, especially spline-based methods. There are several spline model choices such as a regression spline, monotone spline, and smoothing spline. We choose a newer model, called a p-spline regression model, whose complexity is between the regression spline model that is the simplest but requires elaborate choice of knot placement, and the smoothing spline method, a variational method for functional estimation, but which requires solving large linear systems. The physical constraints on a spline-based solution, a linear function of B-spline basis at equally placed knots, include the penalty term of the sum of second-order squared differences on the solution, and non-negativity. A smoothing parameter η is used to control the tradeoff between the penalty term and the prediction fitness (log likelihood) to data. To assess the effect of the smoothing parameter (or the choice of a hyperparameter), we also compute the “perturbed” solution by varying from the chosen smoothing parameter value η_0 , so Figure 3.4 shows three solutions $\hat{f}_{\eta_0}, \hat{f}_{4\eta_0}, \hat{f}_{\eta_0/4}$. This is compared to the “true” profile that is obtained from another independent measurement method. It is a good thing that the solution is relatively robust to the choice of smoothing parameter even in such a wide range. The “noisy” feature in the middle of the solution profile may be traced to the relatively noiser nature in the data (cf. Figure 3.5), where one can see that all three solutions give fairly good predictions, with the chosen one \hat{f}_{η_0} giving the best fit.

In summary, we discussed a Bayesian modeling strategy that should apply to a wide range of inverse problems. We emphasized that model selection cannot be based solely on goodness of fit to data. We discussed the need for prior modeling, for which we proposed some Bayesian nonparametric regression methods. We obtained very satisfactory results on the neutron depth profiling problem. Further work will be needed in order to have more realistic uncertainty assessment due to model selection. We are also interested in statistical comparison of reconstructed profilings using independent measurement methods.

3.1.6 Bayesian Methods for Combining Data: Using Independence, Common Mean, Hierarchical and Partitions Models

Don Malec

Statistical Engineering Division, ITL

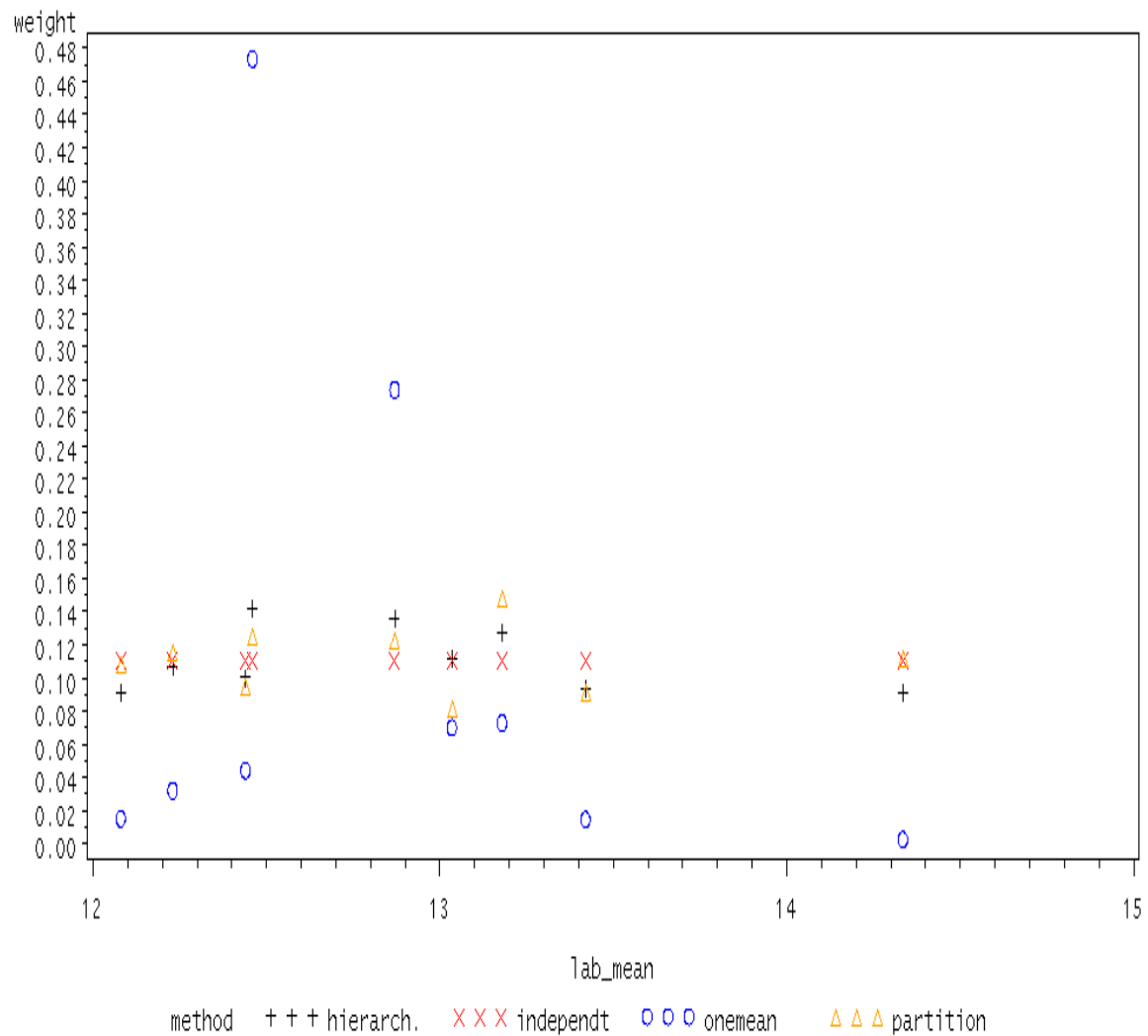


Figure 3.6: Component Weights in the Consensus Mean

In seeking a more accurate estimate of a measurand, its measurements from different laboratories are often combined together and used as a consensus mean. Along with its measure of accuracy, the consensus mean provides the gold standard that accompanies a standard reference material. The data obtained from different laboratories can be combined in a number of ways, resulting in a variety of "consensus means". The simplest method of combining results is to take an average of the laboratory sample means. Next simplest, perhaps, is an average of the laboratory means, weighted inversely by their estimated within lab variance. Both these estimates will be seen to be special cases resulting from a model that postulates a Normal distribution among laboratory effects; referred to as a one-way random effects model. Estimates that are automatically and appropriately weighted over a range of weights can be obtained using this model (e.g., Vangel and Rukhin, Mandel and Paule). Although representing an improvement in estimation methodology, the use of the normally distributed one-way random effects model still includes the major assumption that the lab effects are simply distributed around a central value. This model may not accurately describe what a lab effect actually is, for example unreported type- B errors may cause lab effects to vary in unsystematic ways across labs. A flexible model that would allow labs to vary around a central value in irregular ways may be needed. Such a model is proposed here which includes the one-way random effect model as a special case. The model will provide the ability to automatically downweight a specific lab effect based on its dissimilarity to other lab effects.

The following basic within-lab model is used throughout. Then, four different models to describe the relation between labs effects are presented; each of these represents varying degrees of assumptions.

The basic model accounting for within lab variation assumes that measurements in each lab are an attempt to estimate a true value where

R : number of labs, $i=1, \dots, R$

α_i : the measurand in lab, i , (i.e. the lab effect without replication error),

The targeted consensus mean, $\bar{\alpha}$, will always be defined as :

$$\bar{\alpha} = \sum_{i=1}^R \frac{\alpha_i}{R}$$

As defined here, α_i is an unknown target that can only be estimated from a series of experiments. Let:

n_i = number of observations in lab i , $i=1, \dots, R$.

For lab i , given the population mean and variance, the measurements, y_{ij} , are independent with

$$y_{ij} | \alpha_i, \sigma_i^2 \sim N(\alpha_i, \sigma_i^2), \quad (3.3)$$

In every case examined here, the estimate which minimizes squared error loss (i.e., the posterior mean $E(\bar{\alpha} | \underline{y})$) will be used as the consensus mean.

The statistical relationship among the unknowns, α_i and σ_i^2 , can be independent among laboratories or can exhibit dependencies among the α_i 's. As will be seen, different models for dependencies among the α_i 's will result in different weights being applied to lab averages.

The following outlines four different levels of assumption made on the unknown lab measureands for the purpose of estimating a consensus mean. All priors are chosen to be as non-informative as possible, but are not detailed here.

- The simplest relationship among laboratory measureands is independence; in this case no further structure is placed on the model (3.3).
- Having strong relationships among laboratories that assumes no individual laboratory effect has one mean but possibly different variances:

$$\alpha_i | \mu = \mu \quad i=1, \dots, R,$$

i.e., all labs directly measure the same thing. (As before, no additional structure is specified for the σ_i^2 .)

- Incorporating individual laboratory effects and using the data to ascertain the magnitude of these lab effects leads to:

$$\alpha_i | \mu, \delta^2 \sim N(\mu, \delta^2), \text{ ind.}$$

(As before, no additional structure is specified for the σ_i^2 .)

- Finally, when similar laboratories can be grouped together to form subsets, the groups can be defined in a partition using the following notation:

For the set of integers $\{1, \dots, R\}$ let L be the number of partitions indexed by $g = 1, \dots, L$ (e.g. $\{(1), (2, \dots, R)\}$). Define $d(g)$ to be the number of subsets in partition, g (e.g. if partition g denotes $\{(1), (2, \dots, R)\}$, then $d(g)=2$). A particular subgroup in partition g , will be referred to by $k=1, \dots, d(g)$. Define $S(g)_k$ to be the set of integers in subset k of partition g . (e.g., one could define $S(g)_1 = \{1\}$ and $S(g)_2 = \{2, \dots, R\}$, for the partition $\{(1), (2, \dots, R)\}$). Denote the number of integers in a particular subset, $S(g)_k$, as $m(g)_k$. In order to reference labs in their own, isolated, subset from labs within multi-lab subsets, define (asymmetrically), $U(g)$ to be the set of all integers in an $S(g)_k$ where $m(g)_k=1$ and define $M(g)$ to be all subsets, k , in which $m(g)_k > 1$. Lastly, define the indicator, $k(g, i)$, so that if $i \in S(g)_k$, then $k(g, i) = k$.

In the case of partitioning of the laboratories, for all of the labs in partition g ,

$$\alpha_i \sim N(\mu(g)_k, \delta(g)_i^2) \quad \forall i \in S(g)_k \text{ and } k \in M(g)$$

Define the priors for the parameters of the differential distributions among the lab effects

$$\begin{aligned} \alpha_i &\sim N(\theta(g)_k, \gamma(g)_k^2) \quad \forall i \in U(g) \\ \mu(g)_k &\sim N(\theta(g)_k, \gamma(g)_k^2) \quad \forall k \in M(g) \end{aligned} \quad (3.4)$$

where, $\underline{\sigma}^2 = (\sigma_1^2, \dots, \sigma_R^2)$ and $\underline{\delta}(g)^2 = (\delta(g)_1^2, \dots, \delta(g)_{|M(g)|}^2)$. Note that subsets with only one lab, logically, need one less level of model, since there are no other labs in the subset to compare.

case	distributional assumptions	weight, w_i
independent	α_i distinct σ_i distinct	$\frac{1}{R}$
one-mean	$\alpha_i = \mu$ σ_i distinct	$E \left(\frac{n_i/\sigma_i^2}{\sum_{k=1}^R n_k/\sigma_k^2} \middle \underline{y} \right)$
one-way random effects	$\alpha_i \mu, \delta^2 \sim N(\mu, \delta^2)$ σ_i distinct	$E \left(\frac{(\delta^2 + \sigma_i^2/n_i)^{-1}}{\sum_{k=1}^R (\delta^2 + \sigma_k^2/n_k)^{-1}} \middle \underline{y} \right)$
partition	$\alpha_i g \sim N(\mu(g)_k, \delta(g)_k),$ if $S(g)_{k(g,i)} > 1$ α_i distinct, if $S(g)_{k(g,i)} = 1$ σ_i distinct	$\sum_g \frac{m(g)_{k(g,i)}}{R} E \left(\frac{(\delta(g)_{k(g,i)} + \sigma_i^2/n_i)^{-1}}{\sum_{i \in S(g)_{k(g,i)}} (\delta(g)_{k(g,i)} + \sigma_i^2/n_i)^{-1}} \middle \underline{y}, g \right) p(g \underline{y})$

Table 3.1: Summary of the i-th laboratory weight for the four cases (with vague priors)

For each of the four cases presented above, the posterior mean of $\bar{\alpha}$ takes the form of a weighted average the individual lab means, \bar{y}_i . That is, for each case = C,

$$E(\bar{\alpha} | \underline{y}, \text{case} = C) = \sum_{i=1}^R w(C)_i \bar{y}_i$$

Table 3.1 summarizes the resulting weights, based on vague priors for all the cases.

As can be seen, using the independence model results in a consensus mean consisting of a simple average of all lab means while the one-mean model results in weighting the lab means by their estimated replicate variance. The one-way random effect model results in weights between those of the independence and one mean model with values determined by the relative magnitude of the estimated between lab variability and lab replicate variability. The partition model provides a data-based method to find subsets of the labs where for example the, more powerful, one-mean model can apply. In general, the weights from the partition model take the form of a weighted average of weights from a one-way random effects models that may fit the data better to subsets of labs.

One advantage of the partition model over the one-way random effects model is that the resulting weights can reflect the fact that some labs may be more similar than others. As one example, suppose that the partition $\{(1), (2, \dots, R)\}$ has a high enough posterior probability that the other partitions are negligible. Suppose also that lab 1 has measurements very different from the others but the others are very similar to each other. Assume that the other labs are similar to the extent that $\delta(g)_2$ is essentially zero. In this case the weights are near $1/R$ for lab one and near $\frac{(R-1)*\sigma_i^2/n_i}{R*\sum_{i=1}^2 \sigma_i^2/n_i}$ for the other labs. The first lab has the same weight as the independence model the other labs have weights

lab	replicate measurements
1	12.44, 12.48
2	12.87, 13.20
3	12.21, 12.67
4	12.82, 12.92
5	13.18, 13.66
6	12.31, 11.85
7	13.11, 13.25
8	14.29, 14.38
9	12.08, 12.38

Table 3.2: Apple Fiber Data

proportional to the one-mean model weights.

To illustrate, the assumptions for three of the four cases can be applied to the apple fiber data used by Vangel and Rukhin (1999). Posterior inference about the level and scale of the consensus mean, $\sum_i \alpha_i / R$, cannot be made because there are only two observations in each lab.

In this illustration, the weights in the posterior means as the priors tend to their limit (vague) are used. The weights for these cases are just one over the sample size.

Table 3.2 shows the apple fiber data collected by lab. Figure 3.6 shows the resulting weights based on each model plotted by lab sample mean.

If assumptions of a single mean is correct, the models for both the hierarchical (random) individual laboratory effect and for the partitioning of laboratories should reduce to those of a single mean. As can be seen in the figure this is not true. The partition model, of which all of the models are special cases assigns moderately different weights for some of the labs between a number of different models.

In the partition model, the relative weights will depend on both the within lab variances and on the sample lab means. It appears that the lab with the largest sample mean is an outlier.

By viewing the consensus mean as a simple average of unknown lab measurands, a Bayesian approach can incorporate differing assumptions to describe types of lab similarity. In particular, a simple average of lab means and lab means weighted by their replicate variances are seen to both result from model assumptions that can be relaxed with either the one-way random effects model or the partition model.

The partition model is still an experimental approach for estimating a consensus. More work assessing the effects of prior specifications in various settings is needed to fine-tune the method. Ultimately, the partition model may provide a data-based method for producing estimates of the concensus mean that automatically accounts for outliers and the resulting errors in misidentifying outliers.

3.1.7 Parameter Design for Measurement Protocols by Latent Variable Methods

Walter Liggett

Statistical Engineering Division, ITL

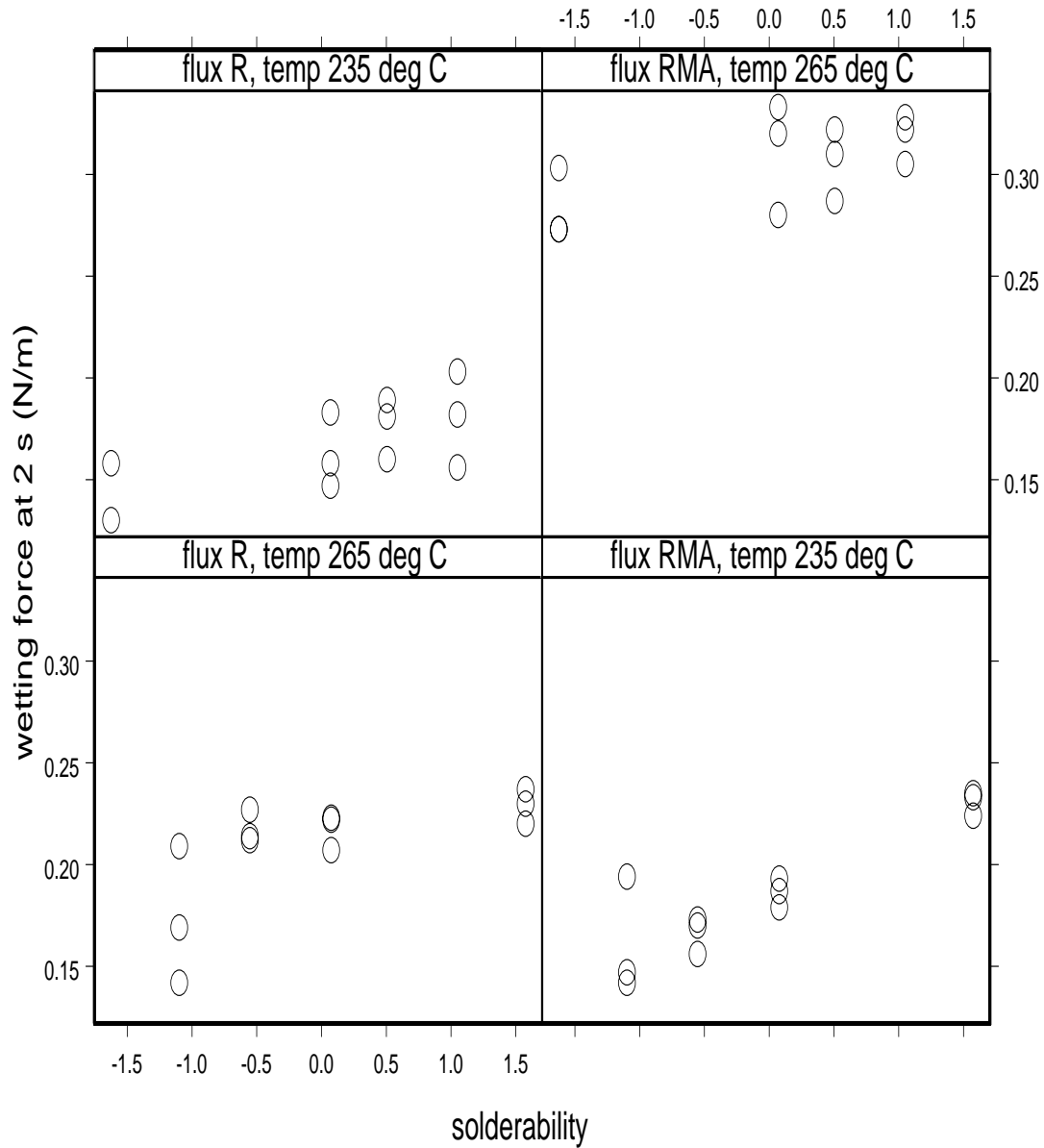


Figure 3.7: Force Measurements Versus Maximum Likelihood Estimates of the Solderability, θ_i .

Engineering projects such as those intended to improve product quality generally start with choice of a measurement system. If what is to be measured is a familiar quantity such as mass or length, then this choice may be between using an instrument at hand or buying a new instrument from a vendor. Gauge studies for choosing instruments to measure familiar quantities are usually easily undertaken. The hardness of a metal specimen and the solderability of the leads on an electronic component are important properties but not ones that are familiar as numerical quantities. The reason is that they are only loosely defined in terms of scientific theory and consequently, not ones for which there are ideal measurement methods, even in principle. Measurement of such properties is based on specification of a protocol, one that gauges indentation of the specimen in the case of hardness and one that gauges wetting of the lead by hot solder in the case of solderability. Because such specification involves many choices, there are in fact many measurement protocols, some of which may be commonly accepted. Measurements based solely on the protocol are sometimes called test methods instead of measurements. The purpose of this project is development of a statistical approach to optimizing test methods.

A distinguishing feature of parameter design for measurement systems is that one must know something about the experimental units in addition to the observed responses they provide. If one knew the value of the property of interest for each experimental unit and if one could choose any unit as an experimental unit, then parameter design for the measurement system would be like other parameter designs. But it is often not true that one can obtain experimental units with known values and even so, there would likely be differences between the available experimental units and the larger population of unknown units. Some authors assume that the values of the property of interest for the experimental units are known. Others assume that the experimental units can be remeasured with different protocols. In this paper, we do not make these assumptions. We do, however, assume that the experimental units come in classes. This weaker assumption is useful in any case and essential when the measurement protocols are destructive and there is no ideal to which the measurement protocols can be linked.

Central to this paper is the model for a protocol response x given by

$$x = \mu + \lambda y + e,$$

where y is a latent variable representing the property of interest, μ and λ are characteristics of the protocol, and e is a zero-mean random variable with variance ψ that is also a characteristic of the protocol.

Protocol optimization focuses on the ratio λ^2/ψ . For a scientifically-defined quantity, the protocol responses are usually scaled so that $\lambda = 1$; then the protocols themselves can be compared in terms of the variances. Generally, however, protocols must be compared in terms of the ratio λ^2/ψ .

In the physical sciences, possible measurement protocols can be defined in terms of factors and their possible settings. For example, hardness measurement involves the indenter shape, the forces placed on the indenter to drive it into the specimen, and the shape

of the anvil used to hold the specimen during indentation. Solderability measurement with a wetting balance involves the temperature of the solder bath and the type of flux applied to the lead before immersion in the bath. Factor setting choices affect the performance of the protocol. The problem at hand is to find the choice that optimizes the performance. This is parameter design.

Experiments with measurement protocols require specimens to be measured, that is, experimental units. Comparison of protocols in terms of the sensitivity λ^2/ψ requires only that the experimental units come in classes that differ in the property of interest but are relatively homogeneous within each class.

When the values of the experimental units are unknown, the primary issue in parameter design is model identifiability. This issue is closely related to identifiability of the normal linear factor model (NLFM) (Bartholomew and Knott 1999). Our model for the measurement protocol responses differs from the NLFM in only one important aspect.

For our model, the response from protocol j applied to an experimental unit with value y is

$$\mathbf{x}_j|y \sim N_{p_0}(\boldsymbol{\mu}_j + \boldsymbol{\lambda}_j y, \boldsymbol{\Psi}_j),$$

where the dimension of the response is p_0 . We consider $p_0 = 1$. For $p_0 = 1$, we have

$$x_j|y \sim N(\mu_j + \lambda_j y, \psi_j).$$

This equation applies to both our model and the NLFM. Note that the sensitivity λ_j^2/ψ_j of one protocol relative to another can be estimated when the model is identifiable.

In the NLFM, responses from the p protocols can be obtained from the same unit. If $p_0 = 1$, stacking the x_j , the μ_j , and the λ_j , and letting $\boldsymbol{\Psi} = \text{diag}(\psi_j)$, we obtain

$$\mathbf{x}|y \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Lambda}y, \boldsymbol{\Psi}),$$

which is equation 3.1 in Bartholomew and Knott (1999). In the physical sciences, measurement protocols are often destructive, that is, only one response can be obtained from each unit. For this case, this equation does not necessarily hold.

We assume that there are classes of experimental units available for the parameter design experiments and that the value of the property of interest for a unit in class i is given by

$$y|\theta_i, \phi_i \sim N(\theta_i, \phi_i), \quad i = 1, \dots, n.$$

When each of the p protocols is applied to a unit selected randomly from class i ,

$$\mathbf{x}|\theta_i, \phi_i \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Lambda}\theta_i, \phi_i \text{diag}(\lambda_j^2) + \boldsymbol{\Psi}).$$

If ϕ_i does not depend on i , this reduces to the NLFM with ψ_j replaced by $\phi \lambda_j^2 + \psi_j$. Because $\lambda_j^2/(\phi \lambda_j^2 + \psi_j)$ is a monotonic function of the sensitivity λ_j^2/ψ_j , we can optimize $\lambda_j^2/(\phi \lambda_j^2 + \psi_j)$ instead of λ_j^2/ψ_j . We assume that ϕ_i does not depend on i , $\phi_i = \phi$. Thus,

$$\mathbf{x}|\theta_i, \phi_i \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Lambda}\theta_i, \tilde{\boldsymbol{\Psi}}), \quad \text{where } \tilde{\boldsymbol{\Psi}} = \phi \text{diag}(\lambda_j^2) + \boldsymbol{\Psi}.$$

So, identifiability for our case parallels identifiability for the NLFM.

But, our case allows the application of the same protocol to several units in the same class; that is each of the p protocols is applied h_i times to units in class i , where h_i does not depend on j .

For the NLFM, a prior distribution is adopted for the latent variables and estimation is based on the resulting marginal distribution for the observed responses. In our case, the prior distribution provides that the θ_i are independent with

$$\theta_i \sim N(0, 1).$$

For the case of $h_i > 1$, maximum likelihood can be used to estimate values of the latent variables, that is, the θ_i . Nevertheless, the use of the prior in estimation may be useful. Thus, we have two maximum likelihood approaches to estimation, which we can compare.

Let \bar{x}_i be the average over the replicates for experimental unit class i . The marginal distribution with θ_i integrated out is

$$\bar{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + h_i^{-1}\tilde{\boldsymbol{\Psi}}).$$

The distribution with θ_i included is

$$\bar{x}_i \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Lambda}\theta_i, h_i^{-1}\tilde{\boldsymbol{\Psi}}).$$

For maximum likelihood estimation in this case, we take

$$\sum_{i=1}^n h_i \theta_i = 0, \quad \sum_{i=1}^n h_i \theta_i^2 = H, \quad \text{where } H = \sum_{i=1}^n h_i.$$

For both cases, the maximum likelihood estimate of $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}} = \frac{1}{H} \sum_{i=1}^n h_i \bar{x}_i.$$

We can substitute this into the likelihood function.

Consider now applying of Markov Chain Monte Carlo (MCMC) techniques to inference for parameter design experiments. Fortunately, the relation between the measured responses and the sources of variation can be represented as a simple directed acyclical graph of the sort that underlies Bayesian inference; and the computation using the BUGS software.

We illustrate with an experiment involving a very simple response surface: a tilted plane. From a posterior distribution on the tilt of the plane and contextual considerations, one can choose a protocol.

Electronic manufacturing involves soldering of components to printed circuit boards. The quality of the soldering depends on the state of the surface of the component leads being soldered, how solder wets the surface. One can measure this wetting by inserting a lead into a solder bath and watching how the force exerted on the lead by the solder changes

with time. Successful wetting causes a force that pulls the lead into the solder bath. A measured response is the force 2 seconds after the lead was first inserted in the bath. (Kil-Won Moon of the National Institute of Standards and Technology provides this example.)

The parameter design experiment involves two factors, the type of flux and the solder bath temperature. Each of these factors has two levels, rosin (R) and rosin moderately activated (RMA) for the flux and 235°C and 265°C for the temperature. Classes of experimental units (leads) are produced by steam aging copper leads under different conditions. In this experiment, one set for two parameter settings and the other set for the other two parameter settings, confounding class effect with the interaction in the response surface model.

The measured responses are plotted versus the MLEs for θ_i in the above figure for each setting of the parameters, with three replicate measurements on each of four classes of experimental units. There is one missing value in the upper left panel and one pair of tied values in the upper right panel. Note that the top and bottom parts have different values of θ_i . In terms of sensitivity λ^2/ψ , the setting RMA, 235°C seems to be the best; but it is unclear whether this is due to real difference or to random variation in the execution of the protocol or the manufacture of the experimental units. Further uncertainty comes from the possibility that the MLEs that determine the horizontal positions in the above figure, might camouflage the effects of the experimental sources of variation.

The model is specified in Table 1: measured responses are x_{ijk} , where i indexes the classes of leads, j indexes the protocols, and k indexes the replicates. The x_{ijk} are taken to be independent conditional on their mean $\mu_j + \lambda_j \theta_{i(j)}$ and variance $\tilde{\psi}_j$.

Table 1. Model for the Wetting Force Data

Protocol	Flux	Temperature	Experimental leads	Sensitivity $\log(\lambda^2/\psi)$
1	R	235°C	$\theta_{i(1)}$	$a_0 - a_1 - a_2$
2	RMA	265°C	$\theta_{i(2)} (= \theta_{i(1)})$	$a_0 + a_1 + a_2$
3	R	265°C	$\theta_{i(3)}$	$a_0 - a_1 + a_2$
4	RMA	235°C	$\theta_{i(4)} (= \theta_{i(3)})$	$a_0 + a_1 - a_2$

Prior distributions on $\theta_{i(1)}$, $\theta_{i(3)}$, μ_j , λ_j , a_0 , a_1 , and a_2 are needed for the Bayesian analysis. We take $\tilde{\psi}_j$ to be a function of these quantities. The WinBUGS code includes the model, data, and initial values where these starting values are drawn from maximum likelihood estimates.

The posterior densities for a_1 and a_2 show that despite the experiment, conclusions that either RMA is better than R and that R is better than RMA are both credible. Similarly, both conclusions about temperature are credible: 235°C is better than 265°C and 265°C is better than 235°C. The evidence for the choice of 235°C over 265°C seems greater in the above figure than in the posterior for a_2 . Lacking textual considerations, such as ease of protocol implementation, one would choose RMA and 235°C.

The development of measurement protocols is a fairly common task in the engineering of materials. An engineer will try to find a response that reflects the property of interest with no obvious sensitivity to unrelated properties of the intended units. Initially, this is an engineering task because it involves scientific models. Such models, however, cannot guide engineers through all choices that they have to make. For this reason, some choices may be uninformed. This paper picks up protocol development at the point where scientific theory has suggested at least one valid protocol and describes experiments for further refinement. These experiments do not lead directly to measurement system improvement in the sense of traceability to widely-accepted measurement standards. Rather, they lead to better performance in local applications such as quantifying variation. An important advantage of these experiments is that they can often be performed without resources from an outside institution.

3.2 Key Comparisons

3.2.1 International Key Comparisons and Uncertainty Principles

Nien Fan Zhang, Jim Filliben, Will Guthrie, Hung-kung Liu, Andrew Rukhin, Nell Sedransk, Blaza Toman
Statistical Engineering Division, ITL

Key Comparisons are special international interlaboratory comparison studies chosen by the Consultative Committees under the International Committee for Weights and Measures (CIPM) to establish the degree of equivalence between national measurement standards. With the recent signing of the Mutual Recognition Arrangement (MRA) by the members of the 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. To assure accurate, efficient assessment of equivalence, the Statistical Engineering Division (SED) has proposed to provide a unified statistical framework and detailed guidance for the Key Comparisons process. This year SED participated in more specific Key Comparisons, in addition to continuing statistical research on the foundations of Key Comparisons.

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. For Key Comparisons, the Consultative Committees under the CIPM 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 Comparisons projects in collaboration with eight out of 10 Consultative Committees under the CIPM.

A Key Comparison database has been developed jointly by NIST and the International Bureau for Weights and Measures (BIPM). However, at present, there is no consensus among the various international labs and consultative committees on the best choice of procedures to be 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 determined, and the corresponding uncertainties are estimated. We see great benefit to the international community in developing a statistical roadmap to clarify the choices and optimize the process. The SED Key Comparisons project promotes a unified approach to Key Comparisons.

Specifically, the data collection phase needs a statistically sound and efficient experimental design. 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 should 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 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 methods used to quantify the degree of equivalence. We believe that it would be beneficial to have a standard process for this task.

In 2002, in addition to participating in specific Key Comparisons, we focused on the foundations of Key Comparisons to provide general guidance and computer tools for their design and analysis. We evaluated different statistical approaches for estimating Key Comparison reference values and their associated uncertainties using parametric, non-parametric, Bayesian, and fiducial methods. We have completed our statistical research on uncertainty analysis for Key Comparisons when the traveling standards have drift effects. These research results have been successfully applied to specific Key Comparisons.

In 2002, we actively participated in meetings held by several Consultative Committees of CIPM on Key Comparisons. We also participated in the Joint BIPM-NPL Workshop on the Impact of Information Technology in Metrology. At the satellite workshop on the Evaluation of Interlaboratory Comparison Data, SED gave a presentation and provided comments on the BIPM Advisory Group's draft document, "Proposed Procedures for the Statistical Analysis of Key Comparison Measurements." Our presentation and comments on that document had great impact on the way the participants from other NMI's are analyzing Key Comparison data.

This project will directly support NIST's new efforts to establish equivalence with other NMI's and RMO's under the MRA. Recent collaborations in this area between SED and staff from other NIST laboratories have clearly identified the desire and need for guidance in carrying out Key Comparisons. In a large 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 Uncertainty Analysis for Key Comparisons with Trends

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

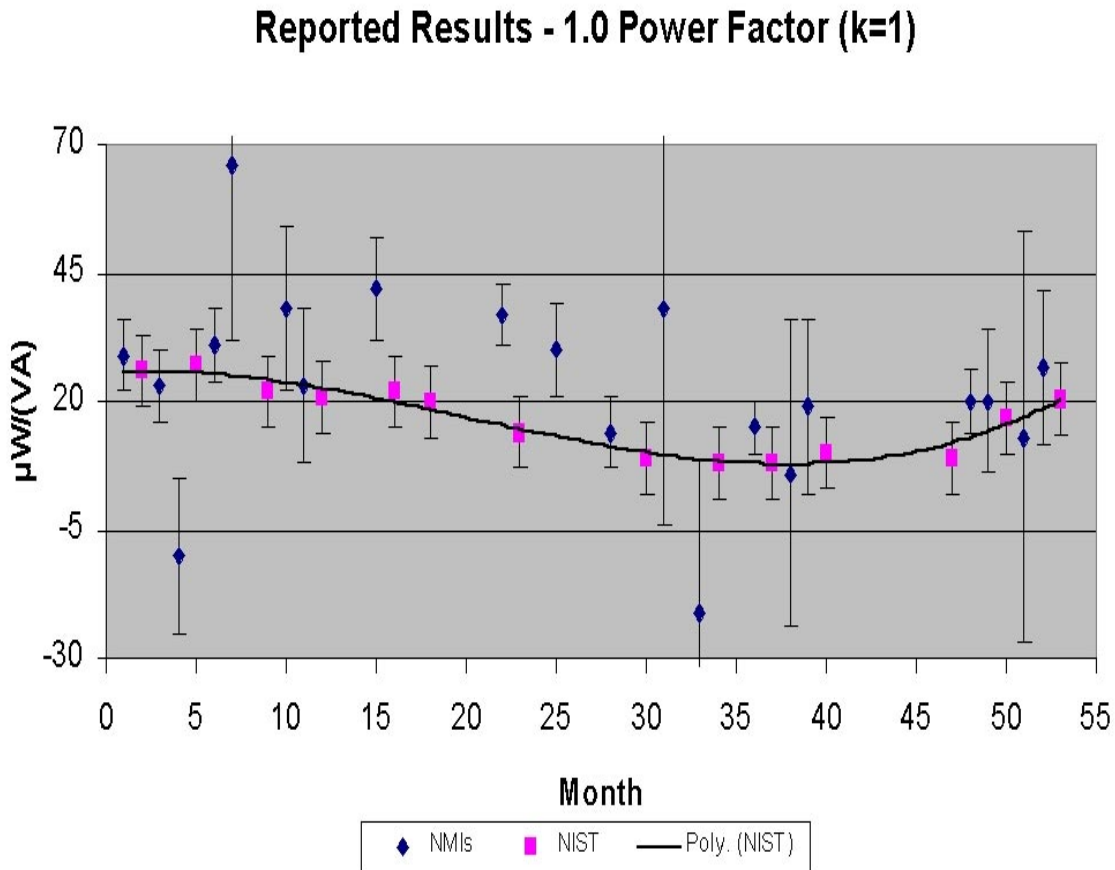


Figure 3.8: The figure above shows the measurement in CCEM-K5 made by the participating NIMs at 1.0 power factor. The measurements made by NIST at 14 time periods during the circulation of the standard demonstrated a drift effect with a fitted cubic regression line. In this paper, we consider the uncertainty analysis for key comparisons with general drift effects.

In most key comparisons, the pilot NMI organizes the circulation and transport of standards or artifacts. The pilot NMI is the only NMI who makes the measurements multiple times during the circulation of the standards. In some key comparisons, the measurements of the transport standards made by the pilot NMI show a drift or a trend. For example, in the key comparison of CCEM-K2, a linear trend was demonstrated. Zhang, Sedransk, and Jarett (2002) proposed a statistical analysis for the uncertainty calculation when a linear drift exists. The analysis was applied to CCEM-K2 and the details can be found in the final report of CCEM-K2. Recently, a non-linear drift was observed in the key comparison of CCEM-K5. CCEM-K5 is the first CCEM-sponsored international comparison of 50/60 Hz electric power. The key comparison began in 1996 and the final results received in 2001. Fifteen NMIs participated in the key comparison and NIST was the pilot laboratory. The comparison was performed at 120 V, 5 A, 53 Hz, and at 1.0, 0.5, and 0.0 power factors. The details of the comparison and the statistical analysis can be found in Draft B of the report by Oldham, Nelson, Zhang, and Liu (2002).

We assume that there are I NMIs participating in the comparison. Without loss of generality, we denote the pilot NMI as the first NMI. We assume that the measurements made by the pilot NMI at K periods are fitted by a Q^{th} order polynomial regression:

$$X_1(k) = \beta_0 + \beta_1 t_1(k) + \dots + \beta_Q t_1^Q(k) + \varepsilon_1(k) \quad (3.5)$$

for $k = 1, 2, \dots, K$ or in a matrix form

$$X_1 = T\beta + E \quad (3.6)$$

with $X_1 = (X_1(1), \dots, X_1(K))'$ denoting a $K \times 1$ vector and for any $k = 1, \dots, K$, $X_1(k)$ is the measurement by the pilot NMI at time period $t_1(k)$. The parameter vector $\beta = (\beta_0, \beta_1, \dots, \beta_Q)'$ is a $(Q+1) \times 1$ parameter vector and T is a K by $(Q+1)$ matrix with the elements of the first column being 1's and the other (k, j) elements (for $j = 2, \dots, (Q+1)$) being $t_1^{j-1}(k)$, with $t_1(k)$ denoting the time period when the pilot NMI measured the standard the k^{th} time. The term $\varepsilon_1(k)$ is the random error for the pilot NMI with zero mean and variance σ_ε^2 . $E = (\varepsilon_1(1), \dots, \varepsilon_1(K))'$. From regression analysis,

$$\hat{\beta} = (T'T)^{-1}T'X_1 \quad (3.7)$$

$$Var[\hat{\beta}] = (T'T)^{-1}\sigma_\varepsilon^2 \quad (3.8)$$

with $(T'T)^{-1}$ a $(Q+1)$ by $(Q+1)$ matrix. Assume that X_i ($i = 2, \dots, I$) is the measurement made by the i^{th} non-pilot NMI at time period of t_i . The predicted value of the i^{th} NMI at t_i based on the measurements by the pilot NMI is denoted by $X_{i,p}$. The difference between the measurement and the prediction or the correction for the i^{th} NMI is defined as

$$D_i = X_i - X_{i,p} \quad (3.9)$$

with its variance

$$Var[D_i] = Var[X_i] + Var[X_{i,p}] \quad (3.10)$$

and $Var[X_i]$ can be estimated from the i^{th} NMI's uncertainty budget. The 2^{nd} term in the right-hand side of (6), i.e., the variance of the prediction (not the fitted value) is in terms of the point prediction with error $\varepsilon(i)$. Namely,

$$Var[X_{i,p}] = Var[\vec{t}_i \hat{\beta} + \varepsilon(i)] = Var[\vec{t}_i \hat{\beta}] + \sigma_\varepsilon^2 \quad (3.11)$$

with the vector \vec{t}_i denoting the row vector $(1, t_i, t_i^2, \dots, t_i^Q)$ and t_i is the time period when the i^{th} NMI made its measurements. It turns out that

$$Var[D_i] = Var[X_i] + \sigma_\varepsilon^2(1 + \vec{t}_i(T'T)^{-1}\vec{t}_i') \quad (3.12)$$

For the pilot NMI, the corresponding correction is calculated as the average of corrections made at $t_1(i)$ for $i = 1, 2, \dots, I$ or

$$D_1 = \frac{1}{K} \sum_{k=1}^K [X_1(k) - X_{1,p}(k)] \quad (3.13)$$

D_1 is usually estimated by 0. We assume that for the pilot NMI, a measurement can be expressed as $X_1(k) = X_{1,A}(k) + X_{1,B}$, with the two components $X_{1,A}(k)$ and $X_{1,B}$ independent from each other and their corresponding uncertainties are those due to the Type A and Type B evaluations for the pilot NMI. The variance of D_1 is

$$Var[D_1] = \sigma_{X,B,1}^2 + \frac{1}{K} \sigma_{X,A,1}^2 \quad (3.14)$$

with $\sigma_{X,A,1}^2$ and $\sigma_{X,B,1}^2$ denoting the Type A and B variance components, respectively, of the measurements made by the pilot NMI. The reference value of the key comparison denoted by X_{krv} in this paper, is defined as the weighted mean of the corrections with the weights determined by their variances

$$X_{krv} = \sum_{i=1}^I v_i D_i \quad \text{and} \quad v_i = \frac{\frac{1}{Var[D_i]}}{\sum_{j=1}^I \frac{1}{Var[D_j]}} \quad (3.15)$$

It turns out that

$$Var[X_{kcrv}] = \frac{1}{\sum_{i=1}^I \frac{1}{Var[D_i]}} + \frac{2\sigma_\varepsilon^2}{(\sum_{j=1}^I \frac{1}{Var[D_i]})^2} \times \sum_{i>k}^I \sum_{k=2}^I \frac{\vec{t}_i(T'T)^{-1}\vec{t}_k'}{Var[D_i] \times Var[D_k]} \quad (3.16)$$

It is clear that the second term is due to the covariances among the D_i 's.

For each NMI, such as the i^{th} NMI, the degree of equivalence with the reference value is defined as

$$D_{i,kcrv} = D_i - X_{kcrv} \quad (3.17)$$

The corresponding uncertainty for the i^{th} non-NIST lab is

$$Var[D_{i,KCRV}] = (1 - 2v_i)Var[D_i] + Var[X_{KCRV}] - 2\sigma_\varepsilon^2 \sum_{k=2, k \neq i}^I v_k [\vec{t}_i(T'T)^{-1}\vec{t}_k'] \quad (3.18)$$

which can be calculated from (8) and (12). For the pilot NMI, similar to (14)

$$Var[D_{1,KCRV}] = (1 - 2v_1)(\sigma_{X,B,1}^2 + \frac{\sigma_{X,A,1}^2}{K}) + Var[X_{KCRV}] \quad (3.19)$$

with v_1 denoting the weight corresponding to the pilot NMI and given by (11).

The degree of equivalence between two NMIs is defined as

$$D_{i,k} = D_i - D_k$$

If neither of the two NMIs is the pilot NMI (i.e., when $i \neq k \neq 1$), it turns out that

$$Var[D_{i,k}] = Var[X_i] + Var[X_k] + \sigma_\varepsilon^2 [2 + \vec{t}_i(T'T)^{-1}\vec{t}_i' + \vec{t}_k(T'T)^{-1}\vec{t}_k' - 2\vec{t}_i(T'T)^{-1}\vec{t}_k'] \quad (3.20)$$

If one NMI is the pilot NMI,

$$Var[D_{1,k}] = \sigma_{X,B,1}^2 + \frac{\sigma_{X,A,1}^2}{K} + Var[X_k] + \sigma_\varepsilon^2 (1 + \vec{t}_k(T'T)^{-1}\vec{t}_k') \quad (3.21)$$

Drift effects or trends exist in many international key comparisons. The statistical methodology presented in this paper has been applied to CCEM-K5 comparison and can be applied to most key comparisons where trends are significant.

3.2.3 A Robust Key Comparison Reference Value in Cases of Dominant Type B Error

Blaza Toman

Statistical Engineering Division, ITL

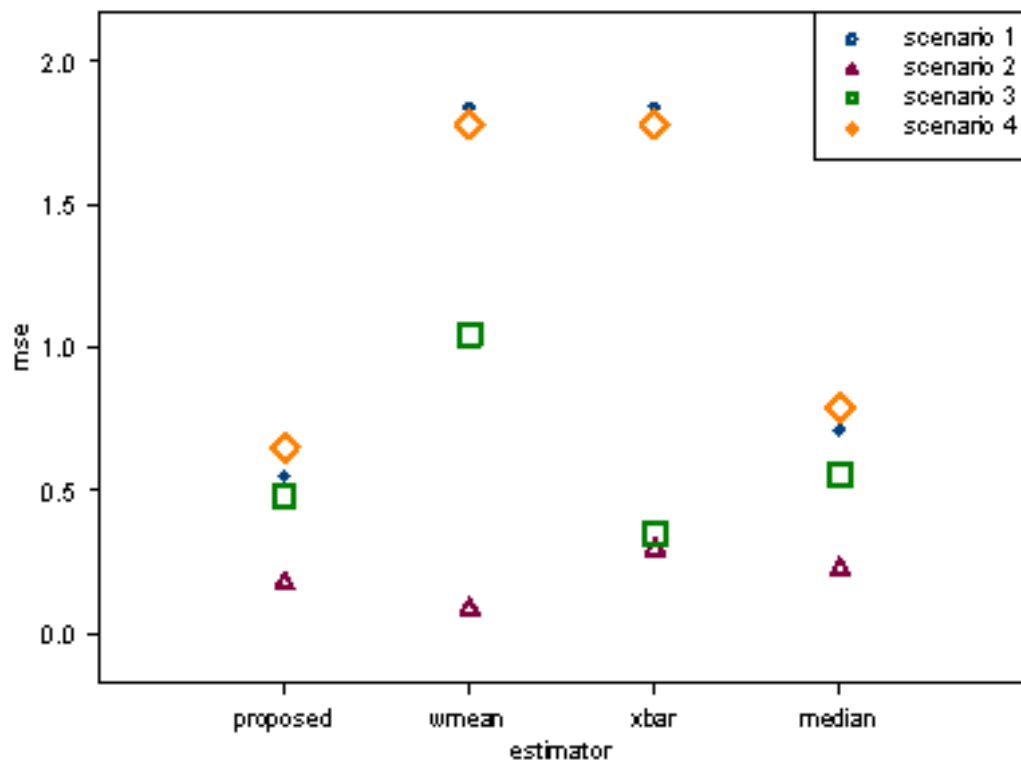


Figure 3.9: The graph shows that for all but scenario 2 (the best outcome scenario), the proposed KCRV outperformed the weighted mean. The new method also outperformed the median but not as significantly. The average did better on scenario 3 but worse for the others. This indicates considerable robustness of the proposed KCRV to flawed reporting of the τ_i^2 s. The cost of such robustness in terms of reduced efficiency when the τ_i^2 are reported correctly (scenario 2) does not seem to be overly large.

There are some Key Comparisons where the total uncertainty of the measurements is almost all due to Type B error obtained by expert opinion. In such cases, the interpretation of each laboratory's results (given in the GUM, 6.2.2) as a probability distribution of the measurand centered at the measurement mean and having a standard deviation equal to the combined uncertainty can be a reasonable approximation. Such a posterior distribution can in fact be obtained by starting with the model

$$Y_i | \mu_i, b_i, \sigma_i^2 \sim N(\mu_i + b_i, \sigma_i^2) \quad i = 1, \dots, k$$

and

$$\mu_i | m_i, \omega_i^2 \sim N(m_i, \omega_i^2)$$

$$b_i | \delta_i, \tau_i^2 \sim N(0, \tau_i^2),$$

where μ_i is the measurand, and b_i a Type B known systematic error, of laboratory i . Allowing for a different measurand for each laboratory is necessary because we must allow for the possibility of an unknown bias for each laboratory. In this model τ_i^2 is the known Type B uncertainty. With a noninformative prior distribution on the μ_i , their posterior distribution is Normal $(y_i, \tau_i^2 + \sigma_i^2)$ and thus if σ_i is approximated by s_i we get the desired result.

In some Key Comparisons, the measurand is a physical constant and a Key Comparison Reference Value (KCRV) is a quantity that needs to be provided. There have been numerous suggestions for the form of the KCRV. The most common are weighted means of the Y_i , with the weights being some functions of $\tau_i^2 + s_i^2$. An important property of such KCRVs is that they are not usually robust to the choice of the τ_i^2 s. As these quantities are not objectively chosen but are based on expert opinion, this is clearly not a desirable property. An alternate solution, which produces a KCRV which is robust to the choice of the τ_i^2 s, can be obtained as follows.

Imagine that a single person with vague prior knowledge of the true physical constant, call it μ_T , consults k laboratories who provide their means y_i and standard deviations $\sqrt{\tau_i^2 + s_i^2}$. The person must combine the k results into a single probability distribution for μ_T . He specifies a normal likelihood to express his opinion about the laboratories' knowledge, namely he specifies the distribution $p(\mu_T | y_1, \dots, y_k, \tau_1^2, \dots, \tau_k^2, s_1^2, \dots, s_k^2)$ as multivariate normal with means $\lambda_i = \mu_T$, and standard deviations $\kappa_i \sqrt{\tau_i^2 + s_i^2}$. The κ_i provide a way to input opinion about the precision of the assessments and thus the desired robustness of the resulting KCRV. Giving a single value for each κ_i may be difficult and/or controversial. A good alternative is to specify a probability model on κ_i , that is let $\frac{v_i c_i^2}{\kappa_i} \sim \chi_{v_i}$. The values of c_i and the degrees of freedom v_i can be best understood from the approximate relationship given by this distribution, which is that $a_i^{-1} c_i < \kappa_i < a_i c_i$, where

$\log a_i = \left(2/v_i\right)^{1/2}$. The resulting posterior distribution of μ_T is a product of student t distributions such that $(\mu_T - y_i)/c_i (\tau_i^2 + s_i^2)$ has a t_{v_i} . The KCRV can be taken as the posterior mean or median of this distribution. The necessary calculations can be easily done using BUGS.

Simulations were performed in order to illustrate the robustness of the proposed estimator. For the purpose of this comparison, the type B error was simulated as a real variance in the data. There were 4 scenarios being simulated. In scenario 1, each laboratory mean was generated from a $N(0, \omega_i^2)$ distribution with $\omega_i^2=10$ for laboratory 1, 2, and 3 and $\omega_i^2=1$ for the remaining laboratories. Further, it was assumed that all 10 laboratories gave their $\tau_i^2 + s_i^2=1$. This is a situation when some laboratories severely underestimate their total variability. In scenario 2, $\omega_i^2=0.1$ for laboratory 1 and $\omega_i^2=1$ for the remaining laboratories. Here it was assumed that laboratory 1 reported $\tau_i^2 + s_i^2=0.1$ and the remaining laboratories reported $\tau_i^2 + s_i^2=1$. This is a case when all laboratories are essentially correct in their variability. Scenario 3 had $\omega_i^2=1$ for all 10 laboratories with laboratory 1 reporting $\tau_i^2 + s_i^2=0.1$ and the remaining laboratories reporting $\tau_i^2 + s_i^2=1$. This is a case of one laboratory severely underestimating its uncertainty. Scenario 4 generated observations for laboratory 1 from a $N(10, 1)$ distribution. All remaining data was generated from a $N(0, \omega_i^2)$ distribution with $\omega_i^2=1$ for laboratories 4 through 10 and $\omega_i^2=10$ for laboratory 2, and 3. It was further assumed that all 10 laboratories gave their $\tau_i^2 + s_i^2=1$. This scenario has both underreported uncertainties for some laboratories and one laboratory whose data are outliers in terms of the mean.

The proposed KCRV was calculated using $1/2 < \kappa_i < 2$ for all 10 laboratories. The graph contains the Monte Carlo root mean square errors associated with the proposed KCRV, the usual weighted mean, the average of the y_i , and their median.

