Statistical Engineering Division
NRC Review 2009

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

Boulder, Colorado
Gaithersburg, Maryland

Dedication

To Joan Rosenblatt,
with highest admiration and respect
— in appreciation

Foreword

Churchill Eisenhart came to NIST (then the *National Bureau of Standards*, NBS) from the University of Wisconsin in 1946 to start the Statistical Engineering Laboratory. Jack Youden was one of his first hires.

In the course of the following 63 years, always under the same name, and executing the same mission, the Statistical Engineering Division (SED) has been a key contributor to the success of many consequential NIST projects related to measurement science.

To achieve this, not only have we applied probabilistic and statistical methods to a wide range of scientific and technological problems successfully, but, at the same time, we have also developed and cultivated long-term relationships with collaborators in the other NIST laboratories.

One of the keys to SED's success has been our having consistently taken steps to become familiar with the areas of science and technology in which we have worked: this commitment has made us *bona fide* scientific partners, and it has made our activities truly inter-disciplinary, and the products of our work relevant.

Purpose & Limitations

This compilation serves to introduce the members of the Statistical Engineering Division (SED) and to highlight some of their recent work.

The authors of the individual contributions solely are responsible for their contents: the technical devices they use, the conclusions they reach, or the opinions they voice, do not necessarily reflect either SED's or NIST's positions. Those listed as collaborators do not necessarily endorse any aspect of the contributions.

Composition

BOULDER, CO			
Buhse	S	Coakley, *PhD*	
Iyer, *PhD*	F	Splett, *MSc*	
Vecchia, *PhD*		Wang, *PhD*	GL

GAITHERSBURG, MD			
Aviles, *PhD*		Bailey	S
Filliben, *PhD*		Guthrie, *PhD*	
Hagwood, *PhD*		Harris, *BSc*	GR
Heckert, *MSc*		Leber, *MSc*	
Leigh, *MSc*		Liggett, *PhD*	
Liu, *PhD*		Lu, *PhD*	
Possolo, *PhD*	DC,GL	Rukhin, *PhD*	
Strawderman, *PhD*	F	Toman, *PhD*	
Rosenblatt, *PhD*	GR	Yang, *PhD*	F
Yen, *PhD*		Zhang, *PhD*	

DC	Division Chief
F	Faculty Appointee
GL	Group Leader
GR	Guest Researcher
S	Secretary

Abbreviations

BFRL	Building and Fire Research Laboratory (NIST)
CNST	Center for Nanoscale Science and Technology (NIST)
CSTL	Chemical Science and Technology Laboratory (NIST)
DHS	Department of Homeland Security
DNDO	Domestic Nuclear Detection Office (DHS)
EEEL	Electronics and Electrical Engineering Laboratory (NIST)
ITL	Information Technology Laboratory (NIST)
MEL	Manufacturing Engineering Laboratory (NIST)
MSEL	Materials Science and Engineering Laboratory (NIST)
NCLSI	National Conference of Standards Laboratories International
NCNR	NIST Center for Neutron Research
NIJ	National Institute of Justice (U.S. Department of Justice)
OLES	Office of Law Enforcement Standards (EEEL)
PL	Physics Laboratory (NIST)
SED	Statistical Engineering Division (ITL)
TS	Technology Services (NIST)

Typesetting

Using LaTeX, processed with Christian Schenk's (2009) `MiKTeX` implementation of TeX and associated utilities (`www.miktex.org`). The fonts are Bitstream Charter, a Matthew Carter design, and Paul Pichaureau's MathDesign; the sans-serif and monospace characters are from the "Bera" version (Malte Rosenau, Ulrich Dirr) of Bitstream Vera, a Jim Lyles design.

Contents

1 Aviles: Biography . . . 8
2 Aviles & Possolo: Body Armor . . . 9
3 Bailey: Biography . . . 11
4 Buhse: Biography . . . 11
5 Coakley: Biography . . . 12
6 Coakley: Neutron Transmission Tomography of Fuel Cells . . . 13
7 Coakley: Radiative Decay of the Neutron . . . 15
8 Filliben: Biography . . . 17
9 Filliben: Measurement Science for Complex Information Systems . . . 18
10 Filliben: Verification and Uncertainty Estimation . . . 20
11 Guthrie: Biography . . . 22
12 Guthrie: Measurements of Blood Lead Levels . . . 23
13 Guthrie: Educational Outreach in Statistical Metrology . . . 25
14 Guthrie & Zhang: Three Statistical Paradigms . . . 27
15 Hagwood: Biography . . . 29
16 Hagwood: Langevin Dynamics for a Nanorod in an Electric Field . . . 30
17 Hagwood: Shape Descriptors for Cell Populations . . . 32
18 Harris: Biography . . . 34
19 Heckert: Biography . . . 35
20 Heckert: e-FITS . . . 36
21 Heckert & Zhang: Scatterfield Microscopy . . . 38
22 Iyer: Biography . . . 40
23 Leber: Biography . . . 41
24 Leber: Experimental Design for Performance Assessment of Radiation Monitors . . . 42
25 Leber: Performance Assessment of Infrared Imaging Systems . . . 44
26 Leber: Standard Reference Materials . . . 46
27 Leigh: Biography . . . 48
28 Leigh: Prediction of Cement Characteristics . . . 49
29 Leigh: Advanced Spectral Portal Testing . . . 51
30 Liggett: Biography . . . 53
31 Liggett: Slice-by-Slice Comparison of Computed Tomography Scans . . . 54
32 Liggett: Control Group Variation in Case-Control Studies of Gene-Expression . . . 56
33 Liu: Biography . . . 58
34 Liu: Alternative Approach to Mass Metrology . . . 59
35 Liu: R Package for Statistical Metrology . . . 61
36 Lu: Biography . . . 63

37 Lu: Background Correction in Single Cell Images **64**
38 Lu: Round-robin Study of Seebeck Coefficient Measurements **66**
39 Possolo: Biography **68**
40 Possolo: Geochemical Atlas of the United States **69**
41 Possolo: Tunable Compression of Wind Tunnel Data **71**
42 Rosenblatt: Biography **73**
43 Rukhin: Biography **74**
44 Rukhin: Weighted Means Statistics in Interlaboratory Studies **75**
45 Rukhin: Linkage Analysis in Interlaboratory Studies **77**
46 Splett: Biography **79**
47 Splett & Coakley: Low-Count Isotopic Ratios **80**
48 Splett: Critical Current Metrology for Nb_3Sn Conductor Development **82**
49 Strawderman: Biography **84**
50 Strawderman: Non-conformity in Interlaboratory Studies **85**
51 Toman: Biography **87**
52 Toman: Computational Models **88**
53 Toman: Uncertainty Analysis in Interlaboratory Studies **90**
54 Vecchia: Biography **92**
55 Vecchia: Pulse Shape Discrimination for Fast Neutron Spectroscopy **93**
56 Wang: Biography **95**
57 Wang: Waveform Metrology **96**
58 Wang: Fiducial Prediction Intervals **98**
59 Yang: Biography **100**
60 Yen: Biography **101**
61 Yen: Estimation of Detection Limits in Explosive Trace Detectors **102**
62 Yen: Motion Imagery Metrics **104**
63 Zhang: Biography **106**
64 Zhang: Allan Variance of Time Series Models for Measurement Data **107**
65 Obituary **109**

1 Ana Ivelisse Avilés

Biography

Ivelisse Avilés holds a B.S. in Industrial Engineering from the University of Puerto Rico, Mayagüez, and a M.S. and Ph.D. in Industrial Engineering and Management Sciences from Northwestern University. From 1995-1997, she worked for Johnson & Johnson. Her experience with the pharmaceutical and medical devices industry motivated her interest in experimental design. Her research focuses on statistical design of physical experiments, linear and non-linear mixed-effects models, and quality improvement and control.

She organized and co-chaired the international Conference on Generalized Linear Models that was held at NIST in April 2002, served as co-Director of the ITL's Summer Undergraduate Research Program (2005-2006), and as Associate Editor for the Journal of the American Statistical Association. Ivelisse spent most of 2008 on Capitol Hill on a special assignment as a Department of Commerce's Science and Technology Fellow serving as primary advisor to Representative Luis Fortuño (R-PR) on science, space, technology, and telecommunications.

Awards

National Science Foundation Graduate Fellow (1997), Grant Mack Memorial Scholarship from the American Council of the Blind (1998), R.A. Freund international scholar from the American Society for Quality (1999), and Summer Research Opportunities Program Alumni Achievement Award from the Committee on Institutional Cooperation (2004).

Selected Publications

Irvine, J.M., Aviles, A.I., Cannon, D.M., Fenimore C., Haverkamp, D., Israel, S.A., O'Brien, G., and Roberts, J. (2007) "Developing an Interpretability Scale for Motion Imagery", *Optical Engineering*, 46(11).

Ferraris, C.F., Hackley, V.A., and Aviles, A.I (2004) "Measurement of Particle Size Distribution in Portland Cement Powder: Analysis of ASTM Round-Robin Studies", *Cement, Concrete and Aggregate Journal*, 26(2).

Ankenman, B.E., Aviles, A.I., and Pinheiro, J.C. (2003) "Optimal Designs for Mixed-Effects Models with Two Random Nested Factors" *Statistica Sinica*, 13: 385–401.

2 Measuring the performance of body armor

AUTHORS Ana Ivelisse Aviles and Antonio Possolo
COLLABORATORS Dennis Leber and Jolene Splett (Statistical Engineering Division, ITL), Kirk Rice, Michael Riley, Amanda Forster, and Diane Mauchant (Office of Law Enforcement Standards, EEEL)

Introduction

NIST research on the performance of body armor (bulletproof vests) is carried out by the Office of Law Enforcement Standards, sponsored by the National Institute of Justice. SED supports some of this research, in particular: estimation of the bullet velocity (v_{50}) for which the probability of bullet penetration is 50 %; characterization of how vest performance varies with the vest's materials and structure, and with its age; determination of a vest's effective lifetime, after which it should be replaced.

The ballistic performance of body armor is tested by firing rounds at a vest mounted against an oil-based modeling clay block, in controlled experiments conducted in a firing range. For each shot, the measured response is either a binary (0/1) indicator of penetration, or a measurement of the depth of the indentation that the round creates on the clay block when there is no penetration.

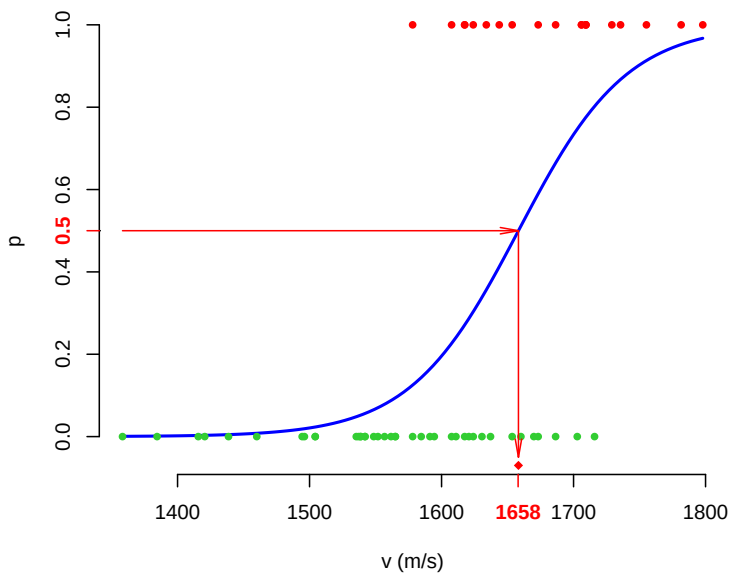
In this overview, we show how we use a logistic regression model to estimate v_{50} and its uncertainty, explain why this approach is advantageous in practice relative to other methods that have been used to estimate the same quantity, and suggest how the approach can be generalized to address other goals.

Logistic Regression

The logistic regression model expresses the probability p of penetration as a function of bullet velocity v according to the relation $\log\big(p/(1-p)\big) = \beta_0 + \beta_1(v - \overline{v})$ where β_0 and β_1 are parameters that characterize the relationship, and $\overline{v}$ denotes the average bullet velocity: the first, β_0, is the log-odds of penetration by a bullet of average velocity; the second, β_1 is the rate of increase of those log-odds with increasing deviations from the average velocity.

Suppose that n bullets, of velocities $v_1, \ldots, v_n$ are fired at the vest, in such a way that the results are like those of n independent trials with probabilities of "success" $p_1, \ldots, p_n$, where $p_i = \exp\big(\beta_0 + \beta_1(v_i - \overline{v})\big)/[1 + \exp\big(\beta_0 + \beta_1(v_i - \overline{v})\big)]$ for $i = 1, \ldots, n$. If the results are $x_1, \ldots, x_n$, where each x_i is either 1 or 0, depending on whether the round penetrated the vest or not, then the maximum likelihood estimates of the parameters are $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that maximize $p_1^{x_1}(1-p_1)^{1-x_1} \ldots p_n^{x_n}(1-p_n)^{1-x_n}$ with respect to β_0 and β_1.

The following figure depicts some of our experimental data, and the logistic regression model that was fitted to them by the method of maximum likelihood. The green dots represent bullets that did not penetrate the vest, and the red dots represent bullets that did penetrate it: their abscissæare their velocities. The blue, sigmoid curve shows how the probability of penetration p varies with bullet velocity v. And v_{50} (marked by a red diamond near the horizontal axis), is found simply by following the arrows, as the velocity that corresponds to $p = 0.5$: analytically, $v_{50} = \overline{v} - \widehat{\beta_0}/\widehat{\beta_1} = 1\,658\,\mathrm{m/s}$. And since the maximum likelihood procedure also produces assessments of the covariance matrix of $\widehat{\beta_0}$ and $\widehat{\beta_1}$, a standard exercise in uncertainty propagation produces the standard uncertainty $u(v_{50}) = 16\,\mathrm{m/s}$, in this case.



The conventional procedure (MIL-STD-662F, V_{50} *Ballistic Test for Armor*) defines v_{50} as the average of an equal number (typically 3) of highest partial penetration velocities and the lowest complete penetration velocities which occur within a specified velocity spread. Since achieving this typically involves considerable trial-and-error, in practice this measurement of v_{50} consumes one or two vest panels and 10–20 shots.

The model-based approach to estimate v_{50} described above simplifies the experimental process considerably, because the only requirement is that the velocities tested should span the range of relevant velocities, and that the ranges of velocities of shots that do, or that do not penetrate, should overlap. Since this assumes that the logistic regression model is appropriate, its adequacy should be assessed in all instances of its application by examination of suitable diagnostics as are commonly employed when fitting generalized linear models.

Extensions

The logistic regression framework can accommodate other variables besides bullet velocity: for example, a model of the form $\log\big(p/(1-p)\big) = \beta_0 + \beta_1(v - \overline{v}) + \beta_2(w - \overline{w})$, where w denotes the number of layers of the microfiber in the vest, allows assessing the effects of bullet velocity (estimating v_{50} in particular), and vest structure simultaneously.

Similarly, survival regression models can be used to express the vest's probability of failure (to stop the class of threats that it was designed for) at any given point during its lifetime, as a function of its physical attributes (number of layers, cross-weaving patterns, polymer in its microfibers, etc.), and of the attributes of the specific threats that it may face (bullet mass, velocity, shape, etc.).

3 Stephany Bailey

Biography

Stephany L. Bailey was born in Indianapolis, IN, then moved to Albuquerque, NM. Several years later she was in the Washington, DC area. 1st National Bank was her first job; M. P. Foley & Co. (steam fitters and plumbers); computer sales at Anderson Jacobson, Inc.; Montgomery County Board of Realtors; Dept. of Defense; and came to NIST in 1989 where she has been in MSEL, Ceramics Division, and ITL, where now she is Division Secretary in the Statistical Engineering Division.

4 Lorna Buhse

Biography

Lorna Buhse was born in England and worked for the British Government as an administrator before immigrating to the United States. She worked for Storage Technology Corporation as an analyst before joining the National Institute of Standards and Technology, where currently she provides secretarial and administrative support to the Boulder wings of the Statistical Engineering Division and the Mathematical and Computational Sciences Division. She is a licensed pilot, an amateur radio technician, and she enjoys swimming, tennis and reading.

5 Kevin J. Coakley

Biography

Kevin J. Coakley earned a B.S. in Physics from Yale University, an M.S. in Physics from the University of Washington (Seattle), and a Ph.D. in Statistics from Stanford University. He joined the Statistical Engineering Division at NIST in 1989. Current research interests include aerosol physics; neutron physics; fast neutron spectroscopy; neutron imaging; astroparticle physics; and computational biology.

Awards

Fellow of the American Statistical Association (elected in 2005).

Japan Air Cleaning Association President's Award for best presentation for paper "Development of an Aerosol Particle Mass Analyzer" given at the 2007 Annual Technical Meeting on Air Cleaning and Contamination Control (May 9-10, 2007, Tokyo). (with N. Fukushima, N. Tajima, K. Ehara, and H. Sakurai).

Selected Publications

W. H. Lippincott, K. J. Coakley, D. Gastler, A. Hime, E. Kearns, D. N. McKinsey, J. A. Nikkel and L. C. Stonehill, Scintillation Time Dependence and Pulse Shape Discrimination in liquid Argon, *Physical Review C*, 78, 035801, 2008.

J.S. Nico, M.S. Dewey, T.R. Gentile, H.P. Mumm, A.K. Thompson, B.M. Fisher, I. Kremsky, F.E. Wietfeldt, T.E. Chupp, R.L. Cooper, E.J. Beise, K.G. Kiriluk, J. Byrne, K.J. Coakley, Observation of the Radiative Decay of the Free Neutron, *Nature*, 444, 1059–1062, 2006.

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

K.J. Coakley, A Cross-Validation Procedure for Stopping the EM Algorithm and Deconvolution of Neutron Depth Profiling Spectra, *IEEE Transactions on Nuclear Science*, 38, 1, 1991.

C.J. Horowitz, K.J. Coakley, and D.N. McKinsey, Supernova Observation via Neutrino-Nucleus Elastic Scattering in the CLEAN Detector, *Physical Review D*, 68, 023005, 2003.

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

6 Statistical Learning Methods for Neutron Transmission Tomography of Fuel Cells

AUTHOR Kevin Coakley
COLLABORATORS Dominic Vecchia (Statistical Engineering Division, ITL, NIST), Daniel Hussey, David Jacobson, Muhammad Arif (Ionizing Radiation Division, PL, NIST)

Introduction

In a proton exchange membrane fuel cell, water is produced when hydrogen and oxygen are combined to produce electricity. Because the reliability of a fuel cell depends on its water transport properties, nondestructive quantification of the spatial distribution of water density in a fuel cell is essential for engineering development. Since water has a high neutron scattering cross section, neutron imaging is an ideal method for measuring water in a fuel cell. For the dry and wet states of the fuel cell, there are spatially varying neutron attenuation images $\boldsymbol{\mu}_{\text{dry}}$ and $\boldsymbol{\mu}_{\text{wet}}$. Ideally, the residual attenuation image, $\boldsymbol{\Delta\mu} = \boldsymbol{\mu}_{\text{wet}} - \boldsymbol{\mu}_{\text{dry}}$, is proportional to the water density in the fuel cell. We reconstruct a residual attenuation image from joint analysis of the wet and dry state projection data collected at the NIST Neutron Imaging Facility with a penalized Poisson likelihood method with a Huber penalty function. We select the parameters in the Huber penalty function by two-fold cross-validation — a statistical learning method.

Reconstruction Method

Given dry and wet projection data, $\boldsymbol{Y}_d$ and $\boldsymbol{Y}_w$, collected during acquisition times $\boldsymbol{T}_d$ and $\boldsymbol{T}_w$, we estimate $\boldsymbol{\Delta\mu}$ by maximizing

$$\boldsymbol{\Phi} = \log \boldsymbol{L}(\boldsymbol{\Delta\mu}, \boldsymbol{Y}_w, \boldsymbol{Y}_d) - \boldsymbol{\beta} R(\boldsymbol{\Delta\mu}, \boldsymbol{\delta})$$

with a parabolic surrogates numerical method where

$$\log \boldsymbol{L}(\boldsymbol{\Delta\mu}, \boldsymbol{Y}_w, \boldsymbol{Y}_d) = -\widehat{\boldsymbol{Y}}_w + \boldsymbol{Y}_w \log \widehat{\boldsymbol{Y}}_w + \text{constant},$$

$$\widehat{\boldsymbol{Y}}_w = \frac{\boldsymbol{T}_w}{\boldsymbol{T}_d} \times \boldsymbol{Y}_d \times \exp(-\int_l \boldsymbol{\Delta\mu} \mathrm{d}l),$$

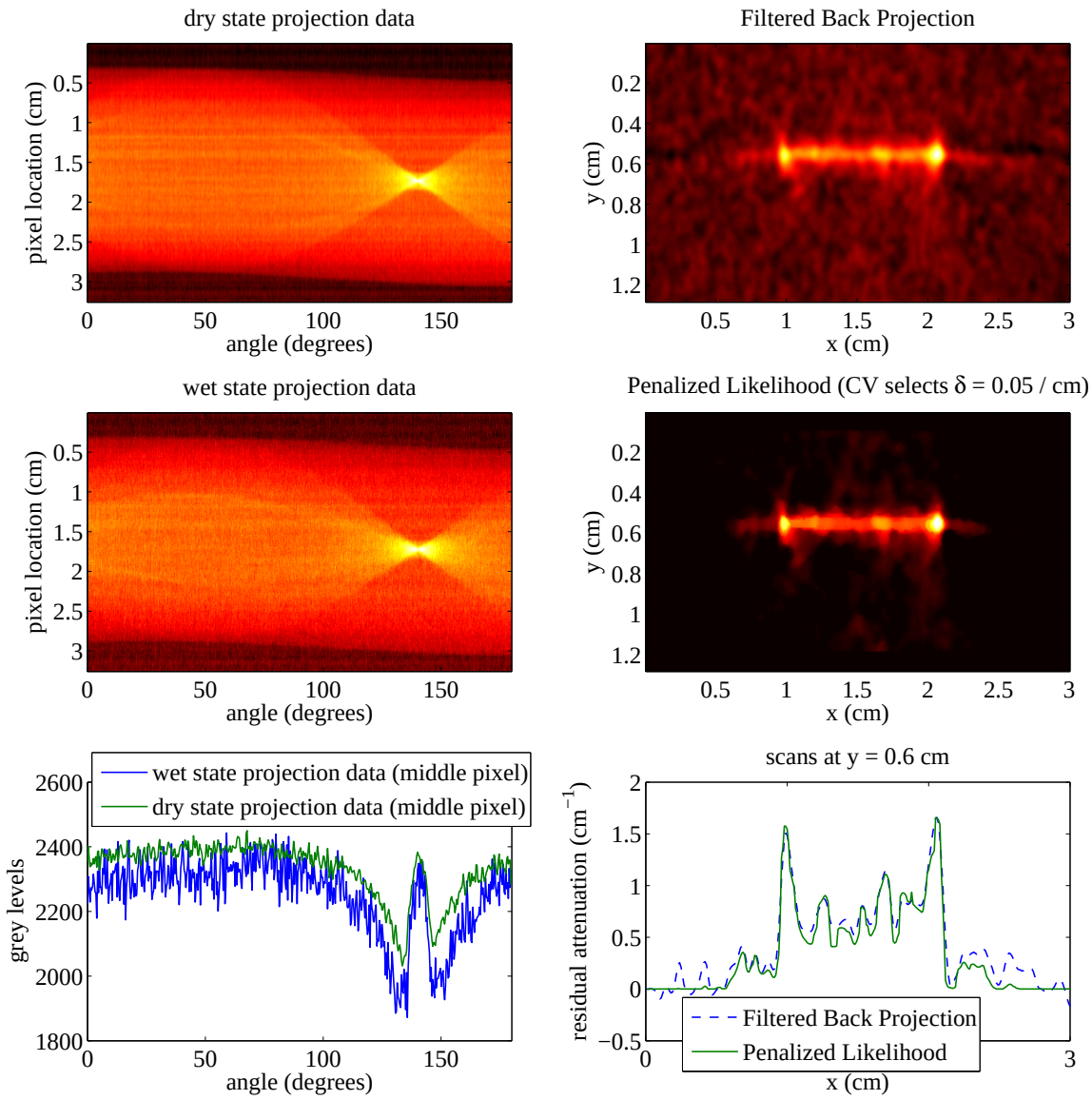
$$R(\boldsymbol{\Delta\mu}, \boldsymbol{\delta}) = \sum w_{kl} H(\boldsymbol{\Delta\mu}_k - \boldsymbol{\Delta\mu}_l, \boldsymbol{\delta}),$$

$$H(x, \boldsymbol{\delta}) = \begin{cases} x^2/2 & \text{if } |x| < \boldsymbol{\delta} \\ \boldsymbol{\delta}(|x| - \boldsymbol{\delta}/2) & \text{if } |x| \geq \boldsymbol{\delta} \end{cases}$$

and

$$w_{kl} = 1, \frac{1}{\sqrt{2}}, 0$$

for adjacent, diagonal, and non-connected pixels respectively. We select the regularization parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ by two-fold cross-validation based on "estimation" and "validation" projection data.



We collect projection data corresponding to both the dry and wet states of the fuel cell. We scale the projection data so that the ratio of the variance to expected value is approximately 1. Based on joint analysis of the projection data, we reconstruct residual neutron scattering attenuation images with a standard Filtered Back Projection method and a Penalized Likelihood method where the adjustable parameters in the Huber penalty function are determined by two-fold cross-validation. The Huber penalty function preserve edges in the reconstruction more effectively than a quadratic penalty function.

7 Radiative Decay of the Neutron

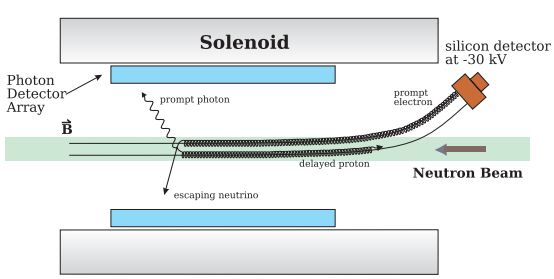
AUTHOR Kevin Coakley

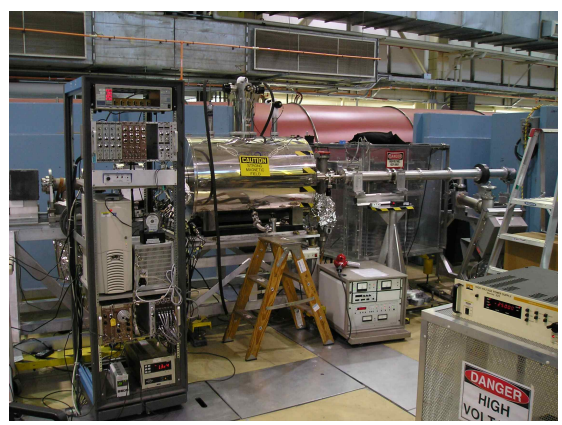
COLLABORATORS R.L. Cooper, T.E. Chupp (University of Michigan), C.D Bass, M.S. Dewey, B.M. Fisher, C. Fu, T.R. Gentile, H.P. Mumm, J.S. Nico, A.K. Thompson (Ionizing Radiation Division, PL, NIST), E.J. Biese, H. Breuer, M. McGonagle (University of Maryland), J. Byrne (University of Sussex), F.E. Wietfeldt (Tulane University)

Introduction

When a free neutron decays, a proton, electron and antineutrino are produced. According to the theory of quantum electrodynamics (QED), there is a small probability that neutron decay produces an inner bremsstrahlung photon with an energy spectrum that falls off rapidly with energy. In 2005, the radiative decay of the free neutron (RDK) was observed for the first time at NIST (`http://www.nature.com/nature/journal/v444/n7122/pdf/nature05390.pdf`).

In this proof-of-principle experiment (RDK I), extraction of the very weak RDK signal from an intense photon background was possible because the RDK photon, electron, and proton produced by each neutron decay event were simultaneously detected. From RDK I, we determined that the branching ratio was $(3.13 \pm 0.34) \times 10^{-3}$ in the energy region between 15 keV and 340 keV. In a ongoing second generation version of the experiment (RDK II), the signal-to-noise ratio is greater because there are twelve scintillator blocks rather than just one scintillator block as in RDK I. Further, due to improved shielding, backgrounds are less intense. We expect to reduce the uncertainty in the branching ratio estimate to 1 percent and get a more precise measurement of the RDK photon energy spectrum.

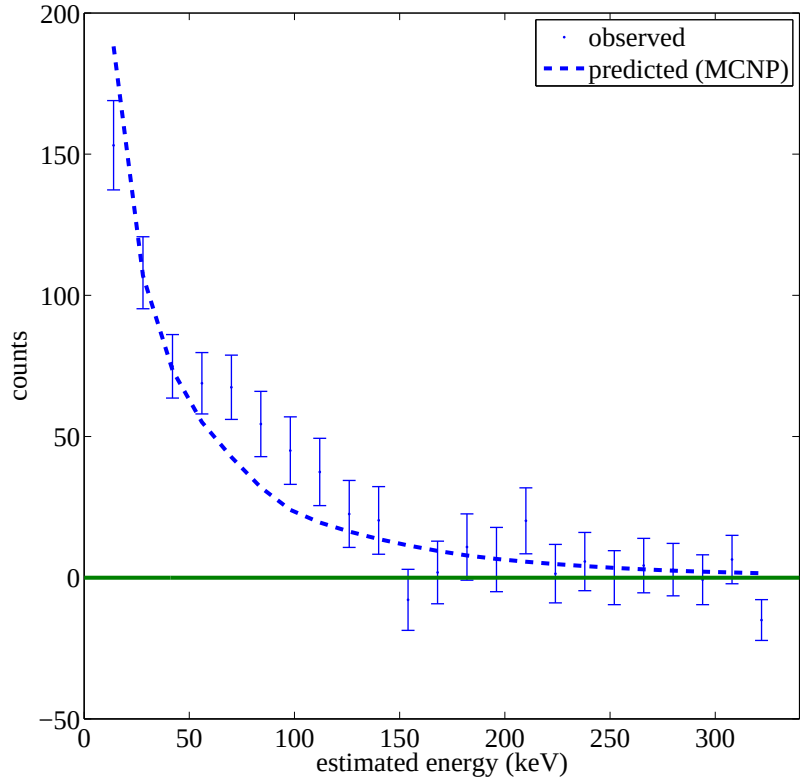




Schematic and photo of RDK II experiment at the NIST Center for Neutron Research. RDK photons that deposit energy in BGO scintillator blocks produce scintillation photons. These scintillation photons are detected by avalanche photodiodes connected to the scintillator blocks.

Modeling and Analysis Efforts

Current SED efforts include: modeling and analysis of calibration experiments and observed data from RDK I and RDK II and development of procedures to test the consistency of QED with our data. The task is challenging because: RDK photons can lose energy due to scattering off materials outside the scintillator; an RDK photon can deposit a fraction of its total energy in the scintillator; the number of scintillation photons produced by a energy deposit is random; and the electrical signal produced by a scintillation photon at a avalanche photodiode depends on amplification noise as well as additive noise.



In a preliminary analysis of RDK I, we simulate the energy deposit spectrum of RDK photons in the BGO scintillator block with a computer code MCNP. Given the simulated energy deposit spectrum, we predict the expected observed energy spectrum based on an assumed probability transition matrix. This transition matrix relates the estimated energy of a photon to the unknown energy deposit and accounts for counting statistics fluctuations in the number of scintillation photons produced by a given energy deposit, quantum efficiency effects, and amplification and additive noise in the avalanche photodiode.

8 James J. Filliben

Biography

Jim Filliben received his B.A. in Mathematics from LaSalle College (1965) and his Ph.D. in Statistics from Princeton University (1969). He then joined NIST for a career that currently spans 39 years.

His research interests include exploratory data analysis, statistical graphics, experiment design, distributional modeling, time series analysis, computational modeling (validation & verification, calibration, and optimization), and sensitivity and uncertainty analysis. He has been an energetic teacher and educator, and an eclectic collaborator in the widest range of scientific and engineering projects. He has published more than 60 technical papers, given more than 250 talks, and taught many courses inside and outside of NIST.

Jim has led SED teams in the production of web-based references that include *StRD: Standard Reference Datasets for Statistical Software Testing*, and the *NIST/SEMATECH e-Handbook of Statistical Methods*. Some of his high-profile NIST projects include Selective Service Draft Lottery, HUD Operation Breakthrough Total Energy Project (BFRL), DOT Daylight Saving Time Study, NIJ Ballistics Database Feasibility Study (MEL & EEEL), and World Trade Center Collapse Analysis (BFRL).

Awards

ASA Fellowship (2003), Department of Commerce Bronze (1981), Silver (2003), and 2 Gold Medals (1984, 2003), and ASA Youden Award (2003).

Selected Publications

(2008). J.G. Hagedorn, J. E. Terrill, A.P. Peskin, J.J. Filliben, "Methods for Quantifying and Characterizing Errors in Pixel-Based 3D Rendering", *NIST Journal of Research*, 113 (4), 221–238.

(2008). H. Kurosaki, R. Radford, J.J. Filliben, K.G.W. Inn, "An Orthogonal Design of Experiment / Exploratory Data Analysis for Plutonium Contamination", *Journal of Radioanalytical and Nuclear Chemistry*, 276 (2), 323–328.

(2007). T.V. Vorburger, J.H. Yen, B. Bachrach, T.B. Renegar, J.J. Filliben, L. Ma, H.G. Rhee, A. Zheng, J. Song, M.A. Riley, C.D. Foreman, S. Ballou, "Surface Topography Analysis for a Feasibility Assessment of a National Ballistics Imaging Database", NISTIR 7362.

9 Measurement Science for Complex Information Systems

AUTHOR James J. Filliben
COLLABORATORS Kevin Mills (Advanced Network Technologies Division, ITL, NIST)

Introduction

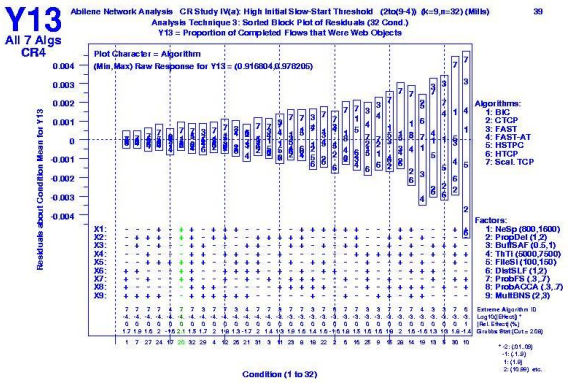
A Complex System is any structure, process, organization, or organism characterized by a large number of interconnected components whose interactions induce collective behavior that is not predictable from observation of its constituent parts alone. Examples of complex systems include physical (phase transitions), biological (ant colonies, slime molds, humans), technological (computer networks), and social (economies, transportation networks) systems.

Since 2007, Kevin Mills has been leading an interdisciplinary research program on "Measurement Science for Complex Informational Systems". The problem being addressed by this project was that there is "no science today that offers the fundamental knowledge necessary to design large complex networks so that behaviors can be predicted prior to building them" (Network Science, 2006 NRC Report).

For a complex system whose definition typically comprises a very large number of factors, and which can generate a correspondingly large number of potential responses, several questions arise: what should be measured, and how should such measurements be analyzed?; how should the results of such analysis be translated into system insight, understanding, and predictability? The research focuses on a paradigmatic example of complex systems: the Internet.

Statistical Framework

In practice, the same generic 5-step statistical approach illustrated in the Figure that has provided the framework for many advances in NIST's physical science projects also provided the framework to address the complex information sciences problem at hand. Further, since complex systems may be viewed from a statistics point of view as multi-response, multi-factor, time series, then tools from all three of these areas may be applied.



Questions

Component 1 in the analytical framework is the elicitation of specific questions, including: What are the important factors that influence Internet congestion? Do these factors interact? Can the common Internet congestion-control algorithm TCP be improved on? Which is the best

among the handful of alternatives that have been proposed for for TCP? Under what conditions is this best? Are our conclusions robust over variations in network parameters and user conditions? The statistical framework for this project also provides an answer to the validation question: are our own Internet simulation models adequate? (The answer is "yes", but the very asking of the question has led to improvements of the simulator).

Experiment Design

Component 2 (experiment design) in the 5-step framework played a particularly important role. It allowed for a common vocabulary, it provided a simplifying 2-element structure (the number k of factors to consider, and the number n of Internet experiments that can be afforded), it forced the specificity of purpose required to translate any of the infinity of possible Internet questions into the specific, concrete question that the "next experiment" was going to address, and it took advantage of the interactive statistical process of starting with an amorphous problem with an "endless" number of (mostly continuous) factors and converging to a finite, workable, scientifically-prioritized subset of factors with well-defined discrete settings. It also opened up a new way of thinking and a new, powerful tool at their disposal, namely the orthogonal fractional factorial experiment design, by which the IT scientist may efficiently, effectively, and systematically probe an information system as complex as the Internet.

The experiments were carried out by running Internet simulation programs under specified conditions. Each simulation run consumes a good deal of "wall clock" time. The project testing became feasible only by combination of the following 2 tools (one statistical and one computational): (i) the fractional factorial experiment designs reduced the decades' worth of running that would otherwise be required into one year; and (ii) the availability of multiple processors allowed distributing the computational load, thereby reducing the one year's worth of processing into a week. Both components were critical.

Data Analysis

Furthermore, the necessities of dealing with component 4 (data analysis) for systems with a large (15 to 50) number of responses — each one sensitive to a different aspect of Internet behavior — led to the design and application of a variety of custom graphical data analysis techniques to extract from the multi-factor and multi-response data the maximal amount of underlying structure and insight into primary factors and interactions alike.

Conclusion

The leadership of world-class IT Internet expert with extensive computational skills, combined with the generic 5-step statistical framework, plus specific, powerful statistical design and analysis tools, has garnered rich insight about the inter net's functioning. This will be summarized in an extensive report, now under preparation. The complex information system results and the corresponding methods for statistical design and analysis will be of considerable interest to the IT measurement science community generally.

10 Design of Experiments Approach to Verification and Uncertainty Estimation of Simulations Based on Finite Element Method

AUTHOR James J. Filliben

Introduction

NIST scientists and engineers deal with physical phenomena and processes (e.g., growth of biological cells), with in-lab physical devices (e.g., scatterfield microscope) and with out-of-lab "real" objects (e.g., building deflection in high winds). In all cases one aims to characterize, measure the sensitivity to influential factors, and optimize such processes, devices, or objects. In some studies, achieving such goals is a major challenge because physical experiments may be difficult to perform, or too expensive or time-consuming (e.g., to optimize a scatterfield microscope); occasionally, physical experiments are impracticable (e.g., to replicate the World Trade Center collapse).

Computational Models

To address this problem, NIST scientists have developed computational models that can serve as surrogates for physical experiments. Such models will either emulate an observable physical phenomenon, or predict a reality that we could not observe. The computational model-building process has three components (Figure 1): (i) the "real" phenomenon itself; (ii) the corresponding mathematical model (e.g., a set of partial differential equations) representing the phenomenon; and (iii) the computational approximation required to implement the mathematical model in practice.

Building such computational models is extremely difficult — major errors can be introduced in the mathematical modeling stage, which may then be exacerbated by approximations, truncations, and algorithmic deficiencies in the computational modeling stage.

Finite-Element Analysis (FEA)

The most common computational model is finite-element analysis (FEA) — it underlies many NIST computational models. For example, BFRL's Fire Dynamics Simulator (FDS) and the Virtual Cement and Concrete Testing Lab (VCCTL); MEL's model for the optimization of a scatterfield microscope; ITL's simulator of Internet traffic and congestion.

Trusted Model

Regarding the use of computational models, two situations arise. First, if the scientist accepts the computational model as being an adequate surrogate, then the data from virtual experiments replace the data from physical experiments, and the scientific goals (characterization, sensitivity analysis, optimization, etc.) remain the same. In this case, the statistical tools for experiment design and data analysis that are used for physical experiments continue to be

appropriate, although the number of factors that the experiment can accommodate is usually larger in virtual than in physical experiments because the effort, cost, and time to run the former usually are smaller than running the latter.

Similarly to many physical experiments, 2-level orthogonal fractional factorial designs have proved to be invaluably efficient and insightful, especially for the typical first goal (when dealing with a large number of factors) of determining which are the most important factors and interactions. Examples of (k, n) where k denotes number of factors and n denotes the number of runs are: World trade center plane impact core column damage ($k = 11$, $n = 17$); scatterfield microscope optimization ($k = 7$, $n = 16$); Internet traffic and congestion simulator ($k = 11$, $n = 32$). Figure 2 gives an example of one of the steps in the usual 10-step graphical procedure that consistently provided valuable insight (main effects plot for the World Trade center core column damage).

Untrusted Model: Verification & Validation (V&V)

The second situation arises when the scientist does not trust the computational model. In this case, the quality of the model needs to be addressed directly, under the general purview of "V&V" (verification and validation) as defined by Oberkampf *et al.* (2002), and engenders two additional questions: (i) Does the computational model match its alleged mathematical description? (verification); and (ii) Does the computational model match reality? (validation).

Verification tends to be the easier of the two problems — especially when benchmarks and standards exist. But even in the absence of standards, not all is lost, as we have found in a comparative study of commercial FEA programs applied to the mechanical performance of a nano-cantilever: a few simple 3-factor designs sufficed to show that they were different from one another.

Regarding validation when no physical data exists, we have found useful to replace the validation question ("Do the results of virtual experiments reproduce those from physical experiments?") temporarily with the sensitivity question ("Which factors do the computer experiments suggest are important?"). The results from such sensitivity analysis allow a subject matter expert to assess the computational model by comparing results with what experience has led her to expect about factor ranking and optimal settings.

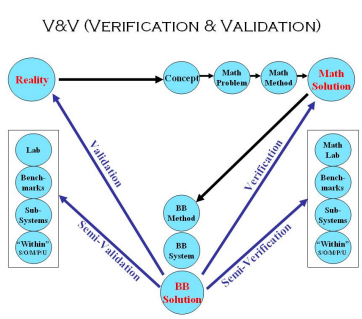


Figure 1

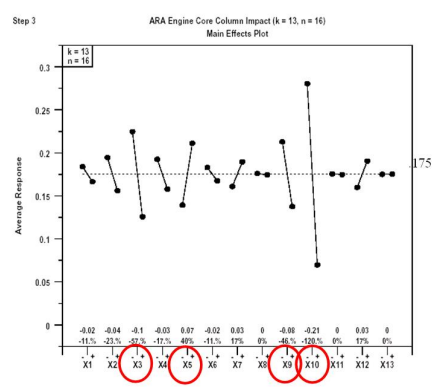


Figure 2

11 William F. Guthrie

Biography

William F. Guthrie received a B.A. degree in mathematics from Case Western Reserve University, Cleveland, OH, in 1987 and an M.S. degree in statistics from The Ohio State University, Columbus, OH, in 1990. He joined the Statistical Engineering Division at the National Institute of Standards and Technology (NIST), Gaithersburg, MD in 1989.

He has collaborated with NIST scientists and engineers in a wide range of areas, applying statistical methods to solve problems in semiconductors and microelectronics, building materials research, and chemical science. His statistical interests include uncertainty assessment, Bayesian statistics, design of experiments, calibration, modern regression methods, and statistical computation.

Awards

NIST Measurement Services Award, 1992, Allen V. Astin Measurement Science Award, 1994, Department of Commerce Bronze Medal, 1997, Department of Commerce Silver Medal, 2003, Department of Commerce Silver Medal, 2004.

Selected Publications

"Combining Data in Small Multiple Method Studies" (with C.R. Hagwood), *Technometrics*, 2006.

"Comparison of SEM and HRTEM CD Measurements Extracted from Test-Structures Having Feature Linewidths from 40 nm to 240 nm" (with M.W. Cresswell, R.A. Allen, C.E. Murabito, R.G. Dixson, A. Hunt), *IEEE Transactions on Instrumentation and Measurement*, 2007.

"Interlaboratory Comparisons," *Encyclopedia of Statistics in Quality and Reliability*, F. Ruggeri, R.S. Kenett, F.W. Faltin (eds.), John Wiley and Sons, 2007.

"A Gravimetric Approach to the Standard Addition Method in Instrumental Analysis," (with W.R. Kelly, B.S. MacDonald), *Analytical Chemistry*, 2008.

"Determination of Sulfur in Biodiesel and Petroleum Diesel by XRF using the Gravimetric Standard Addition Method," (with L.R. Barker, W.R. Kelly), *Energy and Fuels*, 2008.

12 Comparison of Clinical Methods with Isotope Dilution Measurements of Blood Lead Levels

AUTHOR William F. Guthrie
COLLABORATORS Karen E. Murphy, Thomas W. Vetter, Gregory C. Turk (Analytical Chemistry Division, Chemical Science and Technology Laboratory, NIST), Christopher D. Palmer, Miles E. Lewis, Jr., Ciaran M. Geraghty, and Patrick J. Parsons (Wadsworth Center, New York State Department of Health, Albany, New York)

Introduction

Lead is an environmental toxin that can damage several organs of the human body. Acute exposure at blood lead levels (BLLs) $\geqslant 70\,\mu$g/dL can be fatal, whereas, chronic, low-level exposure, at BLLs $\leqslant 10\,\mu$g/dL, is associated with decreased neurocognitive function in young children. Over the last three decades, efforts to reduce or eliminate environmental exposure to lead, by banning the use of leaded-solder in canned food containers, and lead in gasoline and residential paint, have resulted in a substantial decrease in children's BLLs in the U.S. Despite this decrease, however, over 74000 children (3 %) out of the 2.4 million children tested in 2001, had BLLs greater than the $10\,\mu$g/dL threshold, and over 7000 (0.4 %) had BLLs greater than $25\,\mu$g/dL.

Assessment of lead poisoning requires accurate and reproducible analytical measurements of lead abundance in blood over a broad range of concentrations. Regulations require all clinical labs operating in the U.S. to participate in proficiency testing (PT) that is approved by the U.S. Centers for Medicare and Medicaid Services (CMS). Laboratories must report satisfactory results for at least four out of five samples (challenges) in at least two out of three consecutive test events per year. The range of results deemed satisfactory for blood lead PT are broad: $\pm 4\,\mu$g/dL or ± 10 % of the established target value, whichever is greater.

The New York State Department of Health's (NYSDOH) Wadsworth Center has operated a PT program for blood lead analysis for over thirty years. The NYSDOH PT program utilizes whole goat blood containing endogenous lead. Target values of PT samples are established by 15 or more well established reference laboratories using methods routinely applied to blood lead analyses.

Certified reference materials (CRMs) are used for initial method validation in the reference laboratories and to investigate potential sources of bias in discrepant results. Internal quality control (QC) and performance in several external quality assessment (EQA) schemes are also used to track method performance. However, laboratories that rely on internal QC and EQA procedures alone may produce measurements that are reproducible, but biased. CRMs must be measured periodically to re-examine method performance.

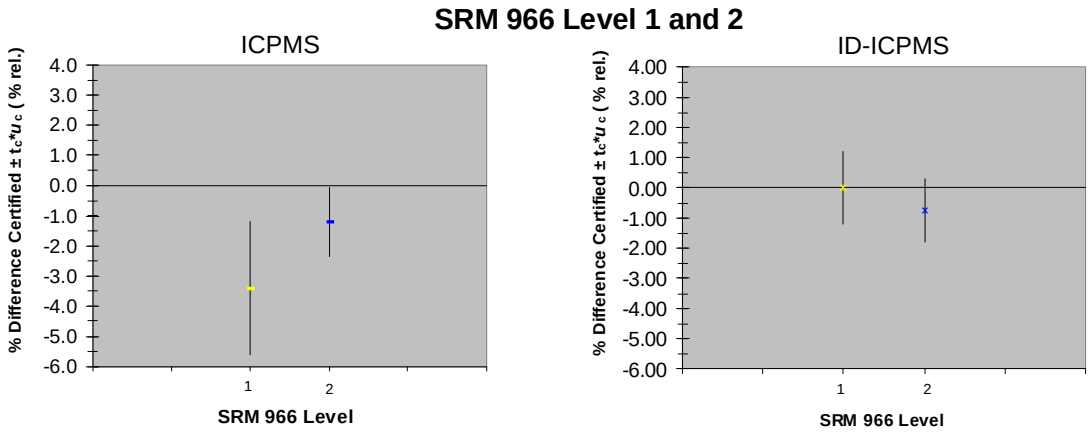
Recently, to provide a new standard with lower BLLs and to have a standard not based on bovine blood, NIST and the Wadsworth Center collaborated on the production of SRM 955c, Lead in Caprine Blood. In addition to using the measurements made at each laboratory for certification of the reference material, this work also offered an opportunity to compare the clinical measurements made using inductively-coupled mass spectroscopy (ICP-MS) with the

more accurate, but slower, isotope-dilution ICPMS method (ID-ICPMS). These comparisons identified a small potential bias in the ICPMS data that requires further investigation and could lead to small improvements in the measurement method(s).