3.2.4 Two New Estimators of the Variance of the Graybill-Deal Estimator of a Common Mean

Nien Fan Zhang

Statistical Engineering Division, ITL

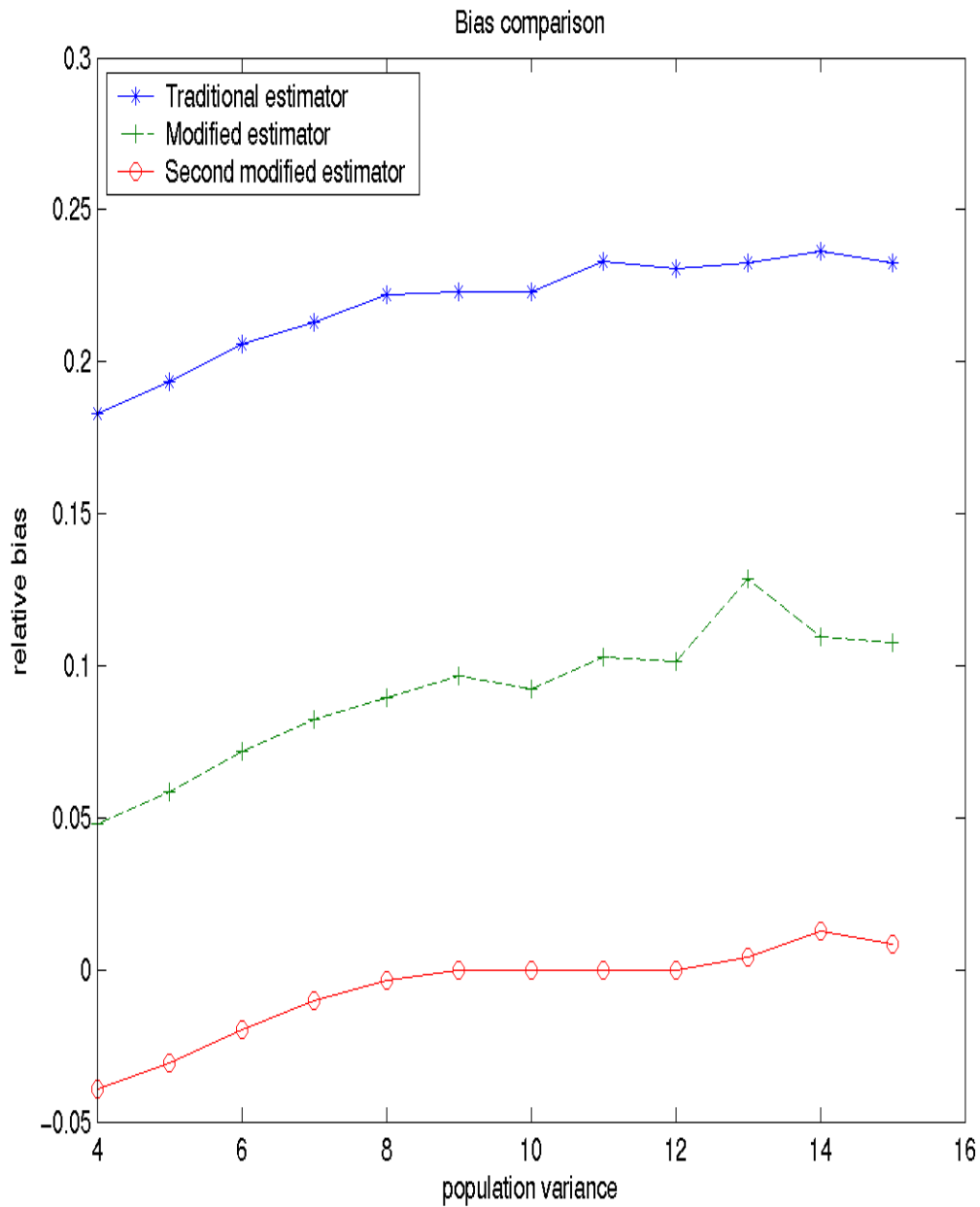


Figure 3.10: This figure shows the relative biases of the traditional estimators and two new estimators of the variance of the Graybill-Deal estimator of a common mean.

Graybill-Deal estimator or the weighted mean based on sample variances has been proposed to estimate the common mean of several normal populations. The usual estimator of the variance of the Graybill-Deal estimator is a biased estimator. In this paper, we propose two new estimators with better properties.

The statistical inference dealing with a common mean of several normal populations with unknown and possibly unequal variances has a long history. Specifically, for a linear model with a common mean such as

$$x_{ij} = \mu + \varepsilon_{ij} \quad (3.22)$$

with μ denoting the common mean and the errors ε_{ij} 's are independent from each other and normally distributed with zero mean and variance of σ_i^2 ($i = 1, \dots, k$) and $j = 1, \dots, n_i$. The common mean, μ , for the k populations can be estimated by the weighted mean:

$$\bar{x}_w = \sum_{i=1}^k w_i \bar{x}_i \quad (3.23)$$

with the weights w_i 's satisfying $\sum_{i=1}^k w_i = 1$ and \bar{x}_i is the sample mean from the i th population, i.e.,

$$\bar{x}_i = \sum_{j=1}^{n_i} \frac{x_{ij}}{n_i} \quad (3.24)$$

From Graybill and Deal (1959), \bar{x}_w is the minimum variance unbiased estimator of μ among all the weighted means when the weights are

$$w_i = \frac{\frac{1}{\sigma_i'^2}}{\sum_{j=1}^k \frac{1}{\sigma_j'^2}}, \quad (3.25)$$

with $\sigma_i'^2 = \sigma_i^2/n_i$. In practice, however, σ_i^2 ($i = 1, \dots, k$) are unknown. Thus, the w_i 's are usually estimated by the following statistic

$$\hat{w}_i = \frac{\frac{1}{S_i'^2}}{\sum_{j=1}^k \frac{1}{S_j'^2}}. \quad (3.26)$$

with $S_i'^2 = S_i^2/n_i$ and S_i^2 is the sample variance from x_{ij} ($j = 1, \dots, n_i$). The corresponding weighted mean

$$\hat{\bar{x}}_w = \sum_{i=1}^k \hat{w}_i \bar{x}_i \quad (3.27)$$

is called the Graybill-Deal estimator of the common mean. In the Proposed International Guidelines for the Evaluation of Key Comparisons Data (2002), the Graybill-Deal estimator of the common mean is proposed to calculate the key comparison reference value (KCRV). Here we restrict the meaning of the uncertainty to a sample standard deviation, or as a Type A uncertainty described in ISO Guide to the Expression of Uncertainty in Measurement (1995)(GUM).

In metrology, it is well-known that the variance of an estimator is as important as the estimator itself. In this paper, we will discuss the estimators of the variance of $\hat{\bar{x}}_w$, the Graybill-Deal estimator, and their properties. We assume that ε_{ij} , or equivalently x_{ij} , are normally distributed.

It is well-known that

$$Var[\bar{x}_w] = \frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i'^2}} \quad (3.28)$$

Although some approximations of the variance of $\hat{\bar{x}}_w$ can be found in the literature, there is no exact formula for this variance. From (3.28), many practitioners proposed to estimate $Var[\hat{\bar{x}}_w]$ by substituting $S_i'^2$ for $\sigma_i'^2$ in (3.28), for $i = 1, \dots, k$. That is,

$$\widehat{Var}[\hat{\bar{x}}_w] = \frac{1}{\sum_{i=1}^k \frac{1}{S_i'^2}} \quad (8)$$

In the Proposed International Guidelines for the Evaluation of Key Comparisons Data (2002), the estimator of the variance of the Graybill-Deal estimator in (8) is recommended. In this paper, we show that the estimator in (8) underestimates the true variance and we will propose two alternatives.

Let us examine the relationship between $\widehat{Var}[\hat{\bar{x}}_w]$ defined in (8) and $Var[\bar{x}_w]$ in (3.28). We can show the following inequalities hold.

$$E[\widehat{Var}[\hat{\bar{x}}_w]] = E\left[\frac{1}{\sum_{i=1}^k \frac{1}{S_i'^2}}\right] \leq Var[\bar{x}_w] = \frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i'^2}} \leq Var[\hat{\bar{x}}_w] \quad (3.29)$$

Form (3.29), $\widehat{Var}[\hat{\bar{x}}_w]$ in (8) underestimates $Var[\bar{x}_w]$ as well as $Var[\hat{\bar{x}}_w]$. Later, we will show that the coverage rates of the interval formed by $\widehat{Var}[\hat{\bar{x}}_w]$ are not desired. We propose a modified estimator of $\widehat{Var}[\hat{\bar{x}}_w]$ for $Var[\bar{x}_w]$ and also for $Var[\hat{\bar{x}}_w]$:

$$\widehat{Var}_Z[\hat{x}_w] = \frac{1}{\sum_{i=1}^k \frac{(n_i-3)}{(n_i-1)} \frac{1}{S_i'^2}}. \quad (3.30)$$

We can show that

$$E \left[\frac{1}{\sum_{i=1}^k \frac{1}{S_i'^2}} \right] \leq Var[\bar{x}_w] = \frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i'^2}} \leq E \left[\frac{1}{\sum_{i=1}^k \frac{(n_i-3)}{(n_i-1)} \frac{1}{S_i'^2}} \right] = E[\widehat{Var}_Z(\hat{x}_w)] \quad (3.31)$$

In particular, when $n_i = n$ for $i = 1, \dots, k$,

$$\widehat{Var}_Z(\hat{x}_w) = \frac{n-1}{n-3} \widehat{Var}(\hat{x}_w).$$

For example, when $n = 20$, the ratio of these estimators is 1.12. When $n = 5$, the ratio is 2. Namely, $\widehat{Var}_Z(\hat{x}_w)$ is twice as large as $\widehat{Var}(\hat{x}_w)$. Thus, even though $\widehat{Var}_Z(\hat{x}_w)$ still underestimates $Var[\hat{x}_w]$ (which seems correct based on a simulation study), it is a better estimator than $\widehat{Var}(\hat{x}_w)$ for reducing the bias, especially for small sample sizes.

We propose another estimator of $Var[\hat{x}_w]$:

$$\widehat{Var}_{Z2}[\hat{x}_w] = \frac{1}{\sum_{i=1}^k \frac{(n_i-3)}{(n_i-1)} \frac{1}{S_i'^2}} \left[1 + 2 \sum_{i=1}^k \frac{1}{n_i - 1} \hat{w}_i (1 - \hat{w}_i) \right] \quad (3.32)$$

with

$$\hat{w}_i = \frac{\frac{(n_i-3)}{(n_i-1)} \frac{1}{S_i'^2}}{\sum_{j=1}^k \frac{(n_j-3)}{(n_j-1)} \frac{1}{S_j'^2}}$$

In particular, when $n_i = n$, $\hat{w}_i = \hat{w}_i$ and

$$\widehat{Var}_{Z2}(\hat{x}_w) = \frac{\frac{n-1}{n-3}}{\sum_{i=1}^k \frac{1}{S_i'^2}} \left[1 + 2 \sum_{i=1}^k \frac{\hat{w}_i (1 - \hat{w}_i)}{n_i - 1} \right].$$

A simulation study has been performed to compare the estimators: $\widehat{Var}[\hat{x}_w]$, $\widehat{Var}_Z[\hat{x}_w]$, and $\widehat{Var}_{Z2}[\hat{x}_w]$. Based on limited simulations, the ordinary variance estimator, $\widehat{Var}[\hat{x}_w]$,

underestimates the variance of \hat{x}_w , the fact we already know from (3.29). $\widehat{Var}_Z[\hat{x}_w]$ reduces the bias while $\widehat{Var}_{Z2}[\hat{x}_w]$ performs better than $\widehat{Var}_Z[\hat{x}_w]$. When $n_i = 15$, $\widehat{Var}_{Z2}[\hat{x}_w]$ overestimates the variance of \hat{x}_w , which is good in a conservative sense.

We also use simulations to compare the coverage rate of the intervals formed by the three estimators of the variance. Specifically, for $\widehat{Var}[\hat{x}_w]$ and a coverage factor of, for example, 2, the 2σ interval is defined as

$$\hat{x}_w \pm 2\sqrt{\widehat{Var}[\hat{x}_w]}$$

The coverage rate is the estimate of the probability with which the true mean μ is covered by the specified interval. The definitions for other two estimators of the variance and other coverage factors are similar. Several cases with different combinations of σ_i^2 and n_i for $i = 1, 2, \dots, k$ and the coverage factors of 2 and 3 are considered in the simulations. Based on the simulations, it is clear that the coverage rates corresponding to $\widehat{Var}_Z[\hat{x}_w]$ and $\widehat{Var}_{Z2}[\hat{x}_w]$ are quite stable from case to case while those for $\widehat{Var}[\hat{x}_w]$ are not. The 2σ coverage rates corresponding to $\widehat{Var}[\hat{x}_w]$ are between 0.74 and 0.88 while those corresponding to $\widehat{Var}_Z[\hat{x}_w]$ are between 0.91 and 0.92, and those corresponding to $\widehat{Var}_{Z2}[\hat{x}_w]$ are between 0.92 and 0.93, respectively. The 3σ coverage rates corresponding to $\widehat{Var}[\hat{x}_w]$ are between 0.88 and 0.96, while those corresponding to $\widehat{Var}_Z[\hat{x}_w]$ are between 0.96 and 0.97, and those corresponding to $\widehat{Var}_{Z2}[\hat{x}_w]$ are between 0.97 and 0.98, respectively. Using the coverage rate as a criterion, $\widehat{Var}_Z[\hat{x}_w]$ and $\widehat{Var}_{Z2}[\hat{x}_w]$ are better than $\widehat{Var}[\hat{x}_w]$. Comparing $\widehat{Var}_Z[\hat{x}_w]$ with $\widehat{Var}_{Z2}[\hat{x}_w]$, the first one has coverage rates closer to the standard normal coverage probabilities than the second one. However, the second one has smaller bias than that of the first one. For many cases, the biases for $\widehat{Var}_{Z2}[\hat{x}_w]$ are positive, which is good in a conservative sense. Both estimators are recommended for consideration.

Two new estimators of the variance of the Graybill-Deal estimator have been proposed. It has been shown that these estimators have smaller biases and better coverage rates comparing with the usual variance estimator of the Gray-bill estimator.

3.2.5 Some Statistical Methods Applicable to Key Comparisons Studies

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

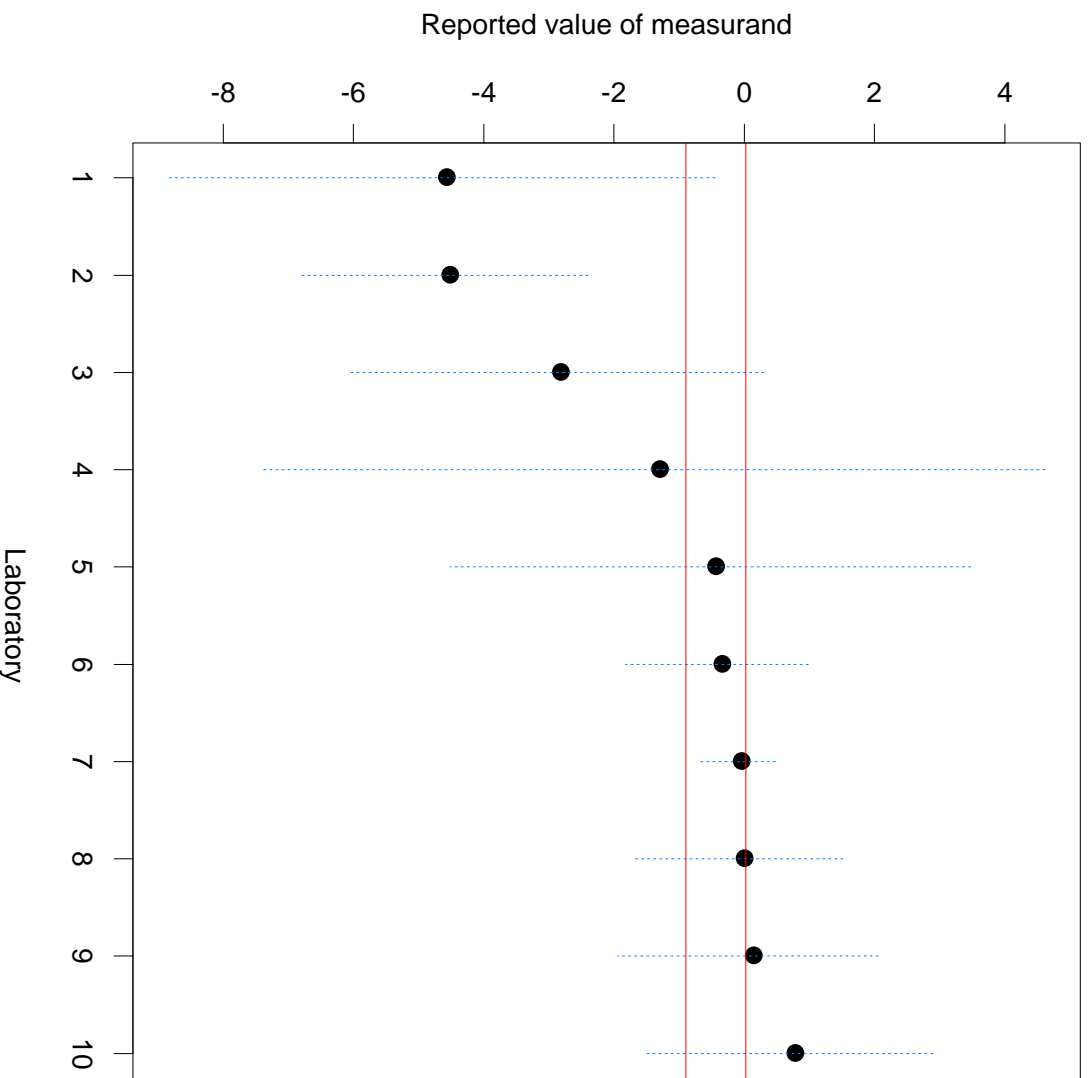


Figure 3.11: Reported values of measurand and their expanded uncertainties (vertical dashed lines). The horizontal lines are the nominally 95% confidence bounds for the KCRV. This example shows that the participating laboratories appear to be inconsistent and hence the calculation of a KCRV may be questionable.

Results of International Key Comparisons of National Measurement Standards provide the technical basis for the Mutual Recognition Arrangement (MRA) formulated by Le Comite International des Poids et Mesures. With many key comparisons already completed and a number of new key comparison experiments currently underway, we now have a better understanding of the statistical issues that need to be addressed for successfully analyzing key comparisons data and making proper interpretations of the results. There is clearly a need for a systematic approach for statistical analyses of key comparisons data that can be routinely implemented by all participating laboratories. We review a number of questions that arise in the context of statistical modeling and analysis of international key comparisons data and propose a systematic approach for answering these questions.

A typical key comparisons study involves two or more participating laboratories. Often one of the laboratories is called the pilot laboratory and this laboratory acts as the coordinator/supervisor of the entire study. The study may involve one or more traveling artifacts and various study designs may be employed to ensure the stability of the traveling artifact, and to detect any drifts in the measurand that may occur during the study period.

Each participating laboratory analyzes its own measurements and reports a final result y after making necessary corrections for known sources of systematic effects. The result is accompanied by a final combined standard uncertainty, $u_c(y)$, and, in some cases, the associated degrees of freedom. In many cases, the final report also gives the full uncertainty budget of each participating laboratory, which lists the various sources of uncertainty, the corresponding standard uncertainties and their types (A or B).

Statistical analyses of these data are conducted to determine the extent to which the different participating laboratories agree, and summary statistics called degrees of equivalence are computed (along with their uncertainties) for each laboratory by examining its deviation from a *key comparison reference value (KCRV)*. A laboratory whose result deviates significantly from the *KCRV* is flagged as a potential outlier and its result – either the reported measurement or its combined standard uncertainty or both – is considered a candidate for further scrutiny.

Suppose the number of participating laboratories is k . Let Y_i denote the result reported by the i^{th} laboratory and let U_i denote the corresponding combined standard uncertainty having ν_i degrees of freedom. We assume that Y_i is normally distributed with mean μ_i and standard deviation σ_i . We also assume that U_i^2 is an unbiased estimator of σ_i^2 and that $\nu_i U_i^2 / \sigma_i^2$ has a chi-square distribution with ν_i degrees of freedom. The value of ν_i is infinity if, in a given instance, σ_i^2 is taken to be equal to u_i^2 and is thus assumed known. Finally, $Y_1, \dots, Y_k, U_1, \dots, U_k$ are assumed to be mutually independent.

Let μ denote the value of the measurand in question. Although each laboratory has made an attempt to determine μ , we have deliberately allowed, in the statistical model, for the possibility that μ_i may be different from μ . The following are some of the questions that need to be answered using the data from the key comparisons study.

1. Scientific validity of the uncertainty budgets. Do the uncertainty budgets provided by the participating laboratories appear to account for all known, nonnegligible sources of uncertainty? If so, do the component standard uncertainties (type-A and type-B) appear to be reasonable?
2. Mutual consistency. Assuming that the combined standard uncertainties for the participating laboratories appear to be reasonable and may be accepted without further discussion, are the results from the different laboratories mutually consistent? That is, are they consistent with the hypothesis that the μ_i are all equal? If the results from the different laboratories appear to be *inconsistent*, then what is the largest number, say L , of laboratories whose results may be considered to be *mutually consistent*? If there are several subsets of L laboratories that may be considered to be mutually consistent, can we identify a subset of L laboratories that may be considered *most consistent*? Which laboratories are not members of this most consistent subset?
3. Computation of a *KCRV*. Suppose we have a consistent set of laboratories whose results are believed to be estimating a common quantity, say μ^* . Ideally, a *KCRV* should then be an efficient estimate of μ^* . How is such a *KCRV* to be computed? What is its uncertainty? It is worth noting that, even if all the laboratories are estimating the same quantity, μ^* , one is unable to conclude, based solely on the results of the key comparison study, that μ^* is equal to μ , the value of the measurand. So, there is no *a priori* reason to expect that the *KCRV* is a valid estimate of μ .

Question 1 is necessarily nonstatistical in nature. Because the final combined standard uncertainties of the different laboratories involve type-B (nonstatistical) evaluations of some of its components, an assessment of the reasonableness of the reported combined standard uncertainties can only be made based on an agreed-upon protocol among the participating laboratories. This will most likely involve careful scrutiny of the uncertainty budgets provided by the participants. We do not further address this issue here.

Question 2 leads to the examination of the statistical hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

versus the alternative that $\mu_i \neq \mu_j$ for at least one pair (i, j) ($i \neq j$). We will refer to such a test as a *consistency test*. Even under Gaussian distributional assumptions, no reasonable exact procedure is available for testing the above hypothesis. An approximate procedure based on the concept of generalized P -values can be used for testing the hypothesis of equal means when variance homogeneity may not hold. If one finds that the results provided by the participating laboratories are not mutually consistent, then it may be necessary to revisit Question 1. One approach to finding consistent subsets of laboratories is to repeatedly apply the consistency test to all possible subsets of laboratories. Let L be the largest integer for which there is a set of L laboratories that pass the consistency test. If this is the only subset of size L that passes the consistency test, then this subset of the laboratories is the one we want. If there are several subsets of size L that pass the consistency test, then it would be useful to provide a complete list of consistent sets of L laboratories and the associated P -values from the consistency test. The subset of L laboratories that yields the highest P -value for the consistency test may be considered the

most consistent subset of L laboratories. Such information is expected to be of assistance in any reconciliation process among the participating laboratories.

An acceptable answer to Question 3 must take into consideration issues that arise when the results from the laboratories participating in the key comparison study appear to be inconsistent. In particular, in our opinion a satisfactory definition of a *KCRV* is not available in this case based solely on data from the key comparison study. In this regard, we note that, although the MRA requires that degrees of equivalence between individual laboratories and a *KCRV* be determined as part of the key comparison study, it does allow for exceptions in which only degrees of bilateral equivalence need to be determined. We propose new approaches, both for computing uncertainty intervals for degrees of equivalence between individual laboratories and a *KCRV*, and uncertainty intervals for bilateral degrees of equivalence. Additionally, we propose a measure of *multilateral degree of equivalence* and suggest an approach for obtaining a confidence bound for this measure.

We have proposed some new statistical methods that may be useful in answering important questions that arise in the context of a key comparison study. The proposed methods are based on a simple model describing the statistical distributions associated with the results and uncertainties provided by the laboratories participating in the study. Clearly, distributional properties ascribed to the standard uncertainties are not rigorously justifiable, particularly since they almost always include type-B components. The same criticism applies to every method that is currently being advocated for analysis of key comparisons data. A more rigorous approach would rely on detailed information from the participating laboratories regarding all of the component measurements and their associated uncertainties used in the computation of the final reported result. In the absence of such information, we believe that the proposed methods offer reasonable approaches for analyzing key comparisons data.

3.2.6 Models and Confidence Intervals for True Values in Interlaboratory Trials

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

T. Mathew
University of Maryland, Baltimore County

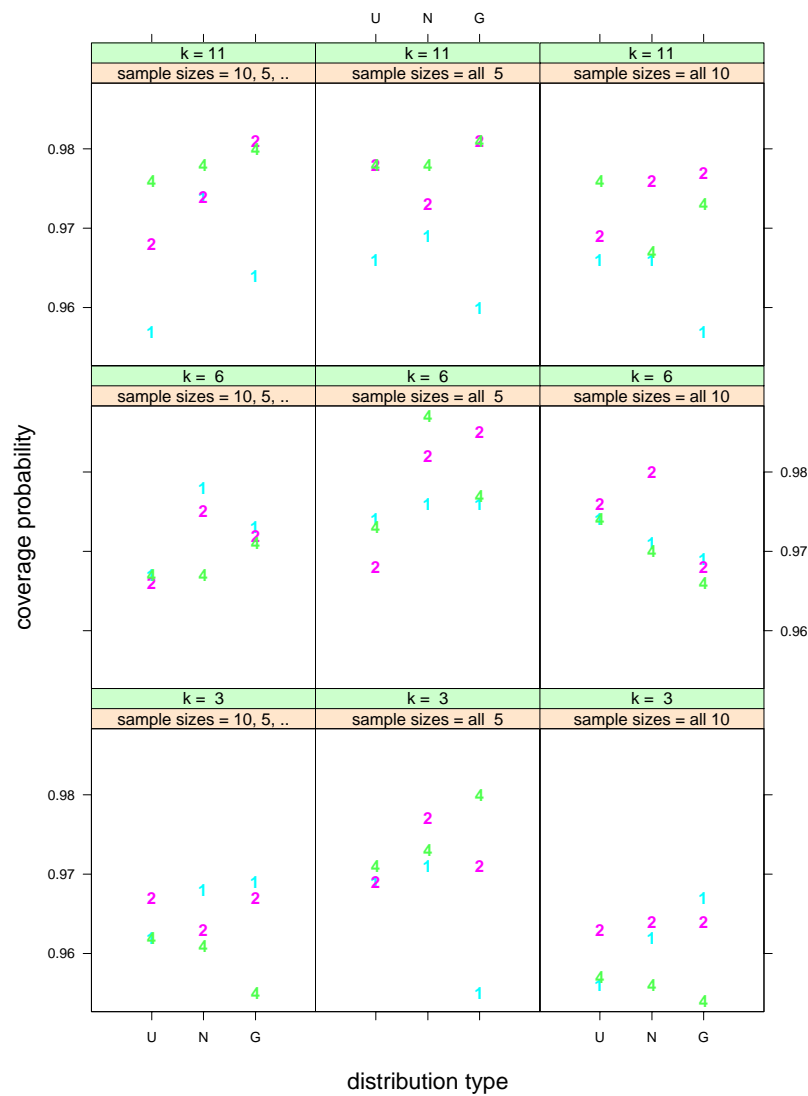


Figure 3.12: Each panel plots the simulated coverage probability of the interval (Y axis) under Model 3 vs. the distribution type – “U” for uniform, “N” for normal, “G” for gamma – of b_i (X axis) for a specific combination of k and n_i configuration. The plotting symbols “1”, “2”, and “4” are used to designate the cases with $\max\{\sigma_i\} = 1, 2,$ and $4,$ respectively.

We consider the one-way random effects model with unequal sample sizes and heterogeneous variances. Using the method of generalized confidence intervals, we develop a new confidence interval procedure for the mean. Additionally, we investigate two alternative models based on different sets of assumptions regarding between group variability and derive generalized confidence interval procedures for the mean. These procedures are applicable to small samples. Statistical simulation is used to demonstrate that the coverage probabilities of these procedures are close enough to the nominal value that they are useful in practice.

We consider the situation in which measurements of an artifact are made by each of k laboratories (or, in some cases, k different measurement methods). The i^{th} laboratory makes n_i independent repeat measurements $Y_{ij}, j = 1, \dots, n_i$. The data from the k laboratories are assumed to follow the model

$$Y_{ij} = \mu_i + e_{ij}$$

with μ_i denoting the mean measured value for laboratory i . If μ denotes the true, unknown measurement of interest, then we write $\mu_i - \mu = b_i$ and call b_i the “bias” of laboratory i . The quantity μ is the parameter that we wish to estimate based on combined information from the different laboratories. The quantities $e_{ij}, j = 1, \dots, n_i$ are random measurement errors associated with the i^{th} laboratory. It is reasonable to assume that $e_{ij}, j = 1, \dots, n_i$ are independent random variables with zero mean and variance $\sigma_i^2, i = 1, \dots, k$. Generally, this error distribution is assumed to be normal.

The particular statistical approach that is appropriate for the estimation of μ depends on what assumptions are made about the b_i or about the relationship of the μ_i to μ . Different sets of assumptions have been considered by various authors. This in turn has led to different analysis methods. Assumption A, given below, encompasses all the different sets of assumptions that have appeared in the literature in connection with a frequentist analysis of the problem.

Assumption A: For $1 \leq i \leq k$, b_i is a random variable whose distribution F_i has a known support, which is the interval $[m_i, M_i]$ (if m_i is negative infinity, then we replace the closed interval at m_i with an open interval; likewise, if M_i is positive infinity).

We now elaborate on various special cases.

- Model 1: Suppose, for each i , F_i is a normal distribution with mean zero and variance σ^2 , $m_i = -\infty$, $M_i = \infty$. We then have the one-way random effects model with unequal sample sizes and heterogeneous error variances. This model has been considered by Rukhin and Vangel (1998), Vangel and Rukhin (1999), Rukhin, Biggerstaff, and Vangel (2000), Paule and Mandel (1982) and others. Methods for estimating μ and for obtaining an approximate confidence interval for μ have been proposed by these authors.

- Model 2: Suppose, for each i , F_i is a completely unspecified distribution and m_i, M_i are known, finite, constants. This case is equivalent to the model considered by Eberhardt, Reeve, and Spiegelman (1989). We will refer to this model as a *bounded-bias model*. Eberhardt et al. used mean-squared error of an estimator as the criterion of goodness and derived a minimax estimator for μ in the class of estimators that are linear functions of the individual laboratory means (or method means). They also proposed an associated approximate confidence interval procedure and evaluated its performance using statistical simulation.
- Model 3: Suppose, for each i , F_i is a completely specified distribution. This case is equivalent to the model described in the ISO Guide to the Expression of Uncertainty in Measurement (ISO GUM) in which the distributions F_i are referred to as *type-B* distributions. One may regard these as *informative* prior distributions on the b_i . Typically, the F_i are assumed to be normal or uniform on a known interval, but other distributions are also sometimes used. For convenience, we will refer to this model as a *GUM type model*.
- Model 4: Suppose, for each i , F_i is a degenerate distribution at b_i , and $\sum_{i=1}^k b_i = 0$. This is equivalent to assuming that a one-way “fixed-effects” model holds for the Y_{ij} and that the true value μ is the average of the k laboratory means $\mu + b_i, i = 1, \dots, k$. This is a standard model and inference about μ is straightforward in this case.
- Model 5: In Model 4, suppose the b_i are all zero. This is the common-means fixed-effects model that has been extensively studied. See, for instance, Jordan and Krishnamoorthy (1996), and Yu, Sun, and Sinha (1999).

We consider Models 1, 2, and 3 for Y_{ij} . Under Model 1, the b_i are iid normal random variables with mean zero and variance σ^2 , and for each $i = 1, \dots, k$, the e_{ij} are iid normal random variables with mean zero and variance σ_i^2 . We develop a generalized pivotal quantity (Weerahandi, 1993) for μ . The confidence bounds for μ are estimated by simulating the distribution of the generalized pivotal quantity. Simulation results show that the coverage probability of the confidence interval so obtained is very close to the nominal value.

For Model 2, it is assumed that, for $i = 1, \dots, k$, the magnitude of the bias b_i is bounded by a positive constant M_i , i.e., $|b_i| \leq M_i$. It is possible that the bias bounds, M_i , associated with the different laboratories are inconsistent with one another and this may result in a situation that the model assumptions fail to define a valid parameter space for μ . We develop a test that can be used to examine whether or not the data are consistent with the specified M_i . We also develop a confidence interval for μ using the method of generalized pivotal quantities. Simulation results show that the coverage probability of the confidence interval under Model 2, although conservative on some occasions, is generally adequate.

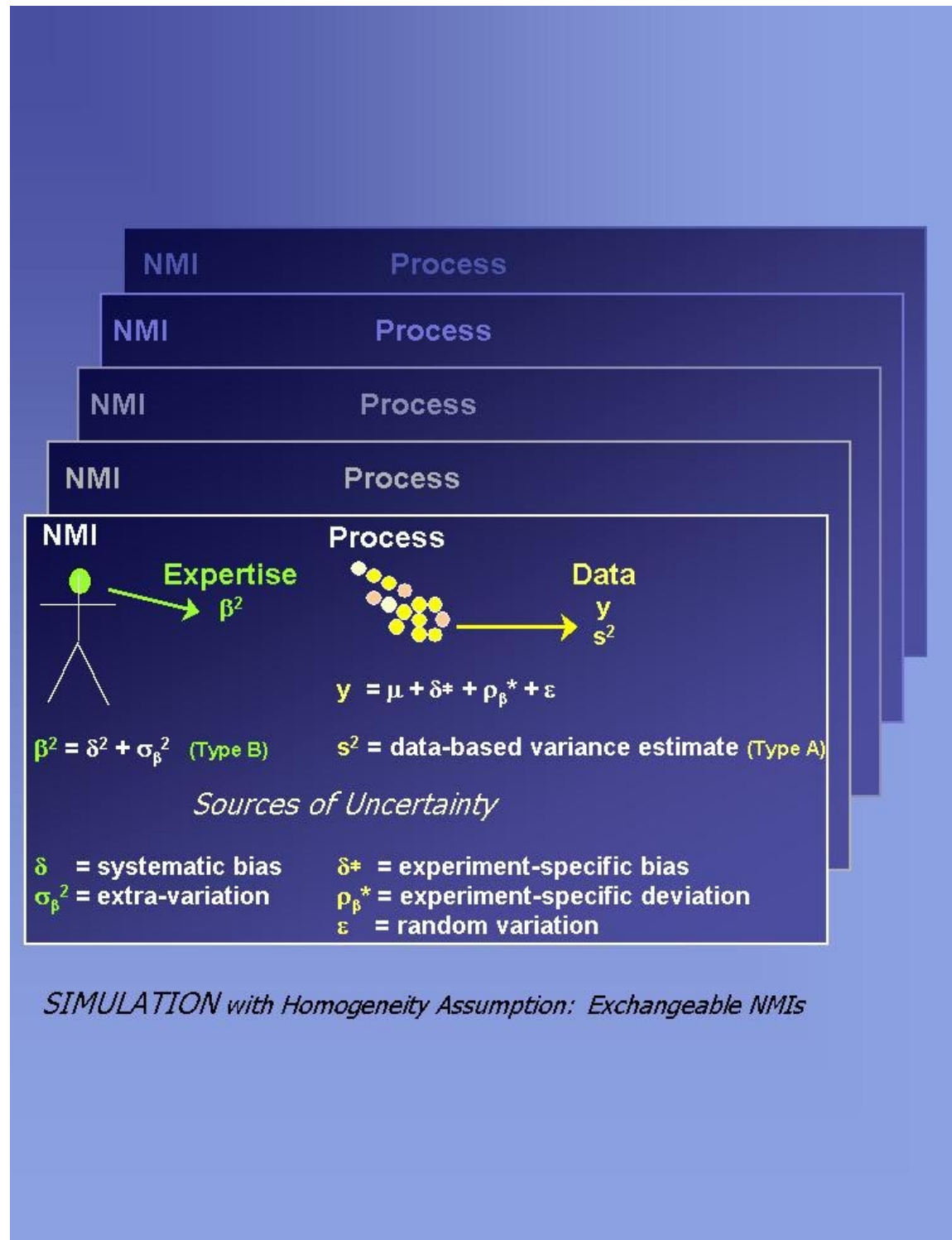
Under Model 3, the b_i are assumed to have known distributions, F_i . These may be regarded as informative prior distributions on the bias constants, b_i , that are postulated based on scientific judgment. This is, in fact, *required* by policy established by the ISO GUM and is followed by most international standards laboratories. Such informative prior distributions are referred to as type-B distributions in the metrology literature.

Here we assume that the b_i are independent random variables with known distributions, F_i . Generally, F_i is assumed to be either normal or uniform, although some other distributions have also been discussed in the GUM. We do not restrict F_i in any way other than that they are fully specified. Although, in practice, it is generally assumed that the b_i are mutually independent, all we need for our procedure to be implementable is that the joint distribution of (b_1, \dots, b_k) is fully specified. We develop a confidence interval for μ using the method of generalized pivotal quantities. Simulation results show that the confidence interval maintains its coverage probability at or above the nominal value.

We propose confidence interval procedures for the mean of measurements of three statistical models that have been used to analyze data from interlaboratory trials. The existing methods either produce no satisfactory confidence intervals when k is small, or have some theoretical shortcomings. The proposed methods provide adequate solutions for a class of problems that are important to NIST activities.

3.2.7 Simulation Study of Estimation Procedures for Key Comparisons

Andrew Rukhin, Nell Sedransk, Blaza Toman
Statistical Engineering Division, ITL



International Key Comparisons give customers a basis for utilizing the measurement services and the calibrated devices and products produced in different nations around the world. These international experiments establish the degree of equivalence between the standard measures from different National Metrology Laboratories (NMIs). From such an international experiment, a Key Comparison Reference Value (KCRV) is often useful; but the methodology for calculating a KCRV remains controversial. Simulation offers a way to study the behavior of calculated KCRV values and their associated uncertainties using different methods.

The statistical issues for the analysis of Key Comparisons data arise from the complicated structure of the uncertainties associated with the data from each NMI that participates. For each single NMI, the process generates observations that are summarized by a single “best” value, y (usually a mean), and a data-based variance estimate, s^2 (Type A uncertainty). The “best” value absorbs the specific experiment’s contribution of both random and systematic variation: experiment-specific bias, experiment-specific deviation, random variation (often called random error).

Because the variation among repeated measurements is usually small in these high-precision laboratories, the uncertainty due to causes not directly observable is often substantial; and expert opinion (Type B uncertainty) is substituted for data-based estimates. This expert opinion may be used to assess bias (systematic effects at that NMI) or to give bounds for extra-variation (random effect) localized to that NMI, but permeating the experiment there. It is important to note that while the expert provides an opinion about these components of Type B uncertainty, the expert’s perception and the actual influence (present in the experimentally observed values for the measurand) can be expected to differ.

Commonly used estimators for the KCRV, denoted by μ , incorporate the “best values” and some or all of the uncertainties in various ways. This extensive simulation examines the vulnerability of eight different KCRV estimators to the several sources of uncertainty. The estimators fall into three groups:

1. unweighted methods for combining data from all the participating NMIs:
 - median, simple mean,
2. data-weighted method:
 - Graybill-Deal estimator,
3. weighted methods where the weights are functions of both data and expert opinion:
 - combined uncertainty-weighted estimator, Mandel-Pauley estimator, DerSimonian-Laird estimator, Bayesian estimator (also its meta-analysis model analogue).

For each estimator, correct calculations for uncertainties are given and the behavior of these uncertainties investigated as well.

In the initial simulations for the case where the NMIs are exchangeable and the random variation is taken to be Gaussian (without contamination or deliberate outliers), the median and the Graybill-Deal estimator provide the least satisfactory answers. All the methods studied by simulation were also applied to data from a Key Comparison with twelve participation NMIs. The results obtained in the simulation were confirmed by data for sinusoidal linear accelerometers over a frequency range from 40 Hz to 5KHz.

The choice of appropriate statistical methodology for international comparisons is central to the success of Key Comparisons conducted by the ten Consultative Committees under the Comité des Poids et Mesures. Correct understanding of the degrees of equivalence between NMIs is the basis for each customer to make their own determinations about interchangeability, depending on the customer's own particular application. Most often that determination depends not only on the degree of equivalence between two NMIs but also on the uncertainty associated with the difference, necessitating a clear and correct understanding of that uncertainty as well.

3.3 IT Performance

3.3.1 Statistical Analysis and Prediction of Extreme Network Performance

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

David Su Doug Montgomery Mark Carson
Advanced Network Technology Division, ITL

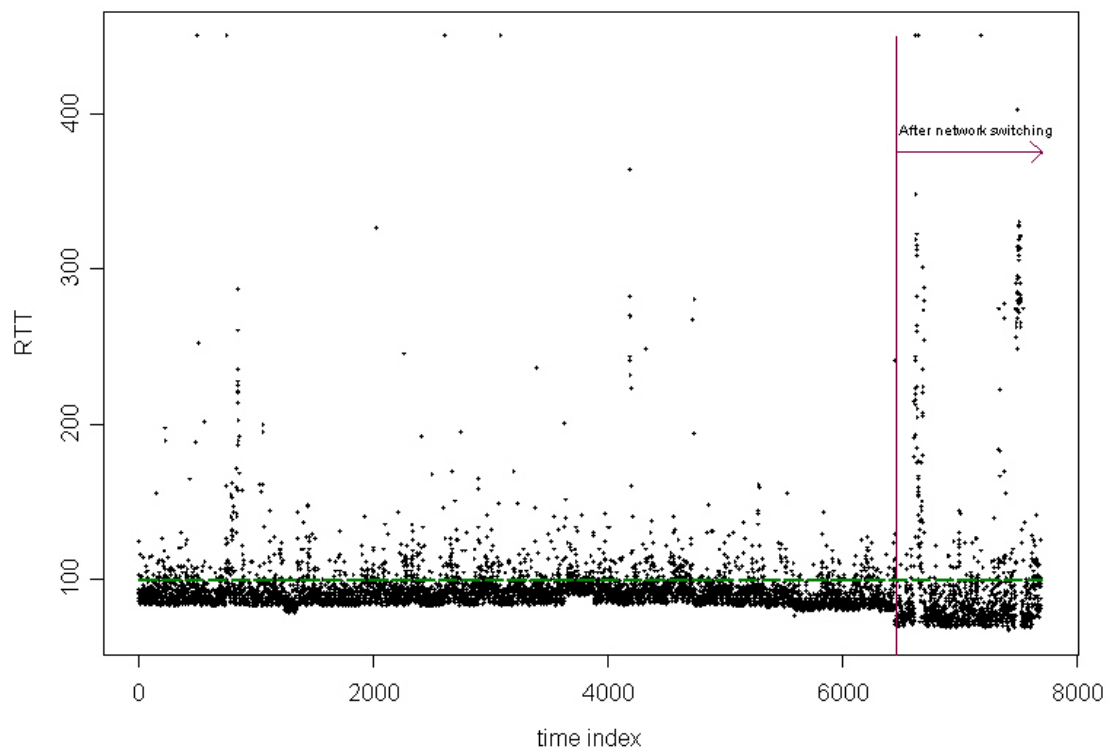


Figure 3.13: : RTT delay time series (in ms) between NIST and HP.

In Internet network performance, extreme performance such as RTT above 100 ms may be a concern to ISP providers and customers alike. How to evaluate and quantify such performance based on available measurement tools such as pingER is a challenging statistical issue because of the sparsity of extreme data. For this reason we propose the generalized Pareto model (GPD), which is an “extreme” form of the limiting distribution of exceedances over a high threshold for “normally-behaving” (aka stationary) network traffic. Thus, it can be used to model the tail behaviors of various heavy-tailed processes proposed in the literature. For non-stationary or anomalous events, we use mixture models such as the mixture of GPD and a uniform. Applications to the performance evaluation of NIST’s network before and after a local network change in 1998 have revealed interesting insights.

Internet network modeling and simulation are characterized by the efforts in “searching for invariant” characteristics of data, in the sense that given the tremendous amount of data collected over different time scales at different locations of the spatial-temporally varying Internet, one is forced to consider only those models that can model observed features that generalize to different data sets. In order for the models or simulation to be relevant, they need to be validated by different data sets over a wide range of temporal and spatial scales (Floyd and Paxson 2001). The two most commonly observed features of data network traffic are the heavy-tailed marginal distribution and long range (or self-similarity) dependence of many network time series, which appear to be found in a wide range of data sets. The two properties are also related, as for example, the self-similar or long range property in observed network measurements may be explained through a heavy-tailed ON-session with regular-varying tails in the so-called ON-OFF process, also called the immigration-death process or $M/G/\infty$ queuing model that is due originally to Cox’s construction (Willinger et al 2002, Crovella and Bestavros 1997). Though mathematically interesting, this line of impressive research is not without insurmountable difficulty in model validation. In particular, there is limited useful data because the Internet is continuously changing and there are limited stationary data at a given time period, as well as sparsity of data in the extreme tails. Consequently, the claim of self-similarity or long-range dependence based on predominantly heuristic statistical techniques has been called into question (Park and Willinger 2000, p.7). In recognition of the uncertainty of statistical models for tail inference, we seek a broader class of parsimonious statistical models that are justifiable on the basis of sound probabilistic/statistical theory, and that can model heavy-tailed processes in different situations at multiple time scales and amplitudes. Based on the more reliable statistical tail models, we also develop new metrics based on extreme quantiles for network performance evaluation that are based on network latency.

Network latency is typically the time taken for a packet to make the round trip from your end-user’s computer to the distant server and back. When measuring latency one should use an application implemented within the server’s IP stack, and which requires the server to perform very little processing to generate a response. Ping is most commonly used for this purpose. ISPs commonly use ping to measure latency, and assume that the server delay is small compared to the forward and reverse delays. The resulting ‘latency’ measure is thus a somewhat coarse indication of Internet performance, but

nonetheless a very widely-used performance metric.

A typical “service latency” specification goes something like this: “Our network latency is the average round-trip time for packets sent between any pair of our backbone routers. It will not exceed 85 ms for pairs of routers located anywhere in Europe.” Note that only average latency is specified, the maximum observed latency can be much higher than the specified value. Besides the median or mean Round Trip Time (RTT) in units of milliseconds (ms, which is one thousandth of a second), performance metrics may also include the median percentage of packet loss, the median unreachability, and related jittering (variability of delays) characteristics.

In this paper, we use the database collected by the Cross-Industry Working Team (XIWT) (<http://www.xiwt.org/>) as an example to illustrate our methodology. The data collection procedure and preliminary findings are reported in the paper “Internet Service Performance: Data Analysis and Visualization” (XIWT 2000). The database consists of RTT series for pairs of sites that participated in this project, and the pingER tool is used. The data were collected in the second half of 1998, with sampling intervals of every half hour. Data values are the mean RTT of the 10 pingER results sent within a given half-hour period. Figure 3.13 shows the time plot of the NIST to HP RTT series (except a few values above 450 ms that were removed to give a visually better plot). It is obvious that the data show a change of network behavior around time index 6453, since the minimum RTTs declined significantly. It turned out that NIST had changed the network during late 1998. Apparently the minimum RTT is much faster after this local network change. But has the upper tail (extreme) network performance improved as well?

Since the class of heavy-tailed distributions is very rich (Goldie and Klüppelberg 1998), one wonders how one can come up with a class of parsimonious parametric models that can be fitted reliably to a broad range of data in the tails. As a compromise, we argue that the generalized Pareto distribution (GPD) model from extreme value theory in time series analysis (Davison and Smith 1990) is the right model. Though the GPD model includes the popular power-law (regular-varying) model as a special case, it is much more flexible and stable since it includes the limiting exponential distribution which corresponds to the asymptotic form for a wide class of heavy-tailed models, namely the subexponential distributions (Goldie and Resnick 1988, Goldie and Klüppelberg 1998).

Mathematically, we assume a time series $X_1, X_2, \dots, X_n, \dots$ to be strictly stationary, in the sense that the joint statistical distribution of any $X_{(i+1)}, \dots, X_{(i+k)}$ does not depend on time shift i , for any k . Our interest is that the marginal distribution of X_1 has CDF $F(x) = P(X_1 \leq x)$ for $0 \leq x < x_F$ or quantile $Q(p) = F^{-1}(p)$ for $0 < p \leq 1$, where $x_F \leq \infty$ is the right endpoint of F , i.e., $x_F = \sup\{x: F(x) < 1\}$. Let $Y = X - u$, the amount of exceedance over a pre-chosen threshold u . Then the conditional cumulative distribution function (CDF) of Y conditioning on $X > u$ is given by:

$$F_u(y) = P(Y \leq y) = P(X \leq u + y | X > u) = \{F(u+y) - F(u)\} / \{1 - F(u)\}.$$

When u is very large, tending to x_F , the functional form of $F_u(y)$ is approximated by the GPD model (Pickands 1975):

$$G(y; a, k) = 1 - (1 - k y/a)^{1/k} \text{ where } k \neq 0$$

$=1-\exp(-y/a)$ where $k=0$,

with

a: scale parameter, $a = a_u$ which depends on the threshold u ,

k: shape parameter

and the range of x is dependent on k : $0 \leq y < \infty$ for $k \leq 0$

and $0 \leq y < a/k$ for $k > 0$.

In particular, when $k \leq 0$, the GPD model provides a theoretically justifiable simple model for the tails of distributions with an infinite right endpoint $x_F = \infty$. When $k < 0$, it is just a reparameterization of the classic Pareto distribution (power-law), and exponential when $k=0$. Figure 3.14 shows the plots of the GPD density functions for various values of k with $a=1$.

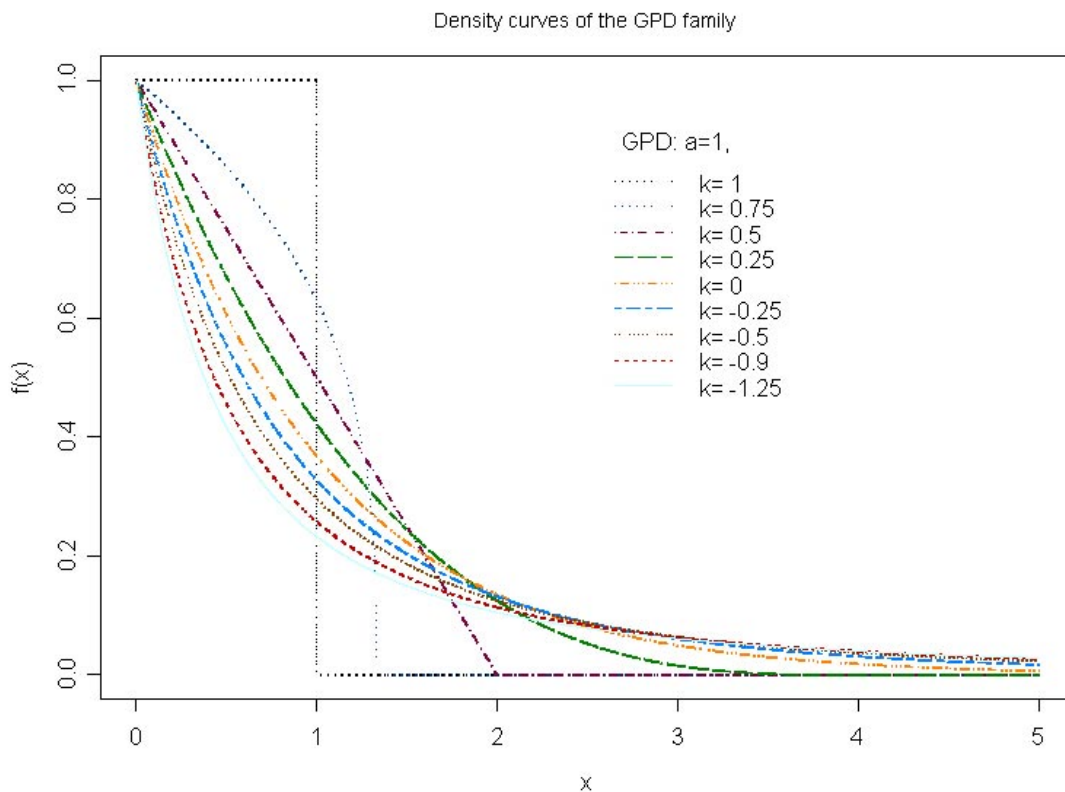


Figure 3.14: An illustration of rich behavior of the GPD density function: for different values of the reciprocal power parameter k (the scale parameter $a=1$ for all cases).

However, the time series stationarity assumption is in doubt for the data after the NIST network switch. Zooming in on the plot of the data after the NIST network change, there are clearly two episodes, or “pulses”, of severe network slowdowns (time indexes 6610-

6700, 7477-7523). Figure 3.15 shows the histograms separately for the normal traffic and pulsatile spurts. The network delays during the “pulsatile” periods appear to be much longer and follow roughly a uniform distribution on the interval of 72 ms to 350 ms.

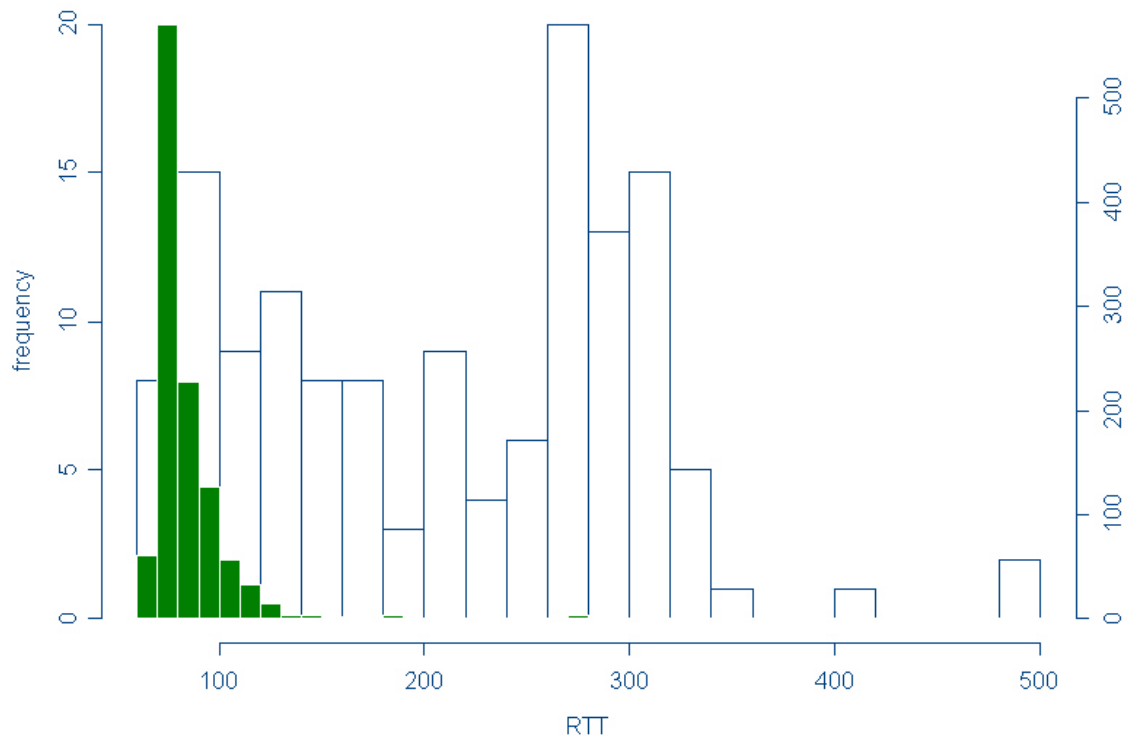


Figure 3.15: Mixture distribution due to anomalous contaminations after network switch: Our mixture analysis identifies two spurts or “pulses” of prolonged network slowdowns that are responsible for causing multimodal mixture behavior in the marginal tail distribution.