Comparison of Results

The measurement methods were compared using several materials, the new SRM 955c and two control materials, because more data was available for the control materials. In particular, SRM 966 Toxic Elements in Bovine Blood, measured over several weeks with the clinical methods, provided the most precise comparison owing to the significant day-to-day variation in the clinical measurements. The figure shows differences between the results for each measurement method and the certified value along with the expanded uncertainties of the differences at the 95 % confidence level. The differences between the ICP-MS measurements and the certified values are statistically significant while the differences between the control measurements of SRM 966 using ID-ICPMS and the certified value were not.

Although the comparisons made with the other materials using the data collected during the certification process were not statistically significant individually, the differences tended to be in the same direction, suggesting that more precise comparisons might also indicate similar small biases. Subsequent comparison of proficiency test results using an expert group of 18 clinical laboratories and all 90 clinical laboratories also indicated that the clinical measurements are likely to be lower than the NIST results for the upper two levels of SRM 955c. In all cases, however, the clinical results and the certified lead levels for each material were still in agreement relative to the mandated clinical guidelines.



Comparison of the certified lead concentration in SRM 966, Toxic Elements in Bovine Blood, with clinical lead measurements made using ICPMS (left) and measurements made using ID-ICPMS method (right). Although there is a small apparent bias in the ICPMS results, they are still well within the guidelines for clinical acceptability.

13 Educational Outreach in Statistical Metrology

AUTHOR William F. Guthrie
COLLABORATORS Ana Ivelisse Aviles, James Filliben, Dennis Leber, Stefan Leigh, Walter Liggett, Hung-kung Liu, John Lu, Antonio Possolo, Andrew Rukhin, Blaza Toman (Statistical Engineering Division, ITL, NIST)

Overview

As part of the Division's mission, many staff members teach short courses and workshops on a range of statistical topics, for a number of different audiences each year.

Internally taught short courses generally target NIST staff, although they sometimes draw local attendees from outside of NIST as well. The short courses are of varying duration and depth, but are designed to cover topics in statistics, probability, data analysis, and statistical computing relevant to NIST scientific staff at levels appropriate for all staff, from technicians to senior scientists. Each short 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 short course is to help researchers recognize opportunities for the use of particular statistical methods and to offer practical guidance in their application.

Externally taught short courses generally target either industrial scientists and metrologists or statisticians. The majority of external short courses are taught at conferences or other events open to the public, but the Division also sometimes teaches courses for individual government organizations as well. Externally-taught short courses generally cover more specialized topics than those taught internally. Uncertainty analysis, both based on the ISO *Guide to the Expression of Uncertainty in Measurement* and using Bayesian methods, is one topic of major interest. Other topics include statistical methods for advanced mass metrology, Bayesian analysis of physical science data using MCMC, and experiment design. Expenses for externally-taught workshops are usually covered in part (travel) or whole (travel and time) by the sponsoring organization.

New Ideas and Directions

New ideas to keep our curriculum fresh and involve more staff members include a new focus on the use of R and the possible addition of a high-level survey course with each topic taught by a different staff member. One of the goals of a survey course would be to use short sessions (1 to 2 hours) to introduce NIST scientists to different analysis methods or types of experiment designs that might not be familiar to them. Based on the interest in the different topics presented, we would plan to offer more in-depth classes that would allow our colleagues outside of statistics to start to use these methods on their own.

Our focus on R is centered on both the development of an R package with metrological tools, described elsewhere in this document, and presentation of classes that will demonstrate R functionality and allow participants to practice using R via various user interfaces in class. For statistical methods that are familiar, these classes will focus only on how the computations and visualizations can be accomplished in R. For less familiar techniques, the computations using R

will be integrated into more general discussions on the proper use and interpretation of those methods.

Another new method that some SED members are trying to reach out to the industrial metrology community is to register as instructors willing to offer training on statistical topics in metrology with NCSLI, one of the main professional societies for metrologists. NCSLI shares their database of registered instructors with its chapters around the country to help facilitate regional training. The first regional training event taught by a registered instructor was a workshop on uncertainty analysis taught by Will Guthrie and given as part of a Twin Cities Regional NCSLI meeting.

Staff of the Department of Homeland Security working on a hands-on exercise to maximize the passage time of a steel ball through an inclined funnel as part of a workshop on experiment design in September 2007 given by Jim Filliben and Dennis Leber.

14 Three Statistical Paradigms for the Assessment and Interpretation of Measurement Uncertainty

AUTHORS William Guthrie and Nien-Fan Zhang
COLLABORATORS Hung-kung Liu, Andrew L. Rukhin, Blaza Toman, Jack C.M. Wang (Statistical Engineering Division, ITL, NIST)

Introduction

The adoption of the ISO *Guide to the Expression of Uncertainty in Measurement* in 1992 has led to an increasing recognition of the need to include uncertainty statements in measurement results. Some of the strengths of the procedure outlined and popularized in the ISO *Guide* are its standardized approach to uncertainty, its accommodation of sources of uncertainty that are evaluated either by statistical data analysis (Type A) or by other methods (Type B), and its emphasis on reporting all sources of uncertainty that have been considered. The main approach to the propagation of uncertainty advocated by the ISO *Guide*, linear approximation of the formula used to obtain a measurement result, is simple to carry out and in many practical situations gives results that are surprisingly similar to those obtained using more sophisticated statistical methods.

Through their work over the years, statisticians have developed various paradigms for statistical inference that are relevant to uncertainty assessment as well: all offer firm probabilistic interpretations of the assessments that they produce, and the assessments themselves may be numerically similar, even though those interpretations are markedly different. To help make some of these methods more generally accessible to the metrology community, SED staff members recently published a comparative discussion of such methods as a chapter in the book *Data Modeling for Metrology and Testing in Measurement Science* edited by Franco Pavese of the National Institute of Metrological Research in Italy and Alistair Forbes of the National Physical Laboratory in the UK. The goals of the chapter were to present different statistical approaches to uncertainty assessment, discuss the interpretations of the uncertainty intervals they produce, and to relate them to the methods that are currently being used or developed within the metrology community.

Approach

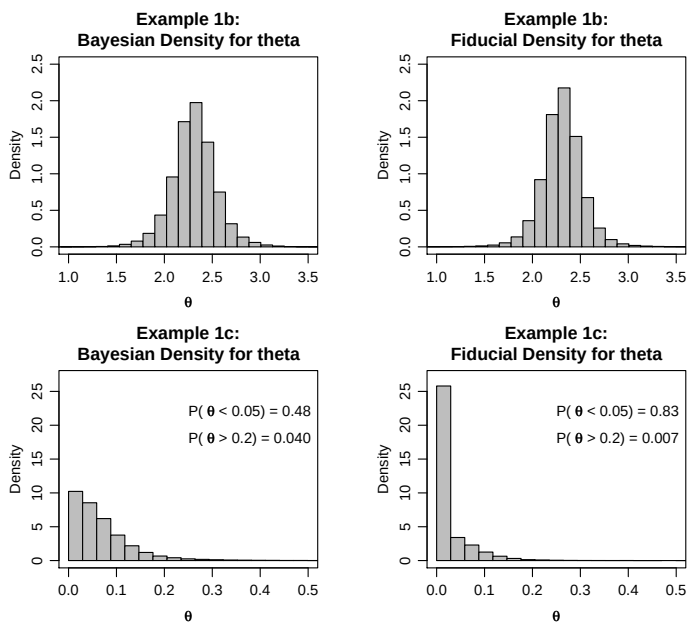
The three statistical approaches considered were the frequentist, Bayesian, and fiducial paradigms. Each was introduced with some general background on the philosophy and underlying assumptions about the statistical models being used, illustrations of the types of computations that might be done (e.g. use of the bootstrap to obtain approximate confidence intervals), and discussion of the interpretation of the results that would be obtained using each approach.

Each of the approaches was also illustrated in detail using two numerical examples. All of the data and computer code necessary for each example were provided so that readers could easily replicate the computations, if desired.

One of the two examples, calibration of an end gauge, was chosen to illustrate the application of each method for a typical uncertainty analysis with many different sources of uncertainty

as might arise in metrological work. The other was a simple, conceptual example chosen to illustrate some of the potential differences between the methods. The goal of this second example was to assess the value of a scalar physical quantity, θ, based on measurements that must be corrected for the presence of additive background interference, β. The data for this example was a series of measurements of the signal plus background, assumed to be normally distributed with mean $\theta + \beta$ and standard deviation σ. The background was assessed using expert judgment (a Type B method) as being uniformly distributed between 1.125 and 1.329.

The figure compares the results of Bayesian and fiducial uncertainty analyses for two different sets of data for this example. In the first row of the figure, the background, β, is well below the signal of interest, θ, and the two analyses give similar results. In the second row of the figure the signal is just above the background and the results are somewhat different from one another. The difference arises from the ways in which the two analyses incorporate the physical constraint that the signal plus background must be larger than the background alone.



Comparison of Bayesian and fiducial probability densities from which uncertainty intervals are to be obtained for two different scenarios, in a simple example where measurements of a signal of interest are corrected for the presence of an additive background assessed using expert judgment.

Unlike the Bayesian or fiducial approaches, frequentist uncertainty intervals are not obtained from a probability distribution for the measurand. Instead intervals with specified confidence levels, or long-run probabilities, are obtained. The table compares approximate frequentist confidence intervals (first 3 columns) with the Bayesian and fiducial results depicted in the figure.

	ISO GUIDE	EISENHART	BOOTSTRAP	BAYES	FIDUCIAL
EXAMPLE 1b	$(1.90, 2.72)$	$(1.78, 2.84)$	$(1.86, 2.64)$	$(1.87, 2.75)$	$(1.87, 2.75)$
EXAMPLE 1c	$(0.00, 0.12)$	$(0.00, 0.20)$	$(0.00, 0.11)$	$(0.00, 0.19)$	$(0.00, 0.14)$

Expanded uncertainty intervals constructed under the three statistical paradigms.

15 Charles Hagwood

Biography

Charles Hagwood received his B.S. degree in mathematics from A&T State University, Greensboro, NC. Afterwards, he attended the University of Michigan, graduating in 1979 with a Ph.D. in mathematics, writing a thesis entitled "Discrete Nonlinear Renewal Theory", under Michael Woodroofe. Between 1979 and 1981, Hagwood was a John Wesley Young Research Instructor in the Mathematics Department at Dartmouth College. During 1981-1987, he was an assistant professor in the Mathematics Department at the University of Virginia. In 1984, he received a Ford Foundation Fellowship and spent one year at Stanford University, in the Statistics Department. He works at NIST since 1987, providing consulting in areas that include reliability, uncertainty, and stochastic processes.

Awards

1990 Andrew R. Chi Prize Paper Award with Grace Yang and Michael Souders, given by the Instrumentation and Measurement Society of the IEEE.

Selected Publications

An Application of the Residue Calculus: The Distribution of the Sum of Non-Homogeneous Gamma Variates. *Am Statistician* 2008.

Combining Data in Small Multiple Method Studies (W. Guthrie), *Technometrics* 2006.

Reliability of Conformance Tests (with Rosenthal, L.S.). *IEEE Trans on Reliab* 2001.

The DMA Transfer Function with Brownian Motion, a Trajectory/Monte Carlo Approach (with Sivathanu, Y. and Mulholland, G.). *Aerosol Sci Tech* 1999.

Exits in Multistable Systems Excited by Coin-Toss Square-Wave Dichotomous Noise A Chaotic Dynamics Approach (with Simiu, E.). *Phy Rev E* 1995.

An Unreliable Server Characterization of the Exp Distr (with Galambos). *J App Prob* 1994.

The Calibration Problem as an Ill-Posed Inverse Problem. *J. Stat. Planning Inf.* 1992.

The Effects of Timing Jitter in Sampling Systems. *IEEE Trans. Instrum. Meas.*, 1990.

A Multidim CLT for Maxima of Normed Sums (with Teicher, H.). *Ann. of Prob.* 1983.

On the Expansion of the Expected Sample Size in Nonlinear Renewal Theory (with Woodroofe, M.). *Ann. Prob. 1982*.

16 Langevin Dynamics for a Nanorod in an Electric Field

AUTHOR Charles Hagwood
COLLABORATORS George Mulholland (Department of Mechanical Engineering, University of Maryland, College Park, and Fire Research Division, BFRL, NIST)

Introduction

Carbon nanotubes have a very broad range of remarkable electronic, thermal and structural properties arising from their unique atomic structure. Industrial applications include electronic devices e.g., nanoelectric motors, and super strength fabrics and materials. These extraordinary properties depend on features of the nanotube, such as its diameter, length and twist.

The lengths of carbon nanotubes have recently been determined based on electrical mobility measurements. In one approach, the data have been analyzed assuming that the orientation of the charged nanorod has a Boltzmann probability distribution. It is of interest to model the orientation and translational motion of the charged rod in an electric field using a Langevin equation and to study the dynamics from this point of view. This approach allows us to compute the trajectory of the rod for short time where the Boltzmann distribution does not apply. After a long period of time, the system eventually reaches equilibrium and as it does, the orientation probability distribution converges to the Boltzmann distribution.

Our goal is to solve the Langevin equation. The rotational and translational motions are coupled, because the friction coefficient depends on the orientation of the rod relative to the electric field. As a first step in this analysis, we consider the model problem of a nanorod constrained to diffuse in only the x and y directions and rotate only about the z axis. The force and torque acting on the nanorod are computed assuming a singly charged nanorod. The friction coefficient is taken to be the free molecular value for a rod and the rotational resistance is also based on free molecular dynamics. This set up reduces to solving the Langevin equation for the motion of the center of mass of the nanorod between two parallel plates distance d apart, one charged with a voltage V and the other grounded, thus an electric field $\vec{E}$ is created between the plates, $|E_y| = qV/d, |E_x| = 0$.

Only in restricted cases can the Langevin equation be solved in closed form. For example,when the linear friction coefficient is independent of orientation, the Langevin equation reduces to the Ornstein-Uhlenbeck process and the transition probability is given in closed form. We have used numerical integration methods for solving Langevin equations and the restricted Ornstein-Uhlenbeck case is used to check our integration. The time for the nanorod to rotate on the order of $\pi/90$ radian sets the time increment for the integration. The basic quantity of interest is the time evolution of the probability distribution of the nanorod as a function x, y, and angle, θ. We are interested in the small time behavior where the Brownian motion has a significant effect, as well as, the long time behavior where the distribution is expected to approach a Boltzmann distribution. We study the effect of the nanorod length, diameter, and the field strength on the time dependent electrical mobility.

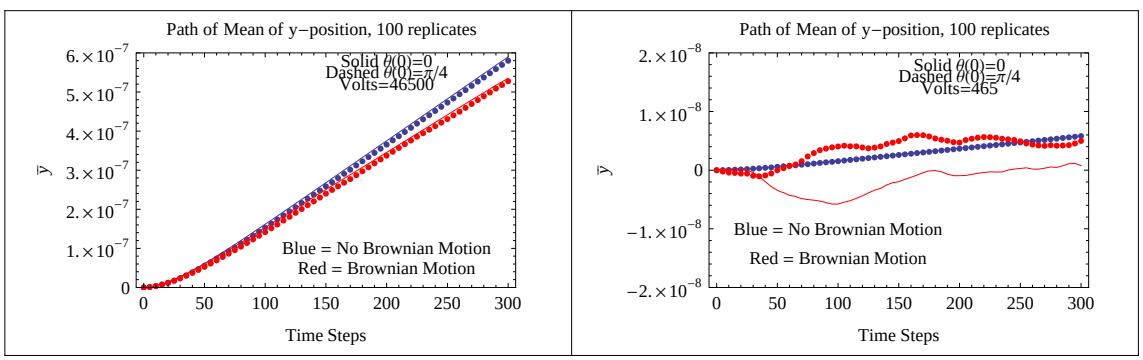
Langevin Dynamics

The electric field $\vec{E}$ and the charge q on the rod induce a torque $\tau = |\vec{r}||\vec{F}_e|\sin\theta = ½L_f q|\vec{E}|\sin\theta$ that orients the rod, where L_f is the length of the rod and θ is the angle the rod makes with the field $\vec{E}$, $0 \leq \theta \leq \pi/2$. Rotational and translational motion are described by the system of stochastic differential equations

$$\begin{aligned} I\ddot{\theta} + R_d\dot{\theta} + \tau &= X_z(t) \\ m\ddot{x} + F_D(\theta)\dot{x} &= X_x(t) \\ m\ddot{y} + F_D(\theta)\dot{y} - q\,|\,\vec{E}\,| &= X_y(t) \end{aligned}$$

where R_d is the angular friction coefficient, $F_D(\theta)$ the linear friction coefficient and $X_x(t), X_y(t)$ and $X_z(t)$ are white noise random forces related to Brownian motion, all having zero means and variances D_x, D_y, D_θ, where these diffusion coefficients are derived according to Einstein's formula ($D = E[X(t)X(t')] = 3m\beta kT\delta(t-t')$), where m is the mass of the rod, T is the temperature of the medium, k is Boltzmann's constant and β is a constant depending on the friction coefficient.

In the graphics below, the effects of initial orientation, voltage and Brownian motion on the y position of the rod are studied. In one case, the rod starts off completely aligned with the field, $\theta(0) = 0$ and in another case its initial orientation is midway between completely aligned and non-aligned, $\theta(0) = \pi/4$. The effects of low voltage 465, high voltage 46500, Brownian motion and no Brownian motion are considered.



Position of the y coordinate with small voltage, with and without Brownian motion

In the case of high voltage the rod aligns itself with the field and its motion is near its equilibrium position, and the effects of Brownian motion and initial position are small. The important point to note in the case of low voltage is, when Brownian motion is present the system is not in equilibrium.

17 Shape Descriptors for Cell Populations

AUTHOR Charles Hagwood
COLLABORATORS Javier Bernal (Mathematical and Computational Sciences Division, ITL, NIST)

Introduction

Biological activity within a cell is important for numerous reasons, e.g., drug discovery, diagnostics or pathology, and gene therapies. One response to activity that can be visualized is a cell's morphology. Fluorescence microscopy provides a means to visualize the effects of biological activity within a cell, for the life cycle of a cell. Features such as cell size, shape, fluorescent intensity, cell concentration often are surrogates for biological processes occurring within the cell and by analyzing them, biological activity can be better understood. Furthermore, a population of identical cells exhibits a distribution of surrogate responses, which for statistical purposes is very important for analyzing and comparing cell colonies. The figure's upper panel shows images of two types of muscle tissue cells.

The Computational Biology Group at NIST, made up of statisticians, cell biologists, computer scientists and mathematicians, was formed to investigate the process of making inference about biological activity from cell imagery. This involves comparing image segmentation and edge detection schemes, as well as, cell tracking schemes, analyzing shape descriptors, and finding appropriate statistical tools to analyze the generated data.

This subproject deals with finding the best shape descriptors for cell imagery data and using them to compare cell colonies. Simple one dimensional descriptors such as area, roundness, curvature are often used, but are not powerful in distinguishing shapes. More advanced methods based on the Fourier transform and the Procrustes metric are investigated here.

The Procrustes Mean Shape and Fourier Descriptors

Part of this investigation required finding the mean shape of a segmented cell. Because shape space is not flat, usual multivariate analysis can not be applied directly, but in a tangent plane about the mean, multivariate analysis can be applied, using tangent coordinates. In the shape analysis literature, there are several possible ways to define mean shape. The Procrustes mean shape is found here. Some of the other procedures for finding a mean shape are closely related to the Procrustes mean.

The Procrustes distance between two curves $y(s)$ and $w(s)$ parametrized by $s \in [0, l]$ is defined by

$$D^2(y, w) = \inf_{a,b} \int_0^l \mid y(s) - bw(s) - a \mid^2 ds. \tag{17.1}$$

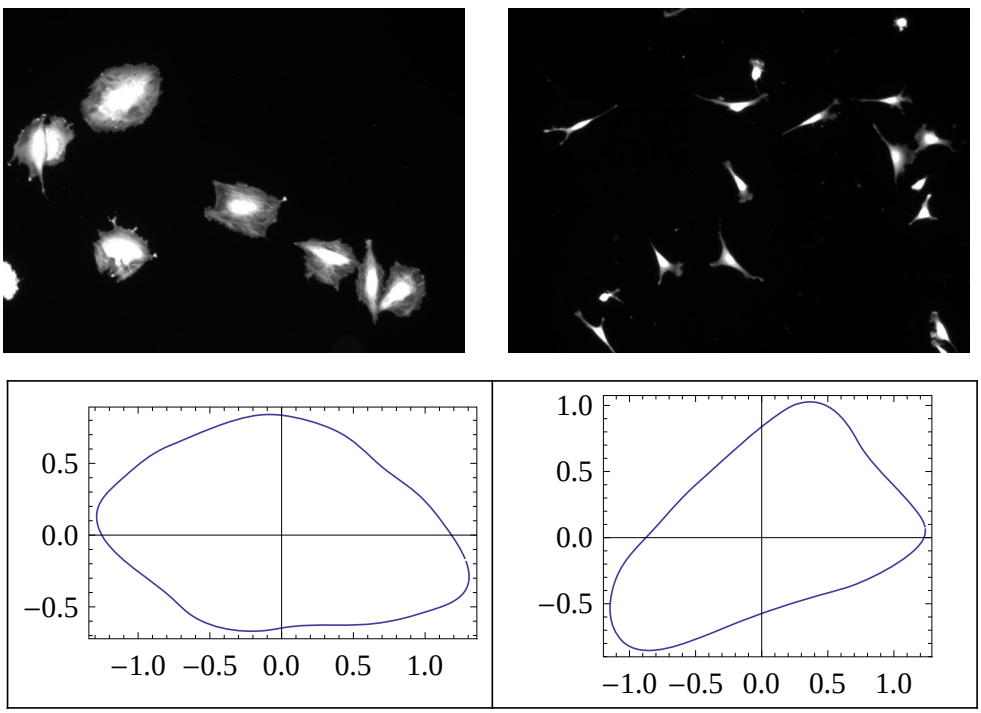
The Procrustes mean shape for a set of curves $w_i(s), i = 1, \ldots, n$ is given by

$$[\mu] = \arg \inf_\mu \sum_{i=1}^n D^2(w_i, \mu) \tag{17.2}$$

where $[\mu]$ denotes uniqueness up to a rotation (R. Larsen, 2005, 14th Scandinavian Conference on Image Analysis). The Canny edge detector was used on each image to determine a discretized boundary of n coordinates, $f_j, j = 1, \ldots, n$. Planar points (x_j, y_j) are denoted by complex numbers $f_j = x_j + iy_j$. Using the discrete Fourier transform, the boundary can be parametrized as

$$y(t) = \sum_{k=-(N-1)/2}^{(N-1)/2} c_k e^{ikt} \qquad c_k = \frac{1}{n}\sum_{j=0}^{n-1} e^{-2\pi jk/n} f_j. \tag{17.3}$$

With this parametrization, the mean curves in the figure's lower panel were determined. For a discretized curve, there are as many Fourier coefficients as boundary points, but not all are needed to give a good approximation to the boundary. Usually, the low frequency terms will give a good approximation. The mean shapes shown in the figure's lower panel are based on using thirty one Fourier coefficients, i.e. in (17.3) $N = 31$. Note: To make $D(y, w)$ symmetric in y and w the curves are normalized to norm 1. So, each Procrustes mean is based on this normalization.



Upper panel: cells from two different lines. Bottom panel: Procrustes means for the two cell lines.

18 David G. Harris

Biography

David Harris grew up around the Chicago area. After high school he attended Harvard University, graduating with a B.S. in mathematics. After graduating, he moved to Maryland to join the National Security Agency (NSA), working there for nine years. There, he worked on cryptology and computer science. He is currently taking a year-long sabbatical to NIST's Statistical Engineering Division.

Awards

2003 — Norman Robert Award for best junior cryptanalyst at NSA
2003 — Gold Bug Team Award for outstanding application of cryptanalytic skills to a high-importance system
2004 — Sir Peter Marychurch Award for Cryptanalytic Excellence
2004 — Cryptomathematics Institute's President's Award for career excellence in cryptology
2008 — Cryptomathematics Institute's Essay Award for best essay about a subject in cryptology

Selected Publications

Harris, David G., 2008, *Simultaneous field divisions: an extension of Montgomery's Trick* IACR Eprint archive

Harris, David G., 2008, *Generic ciphers are more vulnerable to related-key attacks than previously thought* Submitted to Workshop on Coding and Cryptography 2009

Harris, David G. and Oksana Lassowsky, 2002. *Method of Summarizing Text Using Just the Text*

23 major classified papers 2000–2009

19 Alan Heckert

Biography

Alan Heckert joined SED in 1996. He came to NIST in 1985 as member of the consulting group for the NIST supercomputer center. He previously worked for 4½ years for the Statistical Research Division of the Census Bureau. His primary area of interest is statistical computing.

Alan is currently the lead developer for the e-FITS and e-Metrology web projects. He provides computing support for various radiation detection test campaigns conducted by the Department of Homeland Security, and for an MEL scatterfield microscopy project.

Education/Awards

M.S. 1980 (Mathematics with concentration in Statistics), Clemson University
B.S. 1978 (Mathematics), Frostburg State College

Department of Commerce (DoC) Silver Medal, 2003 for the NIST/SEMATECH e-Handbook of Statistical Methods

Selected Publications

T. Kashiwagi, J. Fagan, J. F. Douglas, K. Yamamoto, N. A. Heckert, S. Leigh, J. Obrzut, F. Du, S. Lin-Gibson, M. Mu, K. Winey, R. Haggenmueller (2007), "Relationship Between Dispersion Metric and Properties of PMMA/SWNT Nanocomposites," *Elsevier, Polymer*, Volume 48, No. 16, pp. 4855–4866.

NIST/SEMATECH e-Handbook of Statistical Methods, `http://www.itl.nist.gov/div898/handbook/`, 2003.

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

E. Simiu, N. A. Heckert (1998), “Wind Direction and Hurricane-Induced Ultimate Wind Loads,” *Journal of Wind Engineering and Industrial Dynamics*, 74-76, pp. 1037–1046.

20 e-FITS

AUTHOR Alan Heckert
COLLABORATORS James Filliben, Will Guthrie, Charles Hagwood, Antonio Possolo, Andrew Rukhin, Cameron Rose (SURF student), Bill Strawderman, (Statistical Engineering Division, ITL, NIST)

Introduction

e-FITS is a web-based tool, currently available to NIST staff on an internal server, used to perform the following tasks for over 100 probability distributions.

- Generate graphs of probability functions (probability density, cumulative distribution, inverse cumulative distribution, hazard, cumulative hazard, survival, inverse survival).
- Generate tables for each of these probability functions.
- Generate random numbers from the specified distribution.
- Fit the distribution to user-supplied data, producing parameter estimates, their uncertainties, and diagnostic analysis of the fit.

WERB Review/Validation

To make e-FITS available to the general public, it is undergoing a WERB review: Jim Filliben, Will Guthrie, Charles Hagwood, Antonio Possolo, Andrew Rukhin, and Bill Strawderman, are serving as the division reviewers.

The review will consider the text associated with each distribution, and with the general methods that are used to fit distributions to data, the web interface, the computational engine, and the output. This review will take considerable time and effort. In particular, it involves addressing the issue of what constitutes appropriate validation of results. Due to the extensive nature of e-FITS, our current plan is to release e-FITS in stages.

Implementation of e-FITS

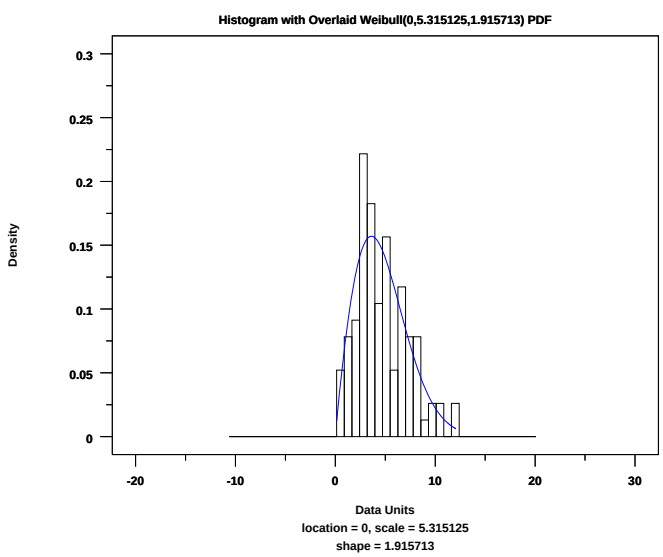
e-FITS is implemented using web forms and the Common Gateway Interface (CGI). The CGI scripts use Perl to process the form and Dataplot (and possibly procedures from other sources that are in the public domain) serves as the computational engine. The user of e-FITS does not need to install or learn the underlying statistical software. CGI scripts perform the computations on the server machine. Our SURF student, Cameron Rose, implemented a subset of e-FITS as JAVA applets (which perform the computations on the client machine).

e-Metrology

The methodology used to develop e-FITS is being extended to the e-Metrology project, which provides forms for common metrology problems encountered by NIST scientists and engineers. Forms are available to address the following problems.

- Uncertainty analysis of output quantities that are functions of input quantities, following the *Guide to the Expression of Uncertainty in Measurement* (GUM, ISO/IEC Guide 98:1995). Utilizes the R-based `gummer` routines written by Hung-Kung Liu, Will Guthrie and Antonio Possolo.
- Consensus means computed by various methods — these are a key component of many SRM analyses.
- Interlaboratory analysis based on ASTM standard E-691, and proficiency testing based on ASTM standards E-2489A and E-2489B.
- Limit of detection analysis based on the proposed ASTM WK 19817, which implements a method developed by Andrew Rukhin, Stefan Leigh and Michael Verkouteren (CSTL).
- Jim Filliben's 10-step analysis of full and fractional factorial designs.
- Linear and quadratic calibration and errors-in-variables regression.

The consensus means has been used for several SRM's. Dale Bentz and Paul Stutzman of BFRL are using the interlaboratory and proficiency analyses. Jeff Fong of ITL is using the 10-step analysis.



Histogram with Overlaid Fitted Weibull Distribution. The graphs shows a histogram of 100 random Weibull numbers overlaid with the Weibull probability density that was fit to this data.

21 Statistical Modeling and Computing Challenges in Scatterfield Microscopy

AUTHORS Alan Heckert, Nien Fan Zhang
COLLABORATORS Rick Silver, Ravi Attota, Ronald Dixson (Precision Engineering Division, MEL, NIST), Thomas Germer (Optical Technology Division, PL, NIST), Bryan Barnes, Hui Zhou (KT Consulting)

Introduction

The technique of scatterfield microscopy uses a reflective optical microscope with an angular scanning capability to probe sub-regions of an image. Advanced electromagnetic scattering models are employed to calculate the expected intensity at the image plane from these structures as functions of incident angle and polarization. The economics of semiconductor manufacturing dictate that scatterometric metrology targets must take up much less area than those currently used to make such measurements.

In the earliest days of integrated circuit manufacturing, optical microscopes were used for the metrology of patterned line widths, the critical dimension (CD) of these lithographically transferred patterns. Line widths have continually decreased with time to sizes much smaller than the wavelengths of visible light, thus eliminating conventional optical microscopy as a viable CD metrology. Scatterometry is an optical method that is relatively cheap and fast compared to alternative methods of line-width metrology.

Simulations

One goal of this project is to quantify, through the use of scatterfield microscopy, the top, middle, and bottom widths of lines that are arranged in an array, the height of the structure, and the resistivity of the material (referred to as the n and k values).

We use simulation to develop a library of reflectivity curves that can be compared to the physical measurements. The factors (top/middle/bottom line widths, height, n and k) are varied in a systematic way. Initial simulations are run to determine which of these factors are the most sensitive and the more extensive simulations are run varying these sensitive factors. The simulations use software run by Thomas Germer that implements a rigorous coupled wave (RCW) analysis.

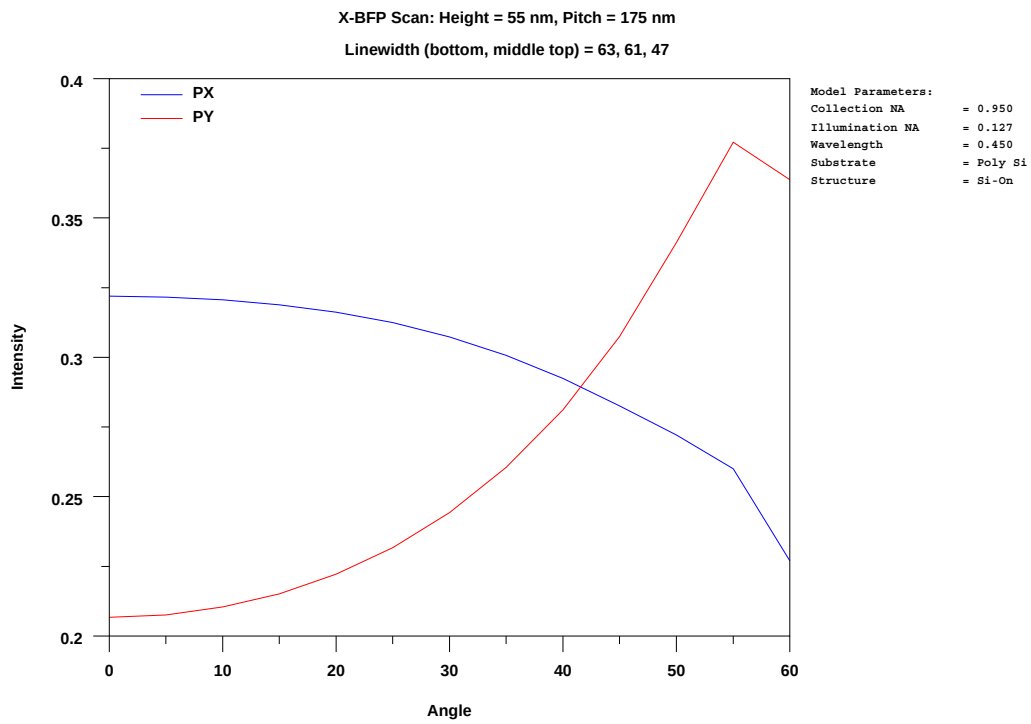
For some of these simulations, attention was restricted to line widths. Optical width characterization is accomplished by comparing the measured reflectivity of the grating as a function of angle position, scan axis, and polarization axis, against a library of simulated reflectivity curves.

The simulations are run on the Linux-based Raritan cluster that is jointly supported by ITL and PL. Although Germer's code can support parallelization for a single case, we implemented the parallelization by running many cases simultaneously. For example, our most recent set of simulations involved running approximately 5,000 simulations in a week's time. This would have been infeasible without the use of the Raritan cluster.

Improving Measurement Accuracy Using Multiple Techniques

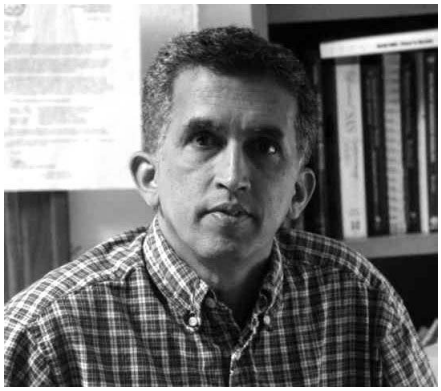
In optical metrology, the experimental signatures are being compared with electromagnetic scattering simulations using a nonlinear generalized least squares approach. When modeling optical measurements, a library of curves is assembled through the simulation of the responses over a multidimensional parameter space. A nonlinear generalized least squares fitting routine is then used to choose the optimum set of parameters that yields the closest agreement between experiment and theory. The corresponding parameter uncertainties are also obtained. This approach assumes that the model adequately describes the physical conditions, and that the goodness of fit achieved with the best set of parameters is acceptable.

To improve optical measurement accuracy, statistical methodologies have been developed to combine the results from optical techniques with information from other reference measurement systems, such as atomic force microscopy (AFM). We have shown that incorporation of the information from AFM measurements reduces the uncertainties of the parameter estimates from the optical measurements.



Sample Simulation Output. This graph shows a sample output for a single simulation. For the x-axis scan direction, an average intensity is computed and plotted for angle positions given in 5 degree increments for two distinct polarizations.

22 Hari K. Iyer



Biography

Hari Iyer was born in Chennai (then known as Madras), India. After graduating from high school he joined St. Xavier's college in Mumbai and completed his B.Sc degree in Mathematics in 1970. Subsequently he attended the University of Notre Dame in Indiana and completed his MS and his PhD degrees in Mathematics (Theory of Finite Simple Groups) under the direction of Professor W. J. Wong. He was an instructor of Mathematics at the University of Utah from 1975 to 1977.

In June 1977, Hari decided to work with Professor Raj Chandra Bose in the field of Experimental Design, at Colorado State University, and received his PhD in Statistics in 1980. Immediately following this he joined the faculty in the Department of Statistics at Colorado State University in 1980.

During the past few years he has collaborated with Jack Wang, Thomas Mathew, Paul Patterson, Jan Hannig, and many of his PhD students, focusing on Fiducial Inference, an inference approach originally introduced by Sir Ronald A. Fisher in the 1930s.

Hari has been a faculty visitor at the National Institute of Standards and Technology (NIST) for nearly 25 years. During the recent years Hari, collaborating with Jack Wang and Jan Hannig, has contributed to research related to quantification of uncertainty as proposed in the Guide to the Expression of Uncertainty in Measurements (GUM). Much of this research is based on Fiducial Inference Methodology.

Awards

2004 – Fellow of the American Statistical Association
2004 – W. J. Youden Award for best paper on Interlaboratory Trials (awarded jointly to Hari Iyer, Jack Wang and Thomas Mathew based on a paper published in JASA).

Selected Publications

Wang, C. M. and Iyer, H. K. (1994). Tolerance Intervals for the Distribution of True Values in the Presence of Measurement Errors, *Technometrics*, 36, 162-q-170.

Iyer, H. K., Wang, C. M. and Matthew, T. (2005). Models and confidence intervals for true values in interlaboratory trials, *Journal of the American Statistical Association*, 99, 1060–1071.

Hannig, J., Iyer H. K., and Paul Patterson, Fiducial Generalized Confidence Intervals, *Journal of the American Statistical Association*, 2006.

23 Dennis Leber

Biography

Dennis Leber joined the Statistical Engineering Division at NIST in January 2001, after five years in the Actuarial Research Department of Prudential Property and Casualty Insurance Company in Holmdel, NJ. Dennis received a B.S. degree in mathematics from Bloomsburg University in 1997, a M.S. degree in statistics from Rutgers University in 1999, a M.S. degree in Mechanical Engineering in 2007, and is currently working towards a Ph.D. degree in the Design, Risk Analysis and Manufacturing division of the Mechanical Engineering department at the University of Maryland. His academic interests include decision making and process modeling and simulation of discrete event systems.

Over the past several years, Dennis has been a steady collaborator of scientific and technical staff of NIST's Ionizing Radiation Physics Division, and the Domestic Nuclear Detection Office of the U.S. Department of Homeland Security, in the design of radiation detection experiments and in the modeling and analysis of the resulting data. Other ongoing collaborations, with researchers in NIST's Office of Law Enforcement Standards (OLES), include studies of performance of ballistic body armor, and the advancement of measurement science for imaging equipment used by firefighters and other emergency personnel.

Selected Publications

Crawdad Analysis Plan, U.S. Department of Homeland Security, Domestic Nuclear Detection Office, Document Number 200-CRAW-106920v2.00, 2008.

Henry Rodriguez, Pawel Jaruga, Dennis Leber, Simon G. Nyaga, Michele K. Evans, and Miral Dizdaroglu (2007) Lymphoblasts of Women with BRCA1 Mutations Are Deficient in Cellular Repair of 8,5'-Cyclopurine-2'-deoxynucleosides and 8-Hydroxy-2'-deoxyguanosine, *Biochemistry*, Vol. 46, pp. 2488–2496.

Andrew Persily, Amy Musser, and Dennis Leber (2006) A Collection of Homes to Represent the U.S. Housing Stock, NISTIR 7330.

24 Experimental Design for Comparative Performance Assessment of Radiation Monitoring Devices

AUTHOR Dennis Leber
COLLABORATORS Leticia Pibida (Ionizing Radiation Division, PL, NIST), Jim Filliben (Statistical Engineering Division, ITL, NIST)

The Statistical Engineering Division supports projects led by the Ionizing Radiation Division that address needs of the Domestic Nuclear Detection Office (DNDO) of the Department of Homeland Security (DHS) related to the development of sensors to detect the presence of illicit radioactive materials in personal luggage and commercial cargo transported in aircraft, boats, and automobiles.

While the purpose of the test and evaluation programs have varied, a majority have focused on comparing the performance of radiation detection systems. Since all levels of government — local, state, and federal — play crucial roles in homeland security, the results from such comparative test and evaluation programs have provided valuable guidance in acquisition decisions and operations development for all involved.

A major contribution from NIST to DNDO's test and evaluation programs has been support to the design of experiments and the detailed planning of test campaigns. While each of the comparative test campaigns provided unique perspectives and challenges, the general experimental design issues and trade-offs remained the same: specificity of the program goals; the desired scope and robustness of conclusions versus the budgetary constraints of the test program; and the use of scientific design principles and techniques versus the limitations of the practical problem being considered.

The inherently binary (Yes/No) response variables of primary interest — correct radionuclide detection and identification — provided additional challenges in the experimental design of such test campaigns given the limited information obtained during each trial. Operational test requirements typically prevent taking steps to guarantee the independence of successive observations, which downstream analyses assume.

The DNDO comparative test campaigns supported by the Ionizing Radiation and Statistical Engineering Divisions, and their general program purposes are:

- Personal Radiation Devices (PRDs) often worn by first responders such as police officers and fire fighters; results were provided to state, local, and federal agencies for decision support regarding device acquisition and appropriate modes of usage.
- Portable radiation detection systems to scan aircraft in international, general aviation environments; the equipment and operating procedures being used by Customs and Border Protection were evaluated and compared to alternatives, to identify potential avenues for improvement.
- Radiation detection systems for maritime use; in comparing both commercially available and developmental solutions for radiation detection in such environment, inputs were defined for a federal pilot program seeking to define an optimal maritime solution; the comparison results of the commercially available equipment were shared with local, state,

and federal agencies to be used in the process of device acquisition, and in the definition of appropriate modes of usage.

- Radiation detection systems used to scan passengers and baggage in international airports; the equipment and operating procedures used by Customs and Border Protection will be compared to alternatives in this test campaign to identify potential avenues for improvement; the results will support a federal pilot program aiming to demonstrate and disseminate the optimal solution that will be identified.

In each of these comparative test campaigns the radiation detection systems were the factor of primary interest. It was also of interest to understand how these detection systems performed in various situations of usage as well as their performance against relevant radionuclides. Furthermore, it was desired to provide a robust set of conclusions over which the comparison of systems was made.

Many factors in these test campaigns are discrete, categorical variables such as radiation source, instrument, and operator. Tolerance limits have been placed on some of the continuous variables — for example, speed and distance — to ensure a valid trial. Uncertainty in other continuous variables has not been addressed, but should be considered in upcoming campaigns.

To emphasize the importance, in a comparative experiment, of exposing all levels of the primary factor to the same test conditions, the experimental designs were structured as Taguchi parameter design layouts. The primary factor of the detection system was assigned to the inner array, while the robustness conditions, which consisted of combinations of factors such as radionuclide, operator, speed, and distance, were assigned to the outer array.

The number of observations made in each cell of the Taguchi parameter design depended on the level of between-system discrimination desired, the specificity at which conclusions would be stated, and the test time available. Careful work was necessary to strike a balance between discrimination ability, scope of conclusions, and available test time. Experimental designs such as fractional factorial and Latin squares were leveraged on occasion sensibly to reduce the number of robustness conditions considered to meet the test campaign constraints.

Aircraft Scanning. An aircraft is being scanned for radioactive material during the international general aviation test campaign. Experimental data is being captured by the data collection system tablet operator in the foreground.

25 Performance Assessment of Infrared Imaging Systems

AUTHOR Dennis Leber
COLLABORATORS Francine Amon, Justin Rowe (Fire Research Division, BFRL, NIST), Nicholas Paulter (Office of Law Enforcement Standards, EEEL, NIST)

With direction and funding from NIST's Office of Law Enforcement Standards (OLES), a collaborative effort was established between the Fire Research Division and the Statistical Engineering Division (SED), to develop a standard to be used to assess the quality of images produced by infrared imaging systems used by firefighters. As traditional approaches to the assessment of images involve subjective assessment by human image analysts, the focus of this work was to develop a program to enable the objective assessment of such systems.

A firefighter uses an infrared imaging system primarily for two recognition tasks:

1. To locate a hidden fire or hot spot in response to a smell of smoke call, or during surveillance after a fire has been extinguished and;
2. To locate an individual within a fire, either a fellow firefighter or a victim.

For a firefighter to be able to perform these tasks successfully when using an infrared imaging system, the image produced by the system must be of sufficiently high quality. The infrared imaging community suggested that the attributes of the image most influential on perceptual quality are contrast, brightness, spatial resolution, and noise. These attributes have been termed Image Quality Indicators, or IQIs. Laboratory methods have been developed by NIST's Fire Research Division to measure and quantify each of these IQIs for a given infrared imaging system.

Data to quantify the image quality level necessary for a firefighter to perform a recognition task successfully was collected by presenting a series of carefully created images to a group of firefighters. The images were created by photographing a scene (wall within a room) that may or may not include a hot spot, or target, using a high quality camera that produced images considered pristine. To provide conclusions valid over a range of conditions, i.e. a large scope, one hundred and eighty pristine images were collected to represent a variety of realistic and potential sightings a firefighter may encounter. These various sightings are referred to as robustness conditions. The robustness factors considered and controlled for were the image scene, the amount of clutter in the scene, the type of target in the scene, the size of the target, and the location of the target in the image.

The IQIs are continuous factors discretized into five levels, and set to define a 25 point design space (image settings) using a Taguchi L25 design. The contrast, brightness, spatial resolution and noise values of each pristine image were then digitally altered according to this 25 point design space to create a total of $180 \times 25 = 4\,500$ images.

Since the same image would be presented to a user multiple times, the images were presented in order of decreasing difficulty in interpretation. That is, the version of an image that was most degraded (blurry, dark, of low resolution, etc.) was presented in the early phase of the test, while the same image with less degradation was displayed later in the test period. This approach was used to help protect against a learning bias.

Using the image laboratory at the Army Night Vision Laboratory, a group of 16 users (who were firefighters) were asked individually to view each of the altered images on a computer screen, and locate the target within the image via a mouse click. The user's mouse click location was compared to the actual target location and the user's response was deemed either correct (1) or incorrect (0). Unfortunately, several users were unable to complete the entire exercise, therefore, rather than $4{,}500 \times 16 = 72{,}000$ data points, only 54,540 were recorded.

A modeling effort is currently underway to relate the infrared imaging system user's probability of successfully completing a recognition task to the given, measurable IQIs of the image. A linear logistic regression model is being considered to describe the relationship between the IQIs, and the probability of successfully completing the recognition task.

Several versions of the model have been considered including main effects only, main effects plus 2-term interactions, main effects plus all possible interactions, and main effects plus squared main effects. From the modeling findings, cross-validation exercises, and consideration of the partial deviance, the main effects plus 2-term interactions model structure was selected initially. Recent changes in the formulation and representation of the IQI values will require further modeling efforts. Generalized additive models will also be considered as possible descriptors for those relationships.

Once a model has been selected, and the IQI values for a particular imaging system have been measured in the laboratory, the predicted probability of a user successfully completing a recognition task using that imaging system can be calculated. This predicted value of success can then be compared to a threshold set by the user community in order to deem the imaging system acceptable or not.

The detailed experimental efforts and initial results from this work have been presented as a November 2008 Master's thesis by Justin Rowe from the University of Maryland's Fire Protection Engineering program (*The Impact of Thermal Imaging Camera Display Quality on Fire Fighter Task Performance*). The final resulting product will be adopted by the National Fire Protection Association (NFPA) in a standard currently being developed (NFPA 1801) as an effort to improve the safety of firefighters and the equipment they use.