The mixture model approach is a quite powerful tool for handling the time-varying and heterogeneous Internet network traffic. For example, the well-recognized “many mice and few elephants” phenomenon in network traffic patterns is just a reflection of the two types of user behaviors, with the “elephant” traffic representing a few fast network connections but very long applications. The mixture model approach is also a very effective mechanism for representing non-Gaussian pulsatile time series data such as river flow and runoff time series (Lu and Berliner 1999).

Figure 3.16 shows the histogram of network time series before and after the NIST network change (the two anomalous “pulses” are removed). It is seen that the NIST network change does indeed result in significantly faster connection times! But the right tail behavior (beyond 100 ms) does not appear to differ much!

Indeed, by fitting the GPD models to the tails, which seem to be good fits in both cases,

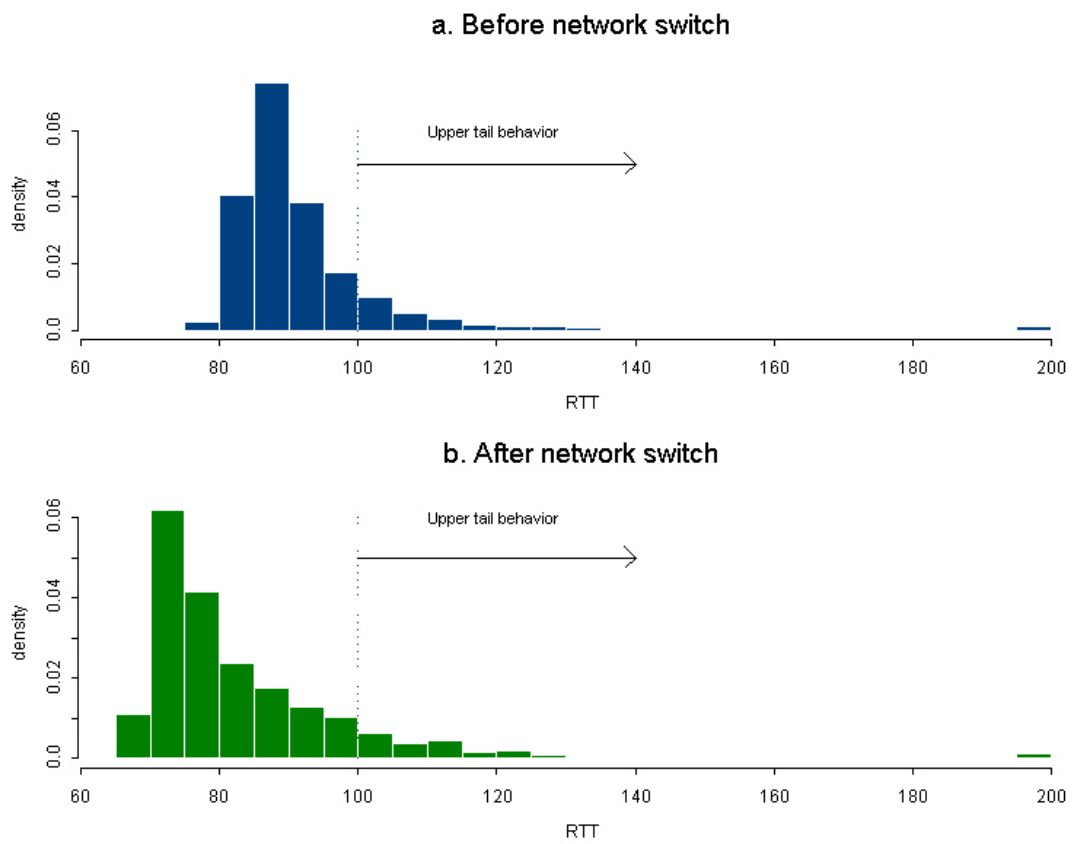


Figure 3.16: Compare the distribution of RTT delays before and after network switching (without the pulsatile events):

the resulting models are very close (see the last figure), and the tail performance after the network switch is only slightly better.

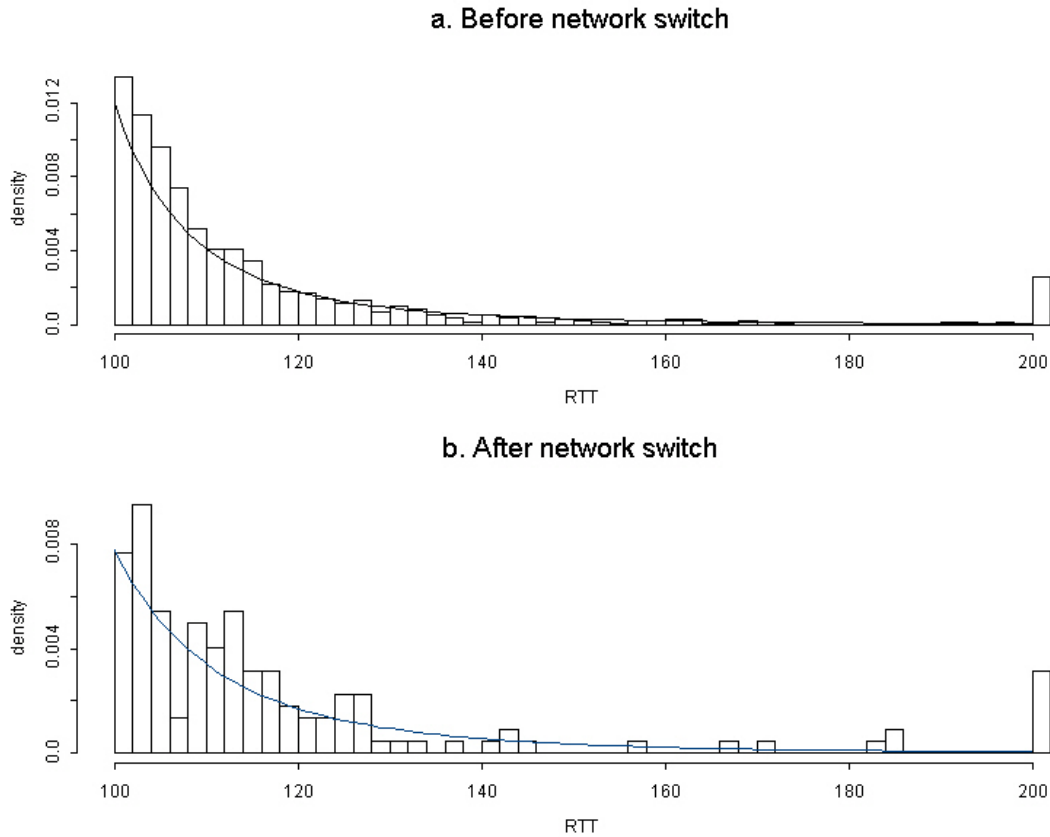


Figure 3.17: Tail histogram and density comparison. a. The solid line denotes the GPD density fit ($u=100$ with exceedance rate of 0.13, $a=10.26$ (0.56), $k=-0.48$ (0.05)). b. The solid line denotes the GPD model fit to the “normal” data after the network switch with $u=100$ (exceedance rate 0.11), MLE fit with $a=12.85$ (1.88), $k=-0.50$ (0.12).

Though we have focused on the NIST-HP RTT data to illustrate our methodology, the same analysis has been tried for other data sets in the XIWT pingER database. In all cases, the GPD model seems to work very well for tail data and the network performance behaviors we have observed seem to persist in other NIST net-related data sets that are collected during the same period. We also anticipate similar analyses and development of tools for other databases that are currently collected elsewhere, such as the pingER RTT database at <http://www.nlanr.net/>. We believe that our methodology should be useful for routine network traffic monitoring and anomaly detection in upper tails. Other potential applications are in network traffic simulation such as the NS-2 network simulator at <http://www.isi.edu/nsnam/ns/> or NIST Net, a Linux-based network traffic emulation tool developed by NIST at <http://snad.ncsl.nist.gov/itg/nistnet/>.

The present work suggests two important directions for further research, with one being to model and incorporate the temporal dependence in the extreme value and mixture models. The growing literature on extreme value theory of stochastic processes is very

relevant. Another direction is to incorporate the hierarchical network structure in the present models, such as user behaviors and protocol/application characteristics. The structural network model may allow us to develop much-needed tools for network performance prediction.

References:

This report is based in part on a paper by Lu and Sedransk (2002) that is under review by IEEE Transactions on Networking. Other cited references are given below.

Crovella, M.E. and A. Bestavros (1997). Self-similarity in world wide web traffic: evidence and possible causes. IEEE/ACM Transactions on Networking, 5(6), 835–846.

Davison, A.C. and R.L. Smith (1990). Models for exceedances over high thresholds (with discussion). J. of Royal Statistical Society, Ser.B, 52, 393–442.

Goldie, C.M. and C. Klüppelberg (1998). Subexponential distributions. In R. Adler, R. Feldman, and M.S. Taqqu (Eds.). A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy Tailed Distributions, 435–459. Birkhäuser, Boston.

Goldie, C.M. and S. Resnick (1988). Distributions that are both subexponential and in the domain of attraction of an extreme value distribution. *Advances in Applied Probability*, 20, 706–718.

Lu, Z.Q. and L. M. Berliner (1999). Markov switching time series models with application to a daily runoff series. Water Resources Research, Vol. 35, No. 2, 523–534.

Park, K. and W. Willinger (2000). Self-similar Network Traffic and Performance Evaluation. Wiley, New York.

Willinger, W., R. Govindan, S. Jamin, V. Paxson, and S. Shenker (2002). Scaling phenomena in the Internet: Critically examining criticality. Proceedings of Natl. Acad. Sci. USA, Vol. 99, Suppl. 1, 2573–2580, February 19, 2002.

In summary, we have demonstrated a flexible and stable model for network traffic data analysis in the generalized Pareto model for tails, which appears to work for multiple time scales and at different data collection sites. We have discussed how to identify and model anomalous and “jamming” network traffic through graphical tools and mixture models. We have developed much needed tail-data based performance metrics and such criteria have shown that the NIST network change in 1998 resulted in only marginally better tail performance.

3.3.2 Modeling the Recovery Process for Notification and Polling

Andrew Rukhin

Statistical Engineering Division, ITL

Kevin Mills, Chris Dabrowski

Advanced Network Technologies Division, ITL

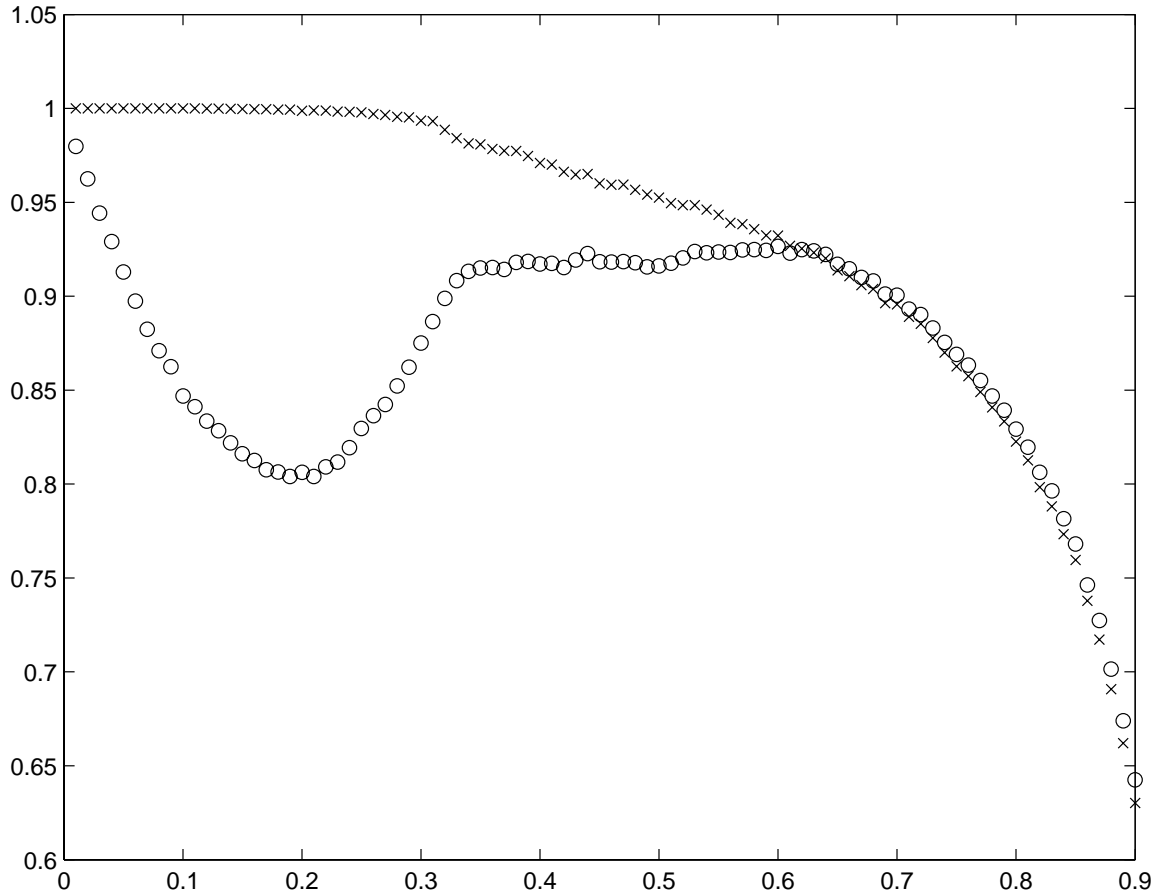


Figure 3.18: Graphs of the update effectiveness for two-party notification ('o') and polling structures (x) for $0 < f < 1$.

This project is to investigate different service-discovery protocols that enable software components to locate available services and to adapt to a stochastic change in the system. These protocols specify alternative architectures and allow a mathematical study of the properties underlying their designs.

In this study we looked at several characteristics of maintaining consistency in a distributed system during catastrophic communication failure. These include: (i) update responsiveness (the average time needed for a service unit to learn about the change in the system), (ii) update effectiveness (the probability distribution of the random time needed to propagate the change in the system), and (iii) update efficiency (the expected number of messages exchanged before the consistency is attained).

Two different architectures (two-party and three-party) were investigated under two consistency-maintenance mechanisms (polling and notification). The behavior of the above characteristics has been studied for these combinations as a function of the failure rate. One probabilistic model includes the uniform distribution for the moment of the directory change and the moment(s) of interface failures. This model was verified by means of an architectural-description language Rapide. Another model (more in the spirit of reliability theory) assumes that the interface failures occur as a renewal process while the directory change is described by an independent Poisson process. The latter model permits the study of multiple interface failures/directory changes.

The resulting formulas demonstrate possible lack of monotonicity (the so-called "saw-tooth effect") for the mentioned characteristics as functions of the failure rate. They indicate the advantages of a three-party architecture in the real-life situations when co-operating software systems may disappear due to physical or cyber attacks, or due to jamming of communications channels, or movement of nodes beyond communications range. These results could also lead to better allocation designs of software components and optimal consistency-maintenance mechanisms.

One node One node fails at random moments T_1, T_2, \dots and recovers after periods D_1, D_2, \dots . The common distribution of T_1, T_2, \dots is a mixture of three uniform distributions. The instants at which the node recovers are $T_1 + D_1, T_1 + D_1 + T_2 + D_2, \dots$ (these are points of recovery); the instants at which the node fails are $T_1, T_1 + D_1 + T_2, \dots$.

The probability that a node is down at a particular moment s can be estimated for large s according to renewal theory for processes with positive repair time by the ratio $\kappa = ET/(ET + ED)$, the so-called *availability* coefficient. The non-functional time then is $(1 - \kappa)D_{run}$, and this is approximately $ED \times EN(D_{run})$ with $N(D_{run})$ denoting the number of failures in the period $(0, D_{run})$. As $ET = t \leq D_{run}$ and this bound is attained when $f = 0$ (f is the failure rate), this coefficient is of little use for small failure rates. The adjusted availability coefficient

$$\kappa_a = \frac{E(T|T < D_{run})}{E(T|T < D_{run}) + ED}$$

gives a better, but still rather poor, approximation to the probability that a node is down at instant s , $s \leq D_{run}$ for small f .

Several nodes With regard to the duration of SM's (Service Manager) failure in the case when there are several, say, N nodes, the average nonfunctional time for the system formed by these nodes is $N(1 - \kappa)D_{run}$. In our setting this number is to be diminished by the simultaneous nonfunctional time of all independent nodes,

$$D_{run}(1 - \kappa_s) = D_{run}(1 - \kappa)^N.$$

Here

$$\kappa_s = 1 - (1 - \kappa)^N$$

is the availability of a system consisting of N nodes working in parallel.

Thus, the total nonfunctional time per SM can be estimated by

$$D_{run}(1 - \kappa) \left[1 - \frac{1}{N}(1 - \kappa)^{N-1} \right].$$

and the probability of failure for a system formed by one (out of four) nodes is

$$P_{ServFail} = 1 - \kappa - \frac{1}{N}(1 - \kappa)^N.$$

The efficient knowledge discovery that enables software units to find available services is crucial for dynamic combination of future software systems. The appropriate mathematical and statistical methodology is vital for the description of this process, and this work provides such a methodology.

3.3.3 Fusion of Biometric Algorithms

Andrew Rukhin

Statistical Engineering Division, ITL

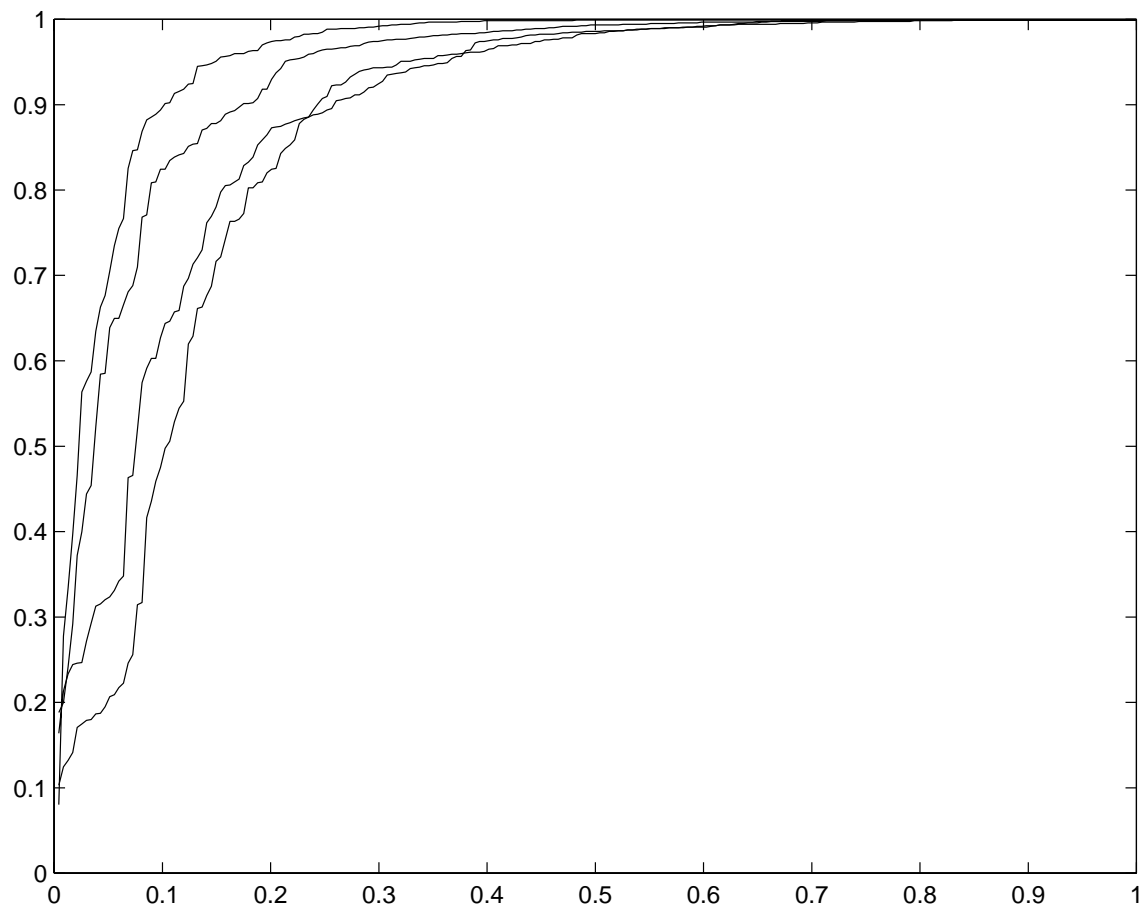


Figure 3.19: Graphs of the cumulative match curves for the algorithms 1–4 The algorithms are ordered like $(2, 4, 3, 1)$.

Biometric systems play an important role in homeland security for the purpose of law enforcement, sensitive areas access, borders and airport control, etc. These systems, which are designed to detect or to verify a person's identity, are based on the fact that all members of the population possess unique characteristics (biometric signatures) such as facial features, eye irises, fingerprints and gait, which cannot be stolen or forgotten. A variety of commercially available biometric systems are now in existence; however, in many instances, there is no universally accepted optimal algorithm. For this reason it is of interest to investigate possible aggregations of two or several different algorithms.

This project was to investigate such a fusion for algorithms in the recognition or identification problem, in which a biometric signature of an unknown person, also known as *probe*, is presented to a system, which compares the new signature with a database of, say, N such signatures of known individuals. On the basis of this comparison, an algorithm presents the similarity scores of this probe to the signatures in the database, called the *gallery*. The gallery items are then ranked according to their similarity scores of the probe. The top matches with the highest similarity scores are expected to contain the true identity.

A common feature of many recognition algorithms is representation of a biometric signature as a point in a multidimensional vector space. The similarity scores are based on the distance between the gallery and the query (probe) signatures in that space (or their projections onto a subspace of a smaller dimension). Because of inherent commonality of the algorithms, the similarity scores and their resulting orderings of the gallery can be dependent for two different algorithms.

As the exact nature of the similarity scores derivation is typically unknown, the use of non-parametric measures of association is appropriate. The utility of statistics such as rank correlation statistics like Spearman's rho or Kendall's tau for measuring the relationship between different face recognition algorithms has been already studied. For common image recognition algorithms, the strongest correlation between algorithms similarity scores happens for both large and small rankings. Thus, in all observed cases the algorithms behave somewhat similarly, not only by assigning the closest images in the gallery but also by deciding which gallery object is most dissimilar to the given image.

The example considered comes from the FERET (Face Recognition Technology) program in which four recognition algorithms each produced rankings from a gallery consisting of $N = 1196$ images and 234 probe images taken between 540 and 1031 days after its gallery match.

It is suggested to think of the action of an algorithm (its ranking) as a permutation π of N objects in the gallery. Thus $\pi(i)$ is the rank given to the gallery item i ; in particular, if $\pi(i) = 1$, then the item i is the closest image in the gallery to the given probe.

If the goal is to combine K independent algorithms whose actions π_j can be considered as permutations of a gallery of size N , then the combined (average) ranking of observed rankings π_1, \dots, π_K can be defined by the analogy with the classical means. Namely, let

the "average permutation" $\hat{\pi}$ of π_1, \dots, π_K be the minimizer (in π) of

$$\sum_{j=1}^k d(\pi_j, \pi)$$

Then one can take $\hat{\pi}$ as the action of the combined algorithm.

A possible model for the combination of dependent algorithms employs a distance $d((\pi_1, \dots, \pi_K), (\sigma_1, \dots, \sigma_K))$ on the direct product of K copies of the permutation group. Then the combined (average) ranking $\hat{\pi}$ of observed rankings π_1, \dots, π_K is the minimizer (in π) of $d((\pi_1, \dots, \pi_K), (\pi, \dots, \pi))$. The simplest metric is the sum $\sum_{j=1}^k d(\pi_j, \pi)$ as above. It is convenient to associate with a permutation π the $N \times N$ permutation matrix P with elements $p_{i\ell} = 1$, if $\ell = \pi(i)$; $= 0$, otherwise. A distance between two permutations π and σ can be introduced as the matrix norm of the difference between the corresponding permutation matrices.

For a matrix P , one of the most useful matrix norms is

$$\|P\|^2 = \text{tr}(PP^T) = \sum_{i,\ell} p_{i\ell}^2.$$

Here $\text{tr}(A)$ denotes the trace of the matrix A .

For two permutation matrices P and S corresponding to permutations π and σ , the resulting distance $d(\pi, \sigma) = \|P - S\|$ essentially coincides with Hamming's metric

$$d_H(\pi, \sigma) = N - \# \{i : \pi(i) = \sigma(i)\}.$$

A useful distance is defined by a positive definite symmetric matrix C as

$$\begin{aligned} d((\pi_1, \dots, \pi_k), (\sigma_1, \dots, \sigma_k)) &= d_C((\pi_1, \dots, \pi_k), (\sigma_1, \dots, \sigma_k)) \\ &= \text{tr}((\Psi - \Sigma)C(\Psi - \Sigma)^T), \end{aligned}$$

with $\Psi = P_1 \oplus \dots \oplus P_k$ is the direct sum of permutation matrices corresponding to π_1, \dots, π_k , and Σ is defined similarly for $\sigma_1, \dots, \sigma_k$.

The optimization problem, which one has to solve for this metric, consists of finding the permutation matrix Π minimizing the trace of the block matrix formed by submatrices $(P_j - \Pi)C_{jm}(P_m - \Pi)^T$, with C_{jm} denoting $N \times N$ submatrices of the partitioned matrix C . Matrix differentiation shows that the minimum is attained at the matrix

$$\Pi_0 = \left[\sum_j P_j C_{jj} \right] \left[\sum_j C_{jj} \right]^{-1}.$$

The matrix Π_0^T is stochastic, i.e., with $e = (1, \dots, 1)^T$, $e\Pi_0 = e$, but typically it is not a permutation matrix, and the problem of finding the closest permutation matrix, say, determined by a permutation π_0 , remains. An efficient numerical algorithm for finding π_0 is based on the so-called Hungarian method for the assignment problem. In this problem with $\Pi_0 = \{\hat{p}_{i\ell}\}$

$$\pi_0 = \arg \max_{\pi} \sum_i \hat{p}_{i\pi(i)}.$$

In this setting one has to use an appropriate matrix C , which should be data-dependent, with C^{-1} being the covariance matrix of all random permutations π_1, \dots, π_k . Because of the necessity of estimating the matrix C and numerical difficulties for large N , one may look for a simpler aggregated algorithm.

Such an algorithm can be defined by the matrix P , which is a convex combination of the permutation matrices P_1, \dots, P_K , $P = \sum_{j=1}^K w_j P_j$. The problem is that of assigning non-negative weights (probabilities) w_1, \dots, w_K , such that $w_1 + \dots + w_K = 1$, to matrices P_1, \dots, P_K . One has $EP_i = \mu$ with the same “central” matrix μ , as in average, for a given probe, all algorithms measure the same quantity, and the main difference between them is the accuracy. Optimal weights w_1^0, \dots, w_K^0 minimize $E\|\sum_j w_j(P_j - \mu)\|^2$.

This optimization problem reduces to the minimization of

$$\sum_{1 \leq j, m \leq K} w_j w_m Etr(P_j P_m^T) - 2 \sum_{1 \leq j \leq K} w_j tr(P_j \mu^T).$$

For $m \neq j$

$$Etr(P_m P_j^T) = E\# \{ \ell : \pi_m(\ell) = \pi_j(\ell) \}.$$

These “covariances” can be estimated from the data by relative frequencies in the previous trials. The same data can be used to estimate μ by the grand mean $\hat{\mu}$ of all available matrices.

Let Σ denote the positive definite matrix formed by the elements $Etr(P_m P_j^T)$, $m, j = 1, \dots, K$. This matrix can be estimated by, say, $\hat{\Sigma}$. The previous data can be used to obtain the estimated optimal weights. After these weights have been determined and found to be nonnegative, one can define a new combined ranking π_0 on the basis of newly observed rankings π_1, \dots, π_k . Let the N -dimensional vector $Z = (Z_1, \dots, Z_N)$ be formed by coordinates $Z_i = \sum_{j=1}^k w_j^0 \pi_j(i)$, representing a combined score of element i . Put $\pi_0(i) = \ell$ if and only if Z_i is the ℓ -th smallest of Z_1, \dots, Z_N . In other terms, π_0 is merely the rank corresponding to Z . In particular, according to π_0 the closest image in the gallery is m_0 such that

$$\sum_{j=1}^k w_j^0 \pi_j(m_0) = \min_m \sum_{j=1}^k w_j^0 \pi_j(m).$$

This ranking π_0 is characterized by the property

$$\sum_{i=1}^N \left(\sum_{j=1}^k w_j^0 \pi_j(i) - \pi_0(i) \right)^2 = \min_{\pi} \sum_{i=1}^N \left(\sum_{j=1}^k w_j^0 \pi_j(i) - \pi(i) \right)^2$$

i.e., π_0 is the permutation that is the closest in the L_2 norm to $\sum_{j=1}^k w_j^0 \pi_j$.

Encouragingly, these weights correspond to the ordering of the algorithms by their cumulative match curves (see Figure I).

In the FERET experiment data the algorithms are : 1 = MIT, March 96 (the smallest weight); 2 = USC, March 97 (the largest weight); 3 = MIT, Sept 96 (the second smallest weight); 4 = UMD, March 97 (the second largest weight). The combined algorithm

behaves better than the best in this group, namely, the algorithm 2 especially for small rankings!

This method can be easily extended to the situation when only partial rankings are available, i.e., when only the several top ranks are given. In this case one has to consider metrics on the coset space of all permutations with respect to the set of permutations that leave the first several ranks fixed.

These results show how to construct new procedures designed to combine several algorithms. Notice that the methods of averaging or combining ranks can be applied to several biometric algorithms, one of which, say, is a face recognition algorithm, and another is a fingerprint (or gait, or ear) recognition device. They can be useful in a verification problem when a person presents a set of biometric signatures and claims that a particular identity belongs to these signatures.

3.4 Process Characterization

3.4.1 Process Characterization - Overview

C.M. Wang

Statistical Engineering Division, ITL

NIST statisticians collaborate with other researchers throughout the NIST Laboratories and also with their industrial partners to characterize complex processes and to address measurement and standards aspects of physical science, engineering, and information technology. Together with subject-matter experts, NIST statisticians develop techniques for evaluating complex physical processes or measurement processes, for tying measurement processes to accepted standards, and for ensuring the quality of measurements.

The primary goal of the project is to assure that appropriate and state-of-the-art statistical planning and analysis is used in NIST work. In some cases, the statistician judiciously selects and implements the most appropriate existing statistical method to analyze experimental data. In many cases, new statistical methods are developed to address unique scientific challenges encountered by the NIST research team. Some SED staff develop theoretical models to augment experimental work done by NIST collaborators. Examples of such work include Monte Carlo simulation of physical processes and stochastic differential equation modeling. Typically, SED staff develop long-term relationships with collaborators in the other NIST Laboratories and develop intimate knowledge of the scientific fields in which they work.

Statisticians develop probabilistic models for physical processes and statistical models for combined uncertainty analysis. Some examples of process characterization include stochastic models for high-speed communications using optical fibers, new measurement methods for characterizing the complex permittivity of dielectric materials (widely used throughout electronics, microwave, communication, and aerospace industries), statistical models for polymer temperature and pressure measurement during fabrication, and characterization of high-speed oscilloscopes for use in optoelectronic device metrology, nonlinear device metrology, high-speed digital circuit design, neutron depth profiling, and subatomic particle lifetimes.

As members of interdisciplinary teams, SED statisticians contribute to process characterization in a variety of ways. SED staff develop appropriate statistical strategies to meet the needs of the research teams and actively participate in the preparation of written records via NIST or archival journal publications. The benefits of process characterization are enjoyed by scientists and engineers, either directly or indirectly, in a myriad of industries.

3.4.2 Characterization of High-Speed Optoelectronic Devices

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

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

D. C. DeGroot
RF Technology Division, EEEL

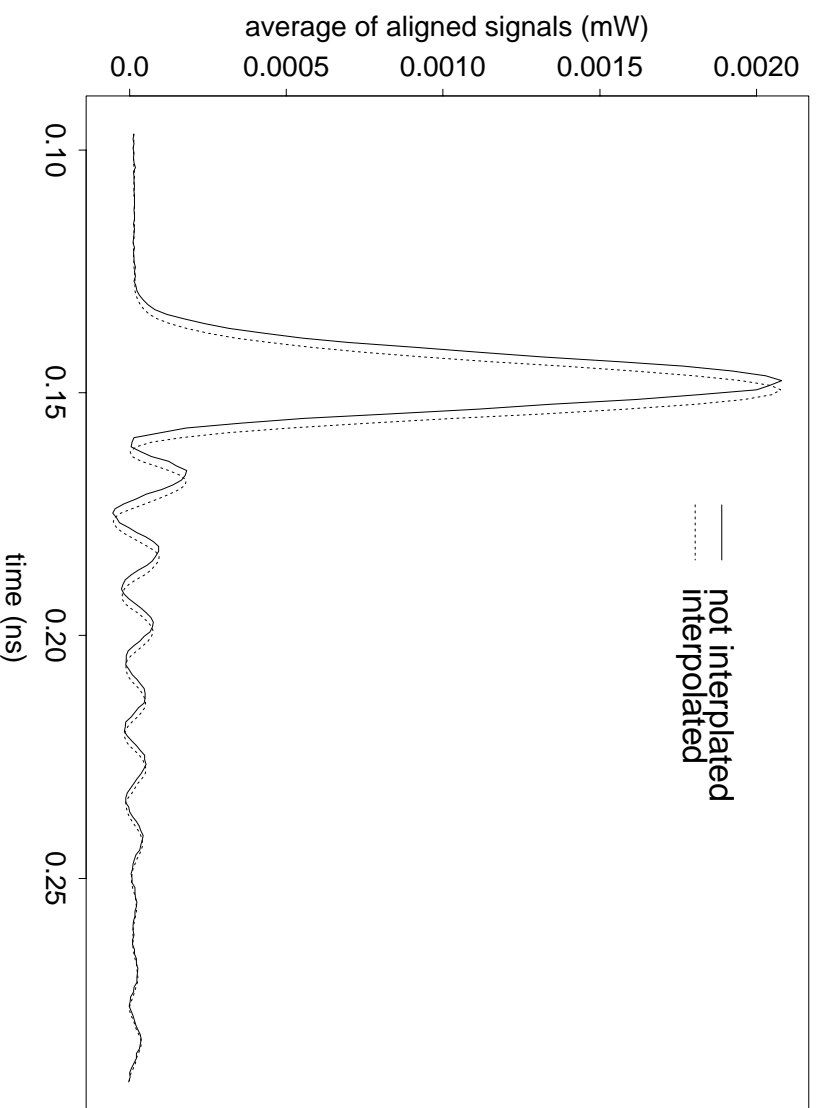


Figure 3.20: A regression spline model is fit to the average of 1000 aligned signals. Based on the estimated TBD, we interpolate onto an equally spaced time grid.

Accurate measurement of high-speed optoelectronic devices, which include the photodiode and sampling oscilloscope, is critical in the design of 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). Research is being done on the next generation of OTDM systems at 40 to 80 gigabits per second in laboratories around the world. To achieve the goal, the industry needs to characterize the impulse and frequency response of high-speed optical reference receivers to at least the third harmonic of the system modulation rate. In support of this effort, the NIST is developing a calibration service for optical reference receivers. This is one of many projects responding to the growing fiber-optic industry's need for standards and calibration where such do not exist.

Optical reference receivers are used for measuring optical waveforms. These optical receivers, however, suffer from several non-ideal properties that must be characterized and compensated for. These effects include timing drift, time-base distortion, timing jitter, and impedance mismatch. In collaboration with the Optoelectronics Division and the Radio Frequency Technology Division of the NIST Electronics and Electrical Engineering Laboratory, SED has developed several statistical signal processing techniques that are being used to correct for the effects of timing drift, time-base distortion, and timing jitter in measurement of optical receivers. Here, we briefly describe some of the SED contributions and accomplishments.

Many waveforms must be averaged to achieve a low noise level. Before averaging, the waveforms are corrected for drift. Relative drifts are estimated from cross-correlation analysis of all distinct pairs of signals. A manuscript on alignment of noisy signals appears in the February 2001 issue of *IEEE Transactions on Instrumentation and Measurement*. In this work, we study the relative performance of various methods for aligning noisy one-dimensional signals. No knowledge of the shape of the misaligned signals is assumed. We simulate signals corrupted by both additive noise and timing jitter noise, which are similar in complexity to nose-to-nose oscilloscope calibration signals collected at NIST. In one method, we estimate the relative shift of two signals as the difference of their estimated centroids. We present a new adaptive algorithm for centroid estimation. We also estimate relative shifts using three different implementations of cross-correlation analysis. In a complete implementation, for m signals, relative shifts are estimated from all $m(m-1)/2$ distinct pairs of signals. In a naive implementation, relative shifts are estimated from just $(m-1)$ pairs of signals. In an iterative adaptive implementation, we estimate the relative shift of each signal with respect to a template signal, which at each iteration is equated to the signal average of the aligned signals. In simulation experiments, for all noise levels, the complete cross-correlation method yields the most accurate estimates of the relative shifts. The relative performance of the other methods depends on the noise levels.

Next, we consider the problem of time-base distortion (TBD) estimation. The model of a discrete time signal is given by

$$s_k = f(t_k) + \epsilon_k,$$

with the k th sample s_k being a function of actual time of sampling, t_k , plus the additive

noise ϵ_k . The actual time t_k can be written as

$$t_k = (k - 1)T_s + g_k + \tau_k,$$

with $(k - 1)T_s$ denoting the ideal sample time and T_s is the sampling interval. Deviations between the ideal and actual times have two components: a deterministic part, g_k , called TBD and a random component, τ_k , called jitter. If left uncorrected, TBD can cause significant errors in pulse width, step transition, and time interval measurements. Discontinuities in the TBD can severely distort a short pulse waveform. Such discontinuities must be detected and avoided in even the crudest measurements. After estimation of the TBD, the measured waveform is interpolated onto an evenly spaced time-grid.

We develop an efficient least-squares algorithm for estimation of TBD and the harmonic distortions simultaneously. The method requires measurements of sinusoidal signals at multiple phases and frequencies. The model of the waveforms of multiple phases and frequencies is given by

$$s_{jk} = \alpha_j + \sum_{i=1}^h [\beta_{ij} \cos(2\pi i f_j t_{jk}) + \gamma_{ij} \sin(2\pi i f_j t_{jk})] + \epsilon_{jk},$$

with s_{jk} denoting the measured signal at time t_{jk} (the k th actual sample time of the j th experiment), f_j is the frequency used in the j th experiment, and β_{ij} and γ_{ij} are the amplitudes of the i th harmonic of the j th experiment. The number of harmonics h is assumed to be finite. The additive noises ϵ_{jk} are assumed to be independently and identically distributed (iid) with zero means and standard deviations $\sigma_\epsilon(j)$. The model allows different additive noise standard deviations for different experiments. The model also assumes that t_{jk} is given by

$$t_{jk} = (k - 1)T_s + g_k + \tau_{jk},$$

with T_s and g_k as defined before, and the τ_{jk} are the random jitters, which are assumed to be iid (and independent of ϵ_{jk}) with zero means and standard deviations $\sigma_\tau(j)$. There will be m experiments with n samples for each experiment; that is, $k = 1, 2, \dots, n$, and $j = 1, 2, \dots, m$.

Let

$$\boldsymbol{\theta} = (g_1, g_2, \dots, g_n, \alpha_1, \beta_{11}, \gamma_{11}, \dots, \beta_{1h}, \gamma_{1h}, \dots, \alpha_m, \beta_{m1}, \dots, \gamma_{mh})^t$$

be the column vector of the unknown parameters of the model. The number of unknowns is $n + m(2h + 1)$. Define

$$z_{jk}(\boldsymbol{\theta}) = \alpha_j + \sum_{i=1}^h [\beta_{ij} \cos(2\pi i f_j ((k - 1)T_s + g_k)) + \gamma_{ij} \sin(2\pi i f_j ((k - 1)T_s + g_k))]$$

and

$$SS(\boldsymbol{\theta}) = \sum_{j,k} (s_{jk} - z_{jk}(\boldsymbol{\theta}))^2.$$

Then the least-squares estimate of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$, is the solution of

$$\min_{\hat{\boldsymbol{\theta}}} SS(\hat{\boldsymbol{\theta}}).$$

A Gauss-Newton type of iterative procedure can be used to solve the minimization problem. If the procedure is implemented directly, it would require $O(n^3)$ operations at each iterative step. This is not acceptable for a large n (in our problems, $n = 4096$). The model has a special structure, however, that can be exploited to obtain an algorithm that requires only $O(n)$ operations at each step. The detailed derivation of the algorithm and other related work appears in the December 1999 issue of *IEEE Transactions on Instrumentation and Measurement*.

The TBD estimation procedure uses weighted nonlinear least squares for parameter estimation. A simulation study showed that the reduction in the root-mean-square (RMS) error of the TBD estimate obtained by using the appropriate weighting is about 20%. The appropriate weighting scheme is to weight each data point proportionally to the inverse of its variance. The variance can be estimated either from independent, repeated measurements or (if we have prior information on the additive and jitter noise variances) from the approximate model

$$\text{var}(s_{jk}) \approx \sigma_\epsilon^2(j) + (f'(t_{jk}))^2 \sigma_\tau^2(j)$$

with $f'(t_{jk})$ denoting the derivative of the measured signal evaluated at $t_{jk} = (k-1)T_s + g_k$. On the other hand, the above expression can be used to estimate $\sigma_\epsilon(j)$ and $\sigma_\tau(j)$ if repeated measurements and the TBD estimate are available. Since estimation of additive and jitter noises requires the knowledge of the TBD to evaluate the time derivative $f'(t_{jk})$, while the TBD estimation routine needs the estimate of $\sigma_\epsilon(j)$ and $\sigma_\tau(j)$ to construct the weights for the least-squares procedure, we use an iterative algorithm. We begin with a set of equal weights, and then estimate the harmonic distortion, TBD, amplitude, and phase parameters. With these estimates in hand, we estimate the jitter and additive noises and use them to form a new set of weights to obtain the TBD and other parameter estimates. This process is repeated until convergence is attained.

The equation for the approximate variance is derived from a first-order expansion of $f(t_{jk})$ on τ . We examined the bias of estimating $\sigma_\epsilon(j)$ and $\sigma_\tau(j)$ using the first-order approximation. The bias is not negligible if the sampling frequency isn't properly chosen and/or the jitter noise is not small. We developed a procedure to adjust for the bias. The procedure is based on a model relating the variance of the measured signal and the additive and jitter noises. Simulations were performed to show the effectiveness of the adjustments. We also showed that the bias of a least-squares TBD estimator obtained from multiple sets of waveforms is small relative to the variance of the estimator, allowing us to compute the uncertainty of the TBD estimate from the standard deviation of individual TBD estimates of each set of waveforms. The proposed uncertainty is integrated over time and ignores any covariance structure in time. We used this uncertainty to monitor the TBD measurements over time. This and other related work appear in the February 2002 issue of *IEEE Transactions on Instrumentation and Measurement*.

In high-speed measurement systems, the target time and actual sampling time may differ because of both systematic TBD errors and random timing jitter errors. For the limiting case in which the signal is sampled continuously and there are no TBD errors or additive noise errors, Gaussian jitter attenuates the power spectrum of the continuously sampled signal by the amount $\exp(-\omega^2 \sigma_\tau^2)$, with the standard deviation of the jitter given by σ_τ . Thus, the power spectrum of a jittered signal is corrected by multiplying it by

$\exp(\omega^2 \sigma_\tau^2)$. At frequencies for which the signal-to-noise ratio is high, this approach generally improves the accuracy of the power spectrum estimate. Earlier, we described how to estimate jitter for the case in which the noise-free signal is a mixture of a sinusoid and its harmonic. Next, we discuss a new method for estimating RMS jitter for the more general case in which the analytical form of the noise-free signal is unknown. A paper based on this work has been submitted to *IEEE Transactions on Instrumentation and Measurement*.

Neglecting TBD errors, we model the j th observed signal at the k th time sample as s_{jk} , with

$$s_{jk} = f(t_k + \delta_j + \tau_{jk}) + \epsilon_{jk}$$

τ_{jk} is a realization of the jitter noise, δ_j is a random time shift (drift) error, ϵ_{jk} is a realization of additive noise and $f(\cdot)$ is an unknown function of time. For each signal, the realizations of the additive noise, jitter noise, and time shift noise processes are assumed to be independent. The realizations of the additive noise and jitter processes have finite variances σ_ϵ^2 and σ_τ^2 . We estimate the relative time shift errors by an all-pairs cross-correlation method. Based on the estimated relative time shift errors, we translate each signal in time using a Fourier method.

We denote the k th time sample of the j th aligned signal as s_{jk}^c . Assuming that we have accurately aligned the signals, a first-order Taylor series argument yields an approximation for the variance of the sampled signal

$$\sigma_s^2(k) = \text{var}(s_{jk}^c) \approx |f'(t_k)|^2 \sigma_\tau^2 + \sigma_\epsilon^2.$$

(Here, we consider the general case in which we do not have an analytic model for the derivative of the noise-free signal at time t_k , $f'(t_k)$. Hence we must estimate this derivative from measurements.) We approximate the jitter variance, σ_τ^2 , as

$$\sigma_\tau^2 \approx \frac{\sigma_s^2(k) - \sigma_\epsilon^2}{|f'(t_k)|^2}.$$

We model the signal average of the noisy signals using B-splines. The B-spline representation at time t is denoted as $s_b(t)$, and the derivative of the B-spline representation at time t is denoted as $s_b'(t)$. To compute our jitter estimate, it is convenient to define the following quantities at the k th time sample

$$a_k = \hat{\sigma}_s^2(t_k) - \hat{\sigma}_\epsilon^2$$

and

$$b_k = |s_b'(t_k)|^2.$$

We expect the ratio of a_k and b_k to be a rough estimate of the jitter variance at all k . Intuitively, we expect more information in (a_k, b_k) data at time samples when the magnitude of the derivative is relatively large. In our studies, the ratio $r_k = a_k/b_k$ had a very large variance at time samples when the magnitude of the signal derivative was very small. Thus, the average of all the r_k values would be a poor estimate of the jitter variance. To reduce the influence of noisy (a_k, b_k) pairs on our estimate, we take two actions. First, we design our estimate so that it depends on (a_k, b_k) values at time samples when the

magnitude of the estimated derivative is greater than a selected threshold. Second, we estimate the jitter noise variance as the ratio of the pooled a_k data and the pooled b_k data. Pooling is a natural way to reduce the influence of highly variable, i.e., noninformative, (a_k, b_k) values on the estimate. Finally, we require that our variance estimate be nonnegative. Thus, our (nonnegative) estimate of the variance of the jitter noise is

$$\hat{\sigma}_\tau = \sqrt{\max\left(0, \frac{\sum_k a_k H(s'_b(t_k), \alpha)}{\sum_k b_k H(s'_b(t_k), \alpha)}\right)}$$

with

$$H(s'_b(t), \alpha) = \begin{cases} 1 & \text{if } |s'_b(t)| > \alpha \max(|s'_b|) \\ 0 & \text{otherwise} \end{cases}$$

and α is an adjustable threshold. Our estimate of the RMS value of the jitter noise is $\hat{\sigma}_\tau$.

By lowering the threshold, we incorporate more of the measured data into our estimate. However, if the threshold is too low, prediction error may increase if we incorporate too much noisy data with little or no additional information content. Since the optimal choice of the threshold is not obvious, we study how the choice of threshold affects results in a Monte Carlo simulation experiment. In general, for any choice of threshold, we expect that the above estimator is biased since it is a nonlinear function of the observed data and nonlinear estimators are generally biased.

We estimate the bias of our estimator using a parametric bootstrap procedure. In the bootstrap simulation model, the noise-free signal is equated to the regression spline model estimate of the average of the aligned observed signals $s_b(t)$. Like the observed data, the synthetic signals are corrupted by time shift errors, additive noise and jitter noise. In the simulation, the time shift parameters are equated to the relative time shift parameters estimated from the observed data. In the bootstrap procedure, we assume that jitter and additive noise are Gaussian random variables with expected values equal to 0 and variances equal to those estimated from the primary “observed” data. The number of signals in each bootstrap set is the same as the number of observed signals. We simulate $m = 30$ bootstrap replications of the observed data.

For each bootstrap replication of the observed data, we estimate relative time shift errors and align the signals using the same algorithms used for the observed data. We estimate a new set of regression spline model parameters, a new RMS additive noise value, and a new RMS jitter noise value $\hat{\sigma}_\tau^*$. The bootstrap estimate of the bias of our jitter estimate is

$$\hat{B}_{boot} = \frac{1}{m} \sum_{j=1}^m \hat{\sigma}_{\tau j}^* - \hat{\sigma}_\tau$$

with $\hat{\sigma}_{\tau j}^*$ denoting the estimate of RMS jitter computed from the j th bootstrap replication. Our bias-corrected estimate of RMS jitter noise is

$$\hat{\sigma}_\tau^* = \hat{\sigma}_\tau - \hat{B}_{boot} = 2\hat{\sigma}_\tau - \frac{1}{m} \sum_{j=1}^m \hat{\sigma}_{\tau j}^*.$$

Provided that the signal is sampled at a sufficiently high rate, we expect our method to be valid for cases in which the noise-free signal is well-approximated as piecewise cubic

polynomials with the first and second derivatives continuous. We recommend that users of our methods perform a stability study to verify that the sampling rate is sufficiently high for the purpose of estimating the RMS value of the jitter noise. We also recommend that users demonstrate that the regression spline has a sufficient number of knots in order to sufficiently model the complexity of the signal of interest. The user should verify that the RMS jitter estimate stabilizes as the number of knots in the regression spline model increases.

In addition to the dissemination of results in refereed journals, we also presented the results in the Automatic RF Techniques Group (ARFTG) Conferences, ASA Spring Research Conference, and many Department of Statistics seminars. We participated in the Telecommunications Industry Association (TIA)/International Electrotechnical Commission (IEC) Standards Committee Working Group 4, TC-86, and presented statistical approaches for development of measurement standards of optical waveforms. Software for implementing these techniques is ready for public distribution.

In the future, we plan to complete a study of the systematic and random errors associated with the statistical signal processing (TBD correction, jitter correction, drift correction) on the power and phase spectrum of the signal of interest.

SEd demonstrated that observed high speed optoelectronic signals can be corrected for the effects of drift, jitter, timebase distortion and impedance mismatch distortion. As a result, the feasibility of a proposed calibration service for a multibillion dollar industry was demonstrated.

3.4.3 Properties of Dielectric Materials

Kevin Coakley, Jolene Splett
Statistical Engineering Division, ITL

Mike Janezic, Raian Kaiser, John Grosvenor
Radio-Frequency Technology Division, EEEL



Figure 3.21: The NIST 60 mm cylindrical cavity resonator.

NIST is developing new methods for characterizing dielectric materials based on measurements of 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.

Background and Completed Work

Permittivity and loss tangent are estimated by placing a cylindrical, cross-linked polystyrene sample in a radio frequency cavity. The estimates depend, in part, on the mean thickness of the sample, the observed Q factor and resonant frequency of the cavity (both with and without the sample), and the electrical length and diameter of the cavity.

We developed two statistical methods for estimating surface roughness and the mean thickness of the samples. The first method models surface roughness as a polynomial surface, while a second technique models surface roughness using a nonparametric method. The estimated mean thickness was in very close agreement for both the parametric and nonparametric models.

To estimate the resonant frequency, f_0 , of the cavity and the corresponding Q factor, we developed a nonlinear parameter estimation algorithm based on the observed resonance curve,

$$T(f) = \frac{T(f_0)}{1 + Q^2(f/f_0 - f_0/f)^2} + BG + \epsilon(f)$$

with BG denoting background and $\epsilon(f)$ is additive noise. In studies involving both real and simulated data, our nonlinear estimation procedure outperformed current state-of-the-art methods used for estimating Q and f_0 .

To estimate the model parameters in an optimal way, we must characterize the frequency-dependent noise in the measurement system. We model the variance of the additive noise as

$$\widehat{VAR}(\epsilon(f)) = \frac{\gamma_1^2}{1 + Q^2(f/f_0 - f_0/f)^2} + \gamma_2^2.$$

We estimate the variance function parameters γ_1 and γ_2 from the residuals computed from a least squares fit of the resonance curve model to the observed data. The use of the selected variance function was justified through analytical derivation for the case of equal variances and no covariance between the real and imaginary components of the scattering parameter.