Test Images. The left panel displays a photo of one of the scenes used in the creation of the test images. The right panel displays an actual test image taken with an infrared imaging camera. The target in this case is the hot spot in the chair to the right.

26 Standard Reference Materials

AUTHOR Dennis Leber

Standard Reference Materials (SRMs) are physical artifacts, mixtures, or compounds that are manufactured according to strict specifications, some of whose chemical or physical properties NIST scientists quantify and certify, employing measurement methods whose uncertainty is characterized fully.

The materials are carefully packaged and include documentation of the assigned certified values with stated uncertainties, the analytical methods used both for the determination of the certified values and for their uncertainties, material safety sheet, and details on use and stability.

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 a key mechanism for supporting measurement traceability.

The Statistical Engineering Division (SED) provides technical support to the SRM program by collaborating directly with the NIST scientists who develop and produce SRMs. NIST's Administrative Manual (Subchapter 5.19, Section 5.19.04), assigns to SED several areas of responsibility:

1. Assist in the design of sampling and measurement strategies for certification of SRMs;
2. Provide technical guidance on the implementation of NIST uncertainty policy;
3. Develop standardized statistical design and analysis templates that can be used by Laboratory personnel to carry out statistical analyses for classes of SRMs that follow fixed approaches;
4. Provide training on the proper use, interpretation, and limitations of these templates;
5. Provide data analysis and uncertainty assessment for SRMs for which appropriate standardized analysis templates are not available;
6. Certify the computation for unit values and stated uncertainties, as appropriate.

The development and production 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.

More than twenty divisions in six NIST laboratories participate in the SRM program producing artifacts that fall into three main categories: chemical compositions, physical properties, and engineering properties. Examples of such SRMs include sulfur in gasoline, organics in whale blubber, human Y-chromosome DNA profiling, and peanut butter. In 2008, more than 250 different SRMs were certified, and SED played a key role in certifying each of them.

SRM 965 Glucose in Frozen Human Serum. This SRM is used to evaluate the accuracy of procedures that determine glucose levels in human serum in the treatment of diabetes. SRM 965 is fundamental in the standardization of direct-reading biosensors for blood glucose.

27 Stefan Leigh

Biography

Stefan Leigh has been with SED at NIST for 30 years. He holds a BA (1967) in mathematics from Princeton University and an MS in mathematical statistics from the University of Maryland (1990). In the late 60's he did graduate work in commutative algebra and group theory under Claude Chevalley at the University of Paris. Prior to joining NIST, Stefan served in the US Army Special Forces (1969-1972), did paralegal work for the District of Columbia, and computer programming for the US General Services Administration. Stefan is an applied statistician engaged in collaborations with NIST scientists. As of January 2009, he has co-authored more than 80 publications.

Stefan has collaborated with hundreds of NIST scientists, and led teams of NIST statisticians, to solve problems in areas that include: mobile home fire standards (HUD), asbestos abatement (EPA), mercury credits (EPA), standard reference materials, extreme winds, *Bremsstrahlung*, DNA fingerprinting (FBI), cryptographic random number generators, face recognition algorithms, and Advanced Spectroscopic Portal Monitors (DHS). Stefan has developed and taught multiple courses on statistical methods for NIST scientists, and co-organized a Conference on Extreme Value Theory and Applications (1993). He has been a mentor for both the NRC postdoctoral and undergraduate SURF programs at NIST.

Selected Publications

Kim, J. H., Leigh, S. D., and Holmes, G. A., E-Glass/DGEBA/m-PDA Single Fiber Composites: The Statistics of Fiber Fragmentation, *Gordon Research Conference on Composites*, Ventura CA, Jan 2006.

Marinenko, R., and Leigh, S., Heterogeneity Evaluation of Research Materials for Microanalysis Standards Certification, *Microscopy and Microanalysis* 10, 491–506, 2004.

Widmann, J. F., Presser, C., and Leigh, S. D., Extending the Dynamic Range of Phase Doppler Interferometry Measurements, *Atomization and Sprays* 12(4), 513–537, July 2002.

Rukhin, A. L., Soto J., Smid, M. E., Leigh, S. D., *et al.*, A statistical test suite for random and pseudorandom number generators for cryptographic applications, *NIST Special Publication 800–22*, 2001.

McKenna, G. B., Vangel M. G., Rukhin, A. L., Leigh S. D., Lotz, B., and Straupe, C., The tau-effective paradox revisited: an extended analysis of Kovacs' volume recovery data on polyvinyl acetate, *Polymer* 40(18), 5183–5205, August 1999.

28 Prediction of Cement Characteristics via Enhanced Material Characterization

AUTHOR Stefan Leigh
COLLABORATORS Paul Stutzman (Materials and Construction Research Division, BFRL, NIST)

Introduction

Ongoing work by Paul Stutzman and other researchers of NIST's Materials and Construction Research Division has been directed towards enhancing performance prediction through improved characterization of cementitious materials. Cements and concretes are the most widely used materials in the world after water. Worldwide, their development and utilization in construction represents significant mensurable fractions of national GDP's. An overarching research goal is the accurate prediction of field performance metrics from accurate characterization of minerological composition and texture of the materials.

From Bogue estimation to X-ray diffraction

Compounding characterization and prediction problems is the industry-wide reliance since the 1930's on the so-called Bogue estimates of mineral composition both for cement clinker, and for cement itself, which are derived from measurements of bulk concentrations of major oxides.

The Bogue method assumes that a particular set of mineral phases are present, and also assumes an idealized chemical composition for each of them. It then solves a linear system of equations that relate the relative abundances of major oxides to the relative abundances of those phases. This method is inaccurate owing both to biases in assumed chemical compositions of the phases, and to the possible presence of mineral phases not specified in the Bogue formulation. NIST is documenting these shortcomings as part of this overall project. In addition, updates to the Bogue approach work poorly for modern "green" cements with limestone, slag, fly ash additives.

X-ray diffraction analysis of cementitious materials (P. Stutzman and S. Leigh, "Phase analysis of hydraulic cements by X-ray powder diffraction: precision, bias, and qualification", *Journal of ASTM International*, 4 (5), May 2007) represents a new state of the art. In particular, X-ray diffraction enables the direct identification and quantification of the mineral phases. X-ray scanning techniques also allow pinpoint determinations of composition over individual mineral grains observed under the microscope. These techniques achieve far greater accuracy than the Bogue method. NIST is in the process of demonstrating this by comparing performance on existing and newly created standard reference materials, certified on the basis of multiple methods.

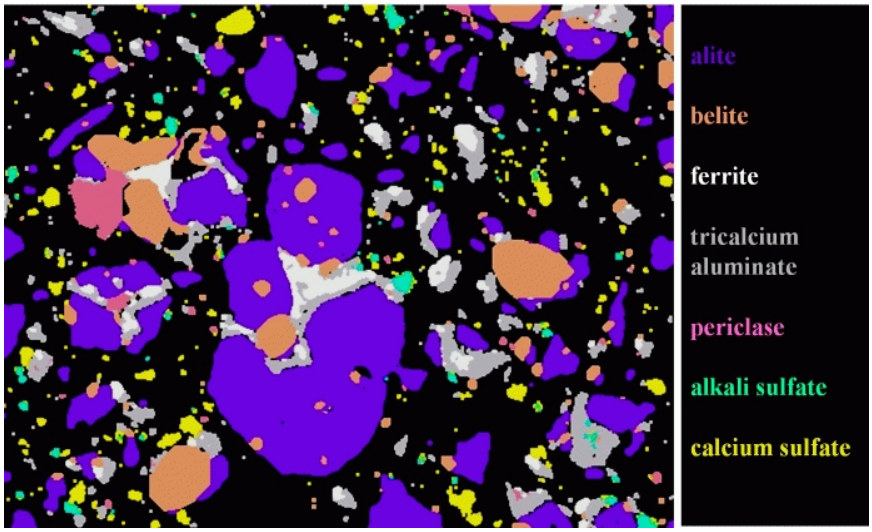
Property Prediction

Heat of Hydration (HOH) is critical when dealing with large masses of cement (e.g., in the construction of the Hoover Dam, where ice was mixed into the setting concrete) or with high temperature curing, where critical instabilities can result when cure conditions don't match

construction environment conditions. Standard construction contracts invoke a HOH clause. Current direct HOH testing is time, staff, resource expensive, and employs noxious reagents.

BFRL and SED are working to develop predictive models for HOH as a function of phase composition determined from X-ray diffraction, and adjunct variables like particle fineness. To find models that improve on existing coarse rules of thumb used in the industry, we use multivariate graphics, PCA, all possible subsets regression coupled with judiciously chosen transformations, and adaptive methods (for example, Leo Breiman's and Jerome Friedman's *Alternating Conditional Expectations*, ACE, and Rob Tibshirani's *Additivity and VAriance Stabilization for regression*, AVAS).

The goal is to develop the quantification and predictive power of the NIST internal techniques further, and to transfer them to ASTM standards, which North American construction specifications rely upon, for example ASTM C1365, *Standard Test Method for Determination of the Proportion of Phases in Portland Cement and Portland-Cement Clinker Using X-Ray Powder Diffraction Analysis*. Such standards will facilitate commerce and strengthen confidence in compositional analyses of these products when they are obtained in conformity with the standards.



Segmented phases merged into a composite image, based on scanning electron microphotographs and X-ray microanalysis: field width, 150 μm. Paul Stutzman, *Cement and Concrete Composites*, 26 (8), 957–966 (2004).

29 Advanced Spectral Portal Testing for the Department of Homeland Security

AUTHOR Stefan Leigh
COLLABORATORS James Filliben, Alan Heckert, Dennis Leber, Antonio Possolo, Andrew Rukhin, William Strawderman, and James Yen (Statistical Engineering Division, ITL, NIST), Leticia Pibida (Ionizing Radiation Division, PL, NIST)

Introduction

The Statistical Engineering Division supports projects led by the Ionizing Radiation Division that address needs of the Domestic Nuclear Detection Office (DNDO) of the Department of Homeland Security (DHS), related to the development of spectroscopic portal monitors used to detect the presence of illicit radioactive materials in commercial cargo containers entering the United States.

Advanced Spectroscopic Portal Monitors

Advanced Spectroscopic Portal (ASP) monitors are produced and tested to enhance DHS's ability to detect and identify radioactive materials in cargo containers.

In December 2004, DHS accelerated the research, development, and prototyping of ASP devices for use in operational environments. DHS made multiple awards to vendors for the development and testing of prototype units for both gamma ray and neutron detection. NIST's analysis and report provided the basis for the choice of the three prototypes that continue to undergo test and development.

The energy spectrum of gamma rays emitted in radioactive disintegrations acts as a fingerprint of the radionuclides that emit them. Sodium iodide crystals, high purity germanium detectors,

and plastic scintillators are sensors used to measure gamma rays. Neutrons generally are more difficult to detect than gamma rays owing to their weak interaction with matter and to their wide range of energies: they are measured using liquid organic scintillators.

Currently deployed portal monitors are able to detect excess radiation above a background threshold in passing vehicles. Such monitors are capable of gross count measurements, but do not provide nuclide identification capabilities. After the monitor signals the presence of significant radioactivity, further investigation is usually necessary to determine if the radiation is a threat. Secondary inspection is time-consuming, especially taking into account the sheer volume of materials transported across the nation's ports and infrastructure on a daily basis. Basic problems include: nuisance (background radiation) alarms, false positives that may unduly interrupt to traffic flow, the risk of false negatives, and logistical challenges associated with deploying sufficient sensors across diverse environments.

Summer 2008 Variant C analyses

In the 3½ years since the completion of the source selection work, we have been involved in design, execution, and analysis of multiple ASP test campaigns.

The field tests comprise a sequence of trials arranged in designed experiments, where multiple cargo configurations pass through the portals, which are variously configured. The responses include the nature of the alarm that is generated, if any, and the corresponding identification of the event that triggered the alarm. There is a large collection of associated covariates that must be taken into account.

We apply conventional procedures to estimate probabilities of correct alarm and identification, and assess the corresponding uncertainties. We also perform other standard analyses (for example, analysis of variance and logistic regression), and examine graphical representations of the data. A very large part of our effort is consumed preparing and validating the data prior to their statistical analysis: for example, we often need to develop and apply interpretive rules to the raw data to produce derivative products that are amenable to analysis.

We have developed non-standard statistical tests to establish equivalence between ratios of frequencies, and for several functions of such ratios. And we also employ bootstrapping and subsampling techniques for uncertainty analysis.

Statistical tests of hypothesis have been employed to establish the equivalence between ratios: we have devised several non-standard tests to answer specific questions that involved more complicated functions of ratios and differences of ratios.

Some of this work is described in entitled "Pass-Fail Testing: Statistical Requirements and Interpretations", co-authored by Andrew Rukhin, Bill Strawderman, Stefan Leigh, and David Gilliam, which has been submitted for publication in the *NIST Journal of Research*: this includes a detailed treatment of sample-size requirements to establish confidence limits for probabilities of detection and false alarm.

30 Walter Liggett

Biography

After receiving his PhD from Rensselaer Polytechnic Institute in applied mathematics, Walter Liggett was employed by Raytheon, the New York City-Rand Institute, and the Tennessee Valley Authority before coming to the Statistical Engineering Division in 1979. Since 2003, the focus of his work has been biological measurements: mass spectral measurements of proteins, microarray measurements of gene expression, and CT scans of lung tumors. he is a member of the board of editors of the Journal of Research of the National Institute of Standards and Technology.

Awards

The Judson C. French Award of NIST, 2002.
The Edward Bennett Rosa Award of NIST, 1996.
Distinguished Achievement Award from the Section on Statistics and the Environment of the American Statistical Association, 1993.

Selected Publications

Walter Liggett (2008) Technical variation in modeling the joint expression of several genes, in *Methods in Microarray Normalization* edited by Phillip Stafford, CRC Press.

Walter S. Liggett, Peter E. Barker, Lisa H. Cazares, and O. John Semmes (2007) An approach to the reproducibility of SELDI profiling, in *Spectral Techniques in Proteomics* edited by Daniel S. Sem, CRC Press.

Walter Liggett (2006) Normalization and technical variation in gene expression measurements, *Journal of Research of the National Institute of Standards and Technology* 111, 361–372.

Walter Liggett and Chris Buckley (2005) System Performance and Natural Language Expression of Information Needs, *Information Retrieval* 8, 101–128.

Walter S. Liggett, Peter E. Barker, O. John Semmes, and Lisa H. Cazares (2004) Measurement reproducibility in the early stages of biomarker development, *Disease Markers* 20, 295–307.

Walter Liggett, Lisa Cazares, and O. John Semmes (2003) A look at mass spectral measurement, *Chance* 16, 24–28.

31 Slice-by-Slice Comparison of Computed Tomography Scans

AUTHOR Walter Liggett

Introduction

This paper presents a method for characterizing the tumor change that occurs in the time between two CT scans and illustrates the method with two CT scans of a slowly changing tumor. The method is based on slice-by-slice matching of the two scans. The computation begins with fitting a thin plate spline to the density values for each slice. Next is registration of each pair of matching density maps followed by differencing of the registered maps. The method presents adjacent difference maps grouped and averaged over limited vertical spans. In the example, this method reduces 12 density maps from each scan to 4 difference maps corresponding to 4 vertical locations in the tumor. These difference maps allow characterization of change with emphasis on changes near the tumor boundary. Such emphasis may have clinical value.

Our example consists of images in the DICOM format obtained from the Public Lung Database to Address Drug Response, which is funded by the Cancer Research and Prevention Foundation. The website `www.via.cornell.edu/crpf.html` is the source of the images. The case identifier is SS0014. There are two scans that we identify by their month, January and the following September.

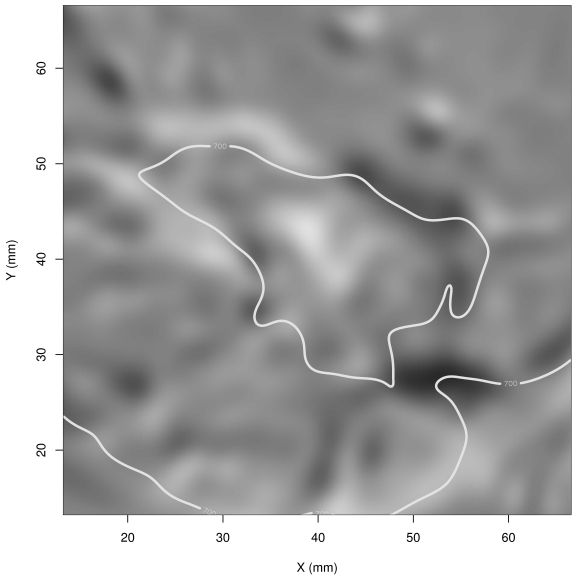
Summarization & Comparison

Consider summarization of CT scans for clinical purposes. In the case of the lung, the result of a CT scan consists of density maps for a sequence of horizontal slices with different vertical positions. Two CT scans that depict a tumor at two different times provide a basis for characterization of tumor change. Summarization is necessary for change characterization because of the large number of density maps that comprise the output of a CT scan. The form of the summarization should be chosen on the basis of its use in clinical discourse. Because such discourse involves the concept of tumor volume, summarization of each CT scan by a volume determination comes to mind initially. However, it is possible that the nuances of clinical discourse cannot be carried by tumor volume alone. For this reason, we seek a summarization method with greater clinical value.

The results of two CT scans might be thought of as density maps arrayed in a two-way table with rows corresponding to the slices and columns corresponding to the two times. This data structure is based on the assumption that there is a correspondence between the slices in one scan and the slices in the other. This assumption is satisfied by the scans in the example explored in this paper. It would seem that it would generally be satisfied if the slice spacing is small enough and the tumor change is not too extreme. Think about two panels, each with a density map from a different scan. If the two density maps were portrayed using only the density values for the original DICOM grid, then the crudeness of the image might interfere with the matching. After interpolating to obtain density values for a finer grid, the matching would be easier. Finally, registration of one density map with respect to the other might make the match even more obvious.

Summarization by tumor volume can be thought of as summarization over the rows of the table followed by differencing the two column summaries. In this paper, we consider differencing each row of the table followed by summarization of the differences over the rows. Differencing two density maps consists of fitting a thin plate spline to the density values on the original DICOM grid and then registering one map with respect to the other. The thin plate spline representation provides a density value for any location on the map and thereby allows use of Euclidean distance between functions for registering images. This representation also allows the registration parameter domain to be continuous. In the case of our example, pairs of registered density maps are so much alike that pinpointing panel-to-panel differences requires attention to detail. To make the differences more obvious, we subtract one density map from the other. This gives what might be called a difference map, one for each row. As the final step in summarization, we divide the difference maps into groups each consisting of maps from three contiguous slices and average over the groups.

The figure shows a difference map upon which is superimposed a constant-level contour that shows the boundary of the tumor and the lung wall. The difference map shows January values with September values subtracted. Thus, bright regions on the map are regions where the January density is higher, and dark regions are regions where the later September density is higher. The constant-level contour comes from the September density map.



Average registered density maps 38-40 (January) minus average density maps 37-39 (September).

We note three regions in the figure, although someone with more training in interpreting CT scans might make different choices. First, in the space between the tumor and the lung wall there is a dark spot indicating an increase in density in this region. Tracing along the boundary of the tumor as given by the contour, we see both regions of increased density and decreased density. On the left side of the tumor, we see bright regions, which might indicate shrinkage of the tumor. On the upper right part of the tumor boundary, we see a dark region, which might indicate growth of the tumor.

32 Control Group Variation in Gene-Expression Studies of Case-Control Differences

AUTHOR Walter Liggett

COLLABORATORS Jean Lozach (Illumina, Inc., San Diego, CA), Anne Bergstrom Lucas (Agilent Technologies, Inc., Santa Clara, CA), Ron Peterson (formerly with Novartis Institutes for Biomedical Research), Marc Salit (Biochemical Sciences Division, Chemical Science and Technology Laboratory, NIST), Danielle Thierry-Mieg (National Center for Biotechnology Information, NIH), Jean Thierry-Mieg, National Center for Biotechnology Information, NIH), Russ Wolfinger (SAS Institute, Inc., Cary, NC)

Introduction

Comparison of a sample of cases with a sample of controls by means of univariate measurement of each individual is a familiar problem. Comparison by means of microarray gene-expression measurement is fraught with unanswered questions because gene expression is a high-dimensional measurement, that is, the number of genes that enter the comparison far exceeds the number of individuals measured. Efron ("Microarrays, empirical Bayes and the two-groups model", *Statistical Science*, 2008) provides a context for formulating these questions. This project has obtained insight into answers to some of these questions on the basis of a specially designed experiment.

Efron (2008) pictures case-control studies for high-dimensional measurements as interpretation of a collection of t statistics, one for each gene in our case. Dependence among these t statistics is an important question. Variation within the control group comes from two sources, biological variation among the individuals in the sample and technical variation among the measurement realizations. Each of these sources may introduce dependence, but, of course, by different mechanisms. To permit insight into the dependence, the experiment considered was designed in a way that allows biological variation and technical variation to be distinguished.

Results

The experimental results include replicate measurements on liver, kidney, and mixtures of these two RNAs in six animals (*Rattus norvegicus*), made with Affymetrix, Agilent, and Illumina platforms. Because the animals formed a control group for a previously run case-control study, biological variation within the control group can be assessed. Because the experiment involves replicates, mixtures, and different measurement platforms, the measurements are expected to satisfy various relations. These relations allow the technical variation to be assessed.

We have obtained insight into the relative size of measurement batch effects and biological variation as represented by the animal-to-animal differences. These differences provide a practical benchmark because the animals were all subject to the same control-group treatment.

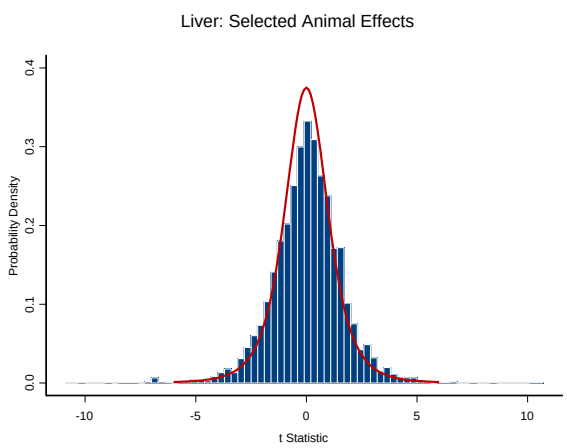
Although calibration curves for individual probes are unknown, a platform-to-platform correspondence identifies probes that measure the same transcripts and allows examination of the

relative sensitivity of probes from different platforms.

For biologists, gene expression microarrays provide an approach to identifying genes with particular properties such as change in expression with experimental treatment. The genes thus identified populate a gene list. Efron (2008) discusses the statistical aspects of gene list identification. Because we have measurements on six animals, we can obtain insight into such gene lists.

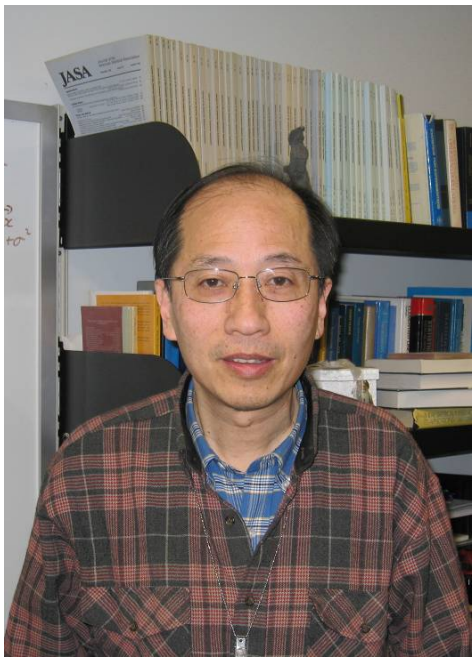
We have obtained some general observations on the metrology issues considered. First, although the animal-to-animal variation is generally larger than the measurement batch effects, our measurements do lead to the conclusion that these effects should not be ignored in experimental design and analysis. It is moreover the case that the measurement batch effects might be larger in a different experiment. Second, over the set of transcripts for which liver expression is appreciably different from kidney expression, no platform is undeniably more sensitive than another. However, the difference in probe sensitivity between two platforms varies appreciably from transcript-to-transcript. That is, one platform seems more sensitive for some transcripts, and the other platform more sensitive for other transcripts. This observation suggests considerations in the interpretation of single-platform studies. Third, we find that gene list reproducibility is likely to be worse than might be expected.