Based on the empirical estimates of the variance function parameters, we determine: (1) Q and f_0 by nonlinear weighted least squares, and (2) the asymptotic standard deviation of the estimates of Q and f_0 .

In the actual experiment, resonance curves are sampled at a fixed number of frequencies; however, the frequency spacing is adjustable. As a first step in determining the optimal

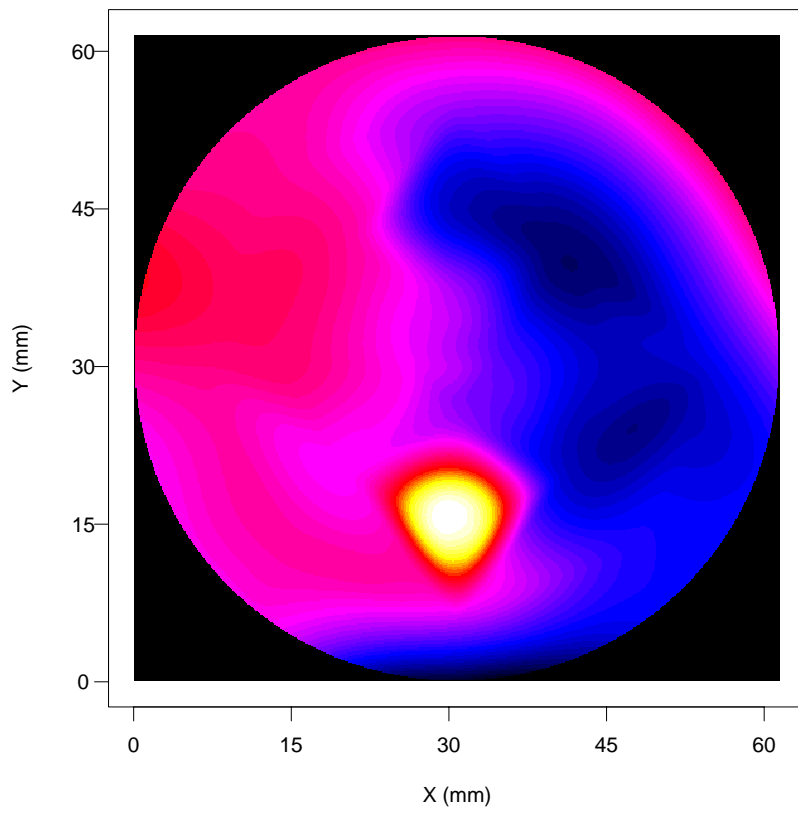


Figure 3.22: A nonparametric model estimate of surface height of a cylindrical sample.

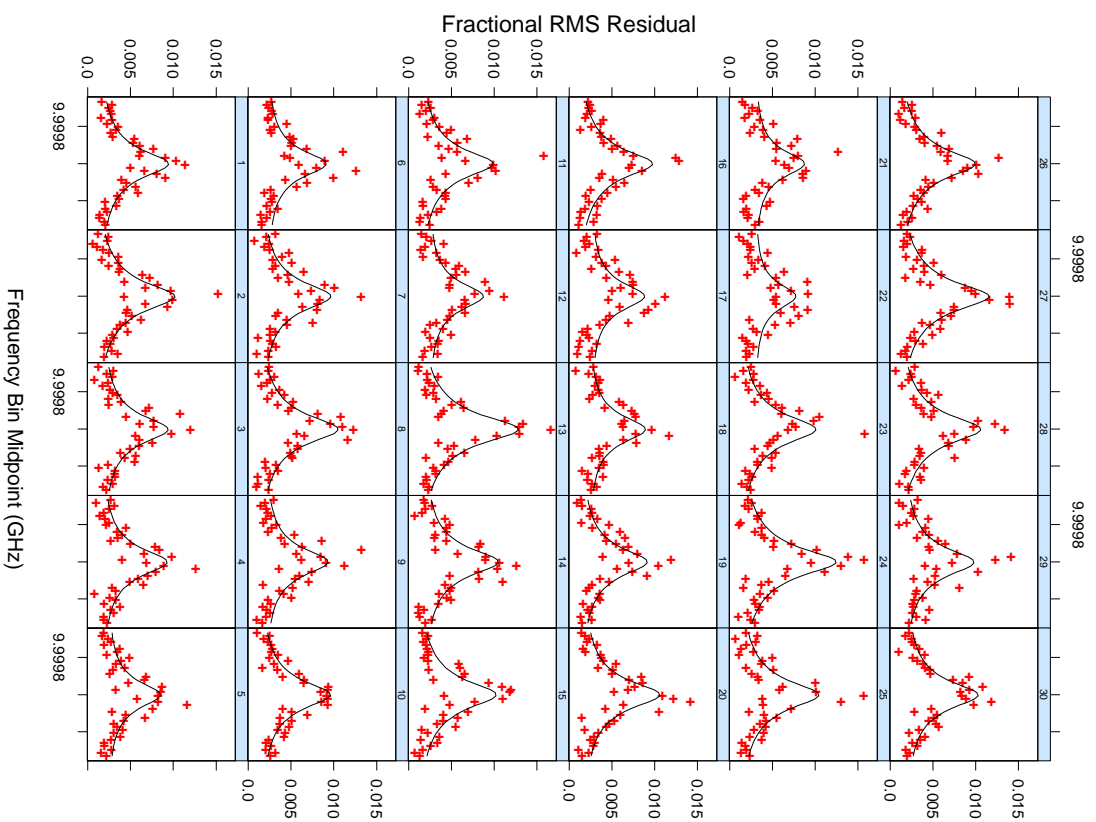


Figure 3.23: Fractional RMS residuals versus frequency bin midpoints for 30 actual data sets. Residuals were assigned to frequency bins to reduce the effect of noise on the fit. The solid line represents the fitted variance cure.

data collection strategy, we compute the asymptotic standard error of the estimators as a function of frequency spacing. In this study, we assume that the additive noise variance is constant over all frequencies. Based on repeat measurements of resonance curves, the additive noise variance clearly depends on frequency.

Estimates of a material's dielectric properties also depend on estimates of the “electrical” length (L) and “electrical” radius (a) of the microwave cavity. We model the measured resonance frequency of the p th transverse electric mode of the cavity as

$$f(p) = \frac{C_{air}}{2\pi} \left[\left(\frac{j_{01}}{a} \right)^2 + \left(\frac{p\pi}{L} \right)^2 \right]^{\frac{1}{2}} + \epsilon(p)$$

with $\epsilon(p)$ denoting additive noise, C_{air} is the speed of light in the air-filled cavity, and j_{01} is the first zero of the Bessel function of the first kind, order one. We developed methods to estimate L and a by nonlinear least squares and nonlinear weighted least squares. For the nonlinear weighted least squares method, we estimated the covariance matrix of the parameter estimators. We also performed analyses that enabled us to determine the optimal number of data points to be used in the fit, and whether or not a skin-depth correction should be applied to the data before fitting. In a repeatability study, we quantified variability due to systematic errors in the measurement system.

The overall uncertainty of the estimated dielectric property parameters depends on random and systematic errors in measured quantities including: temperature, humidity, pressure, Q factors, resonance frequencies, cavity electrical length, cavity electrical radius, etc. To estimate the joint effect of all sources of variability, we developed a Monte Carlo simulation code. Based on this code, we identified the experimental uncertainties that have the most influence on the uncertainty of the estimated permittivity and loss tangent. We quantified systematic uncertainties associated with estimates of the cavity length, cavity diameter, and sample thickness using actual data in conjunction with the Monte Carlo simulation code.

We designed experiments to demonstrate the stability of the measurement process and analyzed the resulting data. We used the repeatability study data and Monte Carlo study results to develop an uncertainty statement. After completing the analysis of the repeatability study data, additional measurements were taken for three samples (two cross-linked polystyrene and one quartz) to verify the stability of the measurement system. Plots of the old and new observations revealed a shift in permittivity for the quartz sample. Further investigation suggested that the shift in the quartz data was related to a problem with the measurement procedure itself. Additional data collected for the three samples after altering the measurement procedure indicate that the measurement system is stable. Interestingly, measurements of cross-linked polystyrene samples did not appear to be affected by the problem in the measurement procedure.

We 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 due primarily 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.

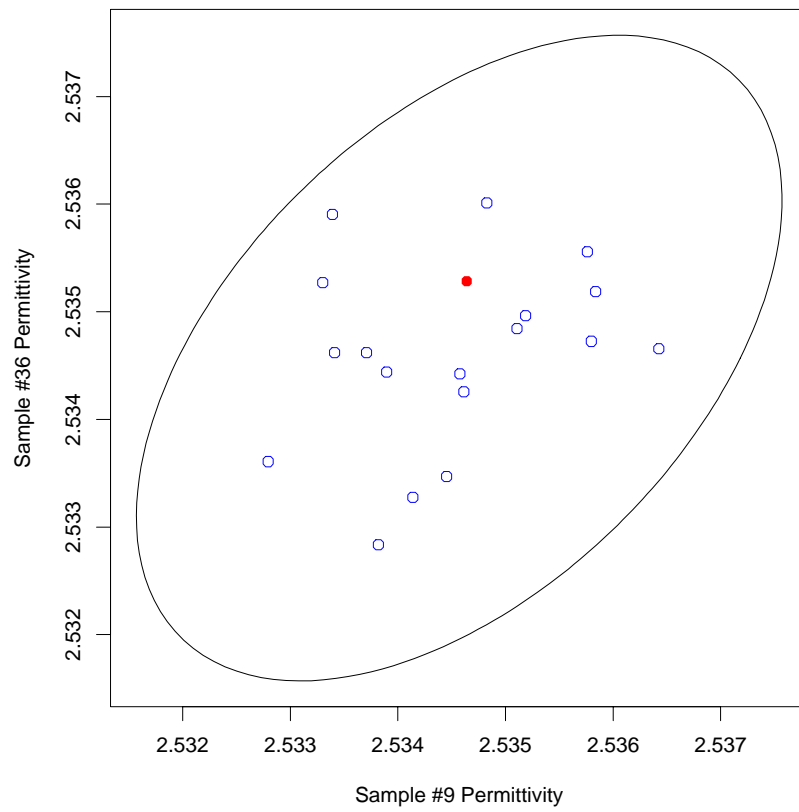


Figure 3.24: An example of a confidence ellipse to monitor two correlated responses. The circles represent historical data while the dot represents a new observation.

FY2002 Highlights

A paper describing procedures for estimating the quality factor and resonant frequency, "Estimation of Q factors and Resonant Frequencies," by K. Coakley, J. Splett, M. Janezic, and R. Kaiser, will appear in the *IEEE Journal of Microwave Theory and Techniques* in 2003.

We developed uncertainty intervals for the SRM certificate and completed draft documentation for the SRM. The draft documentation is currently under review by the EEEL MCOM technical subcommittee and will be published as a NIST Special Publication. The first SRMs will be available in FY2003 pending approval by MCOM.

Software for determining permittivity and loss tangent using a split cylinder measuring system was released by BERB. The software utilizes our algorithm for computing Q and f_0 .

Future Work

We are currently working on a paper in which we will examine the performance of our Q factor and resonant frequency estimation procedure for various levels of Q . Some preliminary data were collected to verify that the algorithm produced acceptable answers for the other measurement systems. An estimation algorithm from the literature (phase versus frequency) will be compared to our algorithm and was added to the estimation software.

We plan to complete a condensed version of the NIST Special Publication for the *NIST Journal of Research*.

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

3.4.4 Residual Resistivity Ratio Metrology for Superconductors

Jolene Splett, Dom Vecchia
Statistical Engineering Division, ITL

Loren Goodrich, Ted Stauffer
Magnetic Technology Division, EEEL

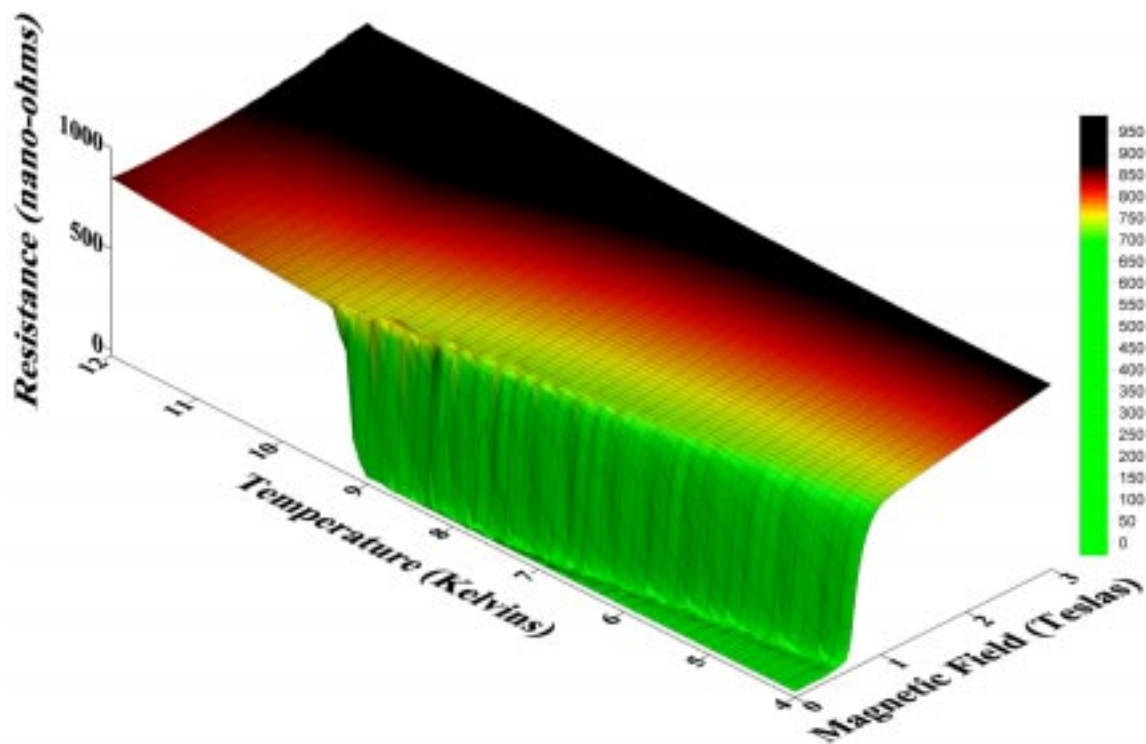


Figure 3.25: Resistance of a high-purity niobium specimen versus temperature and magnetic field.

The U.S. superconductor industry is comprised of many small companies with limited resources for the development of new metrology and standards. The potential impact of superconductivity on electric-power systems makes the technology, and relevant metrology, especially important. NIST serves the industry by advancing the metrology needed to develop large-scale superconductors, by participating in interlaboratory comparisons needed to verify techniques and systems used by U.S. industry, and by developing international standards for superconductivity needed for fair and open competition and improved communication.

NIST has been collaborating with a U.S. company and two U.S. universities on making residual resistivity ratio (RRR) measurements on high-purity niobium (Nb) specimens. Superconducting RF cavities are made with Nb sheets or films, and purity of the niobium is an important factor in the performance of a cavity. The value of RRR is an indicator of the purity (and the low-temperature thermal conductivity) of the Nb and is often used as a material specification.

The RRR is typically defined as the ratio of the electrical resistivities or resistances measured at 273 kelvins (the ice point) and 4.2 kelvins (the boiling point of helium at standard atmospheric pressure). However, pure Nb is superconducting at 4.2 kelvins, so the low-temperature resistance is defined as the normal-state (i.e., non-superconducting state) resistance *extrapolated* to 4.2 kelvins and zero magnetic field.

A resistance surface as a function of temperature and magnetic field is shown in the figure above. When the combination of field and temperature is low enough, the sample is in the superconducting state and the resistance is zero. The transition from superconducting to normal state occurs at lower magnetic fields as the temperature is increased. For temperatures above 9.4 or 9.5 kelvins, the sample is normal at zero magnetic field. The surface was generated with measurements of resistance (R) versus temperature (T) at zero magnetic field and measurements of resistance versus magnetic field (H) at various set temperatures.

There are two methods for obtaining data needed to extrapolate the normal-state resistance of a Nb specimen: (1) measure the normal-state resistance as a function of field at 4.2 kelvins and extrapolate to zero field (field extrapolation), or (2) measure the normal-state resistance as a function of temperature in zero field and extrapolate to 4.2 kelvins (temperature extrapolation). Both methods require the precise measurement of resistance as small as 0.5 micro-ohms on a specimen that resists wetting by solder. Both methods have their difficulties and each would typically be done with a method-specific experimental apparatus. In the NIST experiment, however, both types of measurements are made during a single sequence, with one apparatus, to directly compare methods on a given specimen. Because liquid helium boils near 4.0 kelvins at the atmospheric pressure of our test site, data are reported at 4.0 kelvins rather than 4.2 kelvins.

We are comparing the two methods of measuring the RRR by extrapolation of various statistical models. Empirical and theoretically-based models are being considered for both the temperature-dependence and magneto-resistance methods. For instance, the normal-state resistance previously has been approximated as linear in T^3 . So, for com-

parison to the usual temperature model, we have considered estimation of the exponent ($R = a + bT^c$). The resistance versus magnetic field at 4.0 kelvins indicates that the sample is completely in the normal state at fields above 1.2 to 1.6 teslas, depending upon the RRR of the sample. At first approximation, the magneto-resistance appears to be linear with magnetic field; however, most of the 12 Nb specimens we have measured have slight curvatures with magnetic field. One empirical model that has been considered for the magneto-resistance method is $R = a + b \exp(cH^d)$.

The figure below shows the estimated resistance for repeated measurement curves on one of the Nb specimens. The R vs H data sets (1, 3, and 4) are fit and a value at 4.0 kelvins and zero field is estimated for each set, based on the particular magneto-resistance model above. The R versus T data sets were fit by the cubic and by the variable-exponent model; thus each of these data sets has two estimated resistances at 4.0 kelvins and zero field. The average estimated resistance using $R = a + bT^c$ is 674.9 nano-ohms, and this model is used as a reference since it yields a value that is typically between the other two measured values. Using the ice-temperature resistance of 200433 nano-ohms, the reference RRR would be 297. The average estimated resistance using the cubic temperature dependence model is 678.7 nano-ohms, which gives a RRR that is 0.6 percent lower than the reference. The average predicted resistance using the field dependence model is 671.6 nano-ohms, which gives a RRR that is 0.5 percent higher than the reference. A rigorous analysis of uncertainties associated with various error sources, including the statistical models, will be necessary in order to make quantitative comparisons of the two RRR measurement methods. Preliminary measurement results were reported in a presentation on "Residual Resistivity Ratio Measurements of High-Purity Nb" at the Applied Superconductivity Conference in Houston, Texas, August 4-9, 2002.

Accurate measurement of the RRR of niobium samples is important to assure that critical material-purity specifications are met in the construction of superconducting RF cavities. In its superconducting state, high-purity niobium is used in high-Q resonant cavities for particle accelerators in high-energy physics, nuclear physics, light source, and neutron source applications. One future application of such a neutron source is to transform radioactive waste into shorter-lived, less toxic material.

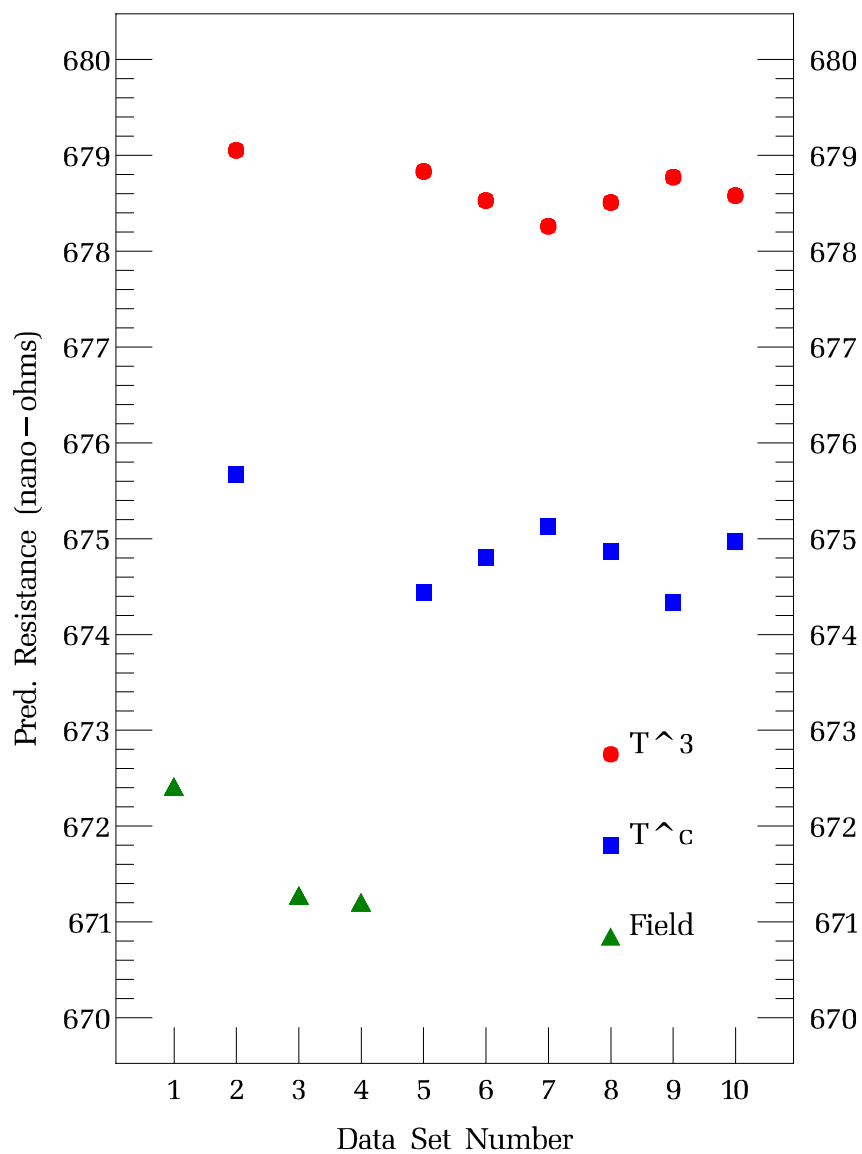


Figure 3.26: Predicted resistance for repeated measurements of a single niobium sample using three different models.

3.4.5 Stochastic Approximation using Twin Processes

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

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

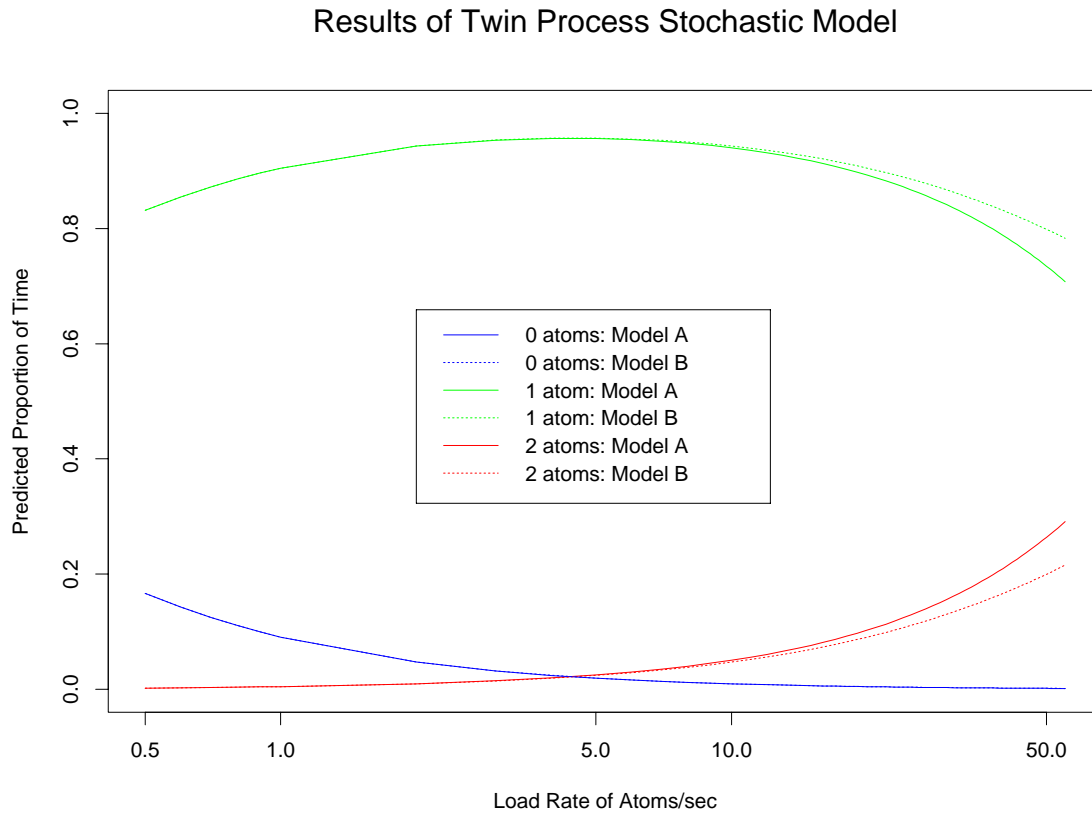


Figure 3.27: This graph shows the modeled proportion of time that a “Twin Process” has a population of 0,1, or 2 atoms as a function of the loading rate R of atoms. Model A is a more exact formula, while Model B is cruder but easier to calculate. The maximum proportion of time that the process has exactly one atom is around 0.97. These results are quite similar to those for the process which the Twin process is approximating.

Researchers in the Physics Laboratory are building a magneto-optical trap (MOT) targeted with a stream of Cr atoms. They want to adjust the loading rate of the stream of atoms so as to maximize the proportion of time that the trap contains only one atom. A modified birth-death process (called a Twin Process) is used to model the effects of imperfect feedback in the physical process. The Twin Process should have a stationary distribution approximately the same as that of the desired process.

Imagine the MOT as having a door that the scientists can open or close to the stream of atoms. Once in the trap, the atom spends an exponentially distributed amount of time there before disappearing. With perfect feedback, 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 follows a very simple birth-death process that implies that increasing the loading rate of the stream of atoms can only increase the desired proportion of time that the atom has only one atom.

However, the number of atoms in the trap cannot be monitored continuously; 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.

There are currently two formulas for the stationary distributions of the twin processes. What we call Model A is a better approximation, while Model B is cruder but easier to calculate. Both assume that the probability of more than 2 atoms in the trap is negligible. For the example in the picture, we have $T=0.01$ sec, and the average lifetime of an atom in the trap is 10 sec. According to Model A, the proportion of time in which the trap has exactly 1 atom, π_1 , is maximized by a load rate of $R = 6.2$ atoms/sec, while in Model B the optimum rate is $R = 6.3$ atoms/sec. Both give a maximum π_1 of around 0.97. These results are quite close to those of computer simulations of the physical process.

Further theoretical work will quantify and bound the differences between the stationary distributions of the Twin process and its target process. Also to be completed are refining and generalizing the models (e.g., accounting for 3 or more atoms in the trap), and possible use of discrete Markov chains.

Consulting with NIST scientists has led to development of a stochastic modeling tool that may prove helpful in analyzing processes in sundry other applications.

3.4.6 Lifetime of Magnetically Trapped Neutrons

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

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

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

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

S.K. Lamoreaux
Los Alamos National Laboratory

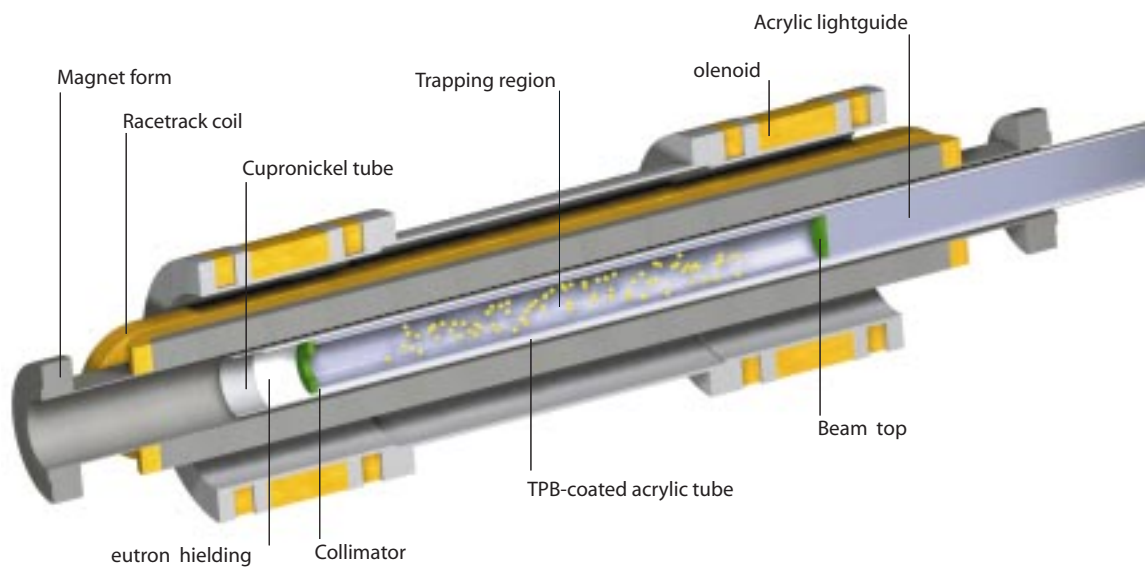


Figure 3.28: Diagram of neutron trapping apparatus.

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

Background

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

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

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

Statistical Analysis

Statistical contributions fall in two general areas. We have developed stochastic models for the experimental data as well as estimation procedures based on either binned or arrival time of the decay data. Based on our stochastic models, we have studied a variety of strategies for estimation of the neutron lifetime. A principle consideration is how to efficiently estimate mean neutron lifetime with neutron decay data that are confounded with background noises. For instance, in one approach, we fit a model to the data from the primary experiment in which neutron decay signals are contaminated by background [4,5].

In another strategy, two separate experiments are performed. One experiment measures pure background signals (to be called the background-only experiment) and the other is the primary experiment of measuring neutron decays that contains unavoidable background signals. Neutron decay data are corrected for background with observations from the background-only experiment before the data are used for mean lifetime estimation. In yet another strategy, we estimate the mean neutron lifetime using the joint likelihood with the data from the primary experiment and the background-only experiment [7].

The primary experiment is composed of two stages of durations T_f and T_d respectively. In the first stage neutrons are generated and trapped magnetically and in the second stage neutron decay signals as well as background noises are recorded. According to our birth-death stochastic model of the trapping process [4,5], the expected number of trapped neutrons is $\lambda\tau(1 - \exp(-T_f/\tau))$ where λ is the rate at which neutrons enter the trap, and τ is the mean lifetime of the neutron. We developed a method to determine the optimal choice of the fill time T_f and the time spent observing decay events T_d . The optimal values, found by simulations, minimize the asymptotic standard error of the lifetime estimate. For the case where a 2 parameter exponential model is fit to background-corrected data, we determined the optimal ratio of “background-only” measurements to primary measurements for various models of the background as well as optimal values of T_f and T_d [6]. For the case where a more complex model is fit to joint likelihood using realizations of the background-only measurement and the primary measurement, we determined T_f, T_d, R [7].

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

Currently, we are studying imperfect background-correction due to alignment errors. In this study, we assume that the background signals that appear in the primary experiment as well as in the the background-only experiment are randomly translated due to uncontrollable delays in timing the experiments.

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

Selected Refereed Publications in Print

1. Magnetic Trapping of Neutrons. P. R. Huffman, C. R. Brome, J. S. Butterworth, K. J. Coakley, M. S. Dewey, S. N. Dzhosyuk, R. Golub, G. L. Greene, K. Habicht, S. K. Lamoreaux, C. E. H. Mattoni, D. N. McKinsey, F. E. Wietfeldt, and J. M. Doyle. *Nature*, 403, 62-64 (2000).
2. Progress Towards Magnetic Trapping of Ultracold Neutrons. P. R. Huffman, C. R. Brome, J. S. Butterworth, K. J. Coakley, M. S. Dewey, S. N. Dzhosyuk, D. M. Gilliam, R. Golub, G. L. Greene, K. Habicht, S. K. Lamoreaux, C. E. H. Mattoni, D. N. McKinsey, F. E. Wietfeldt, and J. M. Doyle. *Nuclear Instruments and Methods A*, 440(3), 522-527 (2000).
3. Magnetic trapping of ultracold neutrons. C. R. Brome, J. S. Butterworth, K. J. Coakley, M. S. Dewey, S. N. Dzhosyuk, R. Golub, G. L. Greene, K. Habicht, P. R. Huffman, S. K.

Lamoreaux, C. E. H. Mattoni, D. N. McKinsey, F. E. Wietfeldt, and J. M. Doyle. Physical Review C, 63, 055502 (2001).

4. Statistical planning for a neutron lifetime experiment using magnetically trapped neutrons. K. J. Coakley. Nuclear Instruments and Methods A, 406, 451 (1998).

5. Likelihood models for two-stage neutron lifetime experiments. G. L. Yang and K. J. Coakley. Physical Review C, 63, 014602 (2001).

6. Neutron lifetime experiments using magnetically trapped neutrons: optimal background correction strategies. K. J. Coakley. Nuclear Instruments and Methods A, 469, 354 (2001).

7. Estimation of the neutron lifetime : comparison of methods which account for background. K. J. Coakley and G. L. Yang, Physical Review C, 65, 064612 (2002)

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. Ultracold neutrons (UCN) have great potential in other major areas of research including neutron reflectometry and Quasi-Elastic neutron scattering. Statistical estimation and planning contributions have applications in other areas.

3.4.7 Cryogenic Detection of Weakly Interacting Particles

K.J. Coakley
Statistical Engineering Division, ITL

D.J. McKinsey
Princeton University

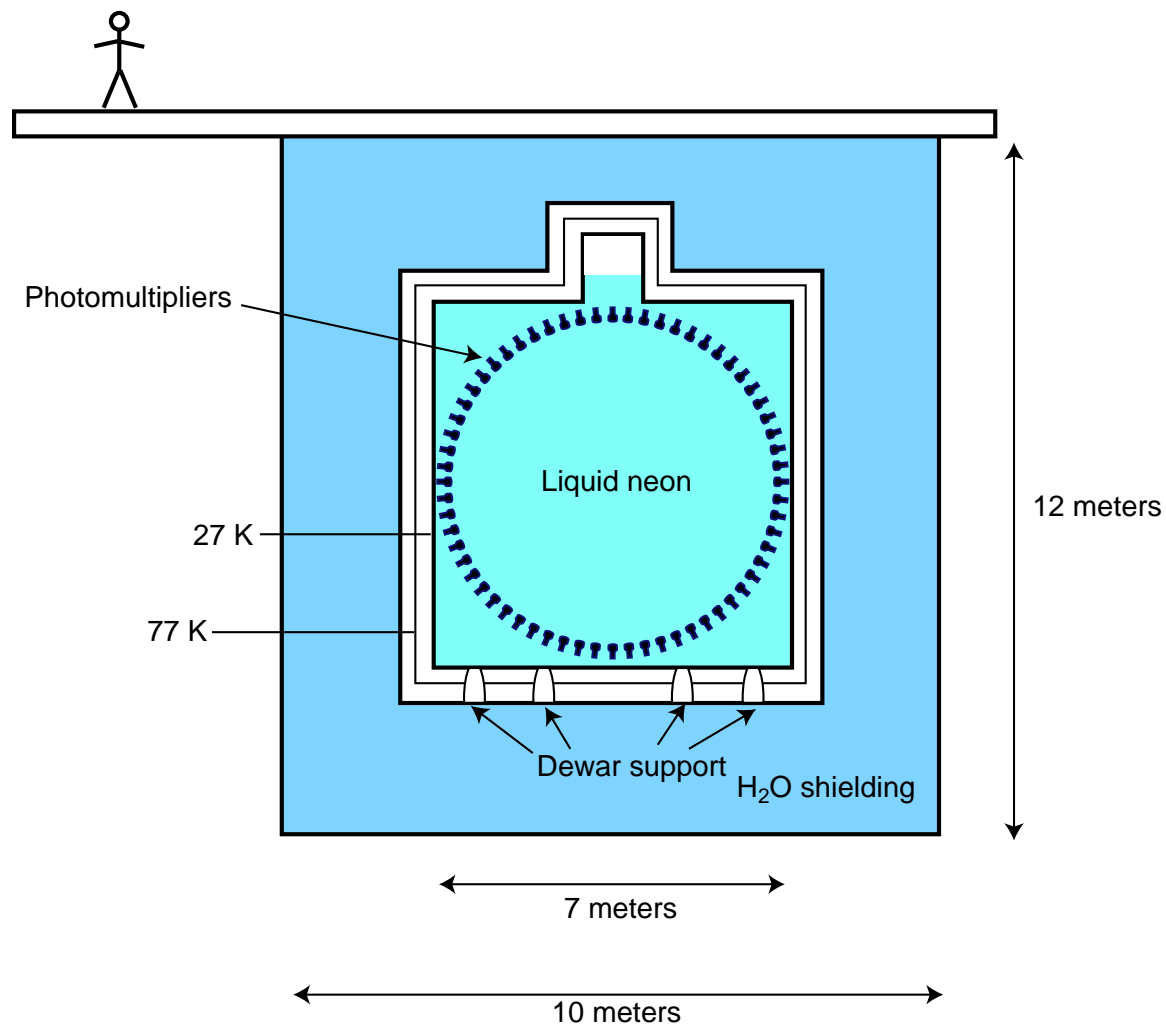


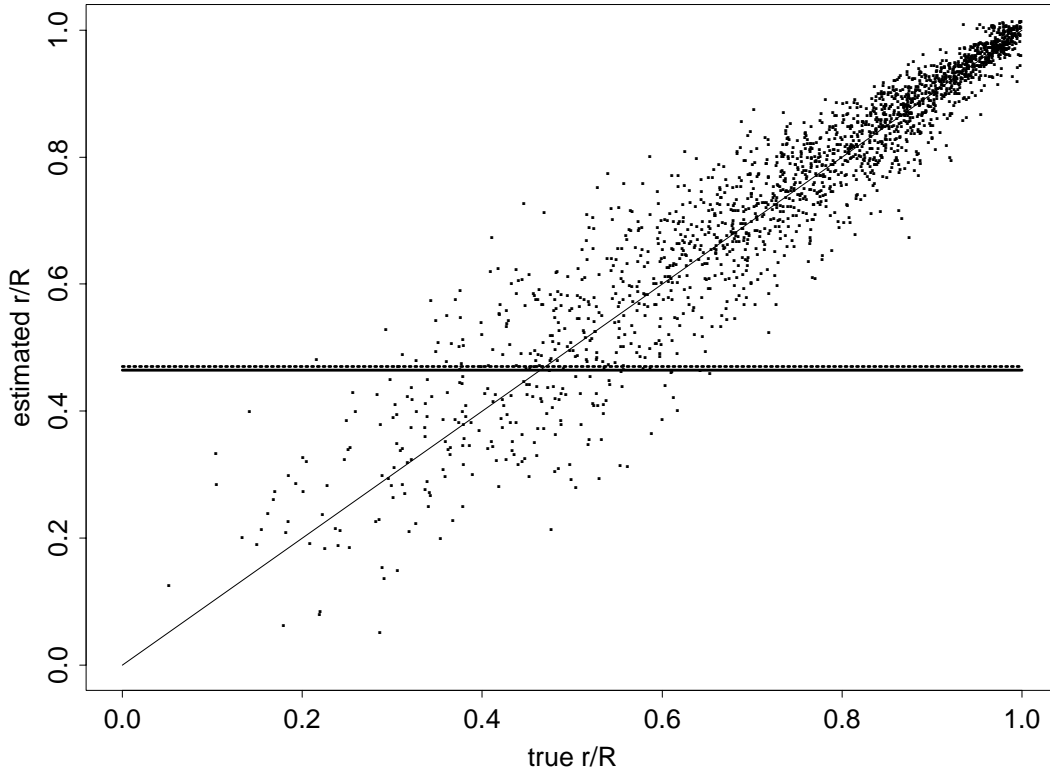
Figure 3.29: Diagram of the proposed CLEAN experiment.

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

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

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

Current and future statistical work includes: stochastic modeling of scattering and transport of gamma ray and scintillation photons, statistical planning, development and testing of statistical background discrimination methods, development of empirical models for calibration of statistical estimates of event location, development of sampling schemes



We simulate events of interest which produces 50 detected photons. The true position of a randomly located event is uniformly distributed throughout a spherical detection volume filled with liquid neon. The photons Rayleigh scatter ($\lambda_s = 0.1 R$). At the detectors, the photons are absorbed and re-emitted at a lower wavelength. The shifted photons are observable. Using a scatter-free transition matrix, we obtain the approximate Maximum Likelihood estimate of the radial location of the event. We correct for bias using a polynomial calibration model, with the calibration coefficients estimated from simulated training data. Ideally, the p th quantile of the estimate of r should equal $p^{\frac{1}{3}}R$. The 0.1 quantile of the estimate (dashed line) is very close to the radial boundary of an inner spherical volume which occupies 10 percent of the total volume of the spherical detection region (solid line). We also show the line of equality corresponding to perfect prediction.

for training data for calibration, energy spectrum estimation, quantification of detector efficiency and false detection rate, uncertainty analysis.

Presentations

With D. McKinsey, gave invited talk “CLEAN” and poster, “Event Location Estimation and Background Discrimination in a Proposed Low Energy Neutrino Experiment” at April 2002 meeting of the American Physical Society.

Publications

C. J. Horowitz, K. J. Coakley, D. N. McKinsey, Supernova Observation Via Neutrino-

Nucleus Elastic Scattering in the CLEAN Detector. Submitted to Physical Review D.

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

3.4.8 Neutron Detector Calibration

K.J. Coakley,
Statistical Engineering Division, ITL

M.S. Dewey
Ionizing Radiation Division, PL

Z. Chowdhuri
NIST Center for Neutron Research, MSEL

W.M. Snow
Physics Department, Indiana University

J.M. Richardson
Science Research Laboratory, Inc, Somerville, MA

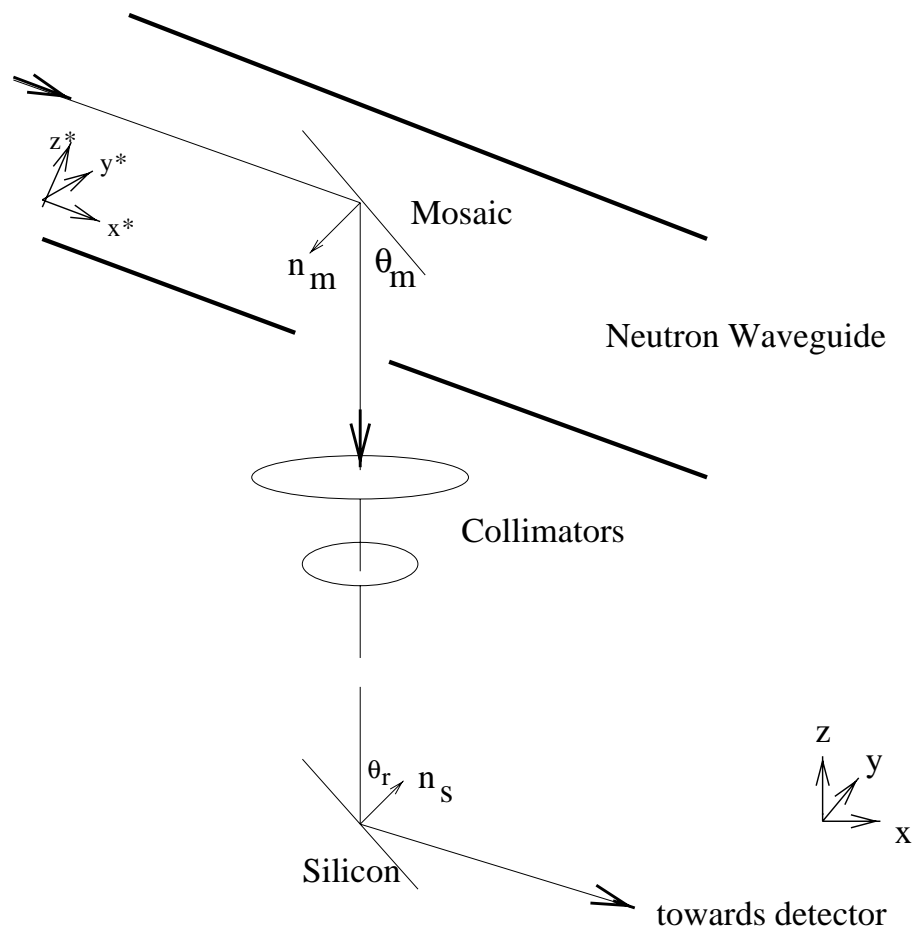


Figure 3.30: In a calibration experiment, neutrons are guided towards a mosaic crystal. Neutrons that pass through collimators are scattered by a silicon crystal. As the silicon crystal is rocked back and forth, the detected intensity of the scattered neutron beam varies. We study the systematic error associated with an estimate, determined from the rocking curve data, of the mean wavelength of the neutron beam.

Stochastic modeling and uncertainty analysis for fundamental metrology.

When a free neutron decays, a proton is created. The currently accepted value of the mean lifetime of the neutron is 885.7 s with a 1-sigma uncertainty of 0.8 s. Along with other experimental data, the mean lifetime of the neutron allows one to test the consistency of the standard model of electroweak interactions. Furthermore, the mean lifetime of the neutron is an important parameter in astrophysical theories. The mean lifetime of the free neutron can be measured using an in-beam technique. At NIST, an improved version of an earlier in-beam experiment is underway. In part of the experiment, a neutron detector is calibrated based on an estimate of the mean wavelength of a reflected neutron beam. Here, we quantify the systematic error of the mean wavelength estimate.

In the in-beam neutron lifetime experiment, a neutron beam passes through a Penning Trap of length L . Some of the neutrons that pass through the trap decay into protons via $n \rightarrow p + e^- + \bar{\nu}_e$. Assuming that each neutron trajectory can be treated classically and is parallel to the long axis of the trap, the amount of time each neutron spends in the trap is L/v , with v denoting the neutron velocity. For $v \gg L$, the probability of decay is

$$P(L, v) = 1 - \exp\left(-\frac{L}{v\tau}\right) \simeq \frac{L}{v\tau}$$

with τ denoting the mean lifetime of the neutron. In the NIST experiment, the fraction of neutrons that decay is very small. Because of this, the mean lifetime is estimated as follows

$$\hat{\tau} = \frac{N_n L}{N_p} < \frac{1}{v} >$$

with N_n denoting the number of neutrons that pass through the trap, N_p is the number of protons that are trapped, and $< \frac{1}{v} >$ is the mean inverse velocity. Since a small fraction of neutrons decay while passing through the trap, the above estimate is biased. However, the magnitude of the bias is negligible compared to other experimental errors. Since the DeBroglie wavelength of the neutron is

$$\lambda = \frac{h}{m_n v},$$

we can write

$$\hat{\tau} = \frac{h}{m_n} \frac{N_n L}{N_p} < \lambda > .$$

The number of transmitted neutrons N_n is measured by counting alpha particles and tritium particles created by $n + {}^6\text{Li} \rightarrow {}^3\text{H} + {}^4\text{He}$. when the neutron beam bombards a lithium film. To within approximately 0.04 %, the cross section for this reaction is proportional to λ . Hence, for an ensemble of neutrons passing through the lithium film, the expected number of detected α particles and tritiums is

$$< N_\alpha + N_T > = \epsilon < \lambda > N_n$$

with $\langle \lambda \rangle$ denoting the average wavelength of the neutrons and ϵ is an efficiency factor. Thus,

$$\hat{\tau} = \frac{h}{m_n} \frac{L}{\epsilon} \frac{N_\alpha + N_T}{N_p}.$$

In the current experiment, the goal is to lower the uncertainty to less than 1.0 s (≈ 0.1 %). The expected reduction in uncertainty is due, in part, to calibration of the neutron detector. Therefore, it is important to demonstrate that systematic error in the neutron calibration experiment is less than the target uncertainty of 0.1 %.

In the calibration experiment, $N_\alpha + N_T$ and N_n and $\langle \lambda \rangle$ are determined for a narrow wavelength beam. From these three measurements, the efficiency ϵ is estimated. Here, the focus is on how accurately the mean wavelength of the beam can be determined from rocking curve data. We estimate the mean wavelength of the beam to be

$$2d_S \sin\left(\frac{1}{2}(\pi - \bar{\theta}^+ - \bar{\theta}^-)\right)$$

with $\bar{\theta}^+$ and $\bar{\theta}^-$ denoting the weighted averages of the positive and negative rocking angles, and d_S is the spacing between scattering planes in the perfect silicon crystal. For the simulated data, the weight is the relative probability of the event.

We simulate the momentum distribution of thermal neutrons produced at the NIST Cold Neutron Research Facility. Based on a stochastic model for the scattering of neutrons off a graphite mosaic crystal, we simulate rocking curve data. We simulate only those events that satisfy a necessary condition to pass through collimators. This importance sampling approach speeds up the Monte Carlo simulation code over a factor of 500. Those neutrons that pass through the collimators then scatter off a perfect silicon crystal. As the crystal is rocked back and forth, we simulate the relative intensity of scattered neutrons. Based on this simulated rocking curve data, we predict the mean wavelength of the neutron beam which passes through the collimators. The statistical bias of the mean wavelength estimate is approximately 0.004 %. When we vary the reflectivity of the mosaic crystal, the mean wavelength of the neutron beam transmitted through the collimators varies. However, the expected value of the predicted value of the mean wavelength of the neutron beam tracked the actual value very well; the accuracy of our predicted value did not vary significantly with reflectivity.

In our primary study, we assume that the perfect silicon crystal is aligned so that its surface normal is orthogonal to the rocking axis. In an additional study, we quantify the additional systematic associated with silicon crystal alignment errors. In our work, we assume that the rocking curve data are not contaminated by a background signal. If rocking curves are contaminated by background, one should correct for background in some manner. If the background correction is imperfect, an additional systematic error could be introduced.

Publication

K. J. Coakley, Z. Chowdhuri, W. M. Snow, J. M. Richardson and M. S. Dewey “Estimation of neutron mean wavelength from rocking curve data,” Measurement Science and

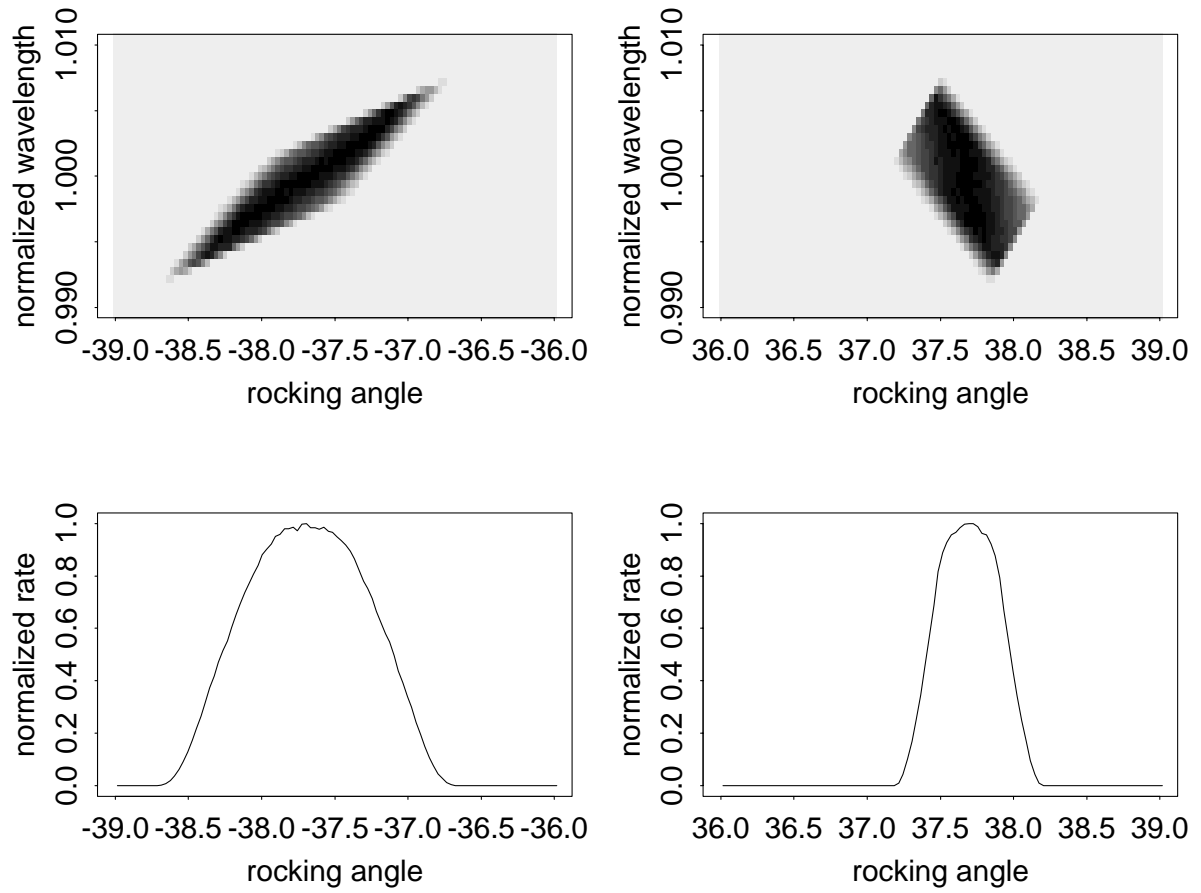


Figure 3.31: Simulated images in wavelength-angle space and simulated rocking curves. $T = 40$ K. One million events.