The experiment described here offers an approach to measurement system insight that could feasibly be part of any substantial gene expression study. There are reasons why one might want to change the design of our experiment. Inclusion of more animals would lead to more insight into gene-list reproducibility. Our investigation provides full coverage only of the probes for which liver expression differs from kidney expression. Inclusion of more animal organs would lead to better coverage of the probes.



Histogram of t statistics and the null density. We consider the 3548 transcripts for which the animal variation dominates the technical variation. For each of these transcripts, we compute the t statistic that compares the first 3 animals with the second 3. The figure compares these t statistics with the 4-degrees of freedom t density. This figure indicates that animal variation within a control group is such that transcript-by-transcript t statistics can be modeled as independent realizations from the null distribution.

33 Hung-kung Liu



Biography

Hung-kung Liu is a statistician with 25 years experience consulting on a wide variety of problems including engineering, physical and biological sciences. He is knowledgeable in various approaches to statistical inference, design of experiments, and computer intensive methods with experience in developing statistical methodology for special problems.

Hung-kung Liu was educated at National Central University in Taiwan and Cornell University. His graduate research work at Cornell, guided by Prof. L. Weiss, investigated testing fit based on an increasing random subset of order statistics. From 1984 to 1991, he was a faculty member of the State University of New York at Stony Brook. He joined the National Institute of Standards and Technology in 1991.

Awards

B.S. in Mathematics from National Central University (Taiwan) 1974, M.S. in Mathematics from Cornell University 1980, Ph.D. in Statistics from Cornell University 1984, ITL Outstanding Contribution Award 2005.

Selected Publications

Sequential confidence regions in inverse regression problems, with J. T. Hwang, *Annals of Statist.* (1990) 18 1389–1399.

Discussion on "Jackknife-after-bootstrap standard errors and influence functions", (with H. Chen) *J. Roy. Statist. Soc. B* (1992) 121–122.

Confidence intervals associated with tests for bioequivalence, (with J. C. Hsu, J. T. G. Hwang, and S. T. Rubreg), *Biometrika*, (1994) 81 103–114.

An ISO GUM Approach to combining results from multiple methods, (with M.S. Levenson *et al.*), *NIST Journal of Research*, 105, 571 (2000)

Bayesian approach to combining results from multiple methods (with N. F. Zhang), *Proceedings of the Section of Bayesian Statistical Science of ASA* (2002) 158–163.

Statistical analysis of key comparisons with trends (with N. F. Zhang, N. Sedransk, and W. E. Strawderman) *Metrologia*, (2004) 231–237.

Empirical modeling methods using partial data (with G. Stenbakken) *IEEE Transactions on Instrumentation and Measurement* (2004) 271–276.

34 Alternative Approach to Mass Metrology

AUTHOR Hung-kung Liu
COLLABORATORS Joe Chalfoun, Patrick Abbott, Zeina Jabbour (Manufacturing Metrology Division, MEL, NIST), Ruimin Liu, Edwin Williams (Quantum Electrical Metrology Division, EEEL, NIST) and James Filliben (Statistical Engineering Division, ITL, NIST)

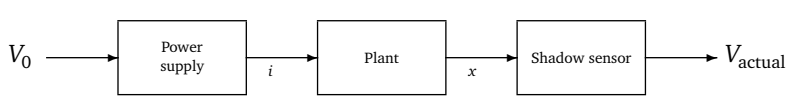
Introduction

In anticipation of the redefinition of the kilogram and to avoid a parallel non-SI dissemination system for mass, an alternative approach to mass measurement is being developed at NIST. Magnetic levitation is utilized to create a mass comparison system in which a test mass artifact in air can be directly compared to a standard mass artifact in vacuum using the same high precision comparator balance. Due to the extreme sensitivity of this comparator balance (in the order of 10 μg), any change in temperature near the balance may affect measurement results. Therefore, permanent magnets will be used in this project to achieve the levitation. A solenoid is added to the system to stabilize the magnetic levitation.

To test the feasibility of this project, a levitation system has been constructed as shown in the figure. A model of the system with permanent magnetic levitation and electromagnetic control was derived. Using the derived model, a feedback control was applied to our proof-of-concept apparatus. The purpose of modeling and controlling such a system is to have an automatic levitation system with a steady balance reading. That is, when the controller is initiated, the levitated body will automatically move from its resting position to a stable levitated position. Furthermore, this controlled system will oscillate at a higher frequency than the balance and with an amplitude low enough that the balance will filter the oscillation and will be able to produce a steady mass measurement result.

Model, Control & Optimization

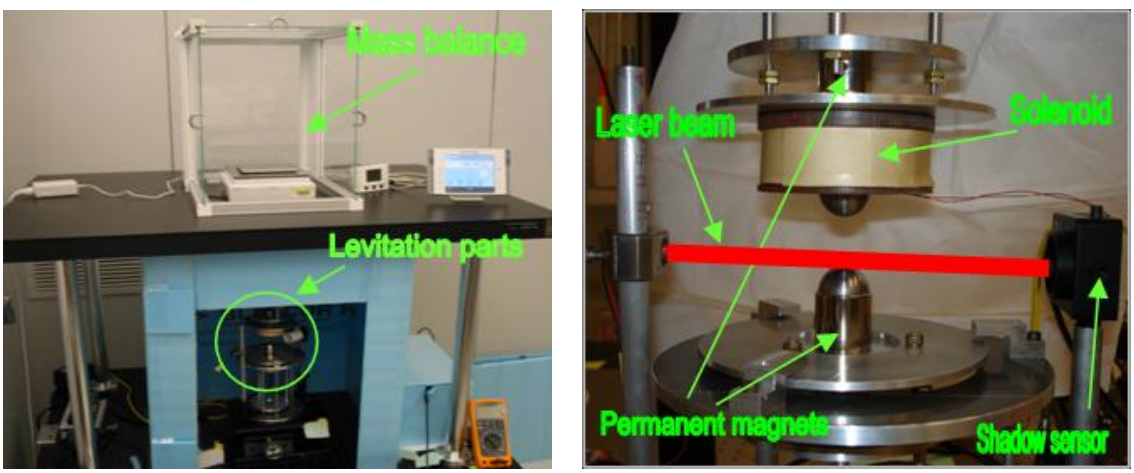
According to Earnshaw's theorem, stable magnetic levitation cannot be achieved by using only permanent magnets. It is necessary to add an active feedback control to regulate the perturbation affecting the system. When the controller is initiated, the vertical position signal V_{actual} is servo controlled about a desired equilibrium position corresponding to an output V_{desired} of the shadow sensor. That is, based on the difference of V_{actual} to V_{desired}, the controller provides an input V_0 for the current power supply to maintain stable levitation. The current power supply provides a current i to the solenoid based on the input voltage V_0. The magnetic field produced by the solenoid exerts an attractive force on the levitated body and raises it to position x. The shadow sensor detects this position and provides an output in the form of a voltage V_{actual}.



Block Diagram of the Open Loop System

The preceding figure shows a block diagram of this system in an open loop with input V_0 and output V_{actual}. We modeled different components of this system separately. Using the first order Taylor series expansion of the equation of motion, plant model was derived to be a second order linear differential equation. Power supply and shadow sensor models are both determined experimentally to be linear. Using these models, a PID (proportional, integral, and derivative position loop) controller was established. This controller was then applied digitally to the levitation apparatus to produce an automatic stable levitation with a vertical position oscillation of 3.4 μm around the desired levitation distance. On this balance of 1 mg accuracy, steady measurement is achieved using the proposed controller.

For the construction of NIST's next generation magnetic levitation balance with bigger mass and stronger magnets, computer experiments were run on a surrogate-FEA computational black boxes MagNet so as to gain insight into dominant factors as well as deduce optimal settings. We utilized important functionalities like scripting, parametrization and interoperability of MagNet to conduct computer experiments efficiently. Verification and Validation were performed to checks that MagNet meets specifications and that it fulfills its intended purpose. We used factorial designs to study multiple factors simultaneous to optimize the magnetic force gradient and the distance between magnetic poles. The factors studied include coercivity of the magnets, the shape of the magnetic poles and the current for the solenoid.



Proof-of-concept Apparatus. The left panel is a photograph of the proof-of-concept apparatus. For simplicity, the proof-of-concept apparatus was constructed with both masses in air at atmospheric pressure. Two attractive permanent magnets are used to make the levitation. A solenoid is added to the system to stabilize the magnetic levitation. The right panel is a photograph of the close-up of the levitation parts in the proof-of-concept apparatus. The vertical position signal is provided by a "shadow sensor" photodiode that detects a laser beam that just skims the top of the magnetic pole that is attached to the levitated assembly.

35 R Package for Statistical Metrology

AUTHOR Hung-kung Liu
COLLABORATORS Antonio Possolo and Will Guthrie (Statistical Engineering Division, ITL, NIST)

Introduction

The techniques for statistical modeling, computation and data analysis that have, particularly during recent years, been developed or adapted for use in metrology often are sufficiently involved to prevent manual application, and instead require that the metrologist employ computer software.

To facilitate statistical and mathematical computation in metrology, in particular for the assessment of measurement uncertainty, we are producing an R package, using the open-source R environment, for statistical computation and graphics that is freely available to all.

R is a GNU project similar to the S language and environment that was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. Owing to its great power and flexibility, it has become the choice environment to prototype and exchange ideas in computational statistics. R provides a high-level programming environment that can incorporate software written in conventional languages (C and Fortran, in particular), and offers a sophisticated packaging and testing paradigm.

Project Goals

The R package for statistical metrology will produce implementations of fast and reliable computational algorithms, based on recent developments in statistics and metrology, and will pay particular attention to making the developed products accessible to scientists within NIST and other NMI's, as well as customers in industry.

The broad goals of the project are:

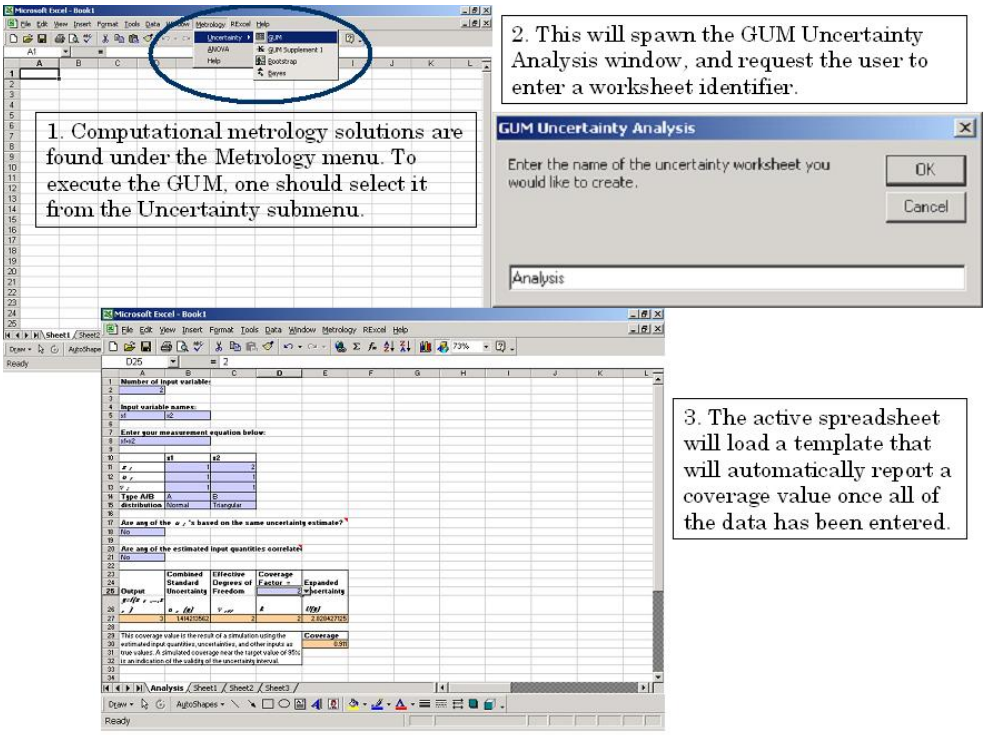
- To provide access to a wide range of powerful statistical and graphical methods for the analysis of metrological data, exploiting the model-oriented constructs that R provides;
- To accelerate the development of extensible, scalable, and interoperable software for metrology;
- To promote the production and dissemination of high-quality documentation that is a key component of reproducible research;
- To provide training in R emphasizing computational and statistical methods for the analysis of metrological data.

User Interfaces

Like R, our package can best be used via a text editor, like emacs, Tinn-R, or WinEdt, that is aware of R's syntax and that can interact with R process. We are also building other user interfaces for our package.

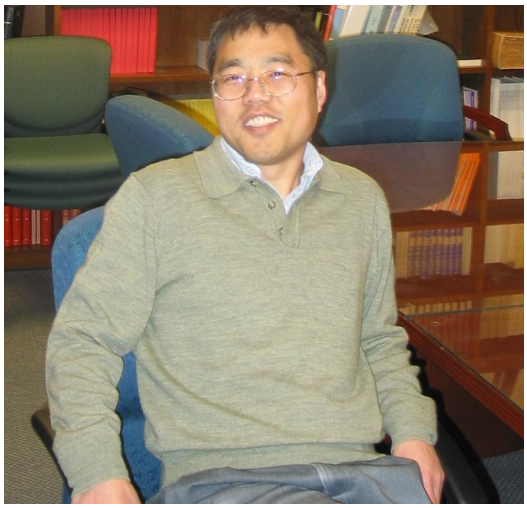
The Rcmdr package by John Fox provides a basic-statistics graphical user interface to R called R Commander. It uses a simple and familiar menu/dialog-box interface. The menus and dialog boxes of the R Commander are used to read, manipulate, and analyze data, and can be modified to suit our specific needs. We can also provide additional classes of statistical models by adding the necessary dialog boxes and menu items, and editing the `model-classes.txt` file in R's `etc` directory.

Since R also has several mechanisms that allow it to interact directly with other software, it enables users to incorporate modules based on other work. As an example, our prototype GUM Uncertainty Calculator demonstrates how R's (D)COM Server (R-Excel Add-in by Baier & Neuwirth) can make R's functionality accessible from within Excel, thus allowing users to use modern complex statistical methodologies that are beyond Excel's statistical capabilities, to widen and enrich the practice of the measurement sciences. Viewed in that context, adopting R as a vehicle for our project does not exclude other development environments and paradigms.



GUM Uncertainty Calculator Interfacing with Excel.

36 Z. Q. John Lu



Biography

John Lu was born to a big family of farmers in a small village in Suxian County of Anhui Province in the middle part of China. He left his hometown for the first time in 1982 to attend The Peking University in Beijing, where he earned a BS in mathematics, with specialization in statistics. After 7 years in Beijing, he came to the US for graduate study in the University of North Carolina at Chapel Hill, where he got a PhD in statistics. He joined NIST in 2001 as a mathematical statistician.

Early on he worked on a number of funded projects and he has always been interested in statistical issues motivated by high-dimensional measurements. In recent years he has been attracted by the challenging statistical problems motivated by the needs in quality assessments of multiplexed biological and biochemical measurements. He has been working on the Gene Expression Metrology competence project, the NIST/ITL new initiative on Medical Imaging (algorithm evaluation), and the joint ITL-CSTL collaboration on Computational Biology (cellular microscopy imaging).

Selected Publications

N. Machkour-Deshayes, J. Stoup, Z. Q. Lu, J. Soons, U. Griesmann, R. Polvani (2006). Form-profiling of Optics Using the Geometry Measuring Machine and the M-48 CMM at NIST. *Journal of Research of NIST*, September-October 2006, Vol. 111, No. 5, pp. 373–384.

M. B. Satterfield, K. Lippa, Z. Q. Lu, M. L. Salit (2008). Microarray Scanner Performance Over a Five-Week Period as Measured With Cy5 and Cy3 Serial Dilution Slides. *Journal of Research of NIST*, September-October 2006, Vol. 113, No. 3, pp. 157–174.

N.D. Lowhorn, W. Wong-Ng, W. Zhang, Z.Q. Lu et al (2009). Round-robin measurements of two candidate materials for a Seebeck coefficient Standard Reference Material. *Applied Physics A* (2009) 94: 231–234.

Z.Q. Lu, N.D. Lowhorn, W. Wong-NG, W. Zhang et al (2009). Statistical Analysis of a round-robin measurement survey of two candidate materials for a Seebeck coefficient Standard Reference Material. *Journal of Research of NIST*, in press.

Z.Q. Lu, C. Fenimore, R.H. Gottlieb, C.C. Jaffe (2008). An Empirical Bayes Approach to Robust Variance Estimation: a Statistical Proposal for Quantitative Medical Image Testing. In WERB.

Z.Q. Lu, M. Satterfield, M.L. Salit (2009). Multiphase and Nonlinear Regression Models for Performance Evaluation of Microarray Gene Expression Measurements. Submitted.

37 Statistical Approaches for Background Correction in Single Cell Green Fluorescent Protein Images

AUTHOR Z.Q. John Lu
COLLABORATORS Kevin Coakley, Katharine Mullen (Statistical Engineering Division, ITL, NIST), Michael Halter, John Elliott, Anne Plant (Biochemical Sciences Division, CSTL, NIST)

Introduction

The green fluorescent protein (GFP) from the jellyfish *Aequorea victoria* has attracted much attention as a tool to study several biological processes. By using DNA engineering, GFP can be used as a label for other interesting proteins that otherwise would be invisible under the microscope. This glowing marker allows visualizing the movements, positions and interactions of the tagged proteins.

The measurement of GFP's fluorescent intensity in cell images gathered with the microscope is affected by background fluorescence: when the GFP intensity is low, correcting for the background fluorescence becomes a critical issue. For example, we find that there is non-uniform background within a single GFP image, and to make the matters worse the non-uniform background intensity can vary over time in a sequence of GFP images of the same targets. This motivated us to develop adaptive and automated background correction procedures, based on data-driven statistical methods, instead of the commonly-used physical standard for flat-field correction.

Non-uniform Background Estimation by Statistical Procedures

We treat an GFP microscopy image as 2D surface data: let $z(x, y)$ denote the pixel intensity at the position with coordinate (x, y) in a rectangular grid. To define a model for the observed image data, it often proves best to analyze the data after applying a transformation f, as $g(x, y) = f(z(x, y))$, and entertain a model where background, signal, and measurement noise combine additively: $g(x, y) = M(x, y) + S(x, y) + \epsilon(x, y)$.

Here $M(x, y)$ is the background noise (large scale, smooth), $S(x, y)$ is the true signal, and $\epsilon(x, y)$ is the high-frequency spotted noise. The goal is to develop spatial models for the background function M and to study estimation methods of the background function based on the observed data g, even though the signal S is unknown.

One approach to this problem consists of employing a robust fitting criterion during model fitting so that the cell signals do not unduly affect the estimate of the smooth, large-scale background. We have considered both global models such as the Zernike polynomial representation and local polynomial approximation using LOESS-like methods.

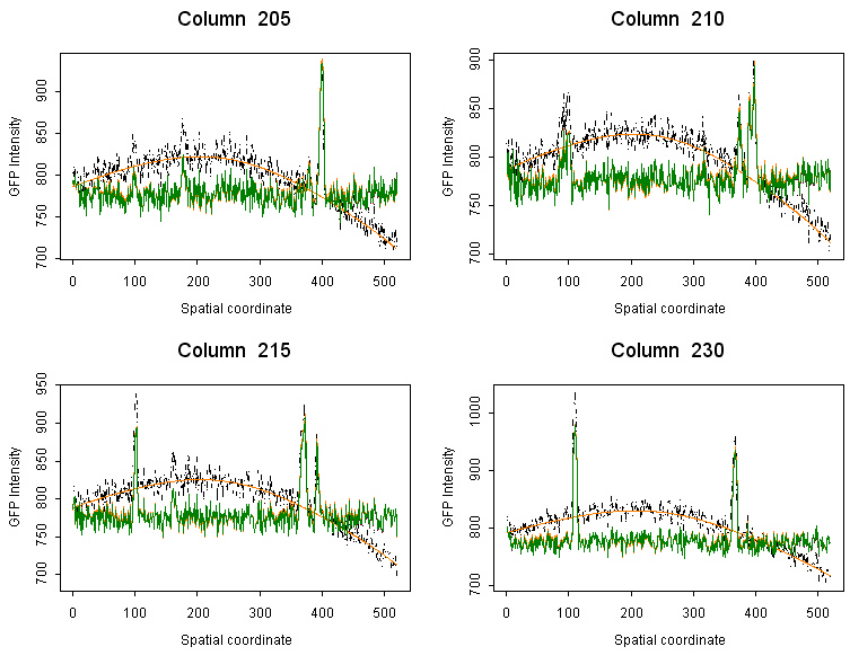
The commonly used method of local background estimation may be regarded as a special case of our second method since it amounts to a locally constant fit. Our statistical approach provides an automated approach to background estimation based on the observed GFP image data alone, without resorting to any physical standards. The figure illustrates typical results. This is a

quick, simple-to-use procedure for flattening the background of GFP images even when the background is non-uniform and may vary between images in a sequence taken over time.

Flat-field Correction

Flat-field correction usually refers to a technique using a physical standard to improve image quality by removing artifacts from 2D images that are caused by variations in the pixel-to-pixel sensitivity of the detector or by distortions in the optical path. Once a detector has been appropriately flat-fielded, a uniform signal will create a uniform output.

For flat-field correction (also called *shading correction*), an image of a blank field (containing no cells) is collected for each channel using the same exposure times and acquisition settings that will be used when collecting images of cells. The "blank" field should be an empty area of a coverslip in the same focal plane as the cells. Flat field adjustment then amounts to rescaling (simple division) the intensity of each pixel. (This assumes, of course, that a multiplicative, instead of an additive, correction is most appropriate for flatfielding.) We believe that our statistical approach to background estimation, and our statistically rigorous approach for background correction will become even more relevant in complex imaging situations, for example in movies of live cells that have been labeled with GFP, because the physical standard is static.



Line profiles of raw images (black) and background-corrected images (green). The red solid lines denote background estimated by the statistical method. The spikes indicate areas in the cells with detectable GFP signals.

38 Functional Data Analysis of Irregularly Sampled Curves in a Round-robin Measurement Study of Seebeck Coefficient Reference Materials

AUTHOR Z.Q. John Lu

COLLABORATORS Winnie Wong-Ng (Ceramics Division, MSEL, NIST), Nathan D. Lowhorn (Former postdoc at NIST, now with the Physics Department at Middle Tennessee State University), Weiping Zhang (Former guest researcher at NIST, now with the Department of Statistics and Finance of the University of Science and Technology of China)

Introduction

The Seebeck coefficient of a material measures the magnitude of an induced thermoelectric voltage in response to a temperature difference across that material. The Seebeck coefficient has units of V/K. If the temperature difference ΔT between the two ends of a material is small, then the thermopower of a material is defined (approximately, see for example, *Thermoelectrics: Basic Principles and New Materials Developments* By G.S. Nolas, J. Sharp, H. J. Goldsmid, 2001, Springer) as $S = \Delta V/\Delta T$ and a thermoelectric voltage ΔV is seen at the terminals.

In practice, real materials often have both positive and negative charge-carriers, and the sign of S usually depends on which of them predominates. Values in the hundreds of $\mu V/K$, negative or positive, are typical of good thermoelectric materials. The efficiency of a thermoelectric material is directly related to the figure of merit $Z = \sigma S^2/\lambda$, where σ is the electrical conductivity, λ is the thermal conductivity, and S is the Seebeck coefficient.

There is a persistent need for a Seebeck coefficient standard reference material (SRM) for ensuring reliable measurements and characterization. Though there are low Seebeck coefficient materials, there is no high Seebeck coefficient SRM. To fill this gap, NIST has initiated a round-robin measurement survey study of two materials, bismuth telluride (Bi_2Te_3), and Constantan (a copper-nickel alloy), in collaboration with 11 other leading research laboratories worldwide.

Round-robin measurement data

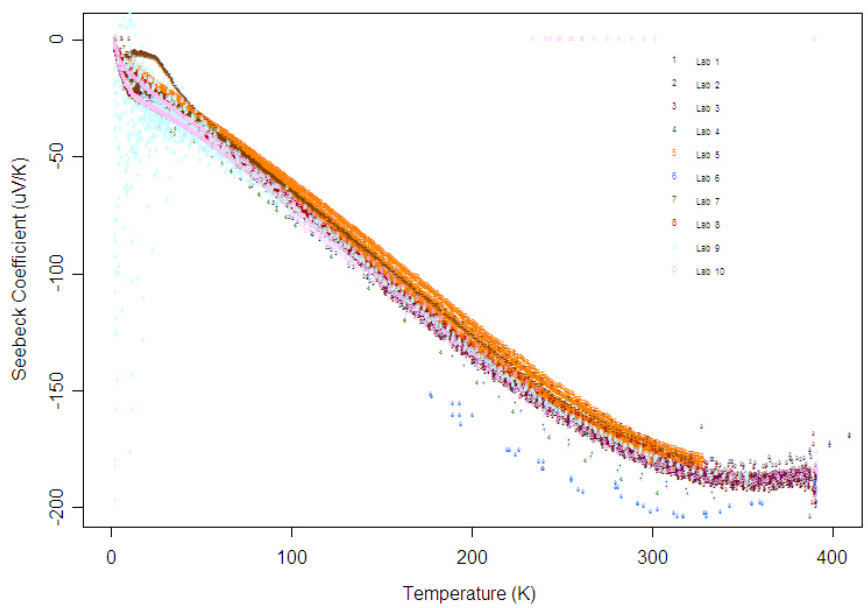
The measurements were conducted in two rounds whereby each sample was measured by 2 different laboratories: this provided a good amount of comparative data while working within the time constraints of the project and the participants. There were considerable differences between the measurement techniques used by the different labs, and between the skill levels of the personnel involved: as a result, the measurements range vary greatly between the labs. The figure shows the measurement survey data on bismuth telluride.

Statistical Analysis and Results

Ideally, one would like to separate the different experimental effects, but this is a highly challenging task in the context of functional data analysis. First, the experimental limitations have

led to a data set with strong confounding between the effects of laboratories and of measurement techniques. Because each sample is measured by only two laboratories, it is difficult to separate sample-to-sample variability from the laboratory effects. Secondly, the fact that the measured data points and the data range vary greatly among the labs, makes comparing the resulting Seebeck coefficient curves very difficult.

To address the second difficulty, considerable effort went into finding an interpolating model that applies to all individual measurement data sets: this turned out to be very fruitful. Owing to the sparsity of the observations, and to differences in the ranges of the measurements in the replicated data sets from a significant number of laboratories, we chose to use ridge regression for data fitting, in order to obtain reasonably stabilized parameter estimates. To compare different data sets pertaining to the same functional relationship, we recommend pointwise comparison at a common set of temperature values in which the Seebeck coefficient values, if not directly observed there, can safely be interpolated to these common temperatures based on the fitted parametric interpolation model. Results of these analyses, including how the choice of bismuth telluride has been made in producing a NIST SRM 3451 for low-temperature Seebeck coefficient are given in a forthcoming paper in the Journal of Research of NIST.



Accepted measurements made on Bi_2Te_3 by 10 labs (possibly using different techniques and different samples). The replicates from the same lab are shown in the same color.

39 Antonio Possolo

Biography

Born in Lisboa, Portugal, a naturalized U.S. citizen, living in the U.S. since 1978, earned a Licenciate in Geology from the Classical University of Lisboa, Portugal, and a Ph.D. in statistics from Yale University (under John Hartigan's guidance).

Pre-doctoral professional engagements: lecturer in the Classical University of Lisboa; geologist of the Geological Survey of Portugal; Teaching Assistant for Frank Anscombe (Yale); Research Assistant to Felix Chayes (Geophysical Laboratory, Carnegie Institution of Washington).

Post-doctoral professional engagements: Assistant Professor in the statistics departments of Princeton University and of the University of Washington in Seattle; Associate Technical Fellow of The Boeing Company (1989-2000); statistician of The General Electric Company (2000-2006); *Chief* of NIST's Statistical Engineering Division since October 1st, 2006.

Professional interests: spatial statistics, point processes, environmental remote sensing, measurement uncertainty, foundations of probability theory and of statistical inference. Other interests: classical music (romantic repertoire and *Bel canto* opera), literature, soccer.

Selected Publications

A. Possolo, M. Kasperski and E. Simiu (2008). Tunable compression of wind tunnel data, *Journal of Engineering Mechanics* (submitted).

M. Harkness, A. Fisher, M. Lee, E. E. Mack, J. A. Payne, J. Roberts, S. Dworatzek, A. Possolo, and C. Acheson (2008). Reductive dechlorination of high levels of TCE in a multi-lab, statistically-based microcosm study, *Environmental Science and Technology* (submitted).

A. Possolo, B. Toman and T. Estler (2008). Contribution to a conversation about the Supplement 1 to the GUM, *Metrologia*, 46, L1–L7.

A. Possolo and B. Toman (2007). Assessment of measurement uncertainty via observation equations, *Metrologia*, 44, 464–475.

J. Silkworth, A. Koganti, K. Illouz, M. Zhao, & S. B. Hamilton (2005). Comparison of TCDD and PCB CYP1A induction sensitivities in fresh hepatocytes from human donors, Sprague-Dawley rats, and Rhesus monkeys and HepG2 Cells *Toxicological Sciences* 87(2): 508–519.

A. Possolo (2005) High-Dosage Vitamin E Supplementation and All-Cause Mortality. *Annals of Internal Medicine* 143: 154.

40 Geochemical Atlas of the United States

AUTHOR Antonio Possolo
COLLABORATORS Andrew Grosz (Eastern Mineral Resources Team, United States Geological Survey, U.S. Department of the Interior), James Yen (Statistical Engineering Division, ITL, NIST)

Introduction

The United States Geological Survey (USGS), in collaboration with other federal and state government agencies, industry, and academia, is conducting the National Geochemical Survey (NGS), `http://tin.er.usgs.gov/geochem/`, to produce a complete geochemical coverage of the U.S. that will (i) enable the construction of geochemical maps, (ii) refine estimates of baseline concentrations of chemical elements in the sampled media, and (iii) provide context for a wide variety of studies in the geological and environmental sciences.

The collaboration between the USGS and NIST has focused on methods for geochemical mapping that can exploit the information in censored data, or *non-detects* (NDs), and that facilitate the identification of promising prospecting regions that may contain mineral deposits of economic value, by focusing on functions of geochemical elemental abundances that are indicative of the underlying mineralogy.

Imputing Non-Detects

Consider the problem of drawing a map depicting values of Fe/Ti, the ratio of the mass fractions of iron to titanium in NGS samples, over the eastern board of the United States. Rather than assigning arbitrary values to NDs, we exploit the only definite information that we have about them: that the corresponding value of Fe/Ti is either below or above its apparent value depending on whether it is iron or titanium that is below the detection limit.

To assign values to these censored observations, we start from a map for Fe/Ti drawn using complete data only, based on a geostatistical model (either kriging or local geographic regression, described below), and compute the expected value of the ratio at each location where either iron or titanium is ND, conditionally on the values of the ratio at sites with complete data, and taking into account the (upper or lower) bound that the ND imposes on the ratio. Then we redraw the map and repeat the procedure until the imputed values, and the map as a whole, no longer change.

Mapping Techniques

Two alternative techniques have proven useful to depict the spatial variation of Fe/Ti over the eastern region, while exploiting the information in NDs. *Local Geographic Regression* (LGR) involves no assumptions about the probability distribution of the data, other than for the acknowledgment that log(Fe/Ti) varies fairly smoothly across the region of interest. *Kriging* is carried out on the assumption that the values of log(Fe/Ti) are like outcomes of a Gaussian random field.

The local geographic regression interpolant is a function' such that log(Fe=Ti) = ' (lon , lat)+ E with suitably small deviations E, where lon and lat denote geographical coordinates, and' is locally quadratic in the sense that, in a small neighborhood of the location with coordinates (lon , lat), where an interpolated value of Fe=Ti is sought, ' (lon , lat) is a second-degree polynomial in the geographical coordinates, tted by weighted least squares, where the weights depend on the distance between the data points and the interpolation location.

For the imputation of values of the ratio corresponding to NDs, we have assumed that log(Fe=Ti) has an approximate Gaussian distribution with mean given by the value estimated by the local regression, and with standard deviation set equal to the local regression's residual scale estimate. This neglects the fact that the deviations E at different locations may be correlated, and likely underestimates that standard deviation.

Ordinary kriging was applied on the assumption that the values of log(Fe=Ti) are approximately like an outcome of a stationary Gaussian random eld, with Matérn's covariance function. The values of Fe=Ti corresponding to NDs were imputed iteratively as the conditional expectation given the values of log(Fe=Ti) at locations with complete data, and given also the censoring value.

Mapping Uncertainty

Statistical cross-validation was employed to ascertain the quality of the geochemical maps. We partitioned the NGS data into one hundred sets with approximately 1% of the samples in each. Then, we tted the model that the map is based on one hundred times, each time leaving one of these sets of samples out as test set and using the other ninety-nine as training set. Then we compared the cross-validation prediction errors with the corresponding estimates of analytical error, which in turn had been derived by comparing replicated measurements that are available for some of the samples in NGS. Since the prediction errors are quite comparable to the analytical errors, both for the kriging and local geographic regression models, both mapping techniques are quite reasonable for this data.

Kriging Interpolant & LGR's Performance. The left panel shows the map produced by the kriging interpolant: the map produced by local geographic regression is similar. The right panel shows the spatial variability of the LGR prediction error (absolute value of the difference between Fe=Ti values that were measured and those that were predicted by cross-validation), expressed as a proportion of the size of the measured values.

43 Andrew Rukhin

Biography

Andrew Rukhin was born in Russia, and graduated from the Leningrad State University in 1967. In 1970 he defended his Ph. D. thesis "Probabilistic and Statistical Problems on Groups" and worked at the Steklov Mathematical Institute until his emigration to the USA in 1976. During his academic career he was a Professor at Purdue University (1977–1987), at the University of Massachusetts, Amherst (1987–1989), and at the University of Maryland at Baltimore County (1989–2008). Now he is a full time employee at SED working on statistical issues of metrology.

Awards

Fellow of American Statistical Association (1998)
Senior Distinguished Scientist Award By Alexander von Humboldt-Foundation (1990)
Youden Prize for Interlaboratory Studies (1998, 2008)
Advisory Editor of "Journal of Statistical Planning and Inference"

Selected Publications

Joint Distribution of Pattern Frequencies and Multivariate Polya-Aeppli Law, *Theory of Probability and its Applications*, 2009.

Confidence Regions for Parameters in Linear Models, *Statistica Sinica*, 2009.

Estimation and Testing for the Common Intersection Point, *Chemometrics and Intelligent Laboratory Systems*, 90, 2008, 116–122.

Statistics in Metrology: International Key Comparisons and Interlaboratory Studies, (with N. Sedransk) *Journal of Data Science* 7, 2007, 393–412.

Statistical Aspects of Linkage Analysis in Interlaboratory Studies, (with W. E. Strawderman), *Journal of Statistical Planning and Inference* 137, 2007, 264–278.

Estimating Common Vector Mean in Interlaboratory Studies, *Journal of Multivariate Analysis* 98, 2007, 435-454.

Distribution of the Number of Words with a Prescribed Frequency, and Tests of Randomness, *Advances in Applied Probability* 34, 2002, 775-797.

44 Weighted Means Statistics in Interlaboratory Studies

AUTHOR Andrew Rukhin

Introduction.

Stochastic modeling and analysis of international key comparisons (interlaboratory comparisons) pose several fundamental questions for statistical methodology. A key comparison is specifically designed to assess the degree of equivalence of calibrations by participating national metrology laboratories at a few, "key", settings for a particular measurement process. Controversy over appropriate choice of statistical model focuses on derivation of the key comparison reference value (KCRV) and its associated uncertainty. Most of the existing procedures are based on weighted means estimators with weights determined from uncertainties reported in the uncertainty budget. These uncertainties are of two types: (i) statistical estimate of standard deviations ("Type A") and (ii) expert scientific judgment ("Type B") of both systematic differences and extra-variation of each laboratory. Usually, the uncertainties involved in their assessment are ignored.

Estimators of key comparison reference value and their uncertainty

Assume there are p laboratories, each measuring the unknown underlying (nonrandom) value μ common to all laboratories. In the simplest random effects model the measurements $\{x_{ij} : i = 1, \ldots, p; j = 1, \ldots, n_i\}$, are of the form $x_{ij} = \mu + \ell_i + e_{ij}$, with independent Gaussian errors $e_{ij} \sim N(0, \kappa_i^2)$ and zero mean between-lab effects ℓ_i with variance σ_{B}^2. All parameters $\mu, \sigma_{\mathrm{B}}^2, \kappa_i^2$, $i = 1, \ldots, p$ are unknown, but the goal is to estimate μ or, more importantly, to provide a confidence interval for μ. The fairly small sample sizes typical in metrology may not allow robust nonparametric inference; out of parametric models ours is the simplest and most widely applicable.

Let $x_i = \sum_j x_{ij}/n_i$ be the within lab means and $s_i^2 = \sum_j (x_{ij} - x_i)^2/[n_i(n_i - 1)]$, (unbiased) estimates of the variances $\sigma_i^2 = \kappa_i^2/n_i$ of x_i. When these variances σ_i^2 are known and $\sigma_{\mathrm{B}}^2 = 0$, the best (in terms of the mean squared error) unbiased estimator of the reference value μ is a weighted means statistic, $\widetilde{x} = (\sum_{i=1}^p w_i x_i)/(\sum_{i=1}^p w_i)$, with $w_i = w_i^{\mathrm{tr}} = \sigma_i^{-2}$. The formula for the variance, $\mathbb{V}(\widetilde{x}) = \mathbb{E}(\widetilde{x} - \mu)^2 = (\sum_i w_i^{\mathrm{tr}})^{-1}$. However, in practice the variances σ_i^2 or the "true" w_i^{tr} are *unknown*. The usual suggestion is to replace σ_i^2 by their estimates s_i^2, i.e. to estimate $\mathbb{V}(\widetilde{x})$ by $1/\sum_{i=1}^p s_i^{-2}$. Although s_i^2 estimates σ_i^2 unbiasedly, this estimate *underestimates* $\mathbb{V}(\widetilde{x})$.

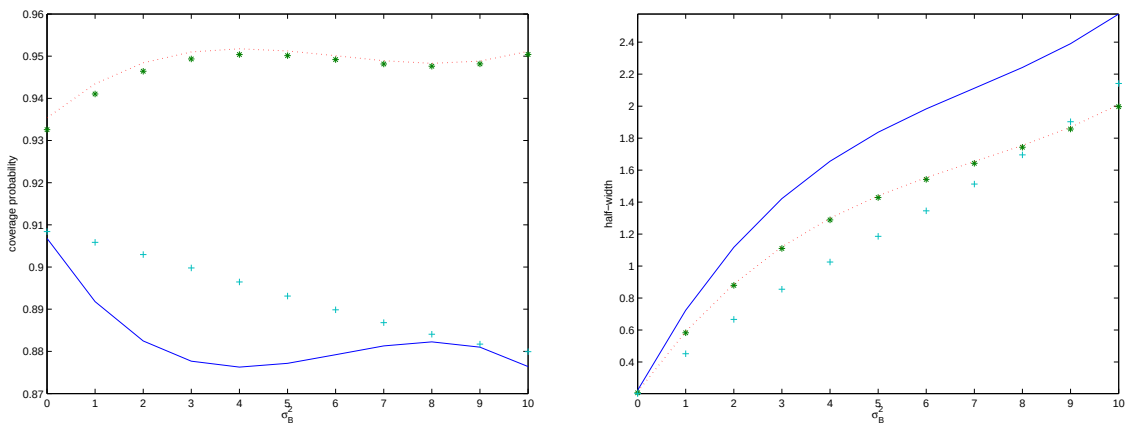
The traditional statistical procedure, the maximum likelihood estimator of μ, does not have an explicit form, although it is a weighted means statistic with the weights inversely proportional to the maximum likelihood estimates of σ_i^2. There are numerical algorithms for its evaluation, in particular in R-language. Alternative simpler procedures in our situation include the sample mean $\bar{x}$ and the so-called Graybill-Deal estimator, $\widetilde{x}_{\mathrm{GD}} = (\sum_{i=1}^p x_i s_i^{-2})/(\sum_{i=1}^p s_i^{-2})$, which merely is the plug-in-version of $\widetilde{x}$. Estimator $\widetilde{x}_{\mathrm{GD}}$ is popular among metrologists. It is used when calculating CODATA recommended values of the fundamental physical constants. However, it has a serious drawback, namely, small values of s_i^2 lead to unjustifiably large weights.

Confidence Intervals

The within-labs variances σ_i^2 can be estimated by the available estimates s_i^2, but the problem of estimating the between-study variance σ_{B}^2 remains. Weighted means statistics with the weights of the form $w_i = (y + s_i^2)^{-1}, y > 0$, are much less sensitive than the Graybill-Deal weights to small values of s_i^2 because of positive y. Indeed presence of y makes it impossible for one laboratory to dominate all others unless all labs produce similar results (in which case σ_{B}^2 is estimated by zero.) The limiting case $y = \infty$ corresponds to the arithmetic (sample) mean with equal weights.

Excellent choices for y can be obtained from the DerSimonian-Laird method, or from the Mandel-Paule algorithm, both of which employ the idea behind the method of moments. A good estimator of $\mathbb{V}(\widetilde{x})$ for any weighted means statistic can be derived from the *almost unbiased* estimator suggested in the context of linear models by Horn, Horn and Duncan. This estimator, $\delta = \sum_1^p \omega_i^2 (x_i - \widetilde{x})^2/(1-\omega_i)$, has weights ω_i proportional to $(y + s_i^2)^{-1}$, where y corresponds to the choice of the weighted means statistic.

Simulations show that confidence intervals of the form, $\widetilde{x} \pm t_{\alpha/2}(p-1)\sqrt{\delta}$, for the Mandel-Paule rule and for the DerSimonian-Laird procedure, outperform the traditional interval based on the Graybill-Deal estimator.



The coverage probability of confidence intervals and their half-widths. The left panel shows the coverage probability of confidence intervals of the nominal confidence coefficient 95 % and $p = 12$ plotted against σ_{B}^2 when the variance estimator is δ. Both the DerSimonian-Laird estimator (line marked by $*$) and the Mandel-Paule procedure (dotted line) sustain this confidence level very well. The Graybill-Deal estimator (continuous line) cannot be recommended as it has the smallest coverage probability giving the widest interval. The maximum likelihood estimator (line marked by $+$) has confidence coefficient smaller than 95 %, but it is shorter.

45 Statistical Aspects of Linkage Analysis in Interlaboratory Studies

AUTHOR Andrew Rukhin
COLLABORATORS Bill Strawderman (Rutgers University), Stefan Leigh (Statistical Engineering Division, ITL, NIST), and David Evans (Manufacturing Metrology Division, MEL, NIST)

Introduction

Modeling and analysis of international interlaboratory data present many fundamental questions for the statistics profession. Particularly challenging is statistical modeling for interlaboratory studies known as Key Comparisons when one has to link several comparisons to or through existing studies. The proposed approach to the analysis of such a data uses Gaussian distributions with heterogeneous variances also employed in meta-analysis. We developed conditions for the existence and uniqueness of uniformly minimum variance unbiased estimators of the contrast parametric functions. When they are not unique, their optimal combinations are derived with estimates of their uncertainty.

Mutual Recognition Arrangement and Key Comparisons

The Mutual Recognition Arrangement (MRA) (1999) for national measurement standards and calibration and measurements certificates is a principal feature of international cooperation for measurement quality assurance. The MRA is realized through Key Comparisons (KC) which typically involve a number of laboratories with several of them (typically National Metrology Institutes), serving as the *pilot* laboratories designed to coordinate the whole study. Each of the regional laboratories analyzes its measurements and reports the results consisting of its estimate of the measurement value along with the combined standard uncertainty. The key comparison reference value (KCRV) and its associated uncertainty are determined on the basis of these characteristics. One of the goals is to establish the degree of equivalence of measurements made by participating laboratories and to quantify this characteristic for all pairs of laboratories.

Commonly one has to link several comparisons to or through existing KC studies. In many situations the laboratories use a transfer instrument to assess the value of a laboratory standard and to compare the relative biases of their measurement processes and standards. Two important parts of such comparisons are estimates of the difference between two artifacts or between two laboratory effects (the mentioned degrees of equivalence.) There are situations where a direct comparison is not possible because the laboratories have not participated in the same KC study or have not measured the same artifact. In the simplest case there is just one linking laboratory which has made measurements on the artifacts common to these two laboratories, but there could be several of such links.

The MRA does not specify exactly how to perform the linkage. Most of the existing proposals treat the statistical estimates of uncertainties as known constants which could lead to artificially small confidence intervals for contrasts.

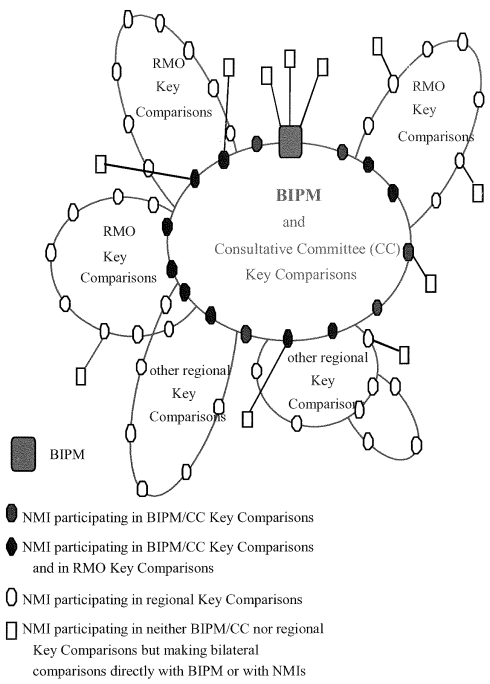
Linking CCEM-K4 and the EUROMET Project 345

A motivating example for this work was the study designed to link two existing Capacitance Standards Key Comparisons: CCEM-K4 and the EUROMET project 345 (10 pF results). These two interlaboratory studies were carried out by the Consultative Committee for Electricity and Magnetism (CCEM) and by the European Metrology Cooperation (EUROMET) in the late nineties. Six national institutes served as linking (pilot) laboratories. These institutes participated in both Key Comparisons, while ten (regional) institutes were additional members of the EUROMET project 345. One can formulate the goal as evaluation of the correction to the measurements of the labs participating only in the EUROMET project 345 to obtain the best estimate of what would have been the result from such a laboratory had it actually participated in CCEM-K4. After this correction has been found, the table of pairwise laboratory contrasts or bilateral equivalences along with associated (combined) uncertainties is determined.

The suggested procedure was implemented later for linking other studies, namely SIM regional comparison SIM.AUV.V-K1 and CCAUV.V-K1, which measured sensitivity of accelerometers.

Statistical Model

We used a natural model, which assumes that several laboratories serve as pilot laboratories measuring two of a given number of artifacts (or participating in two out of several different studies.) Non-pilot laboratories measure only one of these artifacts with data in each laboratory having an additive error structure with unknown (unequal) error variances. One can form a graph with vertexes representing the laboratories, and edges connecting any two vertexes measuring the same artifact. Each of the paths connecting two labs can be used to derive an estimate of the difference between their effects. The statistical issue is how to combine these estimators in the best possible way. The derived procedure leads to an estimator of the artifact contrasts and to the desired degrees of equivalence.



Linkage problem envisioned by BIPM.

46 Jolene Splett

Biography

Jolene Splett graduated from the University of Wyoming at Laramie in 1986 with a Bachelor of Science degree in mathematics and statistics, and with a Master of Science degree in statistics in 1988.

Jolene has been at NIST's Statistical Engineering Division since 1988, except for part of 1995-98, when she was a marketing analyst for Colwell Systems, Inc., in Mendota Heights, Minnesota, and a statistician for the National Marrow Donor Program in Minneapolis, Minnesota. Her areas of interest include statistical software, linear and nonlinear regression modeling, experiment design, and uncertainty analysis.

Awards

NIST Calibration Program Measurement Services Award, 1994.
University of Wyoming Statistics Department Institutional Membership Award, 1988.
Colorado/Wyoming Chapter of ASA Maurice Davies Award, 1986.