More accurate calibration of neutron detectors advances metrology in a fundamental way. Along with other experimental data, the mean lifetime of the neutron allows one to test the consistency of the standard model of electroweak interactions. Furthermore, the mean lifetime of the neutron is an important parameter in astrophysical theories.

3.4.9 Consistency of Nuclear Methods for Thin Film Analysis

K.J. Coakley, J. Lu,
Statistical Engineering Division, ITL

H. H. Chen-Mayer, G.P. Lamaze
Analytical Chemistry Division, CSTL

S.K. Satija
NIST Center for Neutron Research, MSEL

D.S. Simons
Surface and Microanalysis Science Division, CSTL

P.E. Thompson
Naval Research Laboratory

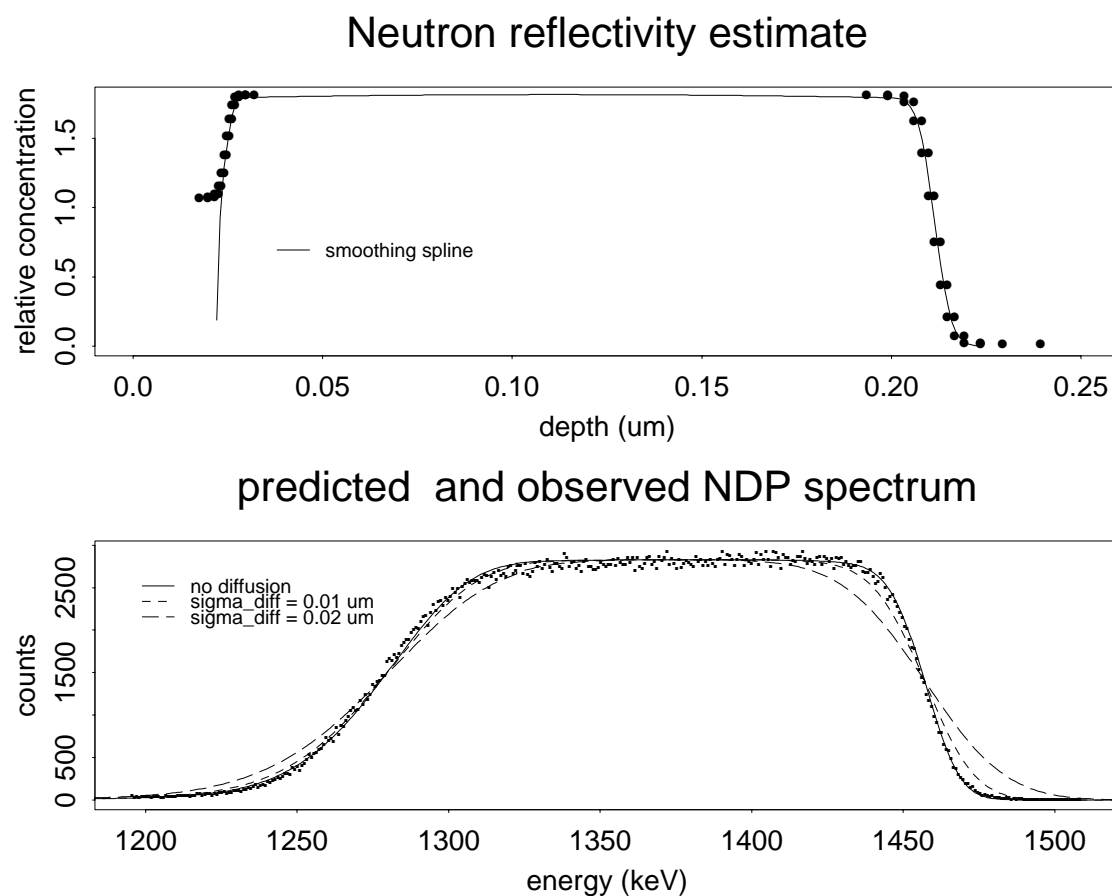


Figure 3.32: Top: Neutron reflectivity estimate of boron concentration profile in BPSG sample assuming no diffusion. Bottom: predicted NDP spectrum based on various models for different diffusion models for boron in BPSG sample.

We calibrate theoretical stopping power model models for metrological applications of Neutron Depth Profiling. We also quantify the consistency of experimental methods for characterizing thin films including Neutron Depth Profiling, Neutron Reflectometry and Secondary Ion Mass Spectrometry.

Neutron Depth Profiling (NDP) is a nondestructive method for analysis of the concentration profile of an element in a material based on the energy spectrum of energetic particles emitted from the material when certain isotopes capture neutrons. Typical applications of NDP include characterization of semiconductor samples, polymers, optoelectronic materials, metal alloys, and most other solids. When a neutron is absorbed by an element, a nuclear reaction produces a particle (e.g., an alpha particle). As the particle travels through a material, it loses energy. The energy loss process is stochastic. The detector response function (DRF) is a probability transition matrix that relates the depth of emission to the expected energy spectrum of the detected particles. The DRF depends on the geometries of the emitter and detector, and assumed models for the stopping power of the material, energy straggling, multiple scattering and alpha particle detector energy resolution [1].

In NDP experiments on samples doped with boron, He nuclei (alpha particles) are produced by either of two nuclear reactions. In the first reaction,
 $^{10}\text{B} + \text{n} \rightarrow ^7\text{Li}(840 \text{ keV}) + ^4\text{He}(1472.6 \text{ keV}) + \gamma(478 \text{ keV})$,
 a 1472.6 keV alpha particle is created. In the second reaction,
 $^{10}\text{B} + \text{n} \rightarrow ^7\text{Li}(1013 \text{ keV}) + ^4\text{He}(1776.73 \text{ keV})$,
 a 1776.73 keV alpha particle is created. The relative probabilities of these two reactions are 0.937 and 0.063. As an alpha particle travels through the silicon, it loses energy. The expected energy $\bar{E}(x)$ after traveling distance x through the material satisfies the following integral equation

$$x = \int_{\bar{E}(x)}^{E_o} \frac{dE}{S(E)}$$

with E_o denoting the initial energy of the particle and $S = \partial \bar{E} / \partial x$ is the stopping power of the material. Using the Stopping and Range of Ions in Matter code SRIM-2000, we predict $\bar{E}(x)$ for each of the two reactions.

Since inferences about the concentration of an element in a material depend on the assumed model for the stopping power of the material, calibration of stopping power models is fundamentally important for metrology. In recent work, we developed a statistical method for calibration of the stopping power model based on independent measurements of a sample by both NDP and Secondary Ion Mass Spectrometry (SIMS) [2]. We prepared a silicon sample by a molecular beam epitaxy method with a well-characterized boron concentration profile. We obtained a high accuracy measurement of the boron concentration profile with Secondary Ion Mass Spectrometry (SIMS). In an NDP experiment, we measured the energy spectrum of emitted alpha particles. Based on the measured boron concentration profile, and our model for the DRF, we predicted the observed NDP data. The locations of the energy peaks in the predicted NDP spectrum were consistently at lower energies than the locations of the observed peaks in the NDP data. From the differences in the locations of the observed and predicted energy peaks, we estimated the

stopping power reduction factor to be 5.06 percent. The associated $1-\sigma$ uncertainty of this estimate was 1.06 percent. Our uncertainty analysis accounted for the spatial variability of the measured boron concentration profile and counting statistics in the NDP data. When the assumed stopping power was reduced by 5.06 percent, the predicted and observed NDP data agreed well.

Microelectronic circuit devices widely employ boron/phosphorus-doped silicate glass (BPSG) thin films that require careful control of the boron concentration in the manufacturing processes. In neutron reflectometry, one measures the angular dependence of the specular reflectivity near grazing incidence. We fit a model to the reflectivity data based on some prior knowledge, to determine parameters representing features such as film thickness, density, and interface roughness. While NDP measures only the boron isotope in the matrix, reflectometry measures the total scattering contribution from the matrix and is not sensitive to the low level of boron. Because NR and NDP are based on entirely different principles, the information obtained is independent and can be used to verify or modify NDP results in film thickness represented by the boron profile.

Both the NDP and the NR experiments were at the NIST Center for Neutron Research. Based on a regression spline representation of the neutron reflectivity estimate of the scattering due to the matrix, we predict a NDP spectrum for different diffusion coefficient values. In the diffusion model, we convolve the NR estimate with a Gaussian kernel with an adjustable standard deviation. The agreement between the measured NDP spectrum and NR estimate (without diffusion) is very good; we conclude that there is no evidence of boron diffusion within the BPSG sample.

Ongoing research projects include empirical estimation of stopping power from multian-gle NDP experiments and statistical methods for profile reconstruction. In the profile reconstruction problem, we wish to estimate the unobserved boron concentration profile from the observed energy spectrum.

Publications and Presentations

[1] K.J. Coakley, R.G. Downing, G.P. Lamaze, H.C. Hofsass, J. Biegel, C. Ronning, "A model for Neutron Depth Profiling Measurements" Nuclear Instruments and Methods in Physics A, pp. 137-144, 366, 1995.

[2] K.J. Coakley, H.H. Chen-Mayer, G.P. Lamaze, D.S. Simons, and P.E. Thompson, "Calibration of a stopping power model for silicon based on analysis of neutron depth profiling and secondary ion mass spectrometry measurements," Nuclear Instruments and Methods in Physics Research B, pp. 349-359, 192, 2002.

[3] H.H. Chen-Mayer, G.P. Lamaze, K.J. Coakley, S.K. Satija, "Two aspects of thin film analysis: boron profile and scattering length density profile," presented at 10th Symposium on Radiation Measurements and Applications - May 21-23, 2002, in Ann Arbor, MI.

[4] H.H. Chen-Mayer, G.P. Lamaze, K.J. Coakley, S.K. Satija, "Two aspects of thin film

analysis: boron profile and scattering length density profile" to appear in Nuclear Instruments and Methods in Physics B.

Our work facilitates metrological applications of Neutron Depth Profiling and Neutron Reflectometry. These methods have broad industrial applications including semiconductor samples, polymers, optoelectronic materials, metal alloys, and most other solids.

3.5 Measurement Services

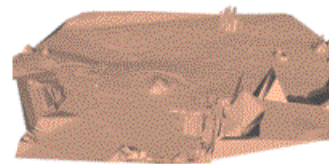
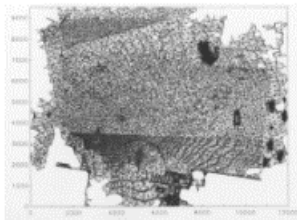
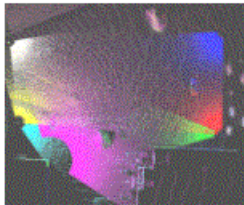
3.5.1 Range Imaging and Registration Metrology

Stefan Leigh, Andrew Rukhin
Statistical Engineering Division, ITL

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

Geraldine Cheok
Structures Division, BFRL

Initial Terrain
Dec. 1999



March 9, 2000

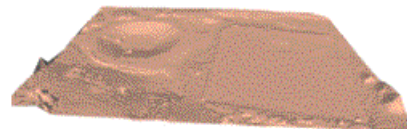
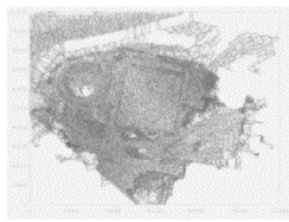
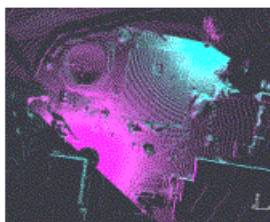


Figure 3.33: Error (bias) as a function of distance for different target colors and reflectance conditions. Individual error bars represent one standard deviation.

Implementations of range imaging sensing, such as LADAR (laser distance and ranging), are already seeing manifold applications in both military and commercial settings. In civil engineering, LADAR can 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 enable visualization and feedback-based corrective measures by onsite or offsite 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 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 algorithms 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 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.

SED work this year focused on analyzing data taken to evaluate the accuracy and precision characteristics of a commercial LADAR sensor. A LADAR is an instrument that can rapidly capture 3-D data of a scene in the form of coordinate (x,y,z) triples, as contrasted to the familiar 2-D projections of standard photography. Generally, LADARs return two pieces of information: range (= distance) and intensity (function of the strength of the reflected signal). Some devices can obtain other spectral data as well, which can be used to aid in object identification.

A series of calibration experiments varying distance, size, color, reflectivity, texture, angle of incidence, and beam divergence of a set of targets were conducted, the intent being

to study degradation of instrument distance estimates as a function of the controlled factors. The results are summarized in the NISTIR 6922 *Calibration Experiments of a Laser Scanner* (Cheok, Leigh, and Rukhin). Broadly speaking, accuracy effects dominate precision effects. Color can induce bias, although not pronounced in these experiments. Reflectivity characteristics of the surface, and its angle of incidence to the beam, can induce dramatic biases. No evidence of significant temporal autocorrelation was observed, but random significant correlation between spatially contiguous measurements was observed, which requires further investigation. Correlative effects are important because the first approach being employed for estimation of uncertainty in scene reconstruction is propagation-of-error.

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 testbed. Riegl, Cyra and Metric Vision are working with NIST to help define test and calibration needs of industry. It is anticipated that other companies will be invited to join this collaboration in FY02.

3.5.2 Sulfate Performance Prediction for Infrastructure Abatement

Stefan Leigh

Statistical Engineering Division, ITL

Paul Stutzman

Materials and Construction Research Division, BFRL

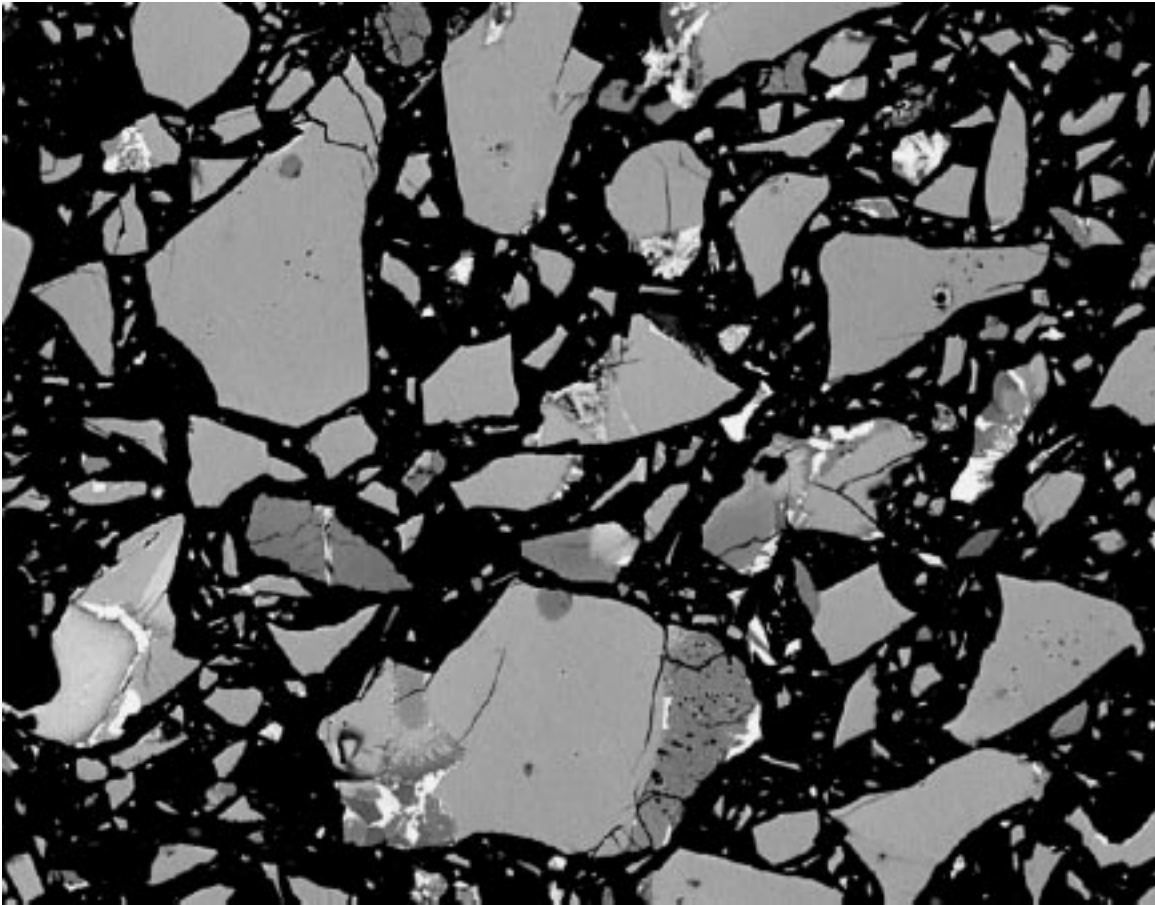


Figure 3.34: Ettringite needles and calcium hydroxide plates. Both are cement hydration products. The *in situ* conversion of the various monosulfate phases to the trisulfate ettringite is accompanied by a significant, destructive volume expansion.

A major objective of the BFRL is to develop computational and experimental materials science based techniques that will enable the prediction and optimization of the initial cost and service life performance and minimize the environmental impact on concrete in existing infrastructure.

The \$100B/year concrete industry includes cement producers, chemical and mineral admixture producers, aggregate producers, ready mixed and pre-cast concrete manufacturers, those who produce the ingredients and the final product, and concrete construction companies, those who build with concrete. The concrete industry depends heavily on the use of consensus standards. The leading standards organization is ASTM, of which the two main committees for the concrete industry are C01 (Cement) and C09 (Concrete). Development of computational techniques for performance prediction are expected to reduce the 6 month - 1 year testing time currently hindering new product development.

Sulfate attack is a widespread form of chemical attack on concrete. Sulfates are often present in groundwater, soil, and seawater. Local high concentrations of sulfates may be associated with industrial wastes. The major cause of expansion is the phase conversion of monosulfate into ettringite. This is accompanied by a large increase in solid volume responsible for creating internal stresses that cause cracking. The abundance and distribution of these phases is related to the cement phase composition. Current ASTM methods for estimating phase compositions of cements are inaccurate and incomplete so limited success has been realized following this approach. X-ray powder diffraction analysis is a direct method for phase identification and measurement of phase composition. The sulfate attack project seeks to develop an accelerated sulfate attack test. This is being attempted in two ways: physical testing of specimens exposed to sulfate ions and analysis of the results of direct phase analysis of the constituent materials of the hydraulic cements. The physical testing is necessary here because it provides what is currently considered the most reliable performance test for comparison.

For the latter approach, a set of cements selected to represent the range of North American cement production with respect to chemistry have been subjected to long-term ASTM testing for sulfate performance. They are also tested for additional performance properties such as heat of hydration (a function of phase composition and surface area), strength development and ultimate strength, time of setting and others. Exposed to sulfate attack, they exhibit a range of performance characteristics from rapid deterioration to no apparent change, based upon expansion measurements. These performance data are being examined particularly as a function of cement phase composition and fineness.

Current SED efforts involve modeling 24-week sulfate-induced expansion as a function of compositional parameters (the so-called cement 'phases'), alkali, gypsum, and sulfate contents, and fineness parameters. Graphical renditions, principal component regression (PCR), and Alternating Conditional Expectation (ACE) are currently being employed. PCR emphasizes a structural component, with contrasted anticorrelating phases (e.g., alite

versus belite), and a fineness component.

Any resulting models are to be used as computational tools for industry performance prediction. In addition, the results will aid in the development of the BFRL Virtual Cement and Concrete Testing Laboratory (VCCTL), enabling it to predict durability from concrete mixture design. This will serve to demonstrate the the kinds of improvements that the VCCTL can bring to bear on the standards process.

3.5.3 Half-life of Arsenic-76

James J. Filliben

Statistical Engineering Division, ITL

Richard Lindstrom

Analytical Chemistry Division, CSTL

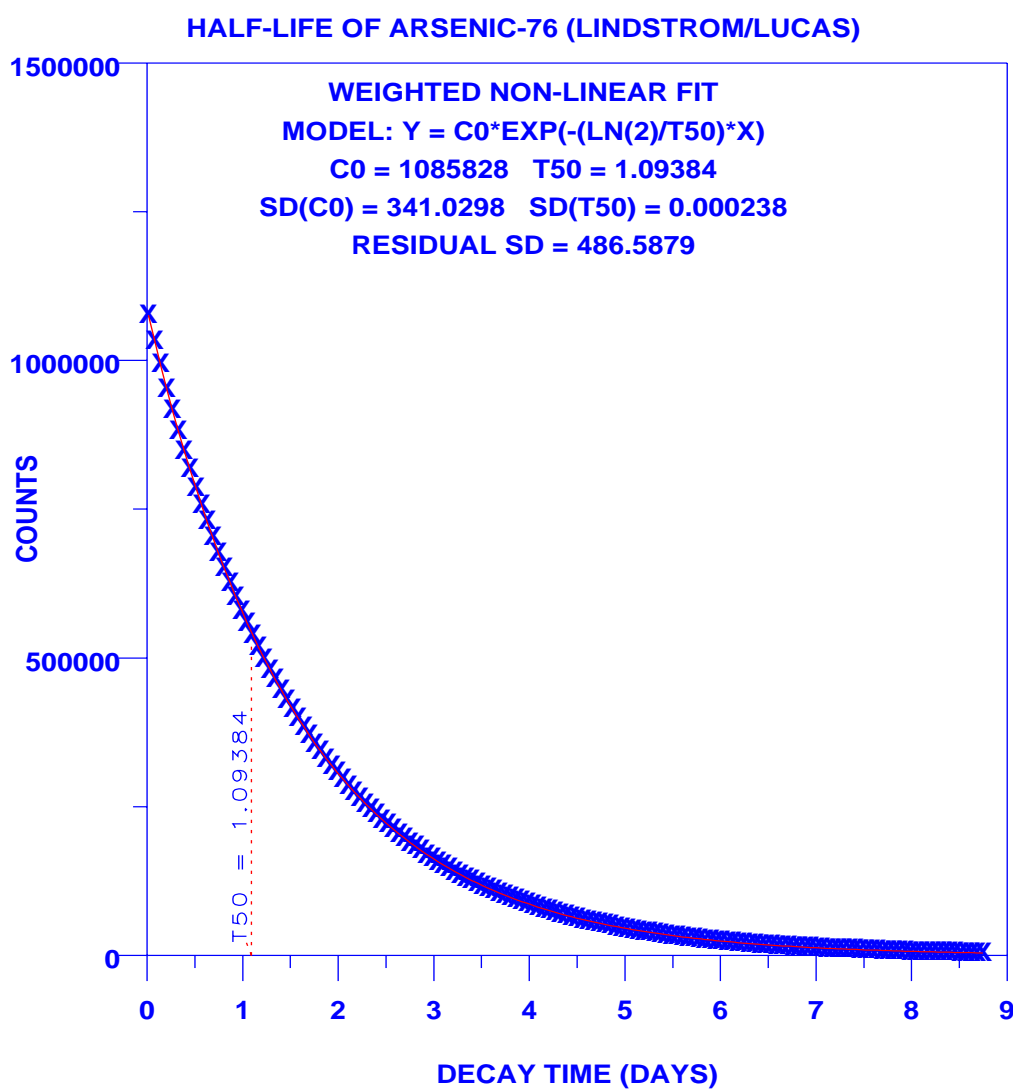


Figure 3.35: This figure shows the 150 Arsenic-76 half-life data points and the superimposed prediction curve from the weighted non-linear fit.

Of the 115 known chemical elements, a given element always has a fixed number of protons, but may have differing numbers of neutrons. The weight of the element is the sum of the protons and neutrons. (Radio)isotopes are variants of an element that have the same number of protons (and hence the same chemical properties), but have differing numbers of neutrons (and hence differing weights). Some isotopes tend to be abundant in nature and are "stable"; other isotopes are "unstable" in the sense that they shed (decay) "extra" neutrons to degenerate into the more stable form. A measure of an isotope's stability is its half-life (the amount of time it takes for half of the atoms to decay into the more stable form). Isotopic half-lives range from millionths of a second (very unstable) to billions of years (very stable). Of the 3500 known isotopes, a relatively small number are naturally stable and abundant. Most isotopes are unstable—many of which are mere lab curiosities but some of which are extremely useful in science, for example:

1. Fire detectors (e.g., Am-241)
2. Agricultural tracers (e.g., P-32)
3. Food irradiation (e.g., Co-60)
4. Pest control (various)
5. Archeological dating (e.g., C-14)
6. Biomedical (many, including As-76)
7. Digital restoration of paintings (several, including As-76)

The choice of which isotope to use for a given application is frequently dictated by availability, and by matching the half-life length with the specific application-dictated time-span. For example, for archeological dating of biologic-based specimens several thousands of years old, Carbon-14 is the isotope of choice due to its half-life of 5700 years. For detection of fractured bones, Technetium-99 may be the isotope of choice for injection since its half life is 6 hours. For digital restoration of artwork, Arsenic-76 (As-76) may be chosen since its half-life is roughly 1 day (after which an autoradiograph image may be taken).

The acknowledged central web-based repository of radioisotope properties (including half-lives) is Lawrence Berkeley National Laboratory's Table of Radioactive Isotopes (<http://ie.lbl.gov/toi>). Referring to the above-mentioned isotope Arsenic-76, which has applications in several areas including painting restoration and biomedical, the LBNL Web Table of Isotopes lists As-76's half-life as 1.0778 +/- 0.0020 (days).

Richard Lindstrom of CSTL, Larry Lucas of PL, and other NIST scientists believe this value is incorrect. This value was derived from work by E.P. Mignonsin in a 1994 paper: "Determination of Half-lives by Gamma-Ray Spectrometry: Improvement of Procedure and Precision", Journal of Applied Radiation and Isotopes 45, 1994, pp. 17-24. It is Lindstrom's belief that Mignonsin's value is in error by over 1%, which is enormous in the context of Mignonsin's stated uncertainty of less than .2%. Nonetheless, 1.0778 is currently the published value of choice for As-76's half-life. Lindstrom and Lucas believe that the "truth" is much closer to an older historical value for the As-76 half-life (1.097

+/- 0.003) that was reported by Emery et al. in a 1972 article in Nuclear Science and Engineering.

To address this problem, Lindstrom, along with colleagues Menno Blaauw (the Netherlands), and Ronald Fleming (University of Michigan), collected carried out an experiment for the purpose of recomputing As-76's half-life.

Compared to many of SED's consulting and collaborative projects, this project required a relatively small SED effort (a consulting session or two), but had a big (supportive) impact. Lindstrom et al. had already carried out the experiment and had in fact arrived at their own updated values for the As-76 half-life. SED's immediate role was 6-fold:

1. to assess the quality of the 150 data points (acceptable);
2. to review fitting process (not perfect, but not bad);
3. to repeat the non-linear fitting process;
4. to arrive at an independent estimate of the As-76 half-life;
5. to independently compute the uncertainty; and
6. to ascertain the robustness of the fitted/estimated value.

As usual, these six steps required a combination of quantitative statistical tools (non-linear fitting) and graphical tools (graphical residual analysis). It is in essence a classical non-linear fitting problem—the only difference being that one of the fitted coefficients turns out to be a very important physical quantity: the half-life of the Arsenic-76 isotope.

The experiment was carried out by Lindstrom as follows:

1. Five sets of measurements were collected;
2. The As-76 was counted at 20 cm fixed-source-detector geometry;
3. 41 to 150 spectra were collected over 4.8 to 8.8 days;
4. 20 million photopeak counts were collected per experiment;
5. Precision pulser corrected for dead time and pileup losses.

The resulting dataset consisted of 150 observations, where an observation consisted of the following triplet:

1. Y : the corrected counts
2. $\text{Var}(Y)$: the variance of the counts

3. X : the decay time (in days)

The x's of Figure 3.34 are the 150 data points. As expected, they clearly show the signature exponential character of radioactive decay. The theoretically appropriate model for this decay is

$$Y = c_0 * \exp(-(\ln(2)/t_{50}) * x)$$

where t_{50} is the desired half-life to be computed.

Given the available variance information in the data set, the appropriate statistical procedure would be to carry out a weighted non-linear least squares fit. Performing such a fit (DATAPLOT), we compared the fitted values with the raw data (see Figure 3.34), and arrived at the following half-life values:

$$t_{50} = 1.093543 \pm 0.000266 \text{ days}$$

where the uncertainty is 1 standard deviation; or using two standard deviations:

$$t_{50} = 1.093543 \pm 0.000532 \text{ days}$$

The 1.093543 ± 0.000532 value is much in line with Emery's 1972 value (1.097 ± 0.003), and is much different from Mignonsin's 1994 published value (1.0778 ± 0.0020). It appears that the original Lindstrom/Lucas concerns were in fact very much justified.

To assess the robustness of the fitting process, 2 additional fits were carried out: an unweighted non-linear fit (not justified per se in this case, but of interest from a robustness point of view), and a variance-stabilizing log-linear fit (with linear model $\log(Y) = a + b * x$). The results were as follows:

Fitting Method	Estimated Half-Life	Standard Error
Weighted Non-Linear	1.09384	0.00023
Unweighted Non-Linear	1.093543	0.00026
Log(Y) Linear	1.09367

The conclusion is that there is a robustness to the fitted NIST result, and under no circumstances is the published LBNL Table of Isotopes result (1.0778) within the range of acceptance.

A manuscript entitled "The Half-life of As-76", co-authored by Lindstrom, Blaauw, and Fleming, has been written and submitted for publication. This will no doubt lead to a correction of the current value in the LBNL web table, and will thus assure that future applications requiring the Arsenic-76 radioisotope half-life will have access to the most accurate, state-of-the-art value available.

3.5.4 Effect of PAC Tube Cooling on Machine Tool Thermal Deformation

Dennis Leber, James Filliben
Statistical Engineering Division, ITL

Mahn-Hee Hahn
Manufacturing Metrology Division, MEL

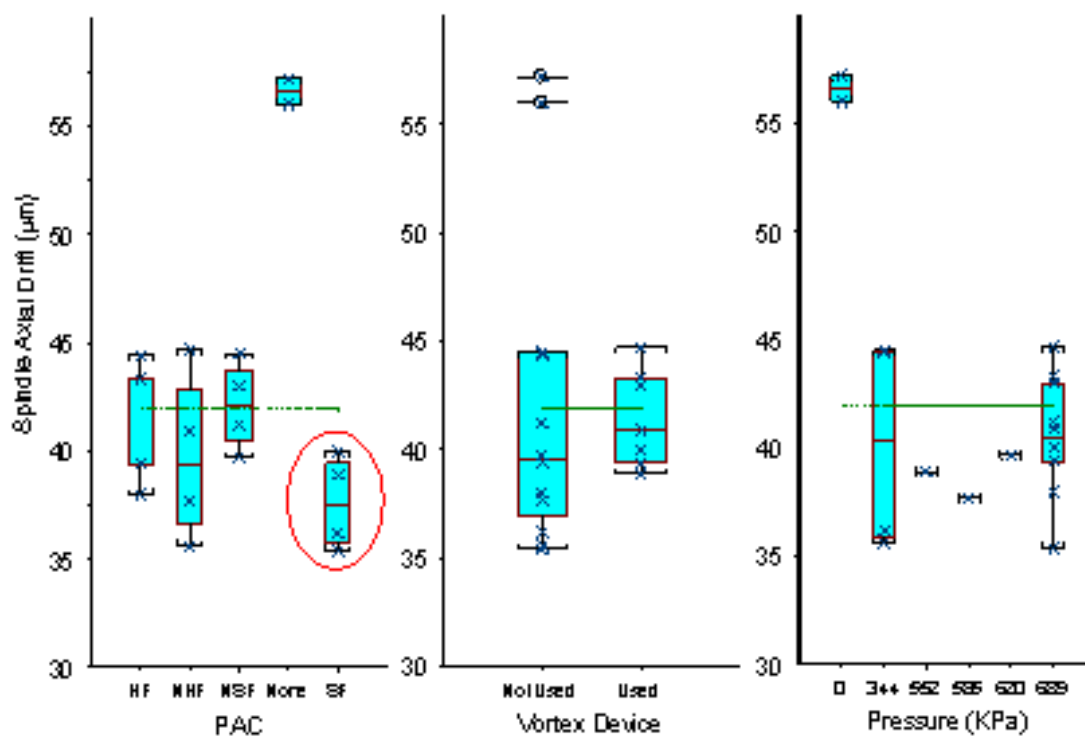


Figure 3.36: Figure 1: Comparing the amount of spindle axial deformation observed based on the various levels of the considered factors of the experiment. The existence of PAC tubing is seen to have a significant favorable effect on spindle axial deformation.

Thermal deformation of machine tool sub-structures has a serious negative impact on the accuracy of the part being machined. The major source of such error is thermal expansion due to heat generated by various sources of electrical power inputs, such as motors, transformers, and mechanical and viscous friction.

In an attempt to minimize the temperature-induced deformation of machine tool components, NIST engineers in the Manufacturing Engineering Laboratory considered an inexpensive cooling device called the Personal-Air-Conditioning (PAC) tubing. Such PAC tubing displays the “Coanda Effect”: tendency of a fluid to cling to the surface that is near an orifice from which the fluid emerges. An important part of the effect is the tendency of a small stream of air or fluid (primary stream) to entrain a large volume of air (secondary stream), and thus cause it to be drawn along with the primary stream. The result is an air-powered fan. A Coanda effect induces the surrounding airflow to increase the cooling effect. Coanda effect cooling devices are described in the literature as a very effective cooling method for some applications.

The PAC tubing is made of a soft silicon extrusion with small-perforated holes manufactured by TEXAN Corporation. When the PAC tubing is pressurized by air, the air escapes through the small holes and clings to the surface of the lip. The escaping air will generate the Coanda effect and, in theory, increase the cooling by agitating the surrounding air. The PAC tubing is marketed for cooling a person working in a hot environment.

In the machine tool, the spindle is considered the largest source of the thermal error and hence was the focus of the experiment to assess the effect of the PAC tubing on thermal deformation. A series of tests were conducted to measure the spindle axial drift, a result of the thermal deformation, applying various PAC tubing in the turning center.

Four types of PAC tubing were considered: SF (Standard Flow), HF (High Flow), NSF (New Standard Flow), and NHF (New High Flow). The original SF and HF tubing are designed to cool a person and perforated air holes are punched at just below two upper lips. To improve PAC tubing efficiency for surface cooling, new tubing has been made by punching the holes close to the lower lip, resulting in PAC tubing types NSF and NHF.

In addition to the four PAC tubing devices considered, the airflow rate applied to the PAC tubing (two flow rate levels) and the existence of a vortex device (existence/non-existence) were considered in the designed spindle experiment. Given the experimental factors of interest and respective levels, the initial experimental plan was a 4x2x2 full factorial design. However, several practical experimental constraints and uncontrollable circumstances led to a modification of the design.

From the above figure of the resulting spindle axial drift observations, we conclude that the use of the inexpensive PAC cooling device was indeed favorable to the reduction in thermal deformation. Although several of the observations with the least spindle axial drift were a result of the SF type PAC tubing, which also appears to perform the best on an overall basis, further investigation showed there is no statistical difference between the four PAC types, only between the existence of a PAC device and not. The additional two factors considered in the experiment, vortex device and airflow rate, were not significant

in reducing the spindle axial drift.

Being able to significantly reduce the thermal deformation in machine tools and in turn increase the accuracy of the parts being machined through an inexpensive PAC cooling device will allow the manufacturing industry to easily increase the quality of the machined parts produced. The results of this study will initially be published in a NISTIR: "Temperature Control of Machine Tool".

3.5.5 Standard Reference Materials

Dennis Leber, James Filliben
Statistical Engineering Division, ITL



Figure 3.37: NIST Health Status Marker Standard Reference Materials

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. These materials are used to perform instrument calibrations as part of overall quality assurance programs, to verify the accuracy of specific measurements, and to support the development of new measurement methods. Industry, academia, and government use NIST SRMs to facilitate commerce and trade and to advance research and development. NIST SRMs are also one mechanism for supporting measurement traceability in the United States.

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 the 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.

Lab Customers

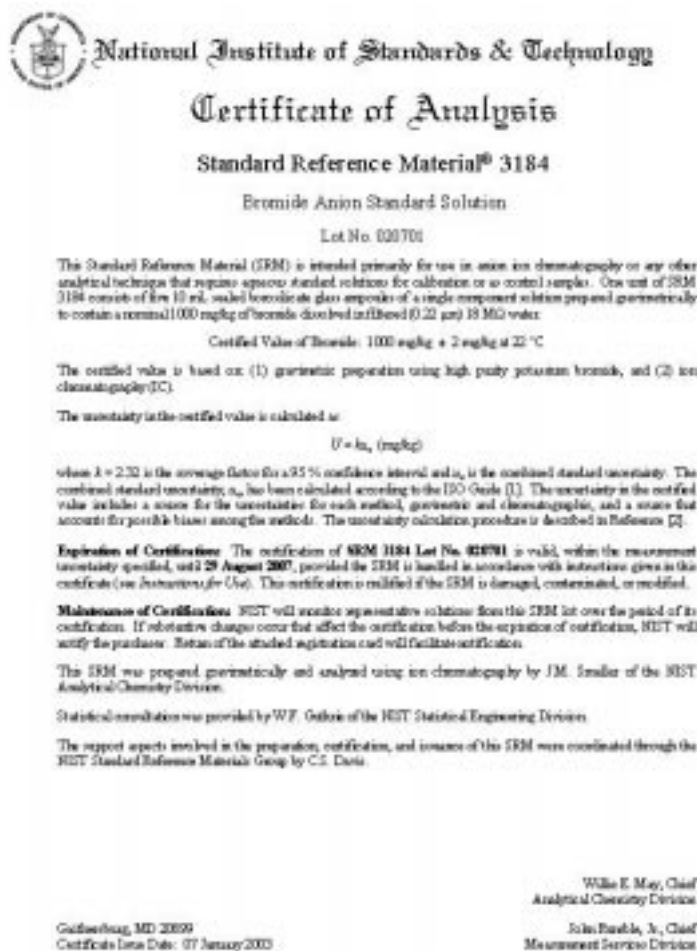
With the exception of the Information Technology Laboratory, all NIST laboratories participate in the production of SRMs and receive SED assistance. Historically CSTL has been, and continues to be, the heaviest consumer of SED SRM services.


Number of SRMs Certified

Although the number of SRMs expected to be completed in 2002 was rather large and optimistic, unexpected circumstances outside of the laboratories' control resulted in less than 20 SRMs being certified in 2002. Due to these delaying circumstances, many of the SRMs expected to have been completed in 2002 have been pushed into 2003.

SRM Accounting

Although the SED SRM accounting and budgeting have been simplified in past years, it still remains complicated – involving 4 separate administrative entities: SED, ITL, the source laboratory, and the SRM Program office. Since SRMs are done on a cost-center basis, such accounting – though burdensome – is nonetheless necessary. The time an SED statistician spends on any given SRM and the phase of the SRM are recorded bi-weekly in a central database. The data from this central database are formulated into quarterly and targeted reports for the laboratories. The laboratories and SED use these reports to monitor and measure the progress of their SRM program and make any necessary budgetary adjustments. In 2002 a web version of this database was created and is expected to be fully utilized in 2003, giving the laboratories continuous access to the SED SRM accounting database. Hopefully the accounting process will continue to be simplified in the future.



 National Institute of Standards & Technology

Certificate of Analysis

Standard Reference Material® 3184

Bromide Anion Standard Solution

Lot No. 020701

This Standard Reference Material (SRM) is intended primarily for use in anion ion chromatography or any other analytical technique that requires accurate standard solutions for calibration or as control samples. One unit of SRM 3184 consists of five 10 mL sealed borosilicate glass ampoules of a single component solution prepared gravimetrically to contain a nominal 1000 mg/kg of bromide dissolved in distilled (0.22 µm) 18 MΩ water.

Certified Value of Bromide: 1000 mg/kg ± 2 mg/kg at 22 °C

The certified value is based on: (1) gravimetric preparation using high purity potassium bromide, and (2) ion chromatography (IC).

The uncertainty in the certified value is calculated as:

$$U = k u_c \text{ (mg/kg)}$$

where $k = 2.58$ is the coverage factor for a 95 % confidence interval and u_c is the combined standard uncertainty. The combined standard uncertainty u_c has been calculated according to the ISO Guide [1]. The uncertainty in the certified value includes a source for the uncertainties for each method, gravimetric and chromatographic, and a source that accounts for possible biases among the methods. The uncertainty calculation procedure is described in Reference [2].

Expiration of Certification: The certification of SRM 3184 Lot No. 020701 is valid, within the measurement uncertainty specified, until 29 August 2007, provided the SRM is handled in accordance with instructions given in this certificate (see Instructions for Use). The certification is nullified if the SRM is damaged, contaminated, or modified.

Maintenance of Certification: NIST will monitor representative solutions from this SRM lot over the period of its certification. If substantive changes occur that affect the certification before the expiration of certification, NIST will notify the purchaser. Retests of the studied solutions and will facilitate certification.

This SRM was prepared gravimetrically and analyzed using ion chromatography by J.M. Smoller of the NIST Analytical Chemistry Division.

Statistical consultation was provided by W.F. Guthrie of the NIST Statistical Engineering Division.

The support aspects involved in the preparation, certification, and issuance of this SRM were coordinated through the NIST Standard Reference Materials Group by C.S. Davis.

Gaithersburg, MD 20899
Certificate Issue Date: 07 January 2003

William E. May, Chief
Analytical Chemistry Division

John Panchik, Jr., Chief
Measurement Services Division

SRM 3184 Page 1 of 2

Figure 3.38: Certificate for SRM 3184

3.5.6 Army CCG Project 474 Gas Mask Verification

James J. Filliben

Statistical Engineering Division, ITL

Robert Fletcher

Surface and Microanalysis Science Division, CSTL

Gas Mask Aerosol Electrometer Calibration

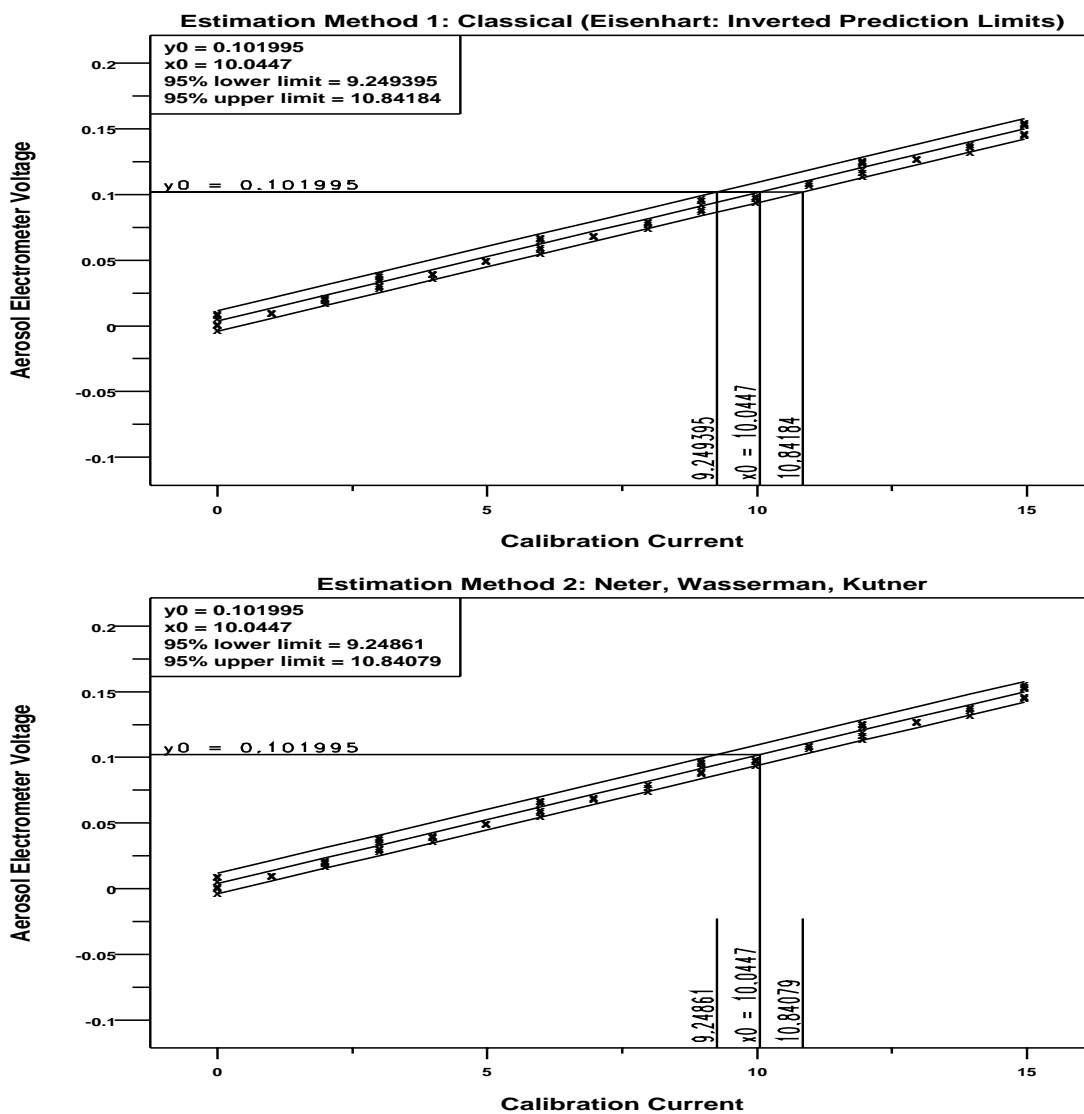


Figure 3.39: Top: Linear calibration limits obtained via the classical (Eisenhart) inverted prediction limits method.

Bottom: Linear calibration limits obtained via the Neter, Wasserman, Kutner method.

With the Iraq conflict pending, and with the possible threat of chemical and biological agents therein, US military personnel are increasingly dependent on having high-quality gas masks as part of their standard gear.

The Army has developed an on-soldier, in-field gas mask fit quality test (the "M41") that measures and compares in situ the ambient aerosol concentration outside of the fitted mask and the aerosol concentration inside the mask. PATS will detect both leaks in the mask and inefficiencies in the filter. This technology is believed to provide a complete diagnostic of the integrity of the mask, the filter, and the fit on the individual.

The M41 core is the CPC (Condensation Particle Counter) which carries out the measuring and reporting of the aerosol concentration values. The accuracy of the CPC's concentration values is obviously of critical importance for the Army testing program and for the safety of our military personnel.

In this regard, the Army has developed a test stand protocol for verifying the concentration accuracy of the CPCs. The basic test stand design consists of a collision aerosol generator, diffusion dryers, an electrical charge neutralizer, a DMA (differential mobility classifier), and the CPC.

Regarding the aerosol generator, a novel approach has been taken by using non-toxic 80 nm solid polystyrene spheres as a surrogate for the poisonous chemical and biological gases. Such micro-spheres are desirable because they behave in a similar manner and have the same fluid dynamical properties in airflow streams.

Regarding the CPC and its accuracy, the usual remedy (an SRM) is precluded in this case because present aerosol technology—especially with the transient nature of aerosols—have made such SRMs unfeasible. In the past, the Army has thus relied on TSI (the CPC manufacturer) to carry out the CPC accuracy check by whatever in-house, company-specific methods were available. More recently, however, due to the very real likelihood of impending use in combat, the Army has declared that the entire CPC calibration process must be upgraded so that it is entirely traceable—to NIST. This is the current focus of Bob Fletcher's work.

NIST employs a dual-method approach to verifying the accuracy of the M41 and PATS CPCs. Method 1 is to utilize a PC (particle counter) filter and carry out raw SEM (scanning electron microscopy) microscopy—a primary method for measuring particle concentration which is traceable back to the NIST SEM pitch standard SRM 484g. Method 2 is to utilize an AE (Aerosol Electrometer)—a method which is ultimately traceable to the NIST electrical standard for current. Hence, during CPC testing at NIST, as a volume of aerosol passes through the test chamber, the CPC will respond with an aerosol concentration, the AE will respond with a (concentration dependent) voltage output, and the PC filter will capture the actual particles (for later counting via the SEM).

To achieve in practice the desired traceability link, the AE must be (pre-)calibrated (before the joint test run) to a femtoamp current source produced by a high precision voltage source and a 100 gigaohm resistor. Both the resistor properties and the voltage were

measured by the Electricity Division at NIST. The NIST voltage source has an accuracy on the order of 100 ppm and the measurements on the resistor showed the resistance to be $1.00269 \pm 0.00070 \times 10^{11}$ ohms.

The net effect is that AE voltage V will have been calibrated against a highly accurate (and NIST traceable) current I . The uncertainties for this linear calibration may be statistically derived in several different ways—the figures show calibration uncertainty intervals for 2 such methods (Figure 3.40, top: Eisenhart's Inverted Prediction Intervals, and Figure 3.40, bottom: Neter, Wasserman, and Kutner's method).

Once this linear calibration has been executed and the uncertainty noted, then it is an easy step to relate the AE voltage reading to a high-accuracy aerosol concentration by invoking the following physical relationship between aerosol concentration N and current I :

$$N = I / (Q \cdot e)$$

where

- N is the aerosol concentration,
- I is the current (NIST traceable),
- Q is the flow rate (known), and
- e is the magnitude of the elementary charge (known).

Thus for a given aerosol exposure, the observed AE voltage may be easily translated into a high-accuracy concentration, and then subsequently compared to the lower (but adequate) accuracy from the original CPC reading.

A world of bio- and chemo- weaponry is a fact of life that will probably not change any-time soon. Given that, the PATS system along with the associated AE voltage-current calibration will bring a new level of accuracy and confidence to the quality and performance of in-field gas masks. The net effect of this improved metrology is literally life-saving—for both military personnel and (subsequently) general citizenry alike.

3.5.7 SRM 2396–Oxidatively-Modified DNA Biomarkers

James J. Filliben

Statistical Engineering Division, ITL

Henry Rodriguez, Miral Dizdar

Biotechnology Division, CSTL

1. MONDAY												
Q1:	7	5	8	1	11	4	12	6	9	2	3	10
Q2:	11	2	12	3	7	10	4	9	1	5	6	8
Q3:	9	11	6	10	3	8	7	4	2	1	5	12
Q4:	2	9	4	12	10	5	6	1	7	8	11	3
2. WEDNESDAY												
Q1:	5	7	2	6	4	3	11	12	8	9	10	1
Q2:	12	8	7	5	1	9	3	10	11	6	4	2
Q3:	1	4	3	2	6	12	5	11	10	7	8	9
Q4:	4	3	9	8	5	11	2	7	12	10	1	6
3. FRIDAY												
Q1:	8	1	10	7	2	6	9	3	4	11	12	5
Q2:	3	6	1	4	9	7	10	8	5	12	2	11
Q3:	6	10	5	11	12	1	8	2	3	4	9	7
Q4:	10	12	11	9	8	2	1	5	6	3	7	4

Figure 3.40: This is the 12-by-12 Latin square design for the DNA Biomarker SRM. The table entry is the primary factor of interest (= bottle/base/compound). The 2 nuisance/secondary factors are quarter-within-a-day (row) and time-segment-within-a-quarter (column). The Latin square design by definition assures that every bottle/base/compound occurs once and only once per column and per row. This Latin square will protect the bottle/base/compound concentration estimates from time contamination between days and within days.

The last few years have seen explosive growth in biotechnology, and NIST's role therein. At the core of biotech is DNA—its measurement, its understanding, and its characterization. In that regard, CSTL's Biotechnology Division is working very hard to develop the technology to:

1. Identify and develop DNA damage measurement techniques;
2. Measure oxidative DNA damage;
3. Characterize DNA repair enzymes; and
4. Produce and provide critical DNA standards.

One such standard is SRM 2396 (Oxidatively-Modified DNA Base Biomarkers) which would allow the research, government (e.g., NIH) and industrial DNA communities to accurately measure and assess "oxidative stress" (DNA lesions caused by the oxidation process). Accurate measurement of such stress is a necessary first step in

1. understanding the DNA damage and repair process; and
2. assessing whether such oxidatively-damaged DNA does in fact serve as an early risk factor for certain age-related diseases.

Henry Rodriguez and Mirad Dizdar of CSTL are in the process of fabricating SRM 2396, which will consist of a 12-vial kit, with each vial containing a different DNA base (compound). The research customer would thus run the entire 12-vial kit through his/her measurement procedure (chromatography (gas or liquid) + UV mass spectrometry) to obtain 12 base concentrations. These 12 customer concentrations would then be compared with the 12 certified SRM 2396 values, and the appropriate customer calibration/adjustment carried out. NIST/CSTL will provide approximately 300 such 12-vial kits for customer distribution.

The experimental starting point for this DNA SRM is 12 (large) bottles—one bottle for each specific DNA base/compound of interest. For each of the 12 bottles/bases/compounds, the CSTL measurement process consists of 2 steps:

1. aliquoting from bottle to vial (to potentially produce 300 vials for each base/compound);
2. subjecting a subset of the (potential 300) vials to NIST's UV mass spectrometry to arrive at a certified value for that given base/compound.

It is desired that this SRM be globally (rather than individually) certified: that is, for a given base/compound out of the 12, the 300 vials should have a single certified concentration (as opposed to having 300 individual certified concentrations). The outcome

from this DNA SRM experiment will thus be 12 (certified) values—one for each bottle/base/compound.

CSTL contacted SED early on, and so SED has had an opportunity to contribute in a significant way via the construction of an appropriate experiment design. Regarding this design, the following questions were posed and answered:

- Q1. Is CSTL interested in bottle to bottle effects? Not per se, since the different bottles are different bases/compounds, but we do want to be assured that the final certified number for a given bottle represents the true bottle value and thus is not contaminated/biased by other factors.
- Q2. What other factors might contaminate the bottle/base/compound readings? As with most experiments, the main potential contamination culprits are "time" factors, such as day and hour within day.
- Q3. How will we protect against day and hour-within-day contamination? We invoke the usual: randomization, blocking, and balancing.
- Q4. How long does it take to carry out a single run (aliquot + UV spectrometer)? It takes about 5 to 10 minutes.
- Q5. How many measurements per day or per hour are possible? About 6 to 12 aliquots per hour is possible, and hence about 50 to 100 or so per day is an upper bound.
- Q6. How many of the 12 bottles/bases per day should be sampled? Since day is a potential confounding factor, drawing from all 12 bottles on each single day would be ideal.
- Q7. How many days should the total experiment take? The practical upper bound on this is 2 weeks (10 work days).
- Q8. For a given bottle/base/compound, how many aliquots should be drawn so as to arrive at the final certified value? This a sample size n question—the most basic of all experiment design questions—whose answer formally depends on 3 issues:
 1. the desired uncertainty for the final computed value;
 2. the raw aliquot-to-aliquot variability that will occur from a given bottle/base/compound; and
 3. the number of runs that is practically affordable (\$/time constraints/reality).

For this DNA biomarker case, no information was available for issues 1 and 2, and issue 3 (affordable sample size) was quite generously large (since a run only takes 10 minutes at most). Given the above, it was decided that an absolute lower bound on the number of runs per bottle would be 10—which is large enough that the 95% probability coefficient is already approximating the asymptotic limit of roughly 2 (1.96), and yet is small enough that it is affordable to run in practice.

The final design is shown in Figure 3.41. This design was constructed by conceptually splitting a single day into 4 time quarters (of roughly 2 hours each), splitting each 2-hour time quarter into twelve 10-minute segments, and then forcing/designing every bottle/base/quarter to occur

1. the same number of times (= 4) within each day;
2. the same number of times (= once) within each 2-hour quarter within a day;
3. the same number of times (= once) within each 10-minute time segment within a quarter;
4. the same number of times (= 12) within each 10-minute time segment across all 3 days.

The net result is that the final design protects against bottle/base/compound contamination that might result from

1. between-day drift;
2. within-day drift;
3. within quarter drift and/or setup effects.

In toto, the final design across the 3 days is a 12-by-12 latin square in which the primary factor is bottle (12 levels), and the 2 nuisance factors are time segment within quarter (12 levels) and quarter (= 4 levels per day x 3 days = 12 levels).

This design has the following advantages:

1. Bias protection: It provides an extremely high level of contamination/bias protection for each 12 bottle/base/compound concentration estimate;
2. Precision: It provides a reasonable number of replicates (= 4 runs per day x 3 days = 12) for each bottle/base/compound concentration estimate; and
3. Practical: This design is comfortably efficient to run (only 48 runs per day, and only 3 days).

This design will be executed by CSTL later on in the year 2003.

This experiment design assures that the resulting 12 SRM DNA biomarker concentration values will be as free from time-contamination as possible. Upon completion, the SRM itself will serve as a vital NIST contribution to advancing the state-of-the-art in the many exciting facets of DNA research.

3.5.8 IAEA Isotope Reference Materials Intercomparison: Carbon and Oxygen

Stefan Leigh, James Yen
Statistical Engineering Division, ITL

Michael Verkouteren, Donna Klinedinst
Surface and MicroAnalysis Science Division, BFRL

EIGENVECTORS

	1	2	3	4	5	6	7
relden	0.3898	0.1214	-0.1628	-0.0123	0.7330	0.0828	-0.1937
instr	0.0622	0.0853	0.2522	0.7252	0.0718	0.3052	-0.2108
refill	-0.1634	-0.3720	-0.4695	0.0409	-0.0430	0.7498	0.1473
evac	-0.4697	0.0501	-0.2820	-0.1172	0.0260	-0.0621	-0.8097
volt	0.2017	-0.4582	-0.1045	-0.2947	0.3283	-0.0976	0.0642
idle	-0.4335	0.2921	-0.1593	0.0670	0.1029	-0.0521	0.3920
intgtm	-0.4135	0.2625	-0.0376	0.0657	0.5104	-0.0206	0.2670
elecen	0.1794	0.2540	-0.5457	0.1253	-0.1062	-0.2524	0.0536
accvolt	0.1186	0.4365	-0.0642	-0.4753	-0.1830	0.3034	0.0108
d18WRG	0.3099	0.4665	0.0077	-0.0628	-0.0476	0.3194	-0.0765
d46WRG	0.2343	-0.0178	-0.5199	0.3424	-0.1710	-0.2520	0.0535

Principal components for a CO₂ explanatory variable array. A δ^{18} response shows high correlation with PC's 1, 4, and 5. PC 1 is a time PC. PC 4 is an instrumental model PC. PC 5 is a relative density PC.

Isotope reference materials are used to relate field measurements to stated references in many international applications of engineering, commerce, and atomic energy regulation. The total combined uncertainties of typical field measurements are in large part due to uncertainties in the realization of the internationally accepted isotope ratio scales. This arises from the uncertainties in value assignments of Reference Materials, but more fundamentally from lack of control of subtle physicochemical and instrumental factors that limit the accuracy and reproducibility of isotopic measurements. Because these factors can be complex combinations of many variables, intercomparison exercises across independent laboratories have historically represented the soundest approach to representing and exploring the broad range of potentially influential variables as well as for establishing credible consensus values.

NIIST assumed the role of pilot laboratory in organizing, collecting data for, and reporting findings for the latest multinational intercomparison sponsored by the International Atomic Energy Agency's (IAEA's) Working Group on Light Stable Isotope Reference Materials. This exercise involved preparation and isotopic measurement of CO₂ derived from

carbonates, waters and pure CO₂ reference materials. The goal was to determine accurate and precise $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ value assignments consistent across materials, and to relate variations in results with discretionary experimental factors to guide future inter-comparisons. Common materials and instructions were shipped from NIST to the participating laboratories and the raw measurement data and factor information reported back to NIST. Data were corrected for cross-contamination and then processed to determine standardized $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ values. Nine laboratories were invited to participate. Data from six were used in the final analysis.

In addition to providing the raw measurement data, each laboratory supplied detailed information on 30 discretionary chemical and instrumental variables that could influence the reproducibility of the measurements. There were three generic types of factors: reported numerical values and settings, values derived from reported measurements, and interpreted procedural differences. Graphics, correlation analysis, and Principal components Analysis (PCA) were used to explore relationships among factors and measurement results.

Examples of specific factor impacts studied include: A specific question going into the exercise was to determine whether the range in specific gravity of the various phosphoric acid preparations used would influence the $\delta^{18}\text{O}$ results: no significant correlations were apparent in any of the modes of analysis for any of the samples, and so the conclusion was negative. A general screen for individual variables, or linear combinations of variables, with high explanatory power for different versions of $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ responses was applied, using Principal Components Regression. Carbonate materials yield no single clearly identifiable factor of major significance. For CO₂ samples, however, uncorrected δ 's exhibit strong negative correlations with eigenvectors heavily weighted by combinations of time-related variables (evacuation time, idle time, integration time) and significant positive correlation with acceleration voltage. In the case of responses corrected for cross-contamination, for $\delta^{18}\text{O}$ the time eigenvector retains its apparent importance, while for $\delta^{13}\text{C}$ instrument 'model' (type) appears most influential.

Isotope reference materials are used to relate field measurements to stated references in many applications of international science, engineering, commerce, and atomic energy regulation. For the carbonate and CO₂ materials, the results of this pilot intercomparison provided value assignments with uncertainties improved by factors up to two over previous assignments. Information gleaned on the impact of discretionary factors will be used to fine tune future intercomparisons.

3.5.9 Charpy V-notch Reference Value Uncertainty

Jolene Splett, Jack Wang

Statistical Engineering Division, ITL

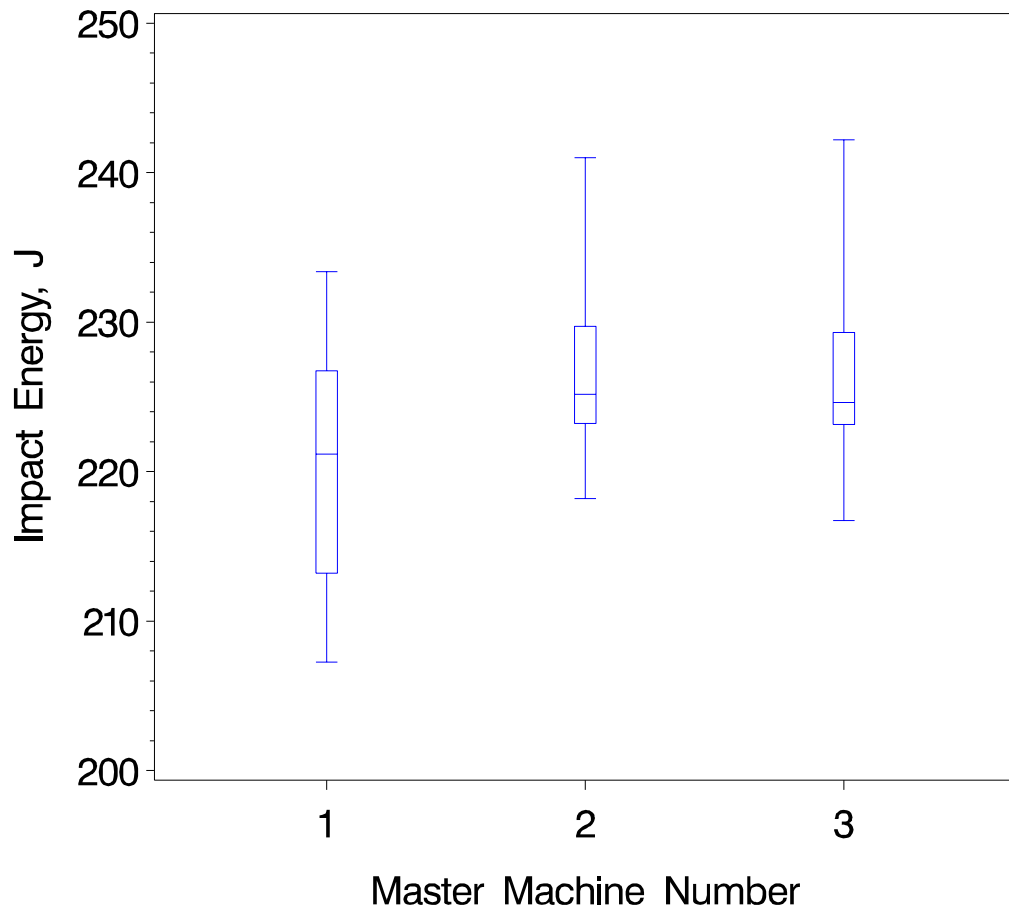


Figure 3.41: Pilot lot results for three master Charpy machines.

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

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

We have completed a paper describing the justification and computation of the certified value uncertainty. The paper, "Uncertainty in Reference Values for the Charpy V-notch Verification Program," has been submitted to the *Journal of Testing and Evaluation*.

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.

3.5.10 Standard Reference Materials for the Food Industry

James Yen, Stefan Leigh, Blaza Toman, Lisa Gill
Statistical Engineering Division, ITL

Kathy Sharpless, Steve Wise, Michele Schantz
Analytical Chemistry Division, CSTL

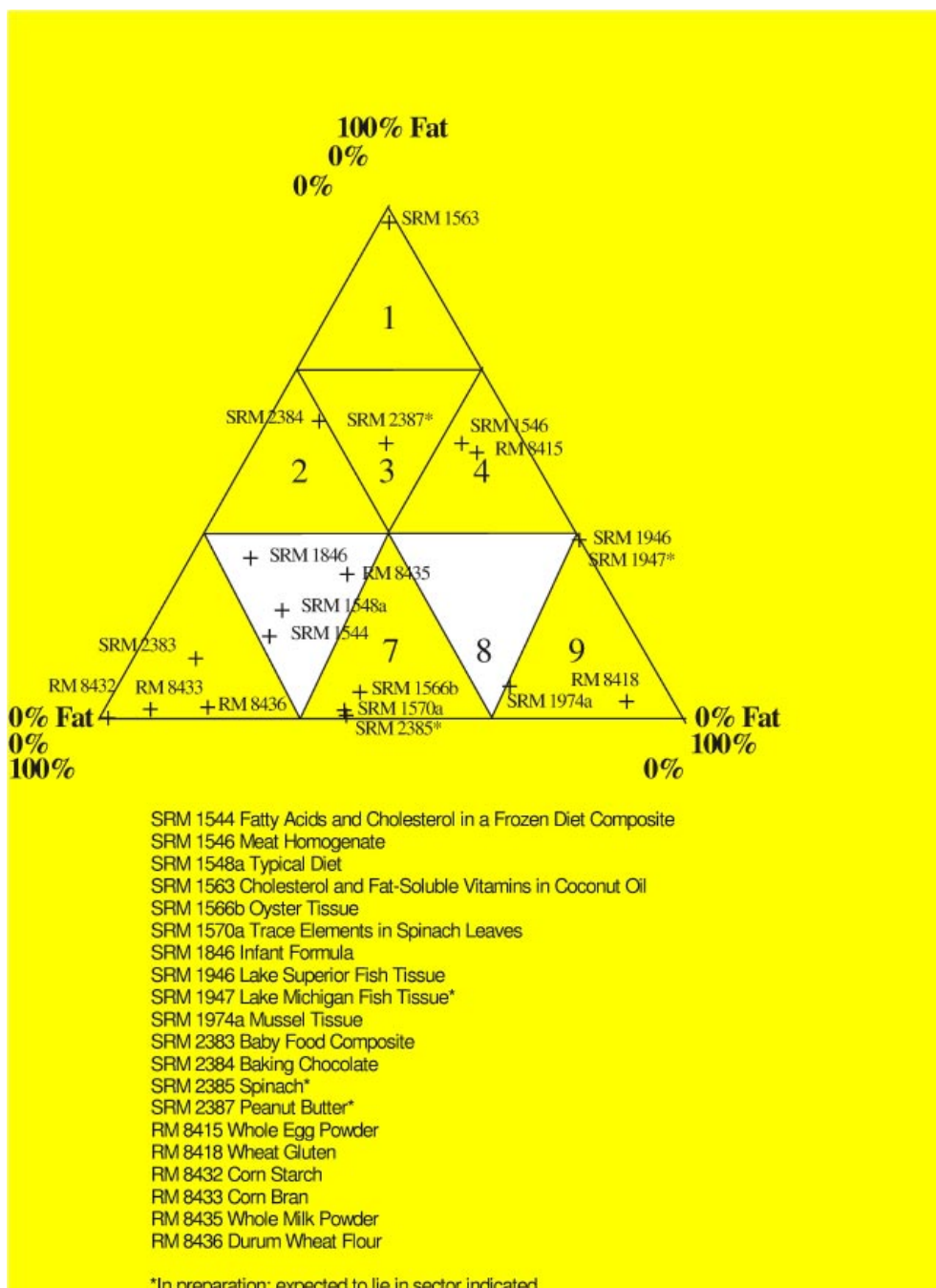


Figure 3.42: NIST food-matrix reference materials and their location on the fat-protein-carbohydrate triangle proposed by the Task Force on Methods for Nutrition Labeling.

The Nutrition Labeling and Education Act of 1990 requires that labels on processed foods distributed in the United States specify the amount of total fat, saturated fat, cholesterol, total carbohydrate, dietary fiber, sugars, protein, vitamin A, vitamin C, sodium, calcium, and iron contained in a single serving. In addition the manufacturer may also provide information about any other vitamin, mineral, or nutrient. To facilitate compliance with this law, well-characterized reference materials are needed by laboratories in the food testing and nutrition communities. NIST has provided and continues to provide reference materials with certified and reference values for vitamins and other nutrients.

AOAC (Association of Analytical Communities) International's Task Force on Methods for Nutrition Labeling has proposed a triangle partitioned into sectors in which foods are placed based on their protein, fat, and carbohydrate content. The accompanying figure shows this triangle along with the location of NIST food-matrix reference materials. AOAC International anticipates that one or two reference materials in a given sector will be representative of other foods in that sector and thus will be useful for method assessment and quality assurance for analyses of those foods.

SED statisticians work with chemists from the Analytical Chemical Division to make sure that each of these food reference materials is issued with an appropriate list of content values and associated uncertainties. Other major participants are industry laboratories associated with the National Food Processors Association (NFPA). These labs participate in Round-Robin measurement studies of the reference materials.

The NFPA Round-Robin estimates of content values are combined with NIST estimates to form certified content values. The long list of certified values with accompanying uncertainties in each food-matrix SRM is the fruit of a long continuing effort at SED to develop appropriate techniques to combine values and uncertainties from multiple methods and multiple laboratories. The method developed by Susannah Schiller and Keith Eberhardt provided sensible uncertainty intervals that overlapped the individual method means. The more recent BOB (Type **B** on Bias) method by Levenson et al provides answers similar to that of the Schiller-Eberhardt method and has the advantage of following the ISO *GUM (Guidelines to the Measurement of Uncertainty)*. Mark Vangel and Andrew Rukhin developed Maximum Likelihood solutions for the multi-method problem and related their results to those of Mandel and Paule. A useful tool for producing these estimates is the function "Consensus Means" developed by Stefan Leigh and Alan Heckert for the DATAPLOT software package.

NIST's program of food-matrix reference materials with reliable content values for various nutrients helps the food industry comply with FDA rules and helps consumers in making dietary choices. Such reference materials also help the food industry by providing measurement traceability for food exports.

3.6 New Methods for Metrology

3.6.1 Errors in Variables for Gas Standard Calibration

S. Leigh

Statistical Engineering Division, ITL

Frank Guenther

Analytical Chemistry Division, CSTL

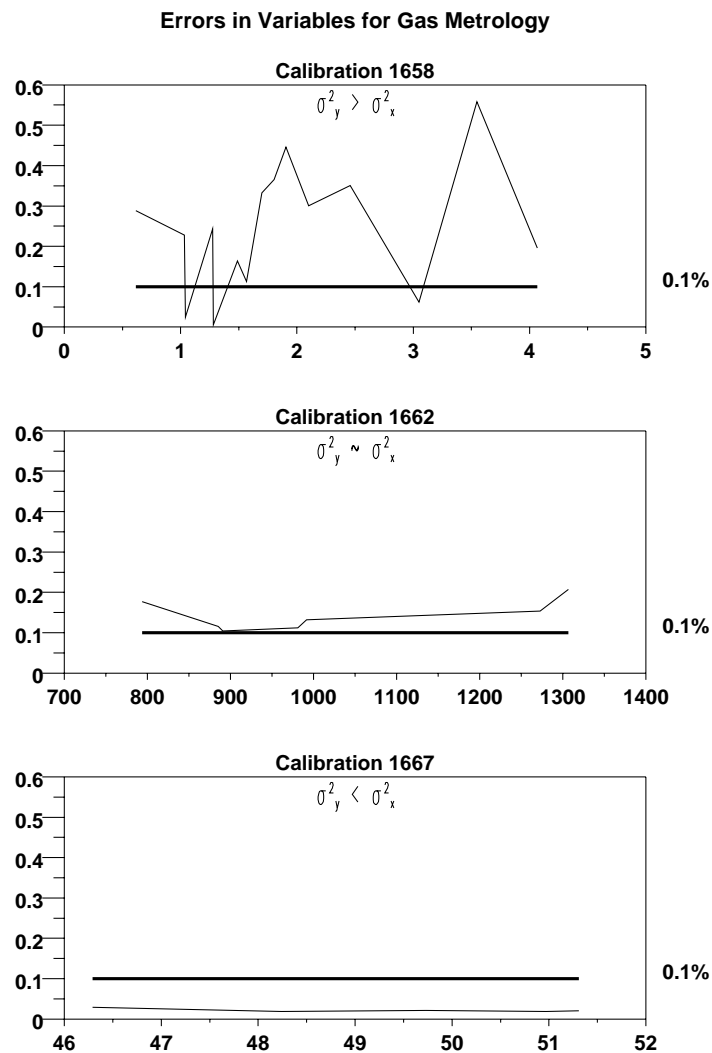


Figure 3.43: Relative standard deviation plots derived from actual calibration lines showing the occurrence of each of the three modes for regression analysis: $\sigma_Y \gg \sigma_X$, $\sigma_Y \approx \sigma_X$, $\sigma_Y \ll \sigma_X$. The solid horizontal 0.1% line represents the assumed relative 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: Ordinary Least Squares (OLS) fitting, with associated Working-Hotelling and Fieller (propagation-of-error based) calibration bounds. But it has been observed that often the basic assumptions of linear least squares are not met: specifically, the crucial assumption $\sigma_Y^2 \gg \sigma_X^2$ may be violated. This fact led the ISO Gas Analysis Working Group TC 158 to develop and adopt a new Gas Analysis Standard (ISO 6143.2) based on the use of Errors in Variables (EiV) methodology for the analysis of gas mixture composition calibrations.

A draft standard was circulated over a period of 3 years prior to formal 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. SED/NIST noted multiple troubling features of the new standard; incorrect language is employed; the use of Errors in Variables methodology, to supplant OLS, is recommended irrespective of the variance ordering in the data; EiV in the Standard is implemented as an unclearly documented ‘black box’ executable code with source code not made available. Finally, and most telling, no reference is made to the large body of hard statistical research in this area, and no reasons given for ignoring classic solutions, such as Maximum Likelihood Estimates for specific variance ordering scenarios.

We are in the process of carefully investigating the EiV literature and NIST Gas Metrology Group’s archival calibration data, with a view to developing a statistically justifiable set of procedures for the analysis. While we are sympathetic to the ISO push for the use of EiV technology in metrological calibrations, it is our belief that the new 6143.2 is untenable: unacceptable in its nontransparency to the community that is expected to use it, indefensible in its substitution of computer code for careful statistical understanding and analysis of data on a case-by-case basis.

Errors in Variables is a complex subject, even for statisticians. Identifiability and estimability problems are acute. Often the existing literature is not clearly expositied, tending to linger over arcane counterexamples rather than presenting practitioners with clearcut guidelines and procedures. We seek to lay down a clear logic for the linear calibration problem with specific reference to NIST Gas Metrology calibration experience. We expect those guidelines to be documented, or documentable, with explicit reference to broadly accepted inferential principles, such as maximum likelihood, method of moments, 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.

Our first goal is to adapt existing Maximum Likelihood approaches, with associated Fieller calibration intervals, to the calibration setting at hand.

EiV typically assumes three underlying equations for the observed pairs X_{ij}, Y_{ij} :

$$X_{ij} = U_i + \delta_{ij}$$

$$Y_{ij} = V_i + \epsilon_{ij}$$

with the linear relation between V_i and U_i

$$V_i = b + mU_i. \quad (3.33)$$

Here $i = 1, \dots, N$ indexes the gas cylinders being used, and $j = 1, \dots, n_i$ indexes the repeated measurements on the i th cylinder.

The U_i denote the true (but unknown) gas standard concentrations, the X_{ij} are the analyst's readings of those concentrations, the V_i represent the true (but unknown) instrumental response, and the Y_{ij} are the analyst's readings of that response.

The assumed underlying stochastic structure is

$$\begin{aligned} \delta_{ij} &\sim N(0, \sigma_\delta^2), \\ \epsilon_{ij} &\sim N(0, \sigma_\epsilon^2). \end{aligned}$$

The errors-in-variables literature distinguishes between two forms of such a model: the Functional model (think F for Fixed) and the Structural model (think S for Stochastic). For the structural model, the U_i 's are considered to be realizations of one random variable U . Estimates then need to be made for μ_U and σ_U^2 .

For the Functional model, one still assumes the above set of equations to be operative, but now the U_i are considered to be “unknown” constants (i.e., not random variables).

The gas metrology calibration set-up, as we understand it, fits the Functional category. In this situation we have $N + 4$ unknown parameters, $m, b, \sigma_\delta^2, \sigma_\epsilon^2$ and U_1, \dots, U_N . The first four of these parameters are *structural* (i.e., they are of interest); the last N unknowns are *incidental* or nuisance parameters.

With respect to the choice of assumptions for σ_δ^2 and σ_ϵ^2 , we can have too much of a good thing. If we assume known or knowable both of the σ 's, then there are 5 sufficient statistics for a 4-parameter estimation problem, which becomes overidentified. It therefore seems not unreasonable to begin by assuming that only σ_δ^2 is known, which we believe is representative of about 80% of the calibrations run at NIST.

To render the estimation problem tractable, one can assume both error variances known, or only one variance known. One can consider - for example - a ratio of the variances ($\lambda = \sigma_\epsilon^2 / \sigma_\delta^2$) to be known, or one can consider both of the variances to be known. In the Functional model, parameters are identifiable without additional assumptions. However, constructing maximum likelihood estimators that exhibit the appropriate “good” behavior with an explicitly parametrizable likelihood that has unambiguous maxima turns out to be more difficult.

Indeed, the log-likelihood function for the Functional model has the form

$$\begin{aligned} &-p \log(2\pi) - p \log(\sigma_\delta) - p \log(\sigma_\epsilon) \\ &-\frac{1}{2} \left[\sum_i \sum_j \frac{(X_{ij} - U_i)^2}{\sigma_\delta^2} + \sum_i \sum_j \frac{(Y_{ij} - (b + mU_i))^2}{\sigma_\epsilon^2} \right]. \end{aligned}$$

with $p = \sum_{i=1}^N n_i$. The problem with this likelihood function is that it may not have a finite maximum. To see this when $n_i \equiv 1$, put $U_i = X_{ij}$, and let σ_δ^2 tend to zero. Then the value of the likelihood approaches infinity, showing that there is no maximum likelihood solution.

If one of σ_δ^2 or σ_ϵ^2 is assumed known, but not the other, then maximum likelihood estimation breaks down (Moberg and Sundberg, 1978), and does not immediately yield sensible estimators. The standard fix is then to use the corresponding MLE's from the Structural model and treat them as method of moments estimators for the Functional model. For σ_δ^2 known,

$$\hat{m} = \frac{s_{XY}}{(s_{XX} - \sigma_\delta^2)}$$

assuming $s_{XX} > \sigma_\delta^2$ and $s_{YY} > s_{XY}^2/(s_{XX} - \sigma_\delta^2)$, with the unknown variance σ_ϵ^2 estimated from $s_{YY} - \hat{m}s_{XY}$.

For σ_ϵ^2 known

$$\hat{m} = \frac{(s_{YY} - \sigma_\epsilon^2)}{s_{XY}}$$

assuming $s_{YY} > \sigma_\epsilon^2$ and $s_{XX} > s_{XY}^2/(s_{YY} - \sigma_\epsilon^2)$. The unknown variance σ_δ^2 can be estimated from $s_{XX} - s_{XY}/\hat{m}$.

In both cases

$$\hat{b} = \bar{y} - \hat{m}\bar{x}.$$

The problems with maximum likelihood for the Functional model relate not to lack of identifiability, but rather to the nuisance parameters U_i , whose number increases with the sample size. In the presence of such nuisance parameters, MLE's need not exist, or if they do (exist), the estimators need not be consistent because the classic good large sample behavior for MLE's requires the number of parameters to be fixed with respect to sample size.

The underlying problem was reviewed before the American Chemical Society (ACS 220th Meeting, Washington, D.C., Abstract 178). An updated presentation is scheduled for PITTCO (March 2003, Orlando) and submission of a "critique with corrections" to *Metrologia* is anticipated for late Summer 2003.

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

3.6.2 Generalized Tolerance Intervals

Jolene Splett, Jack Wang, Hari Iyer, Dom Vecchia
Statistical Engineering Division, ITL

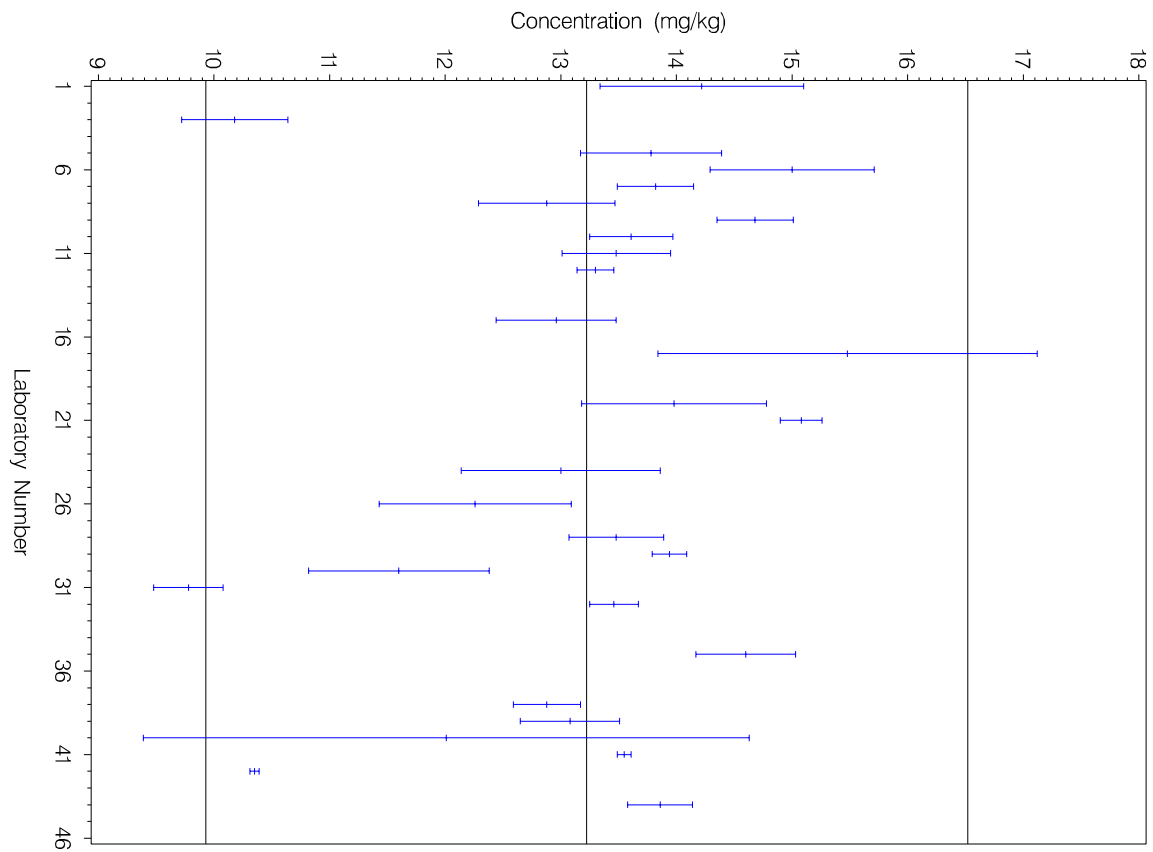


Figure 3.44: Arsenic concentrations in oyster tissue samples (Standard Reference Material 1566a) as measured by 29 laboratories participating in an interlaboratory study. The horizontal reference lines represent the overall mean and the upper and lower generalized tolerance limits for 95 % coverage and 90 % content.

The goal of an interlaboratory study is to determine if measuring systems utilized by different laboratories are providing similar results when measuring a standard artifact. Statistically valid procedures for analyzing the resulting data are crucial for providing meaningful results.

In an interlaboratory study, each laboratory repeatedly measures an artifact. We model the resulting data using a one-way random effects model with possibly unequal sample sizes and possibly heterogeneous variances. Using the method of Generalized Intervals, we have developed a tolerance interval procedure for the distribution of laboratory means. Such information can be useful in deciding whether or not one or more laboratories deviate significantly from the rest.

The accompanying figure displays the results of an interlaboratory study to compare arsenic concentrations in oyster tissue samples (Standard Reference Material 1566a). The vertical “bars” on the plot represent individual laboratory means with one standard deviation limits. The three horizontal lines on the plot represent the overall mean and the upper and lower generalized tolerance limits for 95 % coverage and 90 % content. While two laboratories (3 and 31) appear to have somewhat lower concentrations than the other laboratories, there is not sufficient evidence to eliminate these two laboratories from the study based on the generalized tolerance interval.

The generalized tolerance interval can be somewhat conservative but appears to hold the coverage probability sufficiently close to the nominal value. Thus the generalized tolerance interval is useful for practical applications.

Generalized tolerance intervals provide a statistically valid means of analyzing interlaboratory study data which are often unbalanced and have unequal variances. The generalized tolerance interval procedure is particularly useful since there appears to be no other satisfactory frequentist solution to this problem, except in the large sample case.

3.6.3 Consensus Curve Estimation

James Yen, James Filliben
Statistical Engineering Division, ITL

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

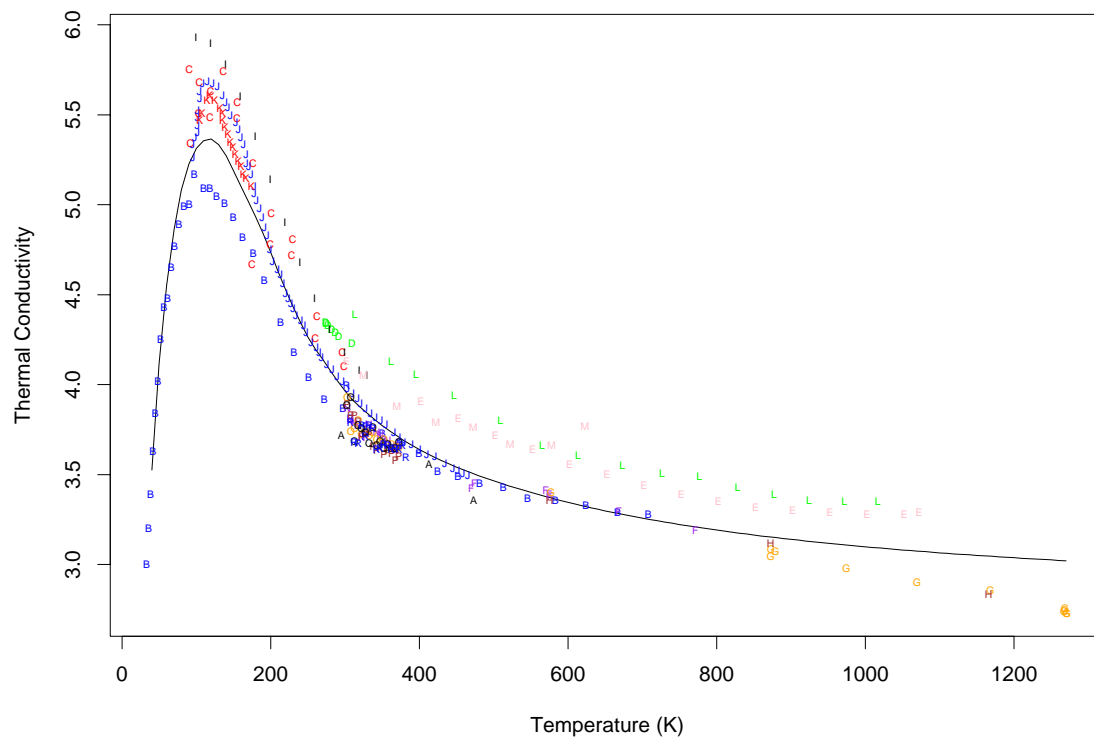


Figure 3.45: The picture shows the data from 18 thermal conductivity data sets; the points from each data set are depicted by different symbols. The solid black line shows an estimated consensus conductivity function.

Corning developed a glass ceramic material, Pyroceram 9606, especially suited for high temperature applications. NIST hopes 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 though, the reference values are given only for a subset of temperatures (e.g., 100K, 200K, 300K, etc.). The picture shows the data from 18 thermal conductivity experiments on this material from around the world; the points from each data set are depicted by different symbols. A way to estimate a “consensus curve” (such as the solid black line) is needed.

The example in the figure highlights some of the problems faced in consensus curve estimation. The data shown from the different labs are of widely differing quality (even though some data sets considered to be from lower quality experiments had already been excluded in a preliminary stage and are thus not included in the picture). 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 an artificial smoothness and arbitrariness of sample size. In addition, many labs provided no or only nominal uncertainty estimates. We need to combine all of these factors to estimate a consensus curve with appropriate uncertainty statements.

Of course, any scientific insight regarding the form of the function should be followed. One can do a weighted nonlinear regression, with the weights taking into account the quality and placement of the data. Spline and locally weighted regression (Loess) methods are very useful for situations in which an explicit function is not required. They are included in many statistical packages and can be used in weighted and unweighted versions. A continuing issue is the attachment of appropriate uncertainties to the estimated functions. The uncertainties provided by the statistical packages are manifestly too small because they treat the collection of data as a single very large data set, resulting in an excessively large sample size. The Residual Sample method assigns uncertainties using an analysis of the samples of residuals from the estimated function.

A simulation study is ongoing to test the properties of various proposed methods of function and uncertainty estimation. In addition, Bayesian methods of analysis will be developed.

There is a definite need for methodological work in estimating consensus curves and their resulting uncertainties from multiple data sets. As just one example, some Reference Materials aim to yield reference values that are temperature-dependent rather than constant.

3.6.4 A Study on the Variance Estimation for a Stationary Process in SPC

Nien Fan Zhang

Statistical Engineering Division, ITL

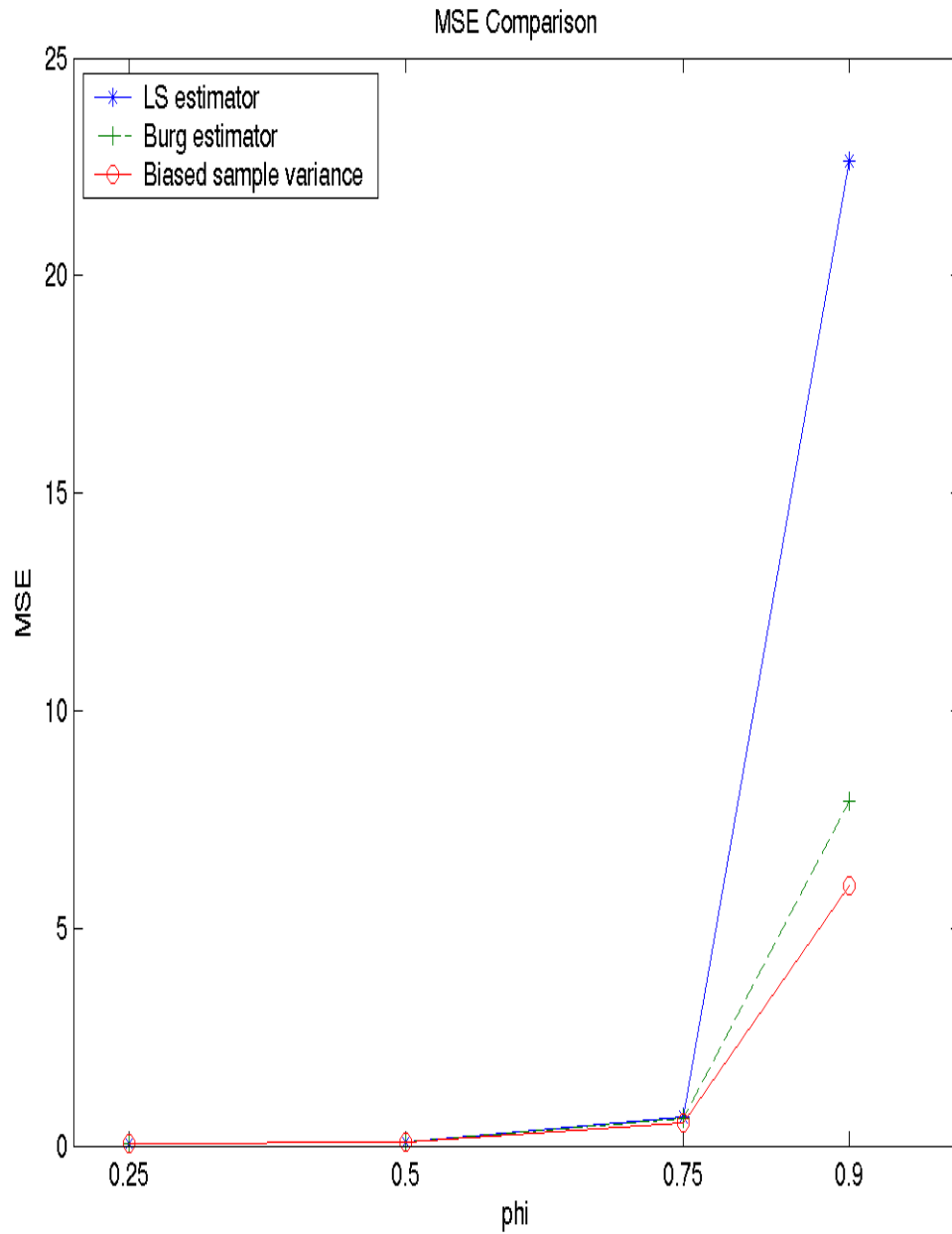


Figure 3.46: This figure shows the MSEs of three variance estimators for AR(1) processes when $n = 60$.

In past years, statistical process control (SPC) methodologies such as process control charts have been widely used in industry for process mean and variability monitoring. When the data are autocorrelated, although the traditional control charts still can be used, their use lacks a solid scientific rationale for ascertaining whether the process is in a state of statistical control, and the charts are often ineffective. A discussion of the impact of autocorrelation on the performance of traditional control charts can be found in Zhang (2000). A direct approach to deal with autocorrelation is to modify the existing SPC charts by adjusting the control limits to accommodate the autocorrelation. Zhang (1998) proposed the EWMAST chart, which is an EWMA chart for stationary processes. Jiang, Tsu, and Woodall (2000) proposed the ARMAST chart as an extension of the EWMAST chart. To construct charts such as the EWMAST chart, the process variance must first be estimated.

For a sequence of independently identically distributed (i.i.d.) random variables, usually the variance is estimated by the sample variance, S^2 . However, when the process data are not from an i.i.d. sequence, different estimators of the process variance might be used. For a weakly stationary process, we can estimate the process variance based on a realization or an ensemble of the process by ergodic theorems (see Priestley (1981), pp. 340-343).

Let x_1, \dots, x_n be a realization from $\{X_t\}$, a stationary process. Like in the case of an i.i.d. sequence, the variance of $\{X_t\}$, σ_x^2 , can be estimated by

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (3.34)$$

with \bar{x} denoting the average of the realization. When the process is i.i.d., it is well known that S^2 is a unbiased estimator of σ_x^2 . However, when a stationary process is not an i.i.d. sequence, S^2 is not unbiased and is only asymptotically unbiased (see Priestley (1981), pp. 321-322). As an alternative, another estimator was suggested

$$S'^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \quad (3.35)$$

The estimator in (3.35) is popular among most time series analysts when the process variance is treated as the autocovariance at lag zero. The comparison of the properties of these two estimators can be found in the discussion on the autocovariance estimators in Priestley (1981), pp. 321-328. To distinguish S^2 from S'^2 , we will stretch the usual terminology and refer to the former as the unbiased estimator and latter as the biased estimator.

However, considering the bias caused by S^2 and S'^2 , some statisticians suggested a model-based approach to estimate the process variance. Many process control engineers also prefer this approach to obtain the estimator of the process variance as a by-product of

process modeling. For example, when $\{X_t\}$ is a stationary AR(3.34) process,

$$X_t - \mu = \phi(X_{t-1} - \mu) + \varepsilon_t \quad (3.36)$$

with μ denoting the process mean and ϕ is the process parameter, with $|\phi| < 1$ and ε_t is white noise with finite variance σ_ε^2 . The process variance σ_x^2 can be expressed as

$$\sigma_x^2 = \frac{\sigma_\varepsilon^2}{1 - \phi^2} \quad (3.37)$$

Given a realization of $\{X_t\}$, ϕ and σ_ε^2 can be estimated by using various algorithms. Then, from (3.37) the corresponding estimates of process variance σ_x^2 can be obtained. We will compare the estimators based on (3.34), (3.35), and (3.37).

Using simulations, we compare the estimators of process variance based on modeling, given in (3.37), with the sample variances consisting of the unbiased estimator and the biased estimator given in (3.34) and (3.35), respectively. We assume that the true process model is an AR(3.34) with known parameters. Two popular AR estimators, the least squares estimator and Burg estimator, are used to obtain the estimates of the process parameter and the white noise variance σ_ε^2 . The least squares and Burg estimators are referred to by Priestley (1981), p. 351 and Brockwell and Davis (1996), pp. 145-146.

In the simulation study, the AR(3.34) parameter $\phi = \pm 0.25, \pm 0.5, \pm 0.75, \pm 0.9$, and the lengths of realizations $n = 60, 100$, and 200 . Without loss of generality, the process mean μ is set to be zero and the white noise variance $\sigma_\varepsilon^2 = 1$.

Specifically, for a given ϕ and a realization length of n (each with 400 warm-ups for an AR(3.34) process), we generated realizations from the AR(3.34) process with process mean $= 0$ and white noise variance $\sigma_\varepsilon^2 = 1$. Then the unbiased and biased estimators of the process variance given in (3.34) and (3.35), respectively, are calculated. From (3.37), the least squares estimator, which is a model-based estimator, is defined as

$$\hat{\sigma}_{x,ls}^2 = \frac{\hat{\sigma}_{\varepsilon,ls}^2}{1 - \hat{\phi}_{ls}^2}$$

with $\hat{\sigma}_{\varepsilon,ls}^2$ and $\hat{\phi}_{ls}$ denoting the estimators of the white noise variance and the process parameter ϕ , respectively, from the least squares estimation algorithm. Similarly, the estimator based on the Burg algorithm is also obtained,

$$\hat{\sigma}_{x,Bg}^2 = \frac{\hat{\sigma}_{\varepsilon,Bg}^2}{1 - \hat{\phi}_{Bg}^2}$$

The algorithms in software MATLAB are used to calculate the least squares and Burg estimates of the process parameters. Minimum *MSE* is used as the criterion to compare

the variance estimators. The MSE of $\hat{\sigma}_x^2$, which is an estimator of the process variance σ_x^2 , is defined as

$$MSE = E[(\hat{\sigma}_x^2 - \sigma_x^2)^2]$$

The MSE values in the study are approximated by the average of the squared errors of the variance estimators. For each of $\phi = \pm 0.25, \pm 0.50, \pm 0.75, \pm 0.90$ and white noise variance $\sigma_\varepsilon^2 = 1$, at least 10000 realizations were generated. In Tables 1 and 2, the approximate mean squared errors for the estimators when $\phi = \pm 0.25, \pm 0.50, \pm 0.75, \pm 0.90$ and $n = 60, 100$, and 200 , are listed. In Tables 1 and 2, for each combination of ϕ and n , the first and second entries are the MSE for the variance estimators based on least squares and Burg algorithms, respectively, while the third and fourth entries are those for the unbiased and the biased sample variances, respectively. Using minimum MSE as the criterion, from Table 1 we conclude that both sample variances are better than the variance estimators based on modeling. In general, the biased sample variance in (3.35) is slightly better than the unbiased sample variance in (3.34). Between the least squares estimator and the one based on the Burg algorithm, the variance estimator based on the Burg algorithm is better. When ϕ is negative, all estimators perform similarly, except the estimator based on the least squares algorithm which performs poorly when $\phi = -0.9$ and $n = 60$. Overall, the biased sample variance performs best.

Table 3 lists the relative increases of MSE when the estimator based on Burg's algorithm is used instead of the biased sample variance. The relative increase of MSE is

$$r = \frac{MSE_{Bg} - MSE_{s'}}{MSE_{s'}},$$

with MSE_{Bg} and $MSE_{s'}$ denoting the MSE 's for the Burg estimator and the biased sample variance, respectively. From Table 3 it is clear that the relative increases of MSE due to using the Burg algorithm are small for all of the cases. In particular, the relative increases when ϕ 's are positive are slightly larger than that when the parameters are negative.

In conclusion, the sample variances, especially the biased sample variance are better estimators than the variance estimators based on the process modeling. This conclusion is made based on the assumption that the true model is known. In practice, the modeling error will cause more of the increase of MSE corresponding to the model-based estimators and make the model-based variance estimators even worse. In addition, a disadvantage of using the model-based estimators is that a practitioner needs to make extra efforts to obtain the variance estimators based on time series modeling.

Table 1: $MSE's$ for the Process Variance Estimators for AR(3.34) Processes with Positive Parameters

n	$\phi = 0.25$	$\phi = 0.50$	$\phi = 0.75$	$\phi = 0.90$
60	0.0435	0.1015	0.6610	22.616
	0.0433	0.1001	0.6209	7.9180
	0.0423	0.0947	0.5317	5.9337
	0.0416	0.0936	0.5319	5.9733
100	0.0265	0.0582	0.3736	7.1384
	0.0264	0.0578	0.3636	5.2805
	0.0260	0.0560	0.3386	4.2094
	0.0257	0.0557	0.3386	4.2170
200	0.0127	0.0302	0.1839	2.6531
	0.0127	0.0300	0.1819	2.5616
	0.0126	0.0297	0.1762	2.3046
	0.0125	0.0297	0.1762	2.3069

Table 2: $MSE's$ for the Process Variance Estimators for AR(3.34) Processes with Negative Parameters

n	$\phi = -0.25$	$\phi = -0.50$	$\phi = -0.75$	$\phi = -0.90$
60	0.0433	0.1033	0.6652	336.163
	0.0431	0.1018	0.6241	8.3497
	0.0435	0.1035	0.6375	8.5049
	0.0421	0.1001	0.6149	8.2154
100	0.0257	0.0602	0.3742	6.4389
	0.0257	0.0597	0.3647	5.1825
	0.0259	0.0601	0.3693	5.2368
	0.0254	0.0589	0.3616	5.1291
200	0.0130	0.0297	0.1875	2.6570
	0.0130	0.0296	0.1856	2.5423
	0.0131	0.0298	0.1868	2.5549
	0.0130	0.0295	0.1848	2.5296

Table 3: Relative MSE Increases due to Using Burg's Algorithm

n	$\phi = 0.25$	$\phi = 0.25$	$\phi = 0.75$	$\phi = 0.90$
60	0.0420	0.0686	0.1674	0.3256
100	0.0270	0.0361	0.0739	0.2522
200	0.0127	0.0131	0.0325	0.1104
	$\phi = -0.25$	$\phi = -0.50$	$\phi = -0.75$	$\phi = -0.90$
60	0.0256	0.0172	0.0163	0.0154
100	0.0096	0.0131	0.0085	0.0104
200	0.0064	0.0044	0.0043	0.0053

3.6.5 SVD-based Structural Approach For Locally Weighted Regression

Z.Q. John Lu

Statistical Engineering Division, ITL

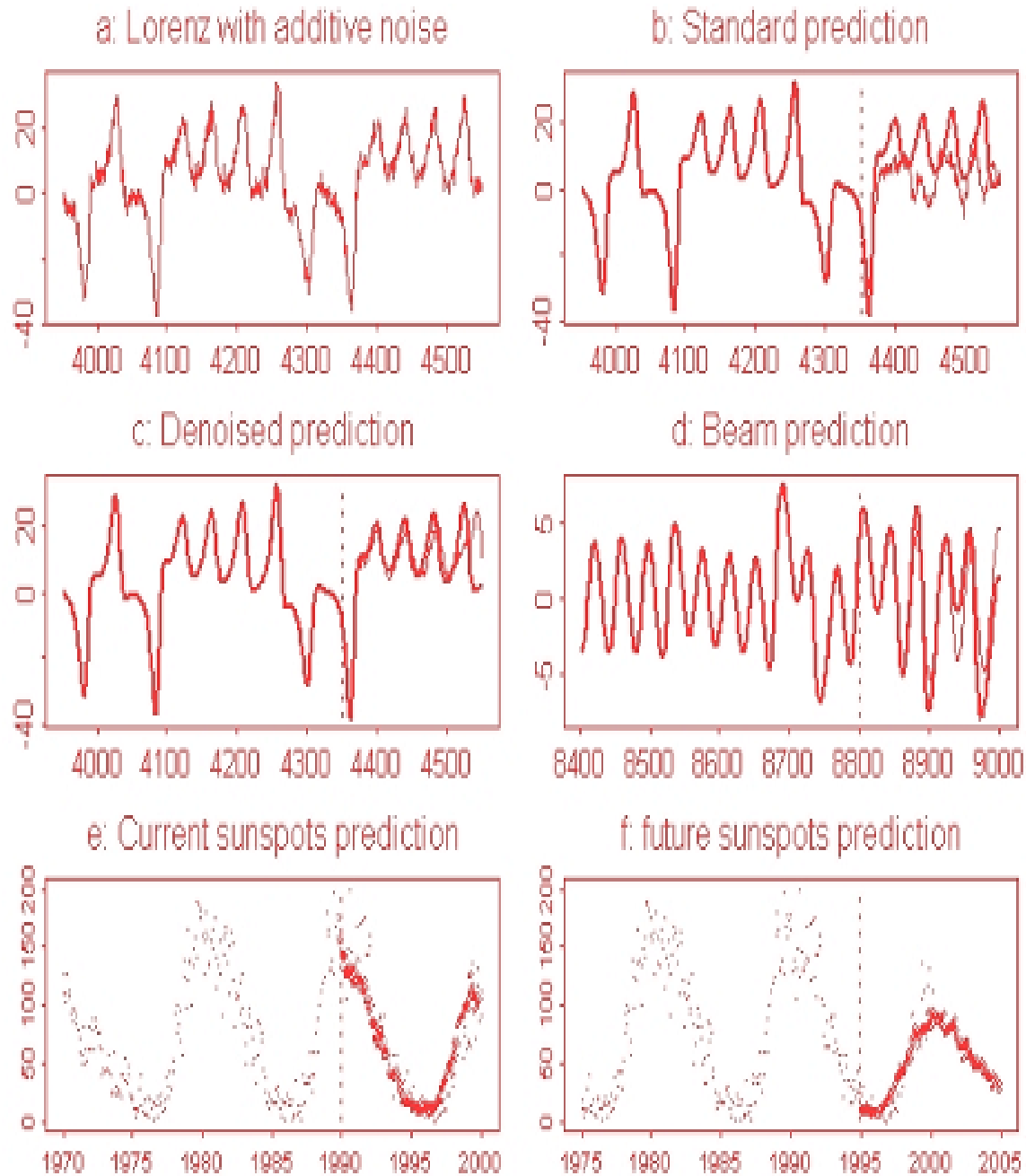


Figure 3.47: (a: data, b: multi-step prediction with standard LWR method, c: local linear method, d: out-of-sample multi-step prediction with SVD-based LWR on a real laboratory data, e,f: multi-year sunspots number prediction).

Though locally weighted regression is among the most popular nonlinear methods for multivariate data, there is a lack of theoretical basis for nonparametric regression in multivariate problems in general. The often quoted “curse of dimensionality” in the statistics literature has not made the earlier efforts in searching for practically useful multivariate nonlinear methods easier. This pessimism seems to go against the adventurous spirit of scientific discovery, especially with the new wave in data mining that various methods are tried as long as they work. In fact, the structure in multivariate data is what separates a “solvable” statistical problem from an “uninteresting” problem. The real-world is full of multivariate problems with plenty of structure. In this report we summarize the work on using the singular value decomposition in elaborating the structure in multivariate data and advocate the SVD-based structural approach for locally weighted regression for situations when there are measurement errors in predictors or there is an underlying lower-dimensional structure in the predictor space. Potential applications to multivariate calibration, nonlinear prediction, and bioinformatics are discussed.

Locally weighted regression (LWR), or local polynomial regression, is perhaps the most widely used nonparametric regression in routine data analysis. For example, it is considered a method of choice for multivariate calibration in chemometrics when data sets exhibit nonlinearity and clustering, and it performs at least as well as neural networks (Chang et al 2001, *Applied Spectroscopy*, Vol. 55, No.9, pp.1199-1206; Center et al 2000, *Applied Spectroscopy*, Vol. 54, No.4, pp.608-623; Despage and Massart 2000, *Anal. Chem.* Vol.72, No.7, pp.1657-1665). However, despite intensive developments in statistical research in the 1990s, theoretical understanding and statistical theory of LWR is still not completely resolved, especially in understanding its success in solving multivariate and apparent high-dimensional problems such as chemometrics. A widely-held misgiving about multivariate nonparametric regression is the “curse of dimensionality”, as it is often thought that any datasets with the number of predictor variables being more than two become extremely difficult. Though this notion comes from a very valid mathematical argument for multivariate data, I argued (Lu 1999, *J. Multi. Anal.* Vol.70, pp.177-201) that it is based on the wrong notion that all multivariate data should have a joint density (absolutely continuous probability measure). Actually, the most interesting multivariate problems are when there is significant “structure” in the data space so that there is an underlying “intrinsic dimension” that is much *lower* than the number of variables. Under a fixed “intrinsic dimension”, or more precisely that defined by the fractal dimension, one can demonstrate that the usual statistical inferences such as classification or prediction can perform reasonably well without the need to worry about the apparent “physical” data space. Furthermore, one can do significantly better by taking advantage of the latent lower-dimensional structure in the data space through use of feature extraction methods such as principal components. We recommend singular value decomposition (SVD)-based methods for such tasks and potential extensions in the context of regression modeling or time series prediction. By performing the SVD operation on the design matrix X , one can then apply LWR on the adaptively selected directions that have the most spread in the potentially nonlinear manifold in the predictor space. These subspace-based neighborhood selection methods apply not only to LWR but also to all kernel/distance-based prediction or classification methods such as kernel estimation or kernel classification. The SVD-based structural approach has the advantages of both de-

noising in the predictor data space and removing irrelevant inputs or redundancy in data. Computationally, SVD-based LWR achieves some kinds of scalability in computation and robustness against model mis-specification and additive data errors.

Some of the most interesting multivariate calibration problems in chemometrics are concerned with the analysis of chemical mixtures. For example, consider the calibration of a near-infrared (NIR) spectrometer, in which the design matrix X denotes the $m \times n$ matrix of spectra, where m is the number of samples in the calibration set and n is the number of wavelength channels, and Y denotes the $m \times 1$ vector of analyte concentrations in the calibration set. Even though it is likely that $n > m$, as long as n and m are greater than the number of independently observable components in the mixtures, the application of SVD to X will yield good calibrations methods using LWR which can perform as well as or better than existing PCR, PLS, and neural networks, among others.

A good testing example is nonlinear prediction of a chaotic time series with measurement errors. The systems theory says that for a wide class of problems defined through simple nonlinear differential equations, the reconstructed state vectors of observed time series lie in a lower-dimensional manifold. State space-based prediction algorithms using local polynomial methods are quite popular, and we have demonstrated in an earlier study (with William Constantine at Insightful Corporation) that the difficult multi-step prediction problem can benefit substantially from SVD denoising and dimension reduction when there are measurement errors and the underlying state space has an intrinsic lower dimension (Figure 3.48).

The SVD-based structural method may also be improved since this dimension reduction or feature extraction approach does not use the information in the response Y and one may be able to use some of the regression variable selections to make more “informed” or “supervised” adaptive choice. Exciting applications are potentially in the emerging bioinformatics arena such as microarray data analysis. High-throughput measurements on the expression levels of thousands of genes are made simultaneously at different experimental conditions, and the goals are to relate this high-dimensional data to known gene or protein functional properties in existing databases, through either clustering or supervised classification. SVD-based nonparametric regression is a promising tool and a viable alternative to some of the more complicated methods such as support vector machines (Brown et al, PNAS Jan.4, 2000, pp.262-267).

In summary, we have developed a new multivariate nonparametric regression method and have discussed some of the newly developed statistical theory of why such methods should work, even when there are many predictor variables. Good methods should always be validated through out-of-sample prediction on real data sets, and we have described some of the successful experiments that have been done and also point out some tantalizing future applications.

3.6.6 A Tutorial Argument for Orthogonal Experiment Designs

James J. Filliben

Statistical Engineering Division, ITL

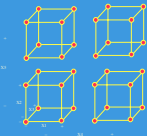
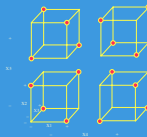
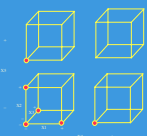
Experiment Design													
Problem: Determine Most Important Factors in a k = 5 Factor Experiment													
Design Name	Design Tableau						Design Geometry			Effect Estimators			
2^5 Full Factorial Design $n = 32$	X1	X2	X3	X4	X5	Y	X1	X2	X3	X4	X5	Y	 $\hat{\beta}_2 = 19.50$ $\hat{\beta}_{34} = 13.25$ $\hat{\beta}_{45} = -11.00$ $\hat{\beta}_4 = 10.75$ $\hat{\beta}_3 = -6.25$ $\hat{\beta}_1 = -1.38$ $\hat{\beta}_5 = -0.63$
	-	-	-	-	-	61	-	-	-	-	-	56	
	-	-	-	-	+	53	-	-	-	-	+	63	
	-	-	-	+	-	63	-	-	-	+	-	70	
	-	-	-	+	+	61	-	-	+	-	-	65	
	-	-	+	-	-	53	-	-	+	-	+	59	
	-	-	+	-	+	56	-	-	+	+	-	55	
	-	-	+	+	-	54	-	-	+	+	+	67	
	-	-	+	+	+	61	-	+	-	-	-	65	
	-	+	-	-	-	69	-	+	-	-	-	44	
	-	+	-	-	+	81	-	+	-	+	-	45	
	-	+	-	+	-	94	-	+	-	+	+	78	
	-	+	-	+	+	93	-	+	+	-	-	77	
	-	+	+	-	-	66	-	+	+	-	+	49	
	-	+	+	-	+	60	-	+	+	+	-	42	
	-	+	+	+	-	95	-	+	+	+	+	81	
-	+	+	+	+	98	-	+	+	+	+	82		
2^{5-1} Fractional Factorial Design $n = 16$	X1	X2	X3	X4	X5	Y	X1	X2	X3	X4	X5	Y	 $\hat{\beta}_2 = 20.50$ $\hat{\beta}_4 = 12.25$ $\hat{\beta}_{24} = 10.75$ $\hat{\beta}_{45} = -9.50$ $\hat{\beta}_5 = -6.25$ $\hat{\beta}_1 = -2.00$ $\hat{\beta}_3 = 0.00$
	-	-	-	-	-	56	-	-	-	-	-	56	
	-	-	-	-	+	53	-	-	-	-	+	53	
	-	-	-	+	-	63	-	-	-	+	-	63	
	-	-	-	+	+	65	-	-	+	-	-	65	
	-	-	+	-	-	53	-	-	+	-	+	53	
	-	-	+	-	+	55	-	-	+	+	-	55	
	-	-	+	+	-	67	-	-	+	+	-	67	
	-	-	+	+	+	61	-	-	+	+	+	61	
	-	+	-	-	-	69	-	+	-	-	-	69	
	-	+	-	-	+	45	-	+	-	-	+	45	
	-	+	-	+	-	78	-	+	-	+	-	78	
	-	+	-	+	+	93	-	+	-	+	+	93	
	-	+	+	-	-	49	-	+	+	-	-	49	
	-	+	+	-	+	80	-	+	+	-	+	80	
	-	+	+	+	-	95	-	+	+	+	-	95	
	2^{5-2} Fractional Factorial Design $n = 8$	X1	X2	X3	X4	X5	Y	X1	X2	X3	X4	X5	
-		-	-	-	-	44	-	-	-	-	-	44	
-		-	-	-	+	53	-	-	-	-	+	53	
-		-	-	+	-	70	-	-	-	+	-	70	
-		-	-	+	+	93	-	-	+	-	-	93	
-		-	+	-	-	66	-	-	+	-	-	66	
-		-	+	-	+	55	-	-	+	+	-	55	
-		-	+	+	-	54	-	-	+	+	-	54	
-		-	+	+	+	82	-	-	+	+	+	82	
-		+	-	-	-	82	-	+	-	-	-	82	
1-Factor- at-a-Time Design $n = 6$	X1	X2	X3	X4	X5	Y	X1	X2	X3	X4	X5	Y	 $\hat{\beta}_2 = -8.00$ $\hat{\beta}_3 = -8.00$ $\hat{\beta}_4 = 8.00$ $\hat{\beta}_5 = -5.00$ $\hat{\beta}_1 = 2.00$
	-	-	-	-	-	61	-	-	-	-	-	61	
	-	-	-	-	+	53	-	-	-	-	+	53	
	-	-	-	+	-	63	-	-	-	+	-	63	
	-	-	+	-	-	53	-	-	+	-	-	53	
	-	-	+	-	+	69	-	-	+	-	+	69	
	-	-	+	+	-	56	-	-	+	+	-	56	
Conclusions: 1-Factor-at-a-Time Designs are Poor. Orthogonal Designs are Excellent.													

Figure 3.48: This is a comparison of 3 orthogonal designs (a 32-run 2**5 full factorial design, a 16-run 2**(5-1) fractional factorial design, and an 8-run 2**(5-2) fractional factorial design versus a 6-run, 5-factor, 1-factor-at-a-time design. One simple comparison criterion is what the 4 designs yield for least squares estimates for the factor 2 effect (= the most important factor in this example). Note from the right-most column that the 3 orthogonal designs yield factor 2 effect estimates that are quite close to the "truth" (19.5) while the all-too-commonly-employed 1-factor-at-a-time design yields a factor 2 estimate of 2.00, which is incorrect by an order of magnitude.

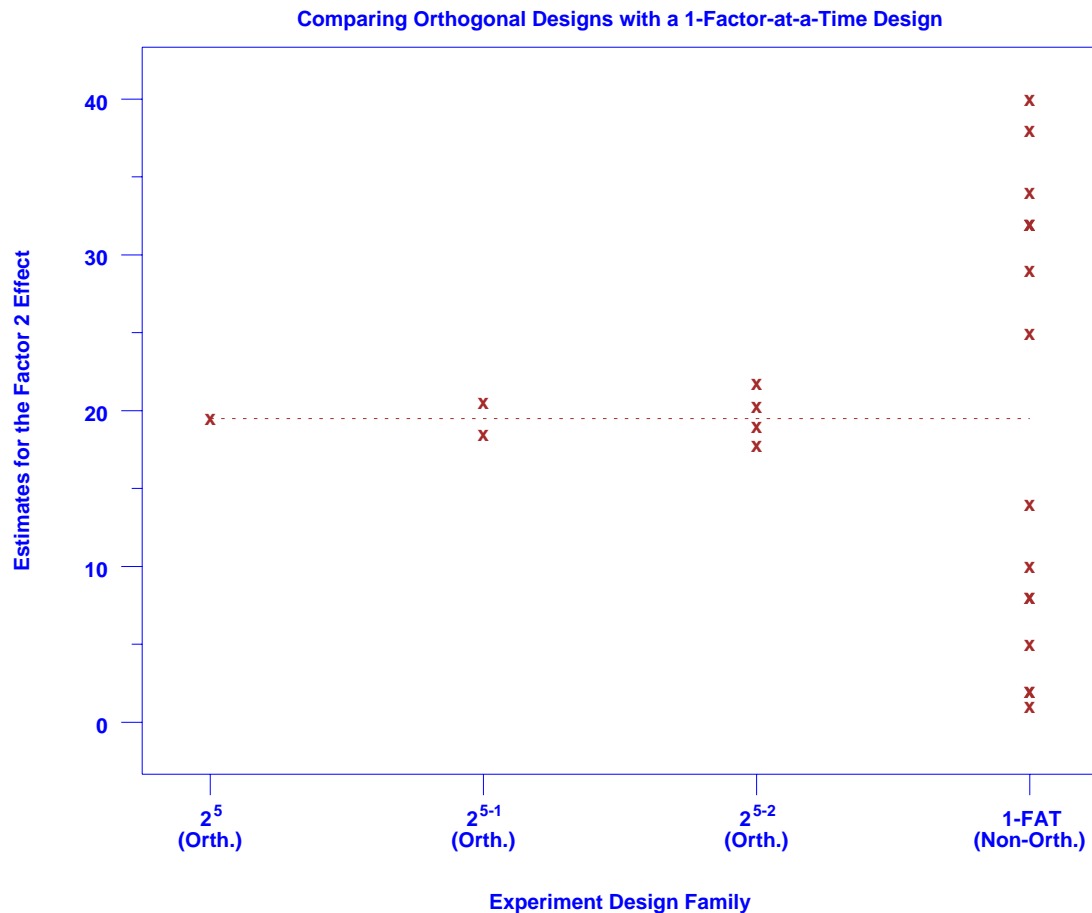


Figure 3.49: This compares factor 2 effect least squares estimates from the 4 design families. Note how the spread of the factor 2 effect estimates for the 2^{5-1} and 2^{5-2} orthogonal fractional factorial designs is relatively tight around the "truth" (19.5), while the 16 1-factor-at-a-time estimates (even allowing for the natural spread of 16 items) are markedly noisy, with very few of the 16 estimates acceptably close to the "truth". In real life, if we had to sample from one 8-run 2^{5-2} effect estimate or one 6-run 1-factor-at-a-time effect estimate, why would the analyst ever choose to sample from the latter's much-too-noisy distribution?

Since the time of Youden in SED, there has been a growing appreciation of the advantages 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 producing effect estimates with maximal precision and minimal bias.

Even with the above advantage, the "average" (NIST and non-NIST) scientist/engineer has still not been convincingly persuaded that orthogonal designs are the automatically preferred choice in practice. One-factor-at-a-time (1-FAT) experimentation still has firm entrenchments in the daily world of research experimentation in science, engineering, and industry. 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.

The present challenge is thus to construct a simple tutorial-level argument in terms that the scientist/engineer can relate to and which will provide compelling proof of the inferiority of 1-factor-at-a-time designs and the superiority of orthogonal designs.

In this regard, many scientists/engineers prefer an argument based on a simple numerical example with a known answer (the "truth"). Further, such scientists frequently prefer to see a specific side-by-side comparison based on criteria that make sense scientifically (like the relative precision and accuracy of the final effect estimates).

To achieve this tutorial end, we "borrowed" a 5-factor, 2-level, 2^{5-1} full factorial example (involving a chemical reaction) from the classic Box, Hunter, and Hunter textbook: "Statistics for Experimenters" (Wiley). Temporarily treating the outcome of this 32-run full factorial design as the "truth", we chose to construct 3 fractional factorial designs:

1. A 16-run 2^{5-1} fractional factorial design (also analyzed by Box, Hunter, and Hunter);
2. An 8-run 2^{5-2} fractional factorial design; and
3. A 6-run 1-FAT (1-factor-at-a-time) design.

The full factorial and the 3 fractional factorials are presented in Figure 3.49; the designs themselves are tabulated in column 2, and a graphical representation is given in column 3. From the graphics, it is clear that each of these 3 fractional designs are (necessarily) subsets of the 2^5 full factorial design. For purposes of developing a concrete, numeric-based comparison, we extract these subsets and carry along the response data from the full factorial. We then compute the usual least squares estimates for the various factor effects, and rank them by magnitude (column 4 of Figure 3.49).

By comparing the various effect estimates in this right-most column, it is seen that the 2 orthogonal fractional factorial designs are quite good in producing estimates close to the full factorial (the "truth"). (Note that the fractional factorial estimates can never be

expected to be identical to the "truth" since only part of the full data set is actually being used in the estimation process.) On the other hand, the bottom design (the 1-factor-at-a-time design) yields factor effect estimates that are grossly poor. Note in particular, how a simple comparison of the factor 2 effect estimate (factor 2 being the most important factor in this specific example with a 2^{**5} "true" value of 19.50) yields a 20.50 for the 16-run $2^{**}(5-1)$ experiment, a 20.25 for the 8-run $2^{**}(5-2)$ experiment, but only a 2.00 for the 6-run 1-FAT experiment.

The net conclusion is that the 3 orthogonal designs (the intrinsically orthogonal 2^{**5} full factorial, the $2^{**}(5-1)$ fractional factorial, and the $2^{**}(5-2)$ fractional factorial) all yield factor 2 effect estimates which are near-equivalent (approximately 20), while the non-orthogonal 1-FAT design yields a factor 2 effect estimate (= 2.00) that is markedly wrong by an order of magnitude.

One may argue that the above results are a matter of chance, that the quality of the factor 2 effect estimate is simply a function of the specific fractional design that was chosen, and could change drastically if other 16-run $2^{**}(5-1)$ designs, other 8-run $2^{**}(5-2)$ designs, and other 6-run 1-FAT designs were chosen. This is a valid point inasmuch as many other such fractional designs do in fact exist.

Limiting ourselves to fractional designs with similar constructs, we note that the 16-run $2^{**}(5-1)$ design (in which the factor 5 settings were generated via $X_5 = X_1 * X_2 * X_3 * X_4$) has an orthogonal, complementary design (in which $X_5 = -X_1 * X_2 * X_3 * X_4$) which could have been used—thus there are a total of 2 such $2^{**}(5-1)$ designs that are possible. Similarly, the 8-run $2^{**}(5-2)$ design (in which the factor 4 and 5 settings were generated via $X_4 = X_1 * X_2$ and $X_5 = X_1 * X_3$) has 3 other "fraternal" designs that are also orthogonal: from $X_4 = -X_1 * X_2$ and $X_5 = X_1 * X_3$, from $X_4 = X_1 * X_2$ and $X_5 = -X_1 * X_3$, and from $X_4 = -X_1 * X_2$ and $X_5 = -X_1 * X_3$.

Finally, for the 6-run 1-FAT design, a multitude of alternate designs are possible and they (unfortunately) change depending on what factor effect is being focused on and estimated (such focusing information is not usually known a priori). For example, for estimating the factor 2 effect only, the given 6-run 1-FAT design in Figure 3.49 utilizes only 2 out of the 6 runs to form that factor 2 estimate:

X1	X2	X3	X4	X5	Y
-	-	-	-	-	61
-	+	-	-	-	63

which in turn yielded the reported (column 4, row 4 of Figure 3.49) factor 2 effect estimate of $63 - 61 = 2$). Thus for estimating the factor 2 effect only, there is not 1 but 16 different 1-FAT 2-run subsets that could have been utilized: 2 possibilities (- and +) for factor 1 x 2 possibilities for factor 3 x 2 possibilities for factor 4 x 2 possibilities for factor 5). The net result is that each one of these 16 designs would possibly yield a different factor 2 effect estimate—depending on what the "background" settings are for factors 1, 3, 4, and 5. For example, for the factor 2 effect estimate (= 2.00) reported in the bottom right cell of Figure 3.49, that estimate was valid only for the background settings of $X_1 = -$, $X_3 = -$, $X_4 = -$, and $X_5 = -$.

Figure 3.50 graphically summarizes and compares the factor 2 effect estimates (vertical axis) from the 4 families of designs (horizontal axis). Proceeding left to right in Figure 3.50, we have:

1. the "true" (factor 2) value (= 19.50) from the 32-run 2^{5-0} full factorial;
2. the 2 estimates (20.5 and 18.5) from the 16-run 2^{5-1} fractional factorial;
3. the 4 estimates (20.25, 21.75, 17.00, and 19.00) from the 8-run 2^{5-2} fractional factorial;
4. the 16 estimates (2, 8, 1, 5, 25, 32, 29, 38, 14, 2, 8, 10, 34, 32, 32, and 40) from the 16 1-FAT designs.

Note how the 2 orthogonal fractional factorial designs both yield factor 2 effect estimates tightly clustered about the 2^5 "truth" (= 19.50). Note also that ALL (2 + 4 = 6) of the orthogonal fractional factorial design estimates are closer to the truth, than ANY of the 1-FAT's 16 estimates. In a related vein, note also that the non-orthogonal 1-FAT factor 2 effect estimates are wildly noisy, with an excess variation that is well beyond the additional natural variation that one would expect by plotting 16 items as opposed to 4 items or 2 items. Rather this excess variation is due to the experiment design fact-of-life:

structurally-poor (non-orthogonal) experiment designs yield statistically-poor (= noisy and biased) effect estimates, while structurally-sound (= orthogonal) experiment designs yield statistically-good (= low variability and low bias) effect estimates.

Specifically comparing the 8-run 2^{5-2} results and the 6-run 1-FAT results, one sees how the similarity in sample sizes (8 and 6) pales in comparison to the dissimilarity in the quality of the factor 2 effect estimates. Given that there is only a 2 run difference between these designs, why would any scientist opt to use the 1-FAT over the near-uniformly better 2^{5-2} orthogonal design? The proof is in the statistics, and this tutorial-level presentation hopefully illustrates how the statistics (the quality of the effect estimates) are uniformly in the scientist's favor if orthogonal designs are chosen over 1-FAT designs.

Parts of this tutorial exposition were used for a poster in the SED Open House, and were well-received by visiting NIST scientists/engineers. A larger portion of this tutorial development was used in a presentation on constructing appropriate designs in connection with the NIST World Trade Center investigation, with the net result that orthogonal designs will have a place in this study, either at a sub-assembly level, or for the full-assembly, or both. The full portion of this tutorial will serve as a basis for a NIST-wide seminar that will be presented as part of the SED Tutorial Lecture Series later in the year 2003. The acceptance by the NIST scientist/engineer of such an argument for orthogonal designs will ultimately yield better effect estimates, more insight into underlying physical mechanisms, and more efficient (less time and money) experimental effort. In the spirit of Youden, this is a goal worth striving for and hopefully this tutorial argument will make that goal attainable for the NIST scientist/engineer.

3.7 Web Products

3.7.1 NIST/SEMATECH e-Handbook of Statistical Methods

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

Barry Hembree, Jack Prins, Pat Spagon, Paul Tobias, Chelli Zey
International SEMATECH and International SEMATECH Member Companies

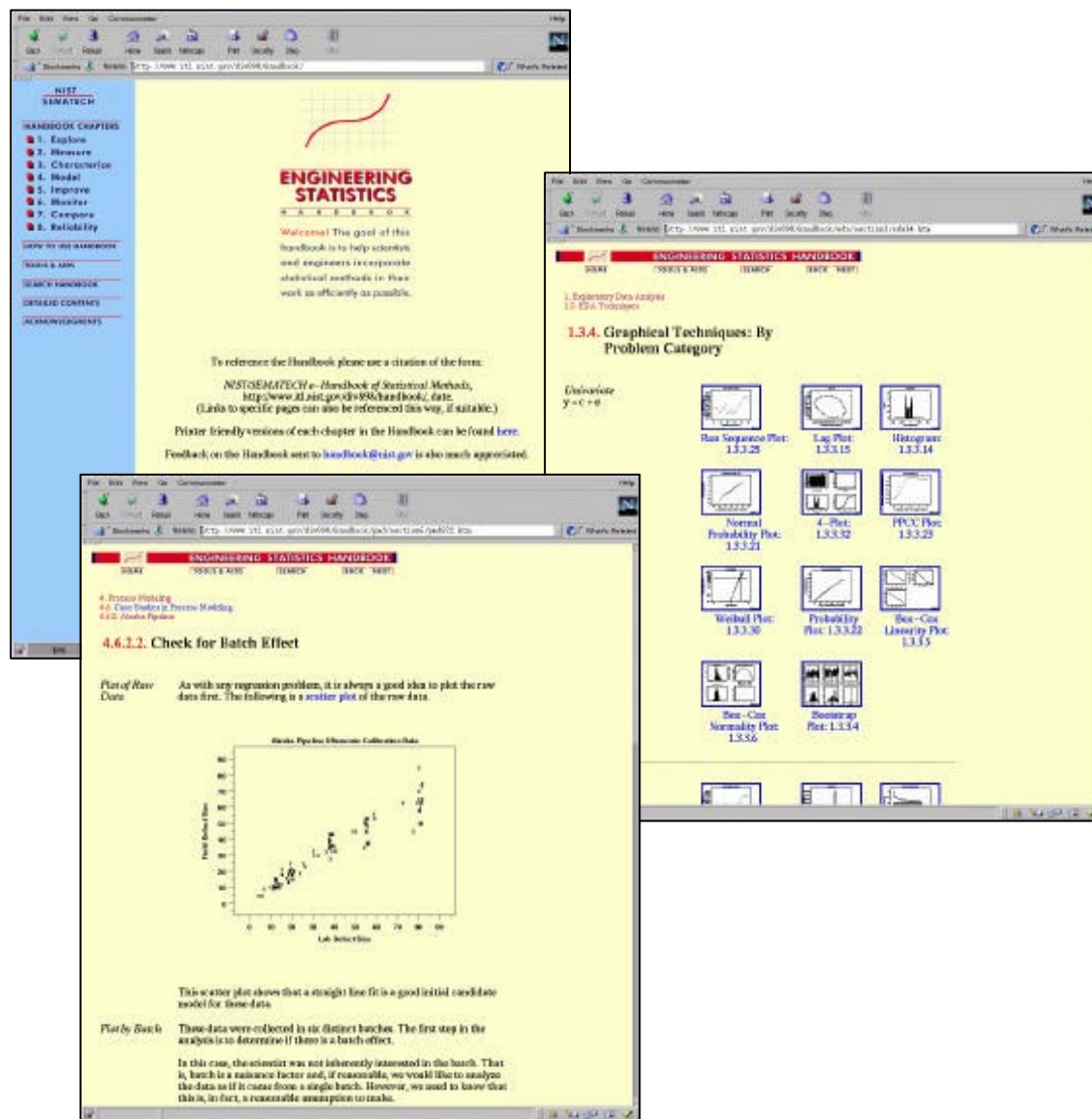


Figure 3.50: Pages from the e-Handbook. The cover, upper left, introduces readers to the site and includes frequently requested information on how to reference the e-Handbook and about printer-friendly files for each chapter. One of the outline pages, right, shows a selection of graphical techniques organized by problem category. A page from a case study, bottom, illustrates step-by-step guidance through an entire problem.

The NIST/SEMATECH e-Handbook of Statistical Methods is a Web-based book written to help scientists and engineers incorporate statistical methods into their work as efficiently as possible. Ideally it will serve as a reference which will help scientists and engineers design their own experiments and carry out the appropriate analyses when a statistician is not available to help. It is also hoped that it will serve as a useful educational tool that will help users of statistical methods and consumers of statistical information better understand statistical procedures and their underlying assumptions, and more clearly interpret scientific and engineering results stated in statistical terms.

The project began with a request from SEMATECH, a consortium of major U.S. semiconductor manufacturers, to update the National Bureau of Standards (NBS) Handbook 91, Experimental Statistics. Handbook 91, written by Mary Natrella of the NBS Statistical Engineering Lab, was a best-selling NBS publication for many years. Engineers and scientists in a variety of fields appreciated it because of its problem-oriented approach to statistics and its detailed examples. The examples of each statistical procedure recommended in the text were also accompanied by fill-in-the-blank worksheets, allowing the reader to quickly and easily repeat the calculations with his or her own data. By the 1990's, however, the emphasis on hand calculations was too dated to be practical and many modern statistical methods were missing from the text, prompting SEMATECH's interest in updating Handbook 91 for the use of their member companies.

As a result of the SEMATECH interest, a joint NIST/SEMATECH project team was assembled to explore the idea, develop a formal project proposal and to carry the project out. With the rapid growth of the Internet when the project proposal was under development, the project quickly evolved from the publication of a new edition of a traditional book to development of an online handbook for distribution via the World Wide Web. The advantages of Web-distribution included easy access by users all over the world, the ability to integrate the software necessary to use the different statistical methods right into the text, and the opportunity to create an easily expandable resource.

The development of the e-Handbook's new format and content was carried out using a top-down approach. The team first laid out the scope of the new handbook and a detailed outline of its content. The outline was designed to lead the user hierarchically from the general topics covered to the specific information needed, avoiding statistical jargon as much as possible. The eight chapters in the top level of the outline include:

1. Exploratory Data Analysis
2. Measurement Process Characterization
3. Production Process Characterization
4. Process Modeling
5. Process/Product Improvement
6. Process/Product Monitoring
7. Process/Product Comparisons
8. Assessing Product Reliability

In addition to the main outline, several other methods of accessing the text were laid out to try to make the information in the e-Handbook as accessible as possible for users

unfamiliar with the traditional organization of information in the statistical literature. Some of these alternative access methods include engineering questions linked to flow charts showing the steps necessary to complete a statistical analysis appropriate to the question, along with indexes of examples and search capabilities.

Since another major goal of the new Handbook was to maintain a practical, problem-oriented approach to statistics, common structures such as a section of detailed case studies using real data from the semiconductor industry and the NIST laboratories were included in each chapter. Standard page formats for each type of page in the Handbook were also carefully developed to improve readability and to make navigation transparent.

Finally, after completing the high-level layout of the entire book, individual team members were assigned for each chapter to fill in the framework developed by the team. Of course, developing a stylistically coherent technical publication with multiple authors, while efficient in some ways, is quite a challenge in others. Fortunately the team found an appropriate editor in Tom Ryan, who diligently read and marked-up the entire text to help ensure that all the chapters of the e-Handbook read with a (reasonably) common voice. Readers of the beta version of the e-Handbook, released about two years ago, also provided many useful comments and corrections.

The approach taken toward integrating statistical software with the Handbook was more bottom-up than that used for updating the Handbook itself. The project team realized from past experience in teaching and consulting that different users like different software. Persuading a user to switch from one package to another generally requires compelling reasons since it costs the user not only money, but more importantly, time. The team also recognized that writing new statistical software that would be universally available across platforms, software written in Java, for example, would add greatly to an already ambitious project. As a result, the vision for software integration focused on development of an open system. Under this model, the project team integrated one statistical package with the Handbook, and established a framework that other software providers can use to integrate their software as well.

The software chosen for integration with the Handbook was Dataplot, a free, downloadable statistical package maintained by the NIST Statistical Engineering Division. One of the primary reasons that Dataplot was chosen as the prototype software was the ability of the Handbook team to easily make any changes in the source code needed to improve integration. Another important consideration was its wide accessibility, allowing almost any user to take full advantage of the Handbook.

Use of Dataplot in this role did require development of a new graphical user interface (GUI) for both Unix and Windows, however. Since Dataplot is now menu-driven it was essential to make its usage transparent so the new users can focus on their own applications right away. The team also produced Dataplot macros to carry out the various analyses used in the case studies.

In addition to facing the same general challenges that were faced in updating the e-Handbook itself, the other main challenges of the software integration proved to be choosing the appropriate tools and methods for software integration that would allow the software to behave identically across operating systems. A variety of approaches had

to be evaluated before hitting on a solution that met all the necessary criteria.

With the release of the first complete version of the e-Handbook this year, the project team has received many gratifying signs of success: positive email feedback from many users of the Handbook and favorable comments from industrial statisticians who report referring their clients to the e-Handbook and teaching short-courses with it. Within a year of its release, the e-Handbook site ranks number 1 and 2 according Google for searches looking for "Engineering Statistics". Some of these users have been from targeted populations (e.g., International SEMATECH member companies), while many others are from other companies and other organizations, within the U.S. and internationally, with whom we have had no previous contacts.

In addition, some vendors of statistical software are developing their own web sites related to the e-Handbook that will allow users to analyze examples and case studies from the Handbook using popular proprietary software. This expands the customer base for the Handbook because many companies standardize on mainstream commercial statistical software packages for in-house use. This also benefits individuals who are already familiar with or who prefer a particular new statistical software package that is being integrated with the e-Handbook.

The e-Handbook will be released on CD in early 2003 so people can use it off-line or create local installations as an alternative to web access. Production of a CD will also ensure that e-Handbook will remain available as a library resource in case the live version cannot be continuously maintained on the web for some reason in the future.

3.7.2 HELP for Missing Data

Hung-kung Liu
Statistical Engineering Division, ITL

J.T. Gene Hwang
Cornell University

Gerard N. Stenbakken, Michael T. Souders
Electricity Division, EEEL

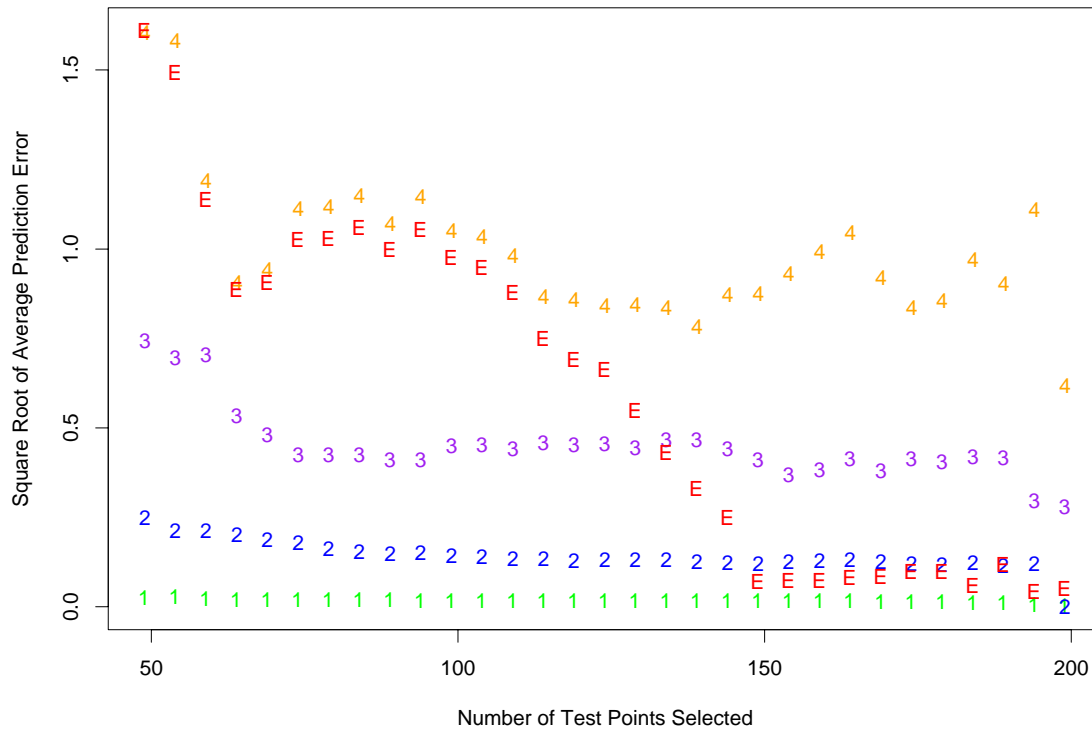


Figure 3.51: The square root of the average prediction error versus the number of test points selected (t_p) are plotted for years 1, 2, 3, and 4, and for “E”, the EM estimates for year 4. We studied the identifiability problem for different t_p . On the lower bound for t_p implying identifiability, our analytical result is 149. The graph reflects this in the fact that the size of the norm of the predicted error of the E’s behaves like the 4’s for small t_p , and comes down to a reasonable small size around $t_p = 149$.

As part of the development of efficient new testing strategies for software-embedded systems, we have collaborated on tests of an extension of the HELP algorithm to devices following a model outside the usual non-software-embedded framework.

Many engineering problems involve high-dimensional observations with mean vectors sitting in a lower dimensional subspace. Exhaustive measurement of all the elements of an observation is often time consuming and expensive. Applying a traditional multivariate linear model, one can combine a small subset of the elements of the observation with a known design matrix to predict the rest of the elements. However, for a complicated engineering system, the design matrix is often hard to fully determine. We investigate an empirical linear model, in which we allow ourselves to use the data to determine the size of the design matrix and to estimate the unknown part of the design matrix. This estimated model is then used to construct point and interval estimates for future observations. This technique is called HELP (High-dimensional Empirical Linear Prediction).

We have extended the HELP algorithm to devices following a model outside the usual non-software-embedded framework for the development of efficient new testing strategies for software-embedded systems. The new device model, while simpler than the models ultimately applicable to software-embedded systems, provides a readily-available starting point for testing an extension of the HELP methodology using the concept of Expectation Maximization (EM) which has potential importance for software-embedded systems. The EM approach is attractive because it would provide an efficient method for extending HELP, or other testing tools, to the more complex device model.

We studied patterns of the observed data to see if they can identify the model being considered. In the plot, the square root of the average prediction error versus the number of test points selected (t_p) are plotted for years 1, 2, 3, and 4, and for “E”, the EM estimates for year 4. We studied the identifiability problem for different t_p . On the lower bound for t_p implying identifiability, our analytical result is 149. The graph reflects this in the fact that the size of the norm of the predicted error of the E’s behaves like the 4’s for small t_p , and comes down to a reasonable small size around $t_p = 149$.

Our results not only help to resolve whether EM works for this situation, but also help engineers properly design their experiment so that the principal components can still be identified when some of the data are missing.

3.7.3 Development of a Web Application for Statistical Analysis

Juan Soto, Alan Heckert, Zuriel Correa
Statistical Engineering Division, NIST

The screenshot shows a web browser window titled "Statistical Uncertainty Analysis of Key Comparisons Input Form - Microsoft Internet Explorer provided by NIST". The address bar shows the URL: <http://p612352/WBC/Clients/AnalysisRelated/key-comparison-form.asp>. The main content area displays a form titled "SUAKC Input Parameters" with the following fields and values:

SUAKC Input Parameters	
Number of NIMs:	14
Number of Artifacts:	3
Non-Pilot Lab Measurement Periods:	1996.81 1997.18 1997.35
Pilot Lab Measurement Periods:	1996.65 1996.92 1997.62
Non-Pilot Lab Type A Uncertainty:	1.88 0.50 0.52 2.3 0.07 0.1
Non-Pilot Lab Type B Uncertainty:	2.29 0.35 0.606 2.19 2.56
Pilot Lab Type A Uncertainty:	0.2
Pilot Lab Type B Uncertainty:	1.51
Regression Coefficient (slope):	30.334 5.992 21.161
Regression Coefficient (intercept):	1.739 1.06 4.529
Residual Standard Deviation (sigma):	1.876 1.066 3.4
Non-Pilot Lab Measurements Filename:	<input type="text"/> Browse...
Data Input Format:	ASCII
<input type="button" value="Execute Analytic >>"/>	

Figure 3.52: Login Page for Web Based Computing

In an effort to maximize both computing and human resources, the Statistical Engineering Division (SED) initiated the development of a web application for statistical analysis in the summer of 2002. A web application dedicated to statistical analysis would provide users with a common interface [web forms] to various analytics built on disparate engines (e.g., Dataplot, MATLAB, S+, and Fortran 77) running on a single web server. Such a system would be advantageous in that it would: (a) eliminate the need for users to learn specialized software products, (b) enable SED staff to increase focus on projects with novel statistical problems, (c) provide a mechanism for the transfer of statistical methodology, and (d) raise statistical competency of NIST scientists.

Zuriel Correa, an Industrial Engineering graduate student from the University of Puerto Rico (Mayagüez), constructed a prototype while visiting the SED this past summer. Since then we have continued to evaluate and extend the system built on multiple technologies, which include: Internet Information Services (IIS) 5, Structure Query Language (SQL) Server 2000, Windows Scripting Host (WSH), Active Server Pages (ASP), VBScript, JavaScript, Hypertext Markup Language (HTML), DATAPLOT, MATLAB, S+, StatServer, and Fortran. To date, the following two analytics have been incorporated into the application.

- consensus means analysis and
- statistical uncertainty analysis for key comparisons

The SED envisions this web application will enable NIST scientists and engineers to gain access to a consolidated array of solutions (analytics) to routine problems in data analysis such as univariate location estimation, consensus means analysis, linear calibration, errors-in-variables regression, and experiment design. Application areas that would benefit from such a system include: Standard Reference Materials (SRMs), Inter-laboratory Studies, Key Comparisons, and Bayesian Metrology.

3.7.4 Development of a Bayesian Software Library

Don Malec, Juan Soto

Statistical Engineering Division, ITL

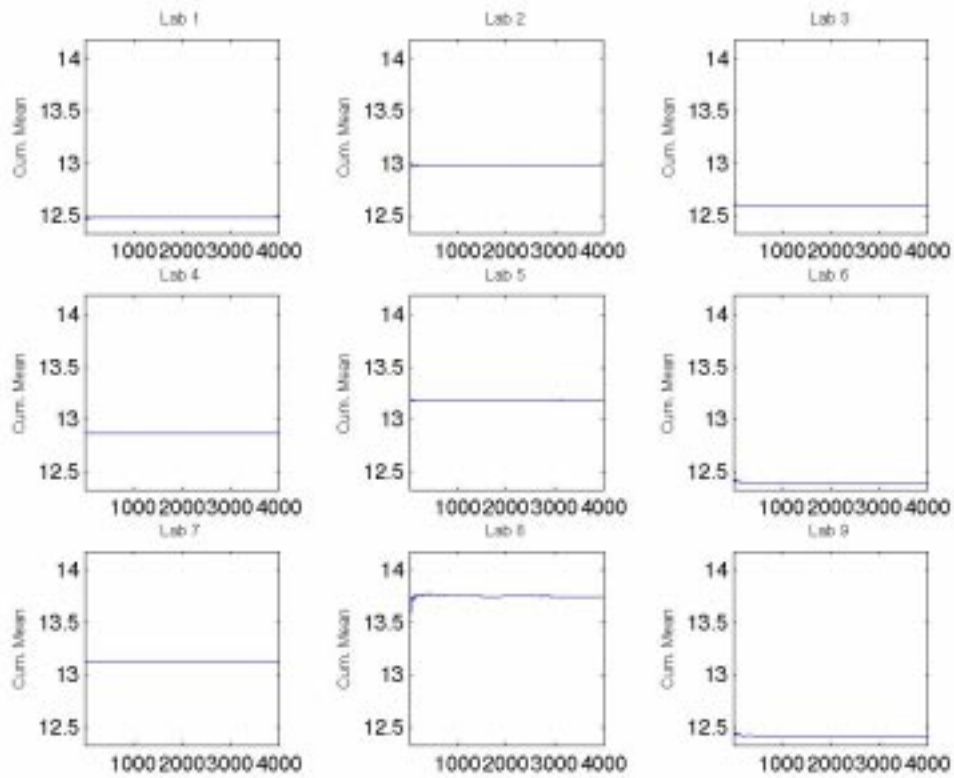


Figure 3.53: Convergence diagnostic for individual Markov chains for each laboratory

As part of a continuing effort to disseminate Bayesian methodology to a wider audience, the Statistical Engineering Division (SED) is developing a software library for the Bayesian analysis of statistical models useful in metrology.

The Bayesian software library will consist of *stand-alone routines and documentation* implemented in Fortran 77 and MATLAB. Thus far SED has implemented a normal hierarchical model in both Fortran 77 and MATLAB. In the future, additional models will be implemented. Potentially, BUGS (Bayesian inference Using Gibbs Sampling) scripts may also be incorporated into the library.

The library will be made publicly available to both NIST and non-NIST staff for use in their own applications via the SED web page.

3.7.5 Web Database Project

Alan Heckert, Nell Sedransk, Ray Miller, David Martin
Statistical Engineering Division, ITL

Y.C. Chang, Jimmy Graham
Information Services and Computing Division

Netscape: SED Web Data Base: Project

File Edit View Go Communicator Help

Bookmarks Location: http://www.cam.nist.gov/~heckert/sed_internal/web_databases/sec/ What's Related

Back Forward Reload Home Search Netscape Print Security Shop Stop

SED Web Database: Project

Introduction This form adds an entry to the SED projects database.

The projects database is used to store any SED related projects. This includes the following:

- Yellowbook articles

The yellowbook article is intended to be rather brief (typically one or two pages). You can include a link to a more extensive write-up of the project. In addition, you may also provide a link to an external image (the image should typically be in GIF or JPEG format).

SED staff members should be entered as SED staff members; non-SED NIST staff should be entered as internal collaborators, and non-NIST staff should be entered as external collaborators. Any non-SED collaborators must first be added to the [person database](#).

Links in the field names provide additional clarification on the information that is being requested for that field.

(Number of SED Staff: Internal Collaborators: External Collaborators: [Get new form](#))

Name of the Project:

SED Staff:

Description:

Optional Field

Name of File: [Browse...](#)

Image File: [Browse...](#)

Image

Figure 3.54: Sample Form for the Web Database

Building Web pages on top of database technology can make Web pages more flexible, dynamic, and easier to maintain. For example, SED staff could update publication lists by filling out a Web form. Persons viewing the SED publications lists would have more flexible searching options.

The first stage was to convert the SRM web pages to a database driven system. The previous web pages were built using static HTML code and were primarily intended to monitor the status of the SRM's. Although this worked reasonably well, it did require some manual editing. The recent change to individual funding of SRM's (as opposed to one large SRM pool) required substantial changes in the information needed on the web pages. In particular, many budget related fields needed to be added. It is also important to have flexible search and extraction capabilities related to these fields. For this reason, SED decided to replace the HTML based pages with a database driven system. This work was performed by David Martin, an SED summer intern. David first converted the Excel file with the current SRM information to an ACCESS database. David then developed a series of web pages using the Cold Fusion (a commercial software program that simplifies managing databases from the web) software program. David developed web pages for inputting and editing the information for an SRM, viewing the SRM's in the database, and for performing various kinds of searches of the SRM's in the database.

In the second stage, 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.

The basic design for the databases is now complete. The system built around the following databases:

- A persons database. In addition to SED staff, this includes any internal and external collaborators.
- A publications database.
- An events database.
- A projects database.
- An activities database.

SED worked with division 896 to define the information that is required for each of these databases. Web input forms have been designed for each of these databases. Ray Miller, a summer intern, assisted in the design of the web forms.

The persons and projects databases are now live for input.

Tasks that remain for the web database project include:

- Activate the remaining three databases.

- Develop "edit" forms for each of the databases.
- Develop "view" and "search" forms for the various databases.
- Develop scripts that will automatically update the static SED web pages on some routine basis. For example, the publication pages could be updated each night.

In summary, making the SED web pages database driven will make extraction of the information contained in them easier and more flexible. In addition, maintenance of the content of the web pages will be easier in that SED staff will be able to enter the relevant information a single time using web forms. Information on the SED web pages will also be more readily accessible to customers outside NIST who now will be able easily to locate linked activities, publications, statistical methodology and data sets.

4. Special Programs

4.1 International Activities

4.1.1 International Organization for Standardization (ISO)

Nien Fan Zhang
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-Simple Replication and Nested Experiments" 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 vote stage in May, 2003.

4.1.2 Hands-On Workshop for SIM Members

William F. Guthrie, A. Ivelisse Avlies, Dennis D. Leber, Hung-kung Liu, and James H. Yen

Statistical Engineering Division, ITL

This year SED staff were invited by the Division Chief of the Analytical Chemistry Division to offer a half-day hands-on workshop on uncertainty estimation as part of an Inter-American Metrology System (SIM) Forum on Quality Systems Implementation held at NIST. Ten scientists from different National Metrology Institutes (NMI's) in the Americas participated. During the workshop NIST staff compared different methods and issues affecting uncertainty computations and shared NIST's views on the correct assessment of uncertainty. After presenting an overview of the uncertainty methods which gave the statistical rationale behind the *ISO Guide to Expression of Uncertainty in Measurement*, which most laboratories use as the basis for uncertainty estimation, the participants and instructors split into subgroups of two scientists and one statistician each and worked together on two uncertainty examples on laptop computers. After going through the two examples as a team, the students worked on a third, similar problem without the statisticians. This hands-on approach, which allowed students to try the methods they learned about before leaving the classroom, was very well received. Opportunities like this to work with our colleagues from other countries 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.

4.1.3 SED Visits to AIST of Japan

Hung-kung Liu, Nien Fan Zhang

Statistical Engineering Division, ITL

In March of 2002, Hung-kung Liu was sponsored by the Advanced Industrial Science & Technology (AIST) of Japan to visit the National Metrology Institute of Japan (NMIJ) for two weeks. The main purpose of this visit was to conduct collaborative research with Dr. Hiroshi Sato of NMIJ on uncertainty analysis for quantized observations, and with Dr. Kensei Ehara, Chief of the Metrological Statistics and Particle Measurement Section (MSPMS) of NMIJ, on constructing background corrected statistical intervals for a Poisson random variable. He also gave a talk on "Uncertainty Analysis with Case Studies", and visited many labs in MSPMS.

In 2002, Nien Fan Zhang was invited to visit MSPMS of AIST in Tsukuba, Japan. During his visit to AIST on July 15 and 16, Nien Fan Zhang gave a seminar titled "Statistical Analysis and Key Comparisons." Many scientists from AIST attended the seminar and joined the follow-up discussion. During his visit, Nien Fan Zhang also discussed some statistical problems with Dr. Kensi Ehara, the Chief of MSPMS, and his staff. The collaboration is continuing.

4.1.4 Clinical Biochemistry

Nien Fan Zhang

Statistical Engineering Division, ITL

During October 15-18, 2002, Dr. Per Winkel, Director of the Department of Clinical Biochemistry at the Central Hospital in Nykobing Falster in Copenhagen, Denmark visited SED. In 2001, Dr. Winkel, a clinical biochemist, contacted and consulted Nien Fan Zhang on analyzing a time series of clinical biochemical quality data and collaboration between them subsequently developed. During his visit, Dr. Winkel and Nien Fan Zhang completed a joint paper, "Serial Correlation of QC-Data-on the Use of Proper Charts." The paper is to be submitted to the *Journal of Clinical Chemistry*.

4.1.5 American Society of Mechanical Engineers

Hari Iyer

Statistical Engineering Division, ITL

During 2002, the ASME PTC (Performance Test Code) 19.1 committee met twice – once in New Orleans and once in Panama City – to finalize the revisions to the product test code. This committee has written a supplement to the ASME performance test codes titled "Test Uncertainty". The next ASME meeting is scheduled for March 4 and 5, 2003, and will be held in Scottsdale, Arizona. The ASME PTC 19.1 committee is on the verge of submitting a revised product test code for review. After the review process is completed, the new version will replace the current document.

4.1.6 CIPM/CCM Working Group for Fluid Flow

James J. Filliben, Will Guthrie, Ivelisse Aviles

Statistical Engineering Division, ITL

George Mattingly

Process Measurements Division, CSTL

On April 11, 2002, SED staff members participated in the Third Meeting of the CIPM/CCM Working Group for Fluid Flow (WGFF) in Arlington, VA. George Mattingly of NIST (CSTL) chaired the meeting attended by working group members and pilot lab representatives from around the world for all six of the WGFF areas:

1. Water flow;
2. Hydrocarbon liquid flow;
3. Air speed;
4. Liquid volume;

5. High pressure gas flow; and
6. Low pressure gas flow.

The purpose of the meeting was to coordinate the implementation of Key Comparisons for each of the six areas, and to present a statistically unified approach that would serve as the framework for all the fluid flow intercomparisons.

To that end, with SED's Will Guthrie and Ivelisse Aviles in support, Jim Filliben presented "Statistical Principles and Techniques for the Design and Analysis of WGFF Key Comparisons". Heavy emphasis was given to the first component (experiment design), showing how critical design principles are in this early planning stage. Adhering to all of the relevant statistical design principles is essential to assure the validity of the final computed KC inter-laboratory equivalency values.

Extensive notes outlined a comparison of various specific experimental plans, and discussion focused on the relative advantages and disadvantages of each. After considerable floor discussion among the statisticians in SED and the working group participants, the relevant issues were "put on the table", leading to statistically sound recommendations. A final specific design—involving a tandem meter arrangement—statistically balanced in the several factors—was proposed for working group consideration.

The sequel to this presentation, to concentrate on the second component in the KC (analysis of KC data based on sound statistical principles and state-of-the-art statistical techniques) will be given to the same WGFF in May, 2003 in the Netherlands.

4.2 Education

4.2.1 Education and Training

Nell Sedransk, SED staff
Statistical Engineering Division, ITL



Figure 4.1: Students from DEX workshop

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

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

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

SHORT COURSES – 2002

- **Exploratory Data Analysis (EDA)**

James Filliben

3/11,12,18,19/2003

Exploratory Data Analysis (EDA) is an approach to analyzing data for gaining insight into the structure underlying data. General EDA principles are discussed, along with a collection of techniques applied to a wide variety of NIST data sets.

- **Introduction to Bayesian Analysis for Scientists and Engineers**

Blaza Toman

11/5,7/2002 and 2/12,14/2002

The course introduces basic concepts of the Bayesian approach to statistical analysis such as subjective interpretation of probability, types of prior distributions, use of Bayes Theorem in updating information and inference procedures such as Bayes estimators and HPD regions. At the conclusion of the class, the students are able to assess their prior knowledge and transform it into a probability distribution for a univariate problem, then combine their prior information with the data and using the software BUGS calculate posterior probabilities of hypotheses of interest and HPD regions for the parameters.

- **Using WINBUGS for Bayesian Analysis of Industrial and Physical Science Data**

Will Guthrie and Richard Evans (Iowa State University),

11/22/2002

This course demonstrates the use of the WINBUGS program to solve statistical problems using a Bayesian approach based on industrial and physical science data. In particular, it analyzes some NIST generated data sets using WINBUGS.

- **Combining Information**

James Yen

9/26,27/2002

This 6 hour workshop used NIST examples to explore the statistical aggregation of information from different experiments or sources. Topics included parameter estimation, the multi-method problem, combined hypothesis tests, simultaneous inference, and function estimation.

- **Analysis of Variance**

Stefan Leigh

6/7,14,21,28/2002

This 4-day course covers the fundamentals of Analysis of Variance, including the analysis of up to 4-way fixed-effect designs, random component, and mixed models. The exposition is strongly dependent on development by example.

- **Functional Data Analysis**

Walter Liggett

3/26/2002, 4/4/2002

The course on functional data analysis covered the material in the book of the same name by Ramsay and Silverman (Springer 1997) except that NIST measurements were used to illustrate the statistical methods presented. The course covered spline smoothing, function registration, replicate functional measurements, functional principal components analysis, linear modeling, and the analytical use of derivatives.

- **Experimental Design for Scientists and Engineers**

James Filliben and Ivelisse Aviles

2/4-8/2002

Experiment design is a systematic, rigorous, data-based approach to scientific and engineering problem-solving. This 5-day workshop covers both the fundamental principles and techniques for the construction and analysis of designed experiments, illustrating its problem-solving role by application to a wide variety of real-world problems.

In addition to short courses, SED provides Web-based training in the form of the NIST/SEMATECH engineering Statistics Handbook. Course notes for some of the short courses in Statistics for Scientists and Engineers may also be available on the Web.

The NIST scientists and engineers with whom SED staff consult belong to a variety of discipline-specific professional societies. To assist in the transfer of methodology to professionals external to NIST who rely on NIST services and NIST-specific expertise, SED staff and NIST scientific personnel have collaborated in presenting a variety of short courses at, or in conjunction with, various professional society meetings. Such courses serve to reach out to the broader scientific community.

WORKSHOPS – 2002

- October 2001: Workshop on Estimating Uncertainties for Chemical Analysis: 1 day
Will Guthrie and Bob Watters (Santiago, Chile) [100 attendees]
- November 2001: Advanced Mass Measurements Workshop: 2 days
Hung-Kung Liu (NIST, RMAP, for Tech. Services) [30 attendees]
- February 2002: Intercomparison Designs and Statistical Optimization: 1/3 day
Jim Filliben (Sante Fe, NM, Low-level radionuclides meeting)
- March 2002: Workshop on Estimating Uncertainties for Chemical Analysis: 1 day
Will Guthrie and Bob Watters (New Orleans, PITTCO) [16 attendees]
- April 2002: Statistical Principles and Techniques for Design and Analysis of WGFF
KCs:
Jim Filliben (Arlington, Va, CIPM/CCM Working Group/Fluid Flow)
- July 2002: Expression of Uncertainty in Measurement: half-day
Hari Iyer (NIST Boulder, EEEL Magn. Tech. Div.)
- July 2002: Expression of Uncertainty in Measurement: half-day
Hari Iyer (NIST Boulder, EEEL Radio Freq. Tech. Div.)
- July 2002: Uncertainty Computation: half-day
Will Guthrie (with others) (NIST, SIM/RMAP/Regional KCs)
- August 2002: Using WinBUGS for Bayesian Analysis: 1 day
Will Guthrie and Rich Evans (New York, Amer. Stat. Assoc.)
- September 2002: Case Studies in Uncertainty Analysis: 1 day
Jack Wang (NIST Boulder)
- September 2002: Data Analysis for Interlaboratory Studies: 1 day
Jack Wang (NIST Boulder)
- March 2003: Workshop on Estimating Uncertainties for Chemical Analysis: 1 day
Will Guthrie and Bob Watters (Orlando, PITTCO)

The SED Seminar Series is of long-standing, and has over the years attracted numerous academic, industrial, and government speakers, representing every level of the profession. The individual seminars are typically one-hour talks, open to SED, NIST staff, and interested public. Speakers generally spend a day as guests of the Division, meeting with SED staff and interested NIST scientific personnel.

SED SEMINARS – 2001 - 2002

- December 11, 2002
Statistical Model Selection and Model Choice
Marianthi Markatou, Columbia University
- November 25, 2002
Fractional Difference Prewhitening in Atomic Clock Modeling
Lara S. Schmidt, U.S. Naval Observatory
- October 9, 2002
Multivariate Exponentially Weighted Moving Average Control Charts Based on the Sign and Signed-Rank Statistics
Alexandra Kapatou, University of Michigan
- September 16, 2002
Detecting Fraud in the Real World
Jose Pinheiro, Biostatistics Division, Novartis Pharmaceuticals
- February 26, 2002
Global Atmospheric Changes: Statistical Trend Analyses of Ozone and Temperature Data
George C. Tiao, The University of Chicago
- February 13, 2002
A Closer Look at Combining a Small Number of Binomial Experiments
Don Malec, The US Bureau of the Census
- January 23, 2002
Data Mining with Stepwise Regression
Dean Foster, Statistics Department, The Wharton School, University of Pennsylvania
- November 14, 2001
Stable Distributions: Models for Heavy-Tailed Data
John Nolan, American University
- August 23, 2001
Computational Statistics with MATLAB
Zhiping You, The Mathworks, Inc.
- June 21, 2001
Bayesian Multi-Use Calibration
Mark Vangel, Dana-Farber Cancer Institute

- February 21, 2001
Optimal Designs for Mixed-Effects Models with Random Nested Factors
Ivelisse Aviles, Northwestern University
- February 5, 2001
Hierarchical Modeling of Supercomputer Reliability
Kenneth Ryan, Department of Statistics, Iowa State University
- January 26, 2001
Self-Modeling Regression for Longitudinal Data
Naomi S. Altman, Chair, Department of Biostatistics, Cornell University

P principal functions of the Statistical Engineering Division are consulting, research, and teaching. The outreach, or teaching function, in all its modes, has a rich fifty-year history which has continuously demonstrated its utility in attracting and serving clients.

4.2.2 Conference on Designs for Generalized Linear Models

Ana Ivelisse Avilés
Statistical Engineering Division, ITL

André I. Khuri
Department of Statistics, University of Florida



Figure 4.2: Conference Participants

Goal and Organization of the Conference

The Conference was sponsored and hosted by the Statistical Engineering Division of NIST in Gaithersburg, MD, April 18-20, 2002. The Organizing Committee consisted of Dr. André I. Khuri, from the University of Florida in Gainesville, FL, and Dr. A. Ivelisse Avilés from NIST.

The goal of the Conference was to address recent experimental design issues that pertain to generalized linear models (*GLM*). The Conference provided a forum for interaction among research scientists working on diverse areas of *GLM*. A total of 16 leading researchers and scholars were invited as speakers. In addition, a total of 16 doctoral students and postdoctoral fellows/junior faculty were invited to attend the Conference. Funding to cover travel expenses for these individuals was provided by a grant from the *National Science Foundation, Division of Mathematical Sciences* (Grant #DMS-0207059). Drs. Khuri and Avilés were the co-principal investigators on this grant, and the University of Florida was the actual recipient of the funding. The *Statistical Engineering Division* has also waived the conference registration fees for several other individuals from the Washington, D. C., area. The travel grants and tuition waivers made it possible for the awardees to attend the Conference and interact with the speakers. The total number of participants at the Conference was 79.

The Main Thrust of the Conference

The main objective of the conference was to discuss design issues that pertain to fitting generalized linear models. In particular, the problem of dependency of a design on the unknown parameters of the model was one of the focus areas. The Conference opened with a talk by Dr. John Nelder who is one of the foremost leaders in the field of *GLM*. In fact, his paper (co-authored with R. W. M. Wedderburn) entitled “Generalized Linear Models”, which appeared in *Journal of the Royal Statistical Society, Series A* in 1972, was the first article on *GLM* and is considered the cornerstone of the methodology that has later led to the development of *GLM*. His talk was followed by a talk presented by Kathryn Chaloner, a leading expert on Bayesian designs for *GLM*. The talks that addressed the design dependency problem included the ones presented by Drs. Randy Sitter, Ben Torsney, Kevin Robinson, and André Khuri. Drs. Malay Ghosh and Joseph Voelkel proposed Bayesian methods for designs used in epidemiological research, and in experiments where the response is binary. The analysis of generalized linear models with random effects was addressed by Drs. Jiming Jiang and Timothy Robinson. Dr. Hari Iyer discussed unbalanced designs for the estimation of variance components in a 4-stage nested random model. Optimal designs for *GLM* were the subject of the talk by Dr. Sergei Leonov. Dr. Jeff Wu discussed design issues useful in quality engineering and reliability improvement. The Conference closed with a panel discussion on present and future directions in the field of *GLM*. The panelists were Drs. John Nelder, Bruce Ankenman, and Blaza Toman. This session was quite lively and several participants were actively involved in the discussions.

Benefits From the Conference

The Conference addressed current design problems and revisited some past design and estimation concerns with regard to *GLM*. This has provided a good review of the subject area, particularly for those who were not quite familiar with *GLM*.

The Conference provided a forum for interaction among the participants. Junior scientists gained from listening and interacting with the more senior researches. Graduate students, in particular, got new ideas that can benefit their doctoral research.

The Conference also addressed some new research ideas that can benefit all those working in the general area of *GLM*. For example, in the panel discussion on Saturday, Dr. John Nelder discussed extensions to the basic assumptions of *GLM* such as the treatment of multiple longitudinal data on each individual in a repeated-measures experiment, as well as extensions to spatial models.

The participants got to know each other on a personal level. This can help future contacts and perhaps collaborations on future research projects.

The participants learned about *NIST* and the research projects it supports. This can be beneficial to both *NIST* and the participants.

The Conference on Designs for Generalized Linear Models was an event that participants will cherish and appreciate for a long time.

4.2.3 Summer Students Program

Charles Hagwood

Statistical Engineering Division, ITL



Figure 4.3: 2002 Summer Students

The Summer 2002 was the second year of the Statistical Engineering Division's reinvigorated Summer Students Program. This year there were seven summer students from universities throughout the U.S. and Puerto Rico. The summer program is an opportunity for university graduates and undergraduates to experience firsthand the statistical workplace. The goal is to attract new students to the field of statistics. Each student is assigned a project and works under the supervision of a staff member. Undergraduates gain a greater appreciation for statistics and graduates use the program to supplement their understanding of their course work. Thus, the Statistical Engineering Division plays a direct role in increasing the pool of well-trained professional statisticians.

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 depend on educational level (freshman - advanced Ph.D student). Students are expected to spend at least two summer months at NIST usually starting in late May or early June. Some students, especially the local ones, may return during semester breaks to work.

The program is coordinated with other student programs at NIST, such as the NIST SURF program and the U.S. Department of Commerce PostSecondary 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 (HBCs) and Hispanic-serving institutions (HSIs). This year over thirty applications to HBCs and HSIs were sent out. Copies of the internship announcement can be found at <http://www.itl.nist.gov/div898/>.

The following letter of appreciation from one of our summer students provides an indication that our program is working.

Dear Nell,

I was so fascinated with the COOP at NIST using the RAVE that I dedicated my last semester researching OpenGL, Virtual Reality (VR) Systems, and the RAVE with two computer science professors at Bowie State University. Using OpenGL, Performer, and C++, I created a virtual tour of our new computer science building. It allows the user to be physically and mentally immersed within a virtual world (the computer science building). Once again, thank you for the opportunity. Thank you for congratulating me on my graduation. It was definitely a challenging semester.

*Sincerely,
Raymond Miller*

P.S. If there are any employment opportunities at NIST, please keep me in mind.

2001 Summer Students

Alisha Sparks, 2002 graduating senior, Voorhees College, Denmark SC
Zuriel Correa, 2nd year graduate student, University of Puerto Rico at Mayaguez
Raymond Miller, senior, Bowie State University, Bowie, MD
Igor Malioutov, senior, Northeastern University, Boston, MA
Margaret Polinkovsky, senior, Case Western Reserve, Cleveland, OH
David Martin, junior, University of Dayton, Dayton, Ohio
Kimball Kniskern, 2002 graduating senior, University of California, Berkeley

Student Projects

IGOR MALIOUTOV

An appropriate title for Igor's project would be NONPARAMETRIC DEPENDENCE CHARACTERISTICS and THEIR USE in AGGREGATED ALGORITHMS in HUMAN ID at DISTANCE. A biometric algorithm produces rankings of elements of the gallery. To understand the dependence between two or several such algorithms, nonparametric dependence characteristics, mainly rank correlation statistics, (like Spearman's rho or Kendall's tau) are helpful. These characteristics were used to construct a new procedure designed to combine several algorithms. This rule is analogous to the weighted average of several dependent observations. The excellent performance of the anticipated procedure was verified by data sets available in the FERET database. This idea of averaging ranks can be extended to several different algorithms, one of which, say, is a face recognition algorithm, and another is a fingerprint (or gait, or ear) recognition device.

Igor worked with Andrew Rukhin.

ALISHA SPARKS

Comparison of the US approach to testing package weights for discrepancies from stated weights with the European approach is of interest to companies that sell products in both markets. The US approach is described in Handbook 133, and the European approach in a document called R87. Alisha Sparks compared these approaches by programming the methods specified in each approach and testing their responses to simulated weight measurements. She found the probability of being declared in violation as a function of the mean of the actual weights for various amounts of variability in these weights. She showed the degree to which the approaches differ, but, of course, she could not decide the implications of the observed differences.

Alisha worked with Walter Liggett.

DAVID MARTIN

In 1997, Alan Heckert developed a web-based system for monitoring the status of SRM's. This system was based on HTML and Perl scripts for updating. With the recent change in the funding of SRM's from a single pool to individual funding for each SRM, a modification of this system was required. Specifically, far more budget information was required for each SRM. Also, there was a requirement for better searching and extraction of SRM's based on this budget information.

For this reason, it was decided to develop a new database-driven web system for SRM's. David implemented this system using the Cold Fusion software program. Cold Fusion is a commercial program that is used to simplify the development of web pages that use databases. David installed a new version of Cold Fusion (we had a rather old version for the e-Handbook project), learned the software, and developed a working system for the web monitoring of SRM's. This involved writing input and edit forms as well as search forms.

In addition, David helped with several miscellaneous tasks over the summer. One example was assisting in the preparation of power point slides for the program review.

David worked with Alan Heckert.

RAYMOND MILLER

Ray assisted in two tasks.

1) SED has contracted with the NIST Web Group to make the content of our web pages more database driven. Ray assisted in the design of the input forms for these databases (there are five databases).

2) Ray assisted Alan Heckert in installing Red Hat Linux on a PC in SED. He installed several software programs that constitute the "DIVERSE" software project. DIVERSE is the software used for high-end visualization. DIVERSE is built on Open-GL and Open-GL Performer. SED would like to investigate the usefulness of this high-end visualization software for statistical applications.

In addition, Ray helped with several miscellaneous tasks over the summer. One example was assisting in the preparation of power point slides for the program review.

Ray worked with Alan Heckert.

ZURIEL CORREA

In an effort to maximize both computing and human resources, the Statistical Engineering Division (SED) initiated the development of a Web Application for Statistical Analyses in the summer 2002. The SED envisions this application will enable NIST scientists and engineers to gain access to a consolidated array of solutions (analytics) to routine problems in data analysis such as univariate location estimation, consensus means analysis, linear calibration, errors-in-variables regression, and experiment design. Application areas that would benefit from such a system include: Standard Reference Materials (SRMs), Interlaboratory Studies, Key Comparisons, Bayesian Metrology, and Training/Education.

Zuriel Correa, an Industrial Engineering graduate student from the University of Puerto Rico (Mayaguez), constructed a prototype while visiting the SED this past summer.

Zuriel worked with Juan Soto.

KIMBALL KNISKERN

Kimball worked on network modeling problems. In order to become better acquainted with the field, he began by reading several papers on the multifractal wavelet models in network time series modeling. He downloaded Rice University's software for the Multifractal Wavelet Model (MWM), which is written for Matlab. This software is used for network traffic modeling and inference. After getting it to run on our system, he performed several simulations. This work is being continued by John Lu in collaboration with Mark Carlson of the Advanced Network Technologies Division. Encouraging simulation results for the NIST PingER data have been produced. This work is part of the NIST Net network traffic simulation project.

Kimball worked with John Lu and Dipak Dey.

MARGARET POLINKOVSKY

The Committee for Weights and Measures, National Metrology Institutes and Regional Metrology Organizations around the world committed all partners to recognize and accept each others measurements based on a standard of equivalence. The degree of equivalence of each national measurement standard is expressed as its deviation from a key comparison reference value (consensus value) and the uncertainty of this deviation at the 95% significance level. The key comparison reference value is based on measurements of a stable circulated object. At present there is no unified statistical procedure for determining the appropriate consensus value in a key comparison. Margaret simulated a multilab study and compared several estimators of the consensus value of the simulation, e.g. the mean, mode, weighted mean, etc.

Margaret worked with Nell Sedransk.

4.2.4 Minority Internship Announcement

Charles Hagwood

Statistical Engineering Division, ITL



Figure 4.4: Alisha Sparks, a student in the Statistical Engineering Division, and Ken Butcher, NIST Technology Services, discuss the statistical analysis of U.S. and proposed European standards for packaging and labeling.

The **Statistical Engineering Division** of the National Institute of Standards and Technology (NIST) announces its **2003 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. levels 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' own 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, or 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: 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.

4.3 New Staff

4.3.1 Juan Soto



BIOGRAPHICAL SKETCH

Juan Soto has been a Computer Scientist in the Statistical Engineering Division (SED) at NIST since May 2002. Juan earned a B.S. in Computational Mathematics from the University of Puerto Rico in 1991, an M.S. in Computational Applied Mathematics from the State University of New York at Stony Brook in 1993, and an M.S. in Computer Science from the University of Delaware in 1996. He is currently enrolled in Statistical Science graduate courses at George Mason University. In November 2001, he was co-recipient of a Department of Commerce Gold Medal for leadership in the development of the Advanced Encryption Standard. His interests include statistical computing, applied mathematics, cryptographic algorithms and computer security.

Prior to joining the SED, he was a senior software developer at Entrust, Inc., where he worked on cryptographic software applications for various Federal government agencies. During the period 1997-2000, he was a computer scientist at NIST's Computer Security Division where he worked on the implementation of a statistical test suite for random number generators, and the development of cryptographic standards. Prior to joining NIST, Juan was a software engineer at Lockheed Martin Management and Data Systems (1996-1997) conducting research in image understanding systems. He also was a mathematics instructor at Catonsville Community College (1993-1994) where he taught undergraduate courses in mathematics and computer science.

ONGOING PROJECTS

- Development of a web application for statistical analyses.
- Development of a Bayesian software library.
- Elicitation of priors software application in S+ for both univariate and multivariate normal linear models.
- Teach a course on Simulation.

4.3.2 Dennis Leber



BIOGRAPHICAL SKETCH

Dennis Leber joined the Statistical Engineering Division at NIST in January 2001. He received his M.S. degree in statistics from Rutgers University in October 1999 and a B.S. degree in mathematics from Bloomsburg University in May 1997. Dennis is currently continuing to strengthen his statistical background and knowledge through graduate coursework at George Washington University. Prior to joining the Statistical Engineering Department, Dennis spent 5 years in the Actuarial Research Department of Prudential Property and Casualty Insurance Company in Holmdel, NJ.

ONGOING PROJECTS

In addition to his work in the Statistical Engineering Division, Dennis spends half of his time in the Economic Assessment Office of NIST's Advanced Technology Program providing statistical and database support.

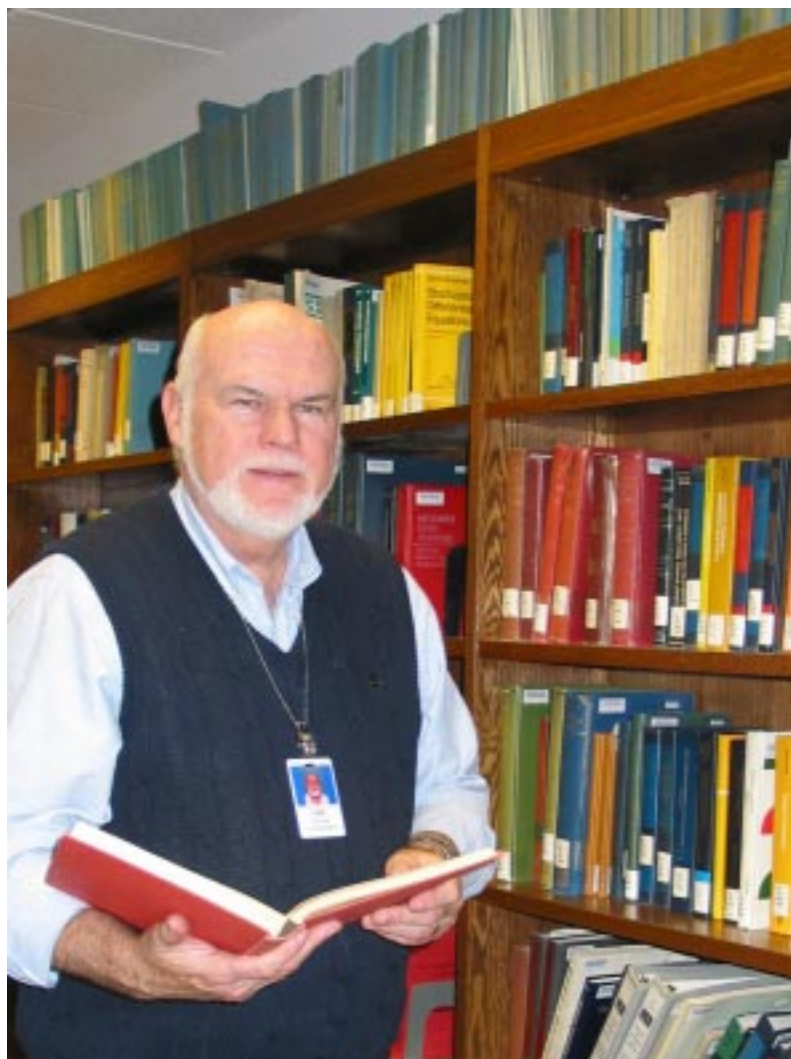
Dennis' projects in the Statistical Engineering Division include:

- HUD Healthy Homes Initiative: Sampling Plan for BFRL's CONTAM Software for Outdoor-Indoor Ventilation Rates.
- Design and analysis of experiment to consider the effect of PAC tub cooling on thermal-expansion-induced lathe deformation.
- DOE Interlab comparison of energy factor measurements in residential hot water heaters.
- Nonlinear regression model of Photovoltaic Cells.
- Develop methodology to assign uncertainty values to batch lots of gaseous NIST Traceable Reference Materials (NTRM).

STATISTICAL RESEARCH

Dennis' main area of statistical interest is experiment design for the physical sciences. Dennis will continue to explore and develop in this area of statistics via consultations with NIST scientists, interactions with SED colleagues, and ongoing graduate-level education.

4.3.3 Bill Strawderman



BIOGRAPHICAL SKETCH

William E. Strawderman is a Professor and former Chair of the Department of Statistics at Rutgers University. He is a fellow of the American Statistical Association and of the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. He received his Ph.D. from Rutgers University and has held visiting positions at Stanford University, Princeton University, Educational Testing Service, and the University of Rouen (France). His research interests include Statistical Decision Theory, and Bayesian Statistics, particularly as related to simultaneous estimation of several parameters. He has published 120 papers in refereed journals.

Currently he is a Visiting Faculty in the Statistical Engineering Division (SED) at NIST. He is working on problems related to combining information in interlaboratory studies.

5. Staff Publications and Professional Activities

5.1 Publications

5.1.1 Publications in Print

1. K. J. Coakley, Z. Chowdhuri, W. M. Snow, J. M. Richardson and M. S. Dewey, Estimation of Neutron Mean Wavelength from Rocking Curve Data, *Measurement Science and Technology* 14 131-139 2003
2. R. S. Cervený and K. J. Coakley, A Weekly Cycle in Atmospheric Carbon Dioxide, *Geophysical Review Letters*, 29, 10 1029/2001GLR13952 2002
3. K. J. Coakley and G. L. Yang, Estimation of the Neutron Lifetime: Comparison of Methods Which Account for Background, *Physical Review C*, 65, 04612-1 2002
4. K. J. Coakley, H. H. Chen-Mayer, G. P. Lamaze, D. S. Simons, and P. E. Thompson, Calibration of a Stopping Power Model for Silicon Based on Analysis of Neutron Depth Profiling and Secondary Ion Mass Spectrometry Measurements, *Nuclear Instruments and Methods in Physics Research B* 192 4: 349-359 2002
5. J. J. Filliben (with Z.C. Lin, A. Berne, B. Cummings, and K. Inn), Competence of Alpha Spectrometry Analysis Algorithms Used to Resolve the Am-24 and Am-243 Alpha Peak Overlap, *Journal of Applied Radiation and Isotopes*, 56(1-2), 2002, pp. 57-63.
6. J. J. Filliben (with H.S. Bennett), A Systematic Approach for Multidimensional, Closed-Form Analytic Modeling: Effective Intrinsic Carrier Concentrations in Gallium 1-x Aluminum x Arsenic Heterostructures, *Journal of Research of the National Institute of Standards and Technology*, 107(1), 2002, pp. 69-81.
7. J. J. Filliben (with K. Gurley, J.-P Pinelli, E. Simiu, and C. Subramanian), Fragility Curves, Damage Matrices, and Wind Induced Loss Estimation, *Proceedings of the Third International Conference on Computer Simulation in Risk Analysis and Hazard Mitigation: Risk Analysis III*, 2002.
8. J. J. Filliben (with K. Inn, et al), The NIST Natural-Matrix Radionuclide Standard Reference Program for Ocean Studies, *Journal of Radionuclide and Nuclear Chemistry*, 248(1), 2001, pp. 227-231.
9. J. J. Filliben (with R.R. Zarr, V. Marineez-Fuentes, and B.P. Dougherty), Calibration of Thin Heat Flux Sensors for Building Applications using ASTM C 1130, *Journal of Testing and Evaluation*, 29(3), 2001, pp. 293-300.

10. 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.
11. 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.
12. J. J. Filliben (with M. Simon (FHWA) and D. Bentz), Concrete Optimization Software Tool: User's Guide, *Federal Highway Administration Report*, March 2001.
13. 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.
14. 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.
15. 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.
16. 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 Comparison, *Proc. Third Workshop on Automatic Identification Advanced Technologies (AutoID02/IEEE)*, March 2002, Tarrytown, NY.
17. S.D. Leigh, (with J. Sieber, J. Yen) Standard Reference Materials for Cements. *Cement and Concrete Research*, 32, 2002, 1899-1906.
18. S.D. Leigh, (with J.F. Widmann, C. Presser) Extending the Dynamic Range of Phase Doppler Interferometry Measurements, *Atomization and Sprays*, 12, 2002, 513-537.
19. 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.
20. 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.
21. 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.
22. H.K. Liu, (with N.F. Zhang) Performance Evaluation of Approaches to Combining Results from Multiple Methods, *Proceedings of the 2002 Joint Stat. Meetings*
23. H.K. Liu, (with N.F. Zhang) Bayesian Approach to Combining Results from Multiple Methods, *Proceedings of the 2001 Joint Stat. Meetings*

24. Z.Q. J. Lu (2002), Local Polynomial Prediction and Volatility Estimation in Financial Time Series, *Modeling and Forecasting Financial Data: Techniques of Nonlinear Dynamics*, (Eds. Soofi, A. and Cao, Ly), Kluwer Academic Publishers, pp 115-135.
25. 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.
26. 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.
27. 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*, 51 (1), 2002, pp. 53-58.
28. J.H. Yen, (with K.E. Sharpless, J. Brown Thomas, B. Nelson, C. Phinney, J. Sieber, L. Wood) Value Assignment of Nutrient Concentrations in Standard Reference Material 2384 Baking Chocolate, *Journal of Agricultural and Food Chemistry*, 50, 2002, pp. 7069-7075.
29. J.H. Yen, (with J.B. Quinn, R.N. Nathan, I.K. Lloyd) Subjective Ceramic Machinability and Material Properties, *Machining Science and Technology*, 6(2), 2002, pp. 1-9.
30. J.H. Yen, Analysis of a Robust Variant of the Weighted Mean, *Proceedings of the Joint Statistical Meetings*, 2002.
31. N.F. Zhang, (with R. Kacker) Online Control Using Integrated Moving Average Model for Manufacturing Errors, *International Journal of Production Research*, 40(16), 2002, pp. 4131-4136.
32. N.F. Zhang, Combining Process Capability Indices from a Sequence of Independent Samples, *International Journal of Production Research*, 39(13), 2001, pp. 2769-2781.
33. N.F. Zhang, What the Generalized Moving Averages Can Do for the Process Monitoring, *Proceedings of Section of Physical and Engineering Sciences of American Statistical Society*, 2001, pp. 730-735.
34. 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.

5.1.2 NIST Technical Reports

1. A.I. Aviles (with C.E. Buchanan, A.V. Hackley, C.F. Ferraris), Analysis of the ASTM Round-Robin Test on Particle Size Distribution of Portland Cement, *NISTIR 6931*, 2002.

2. A.I. Aviles (with C.E. Buchanan, A.V. Hackley, C.F. Ferraris), Analysis of the ASTM Round-Robin Test on Particle Size Distribution of Portland Cement: Phase I, *NISTIR 6883*, 2002.
3. J. J. Filliben (with R. Zarr), International Comparison of Guarded Hot Plate Apparatus Using National and Regional Reference Materials, *NIST Technical Note 1444*, 2002.
4. S.D. Leigh, (with C.R. Schultheisz), Certification of the Rheological Behavior of SRM 2490, Polyisobutylene Dissolved in 2,6,10,14-Tetramethylpentadecane, *NIST IR 01-XXXX*, 2001, 75p.
5. S.D. Leigh, (with P.E. Stutzman), Phase Composition Analysis of the NIST Reference Clinkers by Optical Microscopy and X-ray Powder Diffraction, *NISTIR 1441*, 2001, 44p.
6. S.D. Leigh, (with C.R. Schultheisz, K. Flynn) Certification of the Rheological Behavior of SRM 2491, Polydimethylsiloxane, *NIST SP 260-147*, 2002.
7. S.D. Leigh, (with G.S. Cheok, A. Rukhin) Calibration Experiments of a Laser Scanner, *NISTIR 6922*, September 2002.
8. W.J. Rossiter, B. Toman, (with M.E. McNight, M.B. Anaraki) Factors Affecting Ultrasonic Extraction of Lead from Laboratory-Prepared Household Paint Films, *NISTIR 6834*.
9. 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, 47p.
10. 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*, 2003, 51p.

5.1.3 Publications in Process

1. A.I. Aviles, Robust Experiments with Two Variance Components, under submission.
2. A.I. Aviles, Assembled Designs for Estimating Dispersion Effects, submitted to *Technometrics*.
3. C. J. Horowitz, K. J. Coakley, D. N. McKinsey, Supernova Observation via Neutrino-Nucleus Elastic Scattering in the CLEAN Detector.
4. K. J. Coakley, C. -M. Wang, P. D. Hale and T. S. Clement. Adaptive Characterization of Jitter Noise in High-Speed Sampled Signals.
5. H. H. Chen-Mayer, G. P. Lamaze, K. J. Coakley, S. K. Satija, Two Aspects of Thin Film Analysis: Boron Profile and Scattering Length Density Profile.

6. P. R. Huffman, K. J. Coakley, S. N. Dzhosyuk R. Golub, E. Korobkina, S. K. Lamoreaux C. E. H. Mattoni, D. N. McKinsey, A. K. Thompson, G. L. Yang, L. Yang, and J. M. Doyle, Progress Towards Measurement of the Neutron Lifetime Using Magnetically Trapped Ultracold Neutrons, *Proceedings of Quark mixing, CKM unitarity*, Heidelberg, 19 - 20 September 2002
7. J. J. Filliben (with R. Zarr, et al), Collaborative Thermal Conductivity Measurements of Fibrous Glass and Expanded Polystyrene Reference Materials, submitted to *Proceedings of International Thermal Conductivity Conference*.
8. J. J. Filliben (with R. Zarr, et al), An International Study of Guarded Hot Plate Laboratories Using Fibrous Glass and Expanded Polystyrene Reference Materials, submitted to *ASTM Special Technical Publication*.
9. J. J. Filliben (with E. Simiu, et al), Hurricane Damage Prediction Model for Residential Structures, submitted to *Journal of Structural Engineering*.
10. S.D. Leigh, (with J.M. Smeller) Potassium Bromate Assay by Redox Titrimetry Using Arsenic Trioxide, *NIST Journal of Research*, to appear.
11. S.D. Leigh, (with R. Marinenko) Heterogeneity Evaluation of Research Materials for Standards Certification, *Microscopy and Microanalysis Journal*, to appear.
12. S.D. Leigh, (with M. Schantz, D. Poster, et al) Determination of Polychlorinated Biphenyl Congeners and Chlorinated Pesticides in a Fish Tissue Standard Reference Material, *Analytical and Bioanalytical Chemistry*, to appear.
13. S.D. Leigh, (with A. Rukhin, J. Phillips, P. Grother, et al) Dependence Characteristics of Face Recognition Algorithms, *Pattern Recognition*, to appear.
14. S.D. Leigh, (with C. Beauchamps) Best Measurement Practice Guide: Using Magnetic Methods for the Determination of Nonmagnetic Coating Thickness on Magnetic Substrates, US DOC publication, to appear.
15. S.D. Leigh, (with D.L. Poster, M.M. Schantz, S.A. Wise) Development of Solution (Methanol) and Transformer Oil Standard Reference Materials for Selected Aroclors, *NIST Journal of Research*, to appear.
16. W.S. Liggett, Nonparametric and Semiparametric Models in Comparison of Observations of a Particle Size Distribution, *Journal of the Japanese Society of Computational Statistics*, in press.
17. W.S. Liggett, System Performance and Natural Language Expression of Information Needs, *Information Retrieval*, to be submitted.
18. W.S. Liggett, Parameter Design for Measurement Protocols by Latent Variable Methods, *Technometrics*, to be submitted.
19. H.K. Liu, (with G. Stenbakken) Empirical Modeling Methods Using Partial Data, submitted to the IEEE Transactions on Instrumentation and Measurement.
20. Z.Q. J. Lu, with Nell Sedransk, Generalized Pareto Mixture Models for Network Traffic with Applications to Performance Evaluation, *IEEE/ACM Transaction on Networking*, in review.

21. K.J. Coakley, J.D. Splett, (with M.D. Janezic, R.F. Kaiser), Estimation of Q-factors and Resonant Frequencies, *IEEE Transactions on Microwave Theory and Techniques*, to appear.
22. J.D. Splett, C.M. Wang, Uncertainty in Reference Values for the Charpy V-notch Verification Program, *ASTM Journal of Testing and Evaluation*, submitted.
23. H.K. Iyer, D.F. Vecchia, (with P.W. Mielke, Jr.), Higher Order Cumulants and Tchebyshev-Markov Bounds for P-Values in Distribution-Free Matched-Pairs Tests, *Journal of Statistical Planning and Inference*, to appear.
24. S. Brown, T. Larason, B. Toman, Report on the CCPR-K2.a Key Comparison of Spectral Responsivity over the range from 900 nm to 1600 nm, Draft A.
25. C.M. Wang, (with T.J. Drapela) A Statistical Model for Cladding Diameter of Optical Fibers, *Metrologia*, in press.
26. C.M. Wang, (with F. de Silva) Magnetic Thin Film Interlaboratory Comparison, *NIST Journal of Research*, to appear.
27. C.M. Wang, H.K. Iyer (with T. Mathew) Models and Confidence Intervals for True Values in Interlaboratory Trials, *Journal of the American Statistical Association*, submitted.
28. C.M. Wang, (with C. Fu and K.A. Bertness) Effects of Noise Level in Fitting in-situ Optical Reflectance Spectroscopy Data, *Journal of Crystal Growth*, submitted.
29. C.M. Wang, J.D. Splett, (with T.E. Harvey, K.A. Bertness, and R.K. Hickernell) Accuracy of AlGaAs Growth Rates and Composition Determination Using RHEED Oscillations, *Journal of Crystal Growth*, submitted.
30. C.M. Wang, (with T.J. Drapela, S.L. Gilbert, and W.C. Swann) The NIST Traceable Reference Material Program for Wavelength Reference Absorption Cells, NIST Special Publication, to appear.
31. C.M. Wang, (with D. Williams and U. Arz) An Optimal Multiline TRL Calibration Algorithm, *International Microwave Symposium*, submitted.
32. N.F. Zhang, (with N. Sedransk, D. G. Jarrett) Statistical Uncertainty Analysis of CCEM-K2 Comparisons of Resistance Standards, to appear in IEEE Transactions.
33. N.F. Zhang, Estimation of Process Variance in Using SPC Charts for a Stationary Process, to appear in 2002 Proceedings of Section of Physical and Engineering Sciences of American Statistical Society.
34. N.F. Zhang, Statistical Process Monitoring for Autocorrelated Data, to appear in the Proceedings of 2001 International SCRA Conference.
35. N.F. Zhang, A Study on the Variance Estimation for a Stationary Process in SPC, submitted.
36. N.F. Zhang, The Generalized Moving Averages of a Stationary Process and Their Applications, submitted.

5.1.4 Working Papers

1. K. J. Coakley and D. N. McKinsey, Spatial Methods for Event Reconstruction in CLEAN.
2. C. Hagwood, Dynamic Linear Calibration, 2003.
3. C. Hagwood, Consensus Values in Small Multiple-Method Studies, 2003.
4. C. Hagwood, Estimation of the Waiting Times at Internet Backbone Nodes, 2003.
5. H.K. Liu, (with J.T. Hwang) Does EM Algorithm Work for Identifying Principal components When Massive Data are Missing.
6. D. Malec, Bayesian Inference for a Concensus Mean Using Hierarchical Models: A Review of Current Methods Plus a New, Partition Model Approach.
7. J.D. Splett, K.J. Coakley, (with M.D. Janezic, R.F. Kaiser), Relative Permittivity and Loss Tangent Measurement Using the NIST 60 mm Cylindrical Cavity, *NIST Special Publication*.
8. B. Toman, A Robust Key Comparison Reference Value in Cases of Dominant Type B Error.
9. B. Toman, S Brown, Bayesian Analysis of CCPR Key Comparison on Near-Infrared Spectral Responsivity.
10. B. Toman, Bayesian Models with Type B error.
11. D. Malec, B. Toman, Multivariate Bayesian Model of the SRM 1946, Lake Superior Fish Tissue.
12. J.H. Yen, Approximating Stationary Distributions using Twin Processes.
13. J.H. Yen, Partial Influence functions.
14. N.F. Zhang, (with N. Sedransk) Statistical Analysis for Key Comparisons with Linear Trends.
15. N.F. Zhang, (with H. Liu) Uncertainty Analysis for Key Comparison with Trends.
16. N.F. Zhang, Two New Estimators of the Variance of the Graybill-Deal Estimator of a Common Mean.
17. N.F. Zhang, (with A. Vladar, M. Postek, B. Larrabee) A Kurtosis-based Statistical Measure for Two-dimensional Processes and Its Applications to Image Sharpness.
18. N.F. Zhang, (with P. Winkel) Serial Correlation of QC-data on the Use of Proper Control Charts.

5.1.5 Acknowledgements in Publications

1. K.J. Coakley in: E. Simiu, Chaotic Transitions in Deterministic and Stochastic Dynamical Systems: Applications of Melnikov Processes in Engineering, Physics, and Neuroscience. Princeton University Press, 2002.
2. S.D. Leigh in: C. Elster and A. Link, Analysis of Key Comparison Data: Assessment of Current Methods for Determining a Reference Value, *Measurement Science and Technology*, 12 (2001), p. 1431–1438.
3. S.D. Leigh in: C. Dabrowski, K. Mills and J. Elder, Understanding Consistency Maintenance in Service Discovery Architectures during Communication Failure, 2002.
4. Z.Q. John Lu in: Abdol S. Soofi and Liangyue Cao (eds). *Modeling and Forecasting Financial Data: Techniques of Nonlinear Dynamics*, Kluwer Academic Publishers, 2002.

5.2 Talks

5.2.1 Technical Talks

1. A.I. Aviles, Optimal Design for Mixed-Effects Models with Two Random Nested Factor, Joint Statistical Meetings, New York, New York, August, 2002.
2. D. N. McKinsey and K. J. Coakley, “CLEAN”, April 2002 meeting of the American Physical Society, Albuquerque, New Mexico
3. K. J. Coakley and D. N. McKinsey, Event Location Estimation and Background Discrimination in a Proposed Low Energy Neutrino Experiment, April 2002 meeting of the American Physical, Albuquerque, New Mexico
4. H. H. Chen-Mayer, G. P. Lamaze, K. J. Coakley, S. K. Satija, Two Aspects of Thin Film Analysis: Boron Profile and Scattering Length Density Profile, 10th Symposium on Radiation Measurements & Applications, Ann Arbor, MI, May 21-23, 2002
5. K. J. Coakley, Correcting Optoelectric Signal Measurements for Time Shift Errors, Time Base Distortion and Jitter Noise, Telecommunications Industry Association International Electrotechnical Commission meeting Poipu, Kauai (Hawaii), January 21 to 24, 2002
6. K. J. Coakley, Some Statistical Problems in Optoelectronics, Spring Research Meeting of American Statistical Society, Ann Arbor, MI May, 2002
7. K. J. Coakley, Some Statistical Problems in Optoelectronics, University of Colorado, Denver, May 2002.
8. J. J. Filliben, Flow Measurements for Multi-Meter Transfer Standards, Joint Statistical Meetings 2002, New York, NY, August 13, 2002.

9. J. J. Filliben, Statistical Principles and Techniques for the Design and Analysis of WGFF Key Comparisons, Third Meeting of the CIPM/CCM Working Group for Fluid Flow, Arlington, VA, April 11, 2002.
10. C. Hagwood, Bayesian Calibration, 65th Annual Meeting of the Institute of Mathematical Statistics, July 28-31, 2002, Banff, Canada.
11. C. Hagwood, Calibration a Pressure Guage Using Dynamic Linear Calibration, Annual Meeting of the American Statistical Society, August 11-15, 2002, New York, NY.
12. W.S. Liggett, Comparison of Replicate Functional Measurements: The Physical Sciences Challenge, Developments and Challenges in Mixture Models, Bump Hunting, and Measurement Error Models, Case Western Reserve University, Cleveland, Ohio, June 2-4, 2002.
13. H.K. Liu, Case Studies in Uncertainty Analysis, National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan, April, 2002.
14. H.K. Liu, N.F. Zhang, Performance Evaluation of Approaches to Combining Results from Multiple Methods, Joint Statistical Meetings, New York, New York, August, 2002.
15. Z.Q. John Lu, Tail Metrics for Network Performance based on GPD and Mixture Modeling. DARPA PI workshop, Baltimore, Maryland, April 17-19, 2002.
16. Z.Q. John Lu, SVD-based structured kernel regression for high-dimensional prediction. Spring Research Conference, Ann Arbor, Michigan, May 20-22, 2002.
17. Z.Q. John Lu, Bayesian Approach to Inverse Problems. IMS annual meeting, Banff, Canada, July 2002.
18. N. Sedransk, Statistical Uncertainty Analysis of CCEM-K2 Comparisons of Resistance Standards, IEEE - CPEM 2002 Conference, Ottawa, Ontario, Canada, June, 2002.
19. N. Sedransk, In the Intersection of Statistics and Metrology, University of Connecticut, Storrs, Connecticut, November, 2002.
20. N. Sedransk, Critical Issues for the Analysis of Key Comparison Data, BIPM - NPL Workshop on Statistics of Interlaboratory Comparisons, NPL, Teddington, England, September, 2002.
21. N. Sedransk, Statistical Metrology: Statistics, Standards and Measurement Science, Duke University, Durham, North Carolina, April, 2002.
22. B. Toman, Designs for GLMs: Present and Future Directions, a Panel Discussion, GLM Conference, April 20, 2002.
23. B. Toman, Statistical Analysis of the CCPR Key Comparison, PL NIST, May 14, 2002.
24. B. Toman, Statistical Analysis of the CCPR Key Comparison, NEWRAD 2002 Conference, May 20, 2002.

25. C.M. Wang, SED Recent Efforts on Uncertainty, Symposium on Uncertainty, ASTM E11 Committee Meeting, Glenn Bernie, MD, April 23, 2002.
26. C.M. Wang, A Generalized Confidence Interval for a Consensus Mean with Applications to Interlaboratory Studies, Joint Statistical Meetings, New York, NY, August 13, 2002.
27. J.H. Yen, Trimmed Weighted Means, Mid-Atlantic Probability and Statistics Day, Washington, DC, November 16, 2002.
28. J.H. Yen, Analysis of a Robust Variant of the Weighted Mean, Joint Statistical Meetings, New York, NY, August 2002.
29. N.F. Zhang, Issues for Key Comparisons among National Metrology Institutes, ASTM E11 on Quality and Statistics, Glen Burnie, Maryland, April 23, 2002.
30. N.F. Zhang, Statistical Process Monitoring for Autocorrelated Data, East China Normal University, Shanghai, China, June 13, 2002.
31. N.F. Zhang, Statistical Process Monitoring for Autocorrelated Data, University of Science and Technology of China, Hefei, China, June 18, 2002.
32. N.F. Zhang, Statistical Analysis and Key Comparisons, Metrological Statistics and Particle Measurement Section of National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan, July 15, 2002.
33. N.F. Zhang, Estimation of Process Variance in Using SPC Charts for a Stationary Process, New York, New York, August 2002.

5.2.2 General Interest Talks

1. A.I. Aviles, SURF Experiences at NIST: a Personal Account, University of Puerto Rico, February, 2002.
2. A.I. Aviles, From SURF to Stats: Mixed effect models, NIST, Gaithersburg, MD, August 8, 2002.
3. J. J. Filliben, Basketballs, Funnels, and Designed Experiments, Adventures in Science, NIST, Gaithersburg, MD, November 2, 2002.

5.2.3 Lecture Series

1. J. J. Filliben (with Ivelisse Aviles), Experiment Design for Scientists and Engineers, NIST, Gaithersburg, MD, February 4-8, 2002.
2. S.D. Leigh, *Analysis of Variance for Scientists and Engineers*, NIST, June 2002.
3. B. Toman, Bayesian Analysis for Scientists and Engineers, February 14, 2002.
4. B. Toman, Bayesian Analysis for Scientists and Engineers, November 5, 2002.
5. J.H. Yen, Combining Information, NIST, Gaithersburg, MD, September 26-27, 2002.

5.3 Professional Activities

5.3.1 NIST Committee Activities

1. A.I. Aviles, Member, ITL awards committee.
2. A.I. Aviles, Mentor, 2002 SURF/NSF program.
3. A.I. Aviles, Advisor to the ITL coordinators, SURF/NSF program.
4. A.I. Aviles, Member, NIST employees concerned with disabilities.
5. K. J. Coakley, Boulder Editorial Review Board
6. K. J. Coakley, ITL Awards Committee
7. K. J. Coakley, NIST 2010 strategic planning activity
8. S.D. Leigh (with W. Guthrie), NIST Washington Editorial Review Board (WERB), as of January 2003.
9. W.S. Liggett, Member, NIST Institutional Review Board CSTL/SRMP/CAML Process Improvement Team.
10. N. Sedransk, Member of Measurement Services Group.
11. N. Sedransk, Member of MSAG Task Force on SRM Business Practices.
12. N. Sedransk, Leader of Task Force on Statistical Methodology for Key Comparisons.
13. J.D. Splett, Member, ITL Diversity Committee.
14. J.D. Splett, EEEL MCOM Technical Subcommittee for the Direct Comparison Power System.
15. C.M. Wang, Member, EEEL MCOM subcommittee on relative permittivity and loss tangent SRM.
16. N.F. Zhang, Member, EEEL, MCOM subcommittee on AC-DC Difference of Voltage.
17. N.F. Zhang, Member of Working Group on NIST Quality Manual.

5.3.2 Standards Committee Memberships

1. K. J. Coakley, Telecommunications Industry Association International Electrotechnical Commission, TIA/EIC, Working Group 4, TC-86
2. N.F. Zhang, Project Leader of PDS 21749 of ISO/TC/69 on Application of Statistical Methods.
3. N.F. Zhang, Liaison between ISO/TC69/SC6 and ISO/REMCO on Reference Materials.
4. N.F. Zhang, Member, ASC Z1 Subcommittee on Statistics.

5.3.3 Other Professional Society Activities

1. A.I. Aviles, Review panelist, NSF course curriculum and laboratory improvement program.
2. A.I. Aviles, DOC mentor, National disability mentoring day (October 16, 2002)
3. K.J. Coakley, NIST representative to National Institute of Statistical Science (NISS).
4. W.S. Liggett, American Statistical Association, Section on Statistics and the Environment, Committee on the Distinguished Achievement Award, 2001-2003-.
5. Z.Q. John Lu, Organzier of an invited session, in International Conference On Current Advances And Trends In Nonparametric Statistics, July 15-19, 2002 - Crete, Greece.
6. N. Sedransk, Vice Chair, ASA Publications Committee.
7. N. Sedransk, Member, ASA Subcommittee on Publications Marketing.
8. N. Sedransk, Member, WSS Planning Committee on Statistics for Homeland Defense and Security.
9. N. Sedransk, Member, ASA-ENVR Committee on Fellows.

5.4 Professional Journals

5.4.1 Editorships

1. W.S. Liggett, Board of Editors of the NIST Journal of Research.

5.4.2 Refereeing

1. A.I. Aviles, *Technometrics*.
2. K. J. Coakley, *Biometrics*.
3. S.D. Leigh, *Journal of Polymer Science Part B: Polymer Physics*.
4. S.D. Leigh, *Analytical and Bioanalytical Chemistry*.
5. Z.Q. John Lu, *IEEE Transactions on Signal Processing, Computing in Science & Engineering*.
6. N. Sedransk, *Metrologia*.
7. J. Soto, *Journal of Statistical Planning and Inference*.
8. J. Soto, *ACM Transactions on Modeling and Computer Simulation* (Special Issue on Random Number Generation and Highly Uniform Point Sets).

9. B. Toman, *Technometrics*.
10. B. Toman, *The Institute of Statistical Mathematics*.
11. C.M. Wang, *Psychometrika*.
12. N.F. Zhang, *Technometrics, Metrologia, Mathematical Methods of Statistics*.

5.5 Review Panels

1. N. Sedransk, National Science Foundation.

5.6 Honors

1. W.S. Liggett, Judson C French Award, December 2002.
2. N. Sedransk, Fellow of American Statistical Association.

5.7 Trips Sponsored by Others and Site Visits

1. H.K. Liu, Visit to AIST, Tsukuba, Japan, April, 2002.

5.8 Training & Educational Self-Development

1. A.I. Aviles, Accessibility technology and ITL security awareness training program.
2. A.I. Aviles, Bayesian metrology.
3. S.D. Leigh, Short Course in Cryptography, MATHFEST, Burlington, Vt, Aug 1-2, 2002.
4. S.D. Leigh, Short Course in Public Key Cryptography, Joint Mathematics Meetings, Baltimore, Md, Jan 13-14, 2003.
5. J. Soto, Advanced R Programming, Boston, MA, June 20-21, 2002.
6. J. Soto, STAT 544 Applied Probability, George Mason University, Fairfax, VA, August 27, 2002 - December 10, 2002.
7. J.D. Splett, Guidelines for Evaluating and Expressing Uncertainty of NIST Measurement Results, Boulder, CO, July 2, 2002.
8. J.H. Yen, Project 2000 class, NIST, January 22, 2002
9. N.F. Zhang, Attended 5 management training and other training courses.

5.9 Special Assignments

1. K. J. Coakley, NIST representative to National Institute of Statistical Science (NISS) Affiliates Program.
2. S.D. Leigh, SED liaison for NIST/NRC postdoctoral associateship program.
3. J. Soto, Simulation in Statistical Science Part I, SED Summer Students, July 2, 2002.
4. J. Soto, Simulation in Statistical Science Part II, SED Summer Students, July 9, 2002.
5. J.D. Splett, Organized/coordinated the ITL Diversity Committee's open house for the Boulder Laboratories October 17, 2002.