Selected Publications

McLinden, M. O. and Splett, J. D. (2008), A Liquid Density Standard Over Wide Ranges of Temperature and Pressure Based on Toluene, *NIST Journal of Research*, Vol. 113, No. 1, pp. 29–67.

Splett, J. D., McCowan, C. N., Iyer, H. K., and Wang, C.-M., *NIST Recommended Practice Guide: Computing Uncertainty for Charpy Impact Machine Test Results,* NIST Special Publication 960-18, September, 2007.

Goodrich, L. F. and Splett, J. D. (2007), Current Ripple Effect on n-Value, *IEEE Transactions on Applied Superconductivity*, Vol. 17, No. 2, 2603–2606.

Jargon, J. A., Splett, J. D., Vecchia, D. F., and DeGroot, D. C. (2007), An Empirical Model for the Warm-Up Drift of a Commercial Harmonic Phase Standard, *IEEE Transactions on Instrumentation and Measurement*, Vol. 56, No. 3, pp. 931–937.

Coakley, K. C., Splett, J. D., Janezic, M. D., and Kaiser, R. F. (2003), Estimation of Q-factors and Resonant Frequencies, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 51, No. 3, pp. 862–868.

47 Low-Count Isotopic Ratios

AUTHORS Jolene Splett and Kevin Coakley
COLLABORATORS David Simons (Surface and Microanalysis Science Division, CSTL, NIST)

Introduction

Our work is motivated by the analysis of isotopic ratio data collected at NIST where the minor isotope count is very low and the major isotope count is very large. We consider experiments where instruments yield count data that can be modeled as realizations of a Poisson process with expected value $\mu_S + \mu_B$ where μ_S is the expected contribution due to a signal of interest, and μ_B is the expected contribution of a background process. That is, $n_{obs} \sim \text{Poi}(\mu_S + \mu_B)$. Given the measured value n_{obs} and an estimate of μ_B from an independent background-only experiment, we construct uncertainty intervals and detection probabilities for μ_S. We compare the frequentist coverage properties of various methods for low count Poisson signals contaminated by background.

The statistical problem we study occurs in a variety of application areas including: physics and astroparticle physics, monitoring the processing of nuclear materials for homeland security, isotopic ratio analysis (when the major isotope is large enough so that most of the variability in the ratio is due to the minor isotope), and the detection of low-level radiation.

Methods

There has been great interest in the physics (Mandelkern, 2002) and statistics (Efron, 2003) communities regarding the uncertainty analysis of low count Poisson signals immersed in background. We study three methods that address this problem: the Feldman Cousins (1998) method, the randomized Feldman Cousins method, and a Bayesian method due to Loredo (1993).

The Feldman Cousins (FC) method is an implementation of a frequentist Neyman procedure that utilizes a likelihood ordering approach that automatically selects interval endpoints for the case where the background, μ_B, is assumed to be known.

The FC method was extended by Conrad, et al. (2003) to account for systematic uncertainties in μ_B. We denote this method as the randomized Feldman Cousins (RFC) method because μ_B is treated as a random nuisance parameter. We developed a new implementation of this approach where uncertainty in the background parameter μ_B is accounted for with a parametric bootstrap method. As in the FC method, the upper and lower interval endpoints for the RFC method are determined automatically.

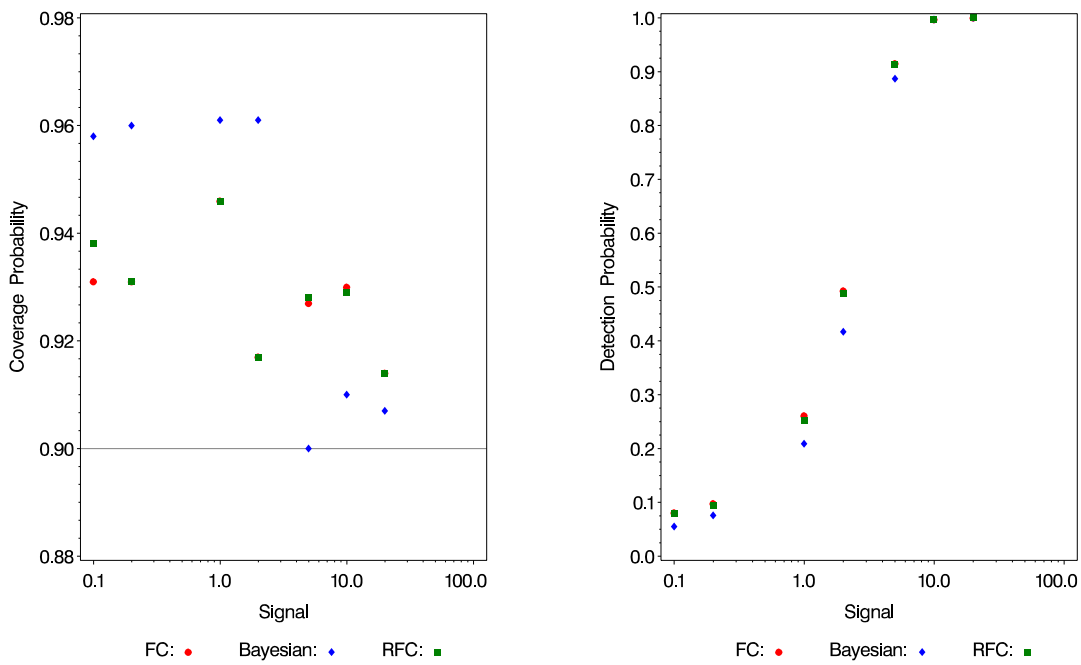
We also determine the posterior probability density function for μ_S with a Bayesian method based on uniform priors. We select Bayesian credibility intervals based on a minimum length criterion.

Discussion

We simulate realizations of data given μ_S and μ_B based on data motivated by NIST experiments in which the signal is weak and the observing time for the background-only measurement is 25 times longer than for the primary experiment. For the cases studied, the fractional uncertainty in the estimate of μ_B ranges from 14 % to 44 % for the FC method.

We quantify the probability that μ_S falls in the uncertainty interval as well as the probability of detecting a signal for each method. The figure summarizes coverage and detection probabilities for the case where $\mu_B = 1$.

Based on our simulations, we found that all methods perform well for the high-signal cases. In general, the FC method has better coverage than the both the Bayesian and RFC methods since the FC coverage was closer to the target coverage in about 66 % of the cases. The FC method also has uniformly higher detection probabilities than the Bayesian method.



Probabilities. Coverage (left panel) and detection (right panel) probabilities for FC, Bayesian, and RFC methods for 90 % intervals using $\mu_B = 1$. For the case where $\mu_S = 0$ (not shown), the coverage probabilities are 0.940 ± 0.005, 1.000 ± 0.000,and 0.945 ± 0.005 for FC, Bayesian and RFC methods, and the detection probabilities are 0.061 ± 0.005, 0.044 ± 0.005, and 0.056 ± 0.005.

48 Critical Current Metrology for Nb_3Sn Conductor Development

AUTHORS Jolene Splett
COLLABORATORS Loren Goodrich, Najib Cheggour, and Jack Ekin (Electromagnetics Division, EEEL, NIST)

Introduction

The main focus of the project is to develop standard techniques for the measurement of critical current of high-temperature and low-temperature superconductors. Critical current is the maximum current that can be carried by a superconductor before the superconductor starts to become resistive. Some applications for which these types of measurements are crucial include: magnetic-resonance imaging, research magnets, fault-current limiters, magnetic energy storage, motors, generators, transformers, transmission lines, synchronous condensers, high-quality-factor resonant cavities for particle accelerators, and superconducting bearings.

Superconductors also have the potential for making a significant impact in enabling practical use of energy derived from nuclear fusion reactions. These are a potential, virtually inexhaustible energy source for the future, do not produce greenhouse gases, and are less likely than fission reactions to endanger the natural environment. Superconductors are used to generate the ultra-high magnetic fields that confine the plasma in fusion energy research. EEEL staff measure the magnetic hysteresis loss and critical current of high-current Nb_3Sn superconductors for fusion and other research magnets.

The Statistical Engineering Division (SED) supports two main aspects of this project: (1) the development of an algorithm to determine the irreversible strain limit of Nb_3Sn superconductors; and (2) fitting non-linear strain scaling, temperature scaling, and unified scaling models for the joint effects of temperature and strain to critical current data.

Irreversible Strain

A superconducting wire is sensitive to many environmental conditions during measurement, including the amount of strain applied to the wire. A small amount of strain may not effect the performance of the wire, however if a wire is exposed to too much strain (compressive or extensive) the damage is irreparable. Knowing the physical properties of a superconducting wire is required to develop high quality devices. SED staff have developed an algorithm that quantifies the strain at which the wire is permanently damaged, called the irreversible strain limit.

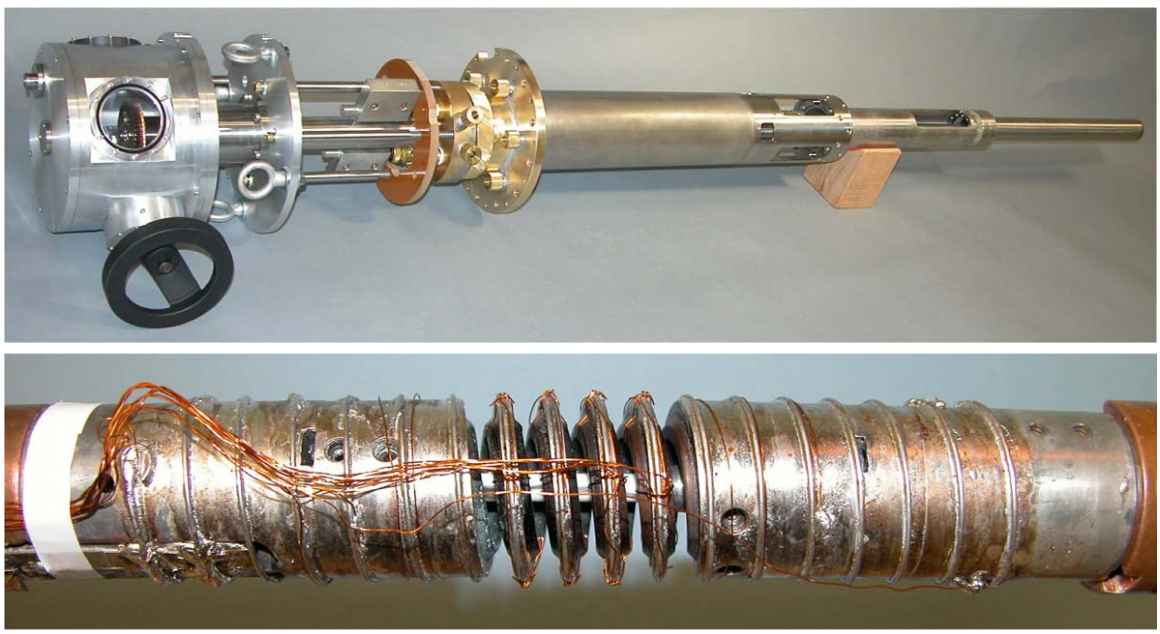
Unified Scaling

EEEL staff have completed the construction and testing of a variable-temperature and variable-strain, or unified, apparatus for measuring critical current. The apparatus combines world class capabilities in variable-temperature and variable-strain measurements and is expected to be the highest-current apparatus of its type in the world. The new apparatus will help answer fundamental questions about the performance of strain sensitive superconductors. The figure

shows the apparatus in the new variable strain and variable temperature measurement system (top) and the spring that fits inside the apparatus with a mounted sample (bottom).

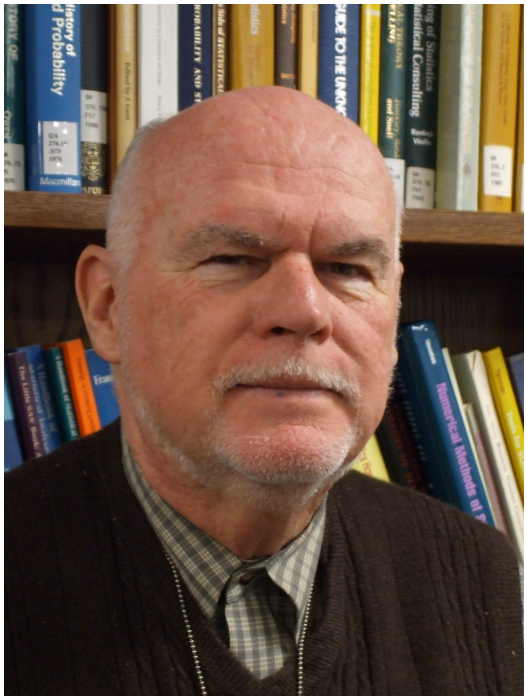
Measurements taken on the new apparatus facilitate the investigation of empirical scaling models for the joint effects of magnetic field, temperature, and strain on critical current. Unified scaling models are the product of individual, nonlinear models for temperature, strain, and field. Because of the modular nature of unified scaling models, there are many different models currently in use. A long-term objective of this project is to provide some guidance to the superconducting community regarding the best scaling models. Scaling models will be fit using nonlinear least squares and evaluated based on the usual regression diagnostics. We will also examine the physical interpretation of estimated parameters.

SED staff have succeeded in fitting three scaling models (temperature, strain, and combined temperature and strain) to critical current data. The data and subsequent model fits will be used to verify or determine the limits of scaling laws. Such information would greatly reduce the amount of data and liquid helium required to measure new samples in the future.



Apparatus. The top photograph shows the new high-current apparatus constructed at NIST to measure the critical-current dependence on strain, temperature and magnetic field. The worm-wheel that torques the spring can be seen through the small, round window. The lower photograph shows the CuBe spring with a helical sample soldered to the spring. Three pairs of voltage taps cover the three central turns of the spring. The current contacts are made at each end of the spring.

49 Bill Strawderman



Biography

Bill Strawderman was born in Westerly, Rhode Island in 1941. He received a BS in Engineering from the University of Rhode Island in 1963, an MS in mathematics from Cornell in 1965 and an MS in 1967 and PhD in 1969 in statistics, both from Rutgers. He was a member of technical staff at Bell Labs (Holmdel) from 1965 to 1967, and taught at Rutgers from 1967 to 1969. He taught at Stanford from 1969 to 1970, and returned to Rutgers in 1970, where he has since remained. He served as Chair of the Department of Statistics at Rutgers for 9 years. He has held visiting posts at Princeton, the University of Paris, the University of Rouen (France) and the University of Rome, and served as an Adjunct Professor of The University of Medicine and Dentistry of New Jersey.

He is a member of IMS, ASA, and ISI, has served as associate editor for JASA, The Annals of Statistics, and The IMS Lecture Notes Series. He has served two terms on the Council of the IMS, been Vice President and President of the New Jersey Chapter of ASA and Chair of the Bayes section of ASA.

Awards

Fellow of IMS and ASA; Rutgers University Graduate School Distinguished Alumnus Award in Science; Distinguished Alumnus Award of the Rutgers Statistics Department; 2008 Youden Award for Interlaboratory Testing (joint with Andrew Rukhin).

Selected Publications

Statistical aspects of linkage analysis in interlaboratory studies, JSPI (2007), 137, 264–278 (with Andrew Rukhin)

Statistical analysis of the multiple artifact problem in key comparisons with linear trend, *Metrologia* (2006), 23, 21–26 (with Nien Fan Zhang, Hung-kung Liu and Nell Sedransk)

A new class of generalized Bayes minimax ridge regression estimators, *Annals of Statist.* (2005), 33, 1753–1770 (with Yuzo Maruyama)

On the construction of Bayes minimax estimators, *Annals of Statistics* (1998), 26, 660–671. (with Dominique Fourdrinier and Martin T. Wells)

50 Simultaneous Estimation and Reduction of Non-conformity in Interlaboratory Studies

AUTHOR Bill Strawderman
COLLABORATORS Andrew Rukhin (Statistical Engineering Division, ITL, NIST)

Introduction

A goal in combining information across several studies is often the determination of a common reference value, which is frequently calculated as a weighted average of the individual laboratory means. Occasionally, however some of the individual means are sufficiently far from the others that it is desirable that they be excluded from, or have less influence on, the reference value.

We study procedures that shrink the vector of individual means toward a weighted mean. The resulting vector valued estimator does not necessarily give a consensus value, but does result in a reduction of the nonconformity in the individual means, and does in fact give a consensus value when the individual means are sufficiently. In the fixed effects case, our estimators are minimax under weighted quadratic loss. Our development allows the inclusion of type B error. In the random effects case, the estimators we study are not known to be minimax, but do have the advantage of further decreasing nonconformity. The methods are illustrated on a series of experiments measuring Newton's gravitational constant..

James-Stein Type Estimators

Let X be a p dimensional multivariate normal vector (of lab means) with mean vector μ and diagonal covariance matrix $\Sigma = \Sigma_1 + \Sigma_2 = \text{diag}(\sigma_1^2 + \tau_1^2, \ldots, \sigma_p^2 + \tau_p^2)$. Also, let $S = \text{diag}(s_1^2, \ldots, s_p^2)$, and $T = \text{diag}(t_1^2, \ldots, t_p^2)$, be such that the s_i's and t_i's are independent of one another and of X, and such that s_i^2/σ_i^2 has a chi square distribution with m_i degrees of freedom, and t_i^2/τ_i^2 has a chi square distribution with n_i degrees of freedom. In this setting, the σ_i^2's correspond to Type A error variances and the τ_i^2's to type B error variances. The main theoretical result is that the estimator (50.1) given below dominates the "usual" estimator of the mean vector μ under weighted squared error with weights proportional to the inverses of the population variances, $\sigma_1^2 + \tau_1^2$, when $p \geqslant 4$. This result advances the known results in shrinkage theory in that it is the first (to our knowledge) to shrink towards a weighted mean, X^*, with estimated weights, and the first to accommodate both estimated type A and type B errors. The estimator is given by

$$\delta(X, S, T) = X^* e + \left(1 - \frac{a}{(X - X^* e)(S^* + T^*)^{-1}(X - X^* e)}\right)(X - X^* e) \tag{50.1}$$

with $0 < a < 2(p-2)$, $X^* = \sum_i wiX_i$, $S^* = \text{diag}\left(s_1^2/(n_1+2), \ldots, s_p^2/(n_p+2)\right)$, and $T^* = \text{diag}\left(t_1^2/(m_1+2), \ldots, t_p^2/(m_p+2)\right)$. The weights, w_i, sum to 1 and are proportional to the elements of $(S^* + T^*)^{-1}$. The p-vector e has all components equal to 1. The positive-part

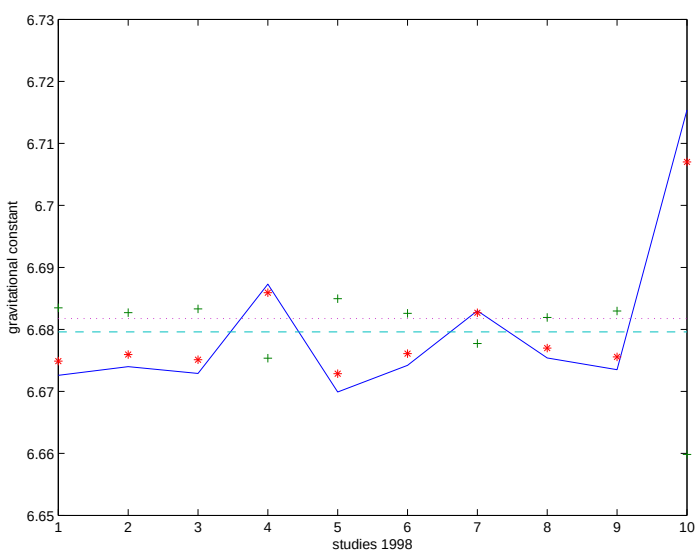
version of estimator (50.1) also is minimax and dominates the usual estimator. The positive-part estimator has the added advantage that when it differs from (50.1), all coordinates are estimated to be equal to X^*, and hence X^* gives a consensus value.

Random Effects Models

In this version of the model, $\mu_i = \mu + \lambda_i$, where μ represents the unknown consensus value and λ_i is a random effect assumed to be normally distributed with mean 0 and unknown variance σ_L^2. For reasons of identifiability, we assume all τ_i^2's to be 0. The DerSimonian -Laird estimator, X_{DL}, is a useful alternative to the Graybill-Deal-like estimator X in this setting. It uses weights that are inversely proportional to the inverses of $(s_i^2 + y_{\mathrm{DL}})$, where y_{DL} is an estimate of σ_L^2. The estimator we propose is estimator (50.1) with X^* replaced by X_{DL}, and with $(S^* + T^*)^{-1}$ replaced by $\mathrm{diag}\left((s_1^2/n_1 + y_{\mathrm{DL}}), \ldots, (s_p^2/n_p + y_{\mathrm{DL}})\right)^{-1}$. As above, the positive part version is recommended in practice. The new estimator typically shrinks the components more than estimator (50.1), and hence tends to reduce nonconformity more. However it is not known whether it is minimax.

Determination of Newton's gravitational constant

Newton's gravitational constant, G, is remarkable by the absence of any known theoretical relationship to other fundamental physical constants, so it cannot be determined by indirect measurements of other quantities. The figure gives results for 10 experimental measurements of G discussed in Mohr and Taylor (2005). It also shows the values of the version of the estimator (50.1) that shrinks toward the DerSimonian-Laird estimator, as well as an estimator proposed by Weise and Woger is also illustrated. The values of X^* and X_{DL} are also shown. It is observed that this version of the estimator (50.1) is on the opposite side of the dashed line indicating that the positive part version of the estimator is just X_{DL} itself for all coordinates.



Gravitational Constant. Measured data (solid line), the Stein Estimator (50.1) shrinking to the DerSimonian-Laird Estimator (+), the Weise-Woger Estimator (∗), the weighted mean X^* (dotted line), and X_{DL}DL (dashed line).

51 Blaza Toman

Biography

Blaza Toman studied Mathematics and Statistics, earning a Ph.D. in Statistics from the Ohio State University in 1987 with a specialization in Bayesian Optimal Design. She taught statistics at the graduate and undergraduate level at Rutgers University and at the George Washington University and was a consultant to several medical device companies on Bayesian clinical trial design and analysis. In 2000 she became a member of the Statistical Engineering Division at NIST. Her main research interests remain Bayesian statistical methods.

At NIST, she became interested in uncertainty assessment for measurements in the physical sciences, and more generally in statistical methods relevant to metrology. Here she collaborates with scientists in several fields, for example: in BFRL, analyzing data from fire models and virtual cement models; in CSTL, analyzing gas chromatography and mass spectrometry data; and in PL, analyzing data on spectral responsivity of photodiodes. She also collaborates with her colleagues in SED to develop statistical methods for interlaboratory studies and key comparisons, and for uncertainty analysis of virtual measurements.

Selected Publications

Bayesian Experimental Design for Multiple Hypothesis Testing, Journal of the American Statistical Association, 1996

Bayesian sample size calculations for binomial experiments, with A. Katsis, Journal of Statistical Planning and Inference, 1999

New Guidelines for $\delta^{13}C$ Measurements, with T. B. Coplen, W. A. Brand, M. Gehre, M. Groening, H. J. Meijer and, R. M. Verkouteren , Analytical Chemistry, 2006

Bayesian Approaches to Calculating a Reference Value in Key Comparison Experiments, Technometrics, 2007

Uncertainty Due to Finite Resolution Measurements, with S. D. Phillips, and W. T. Estler, J. Res. Natl. Inst. Stand. Technol., 2008

Contribution to a Conversation about the Supplement 1 to the GUM, with A. Possolo, and W. T. Estler, Metrologia, 2009

Calibration and Uncertainty Analysis of Predictions from Computational Models NCSLI Measure, 2009

52 Calibration and Uncertainty Analysis of Predictions from Computational Models

AUTHOR Blaza Toman
COLLABORATORS Kevin McGrattan and Anthony Hamins (Fire Research Division, BFRL, NIST)

Introduction

Computer experiments are simulations of physical experiments performed by exercising a mathematical model for a physical or chemical process, to produce model outputs corresponding to sets of values of inputs to the model. They are especially useful when the corresponding physical experiments are difficult or expensive. Quantification of uncertainty for the outputs of such computer experiments is of great interest. The sources of uncertainty are in part due to the experimental measurement uncertainty in the inputs, and in part due to inadequacies of the underlying mathematical model. This paper presents methods for calibration and assessment of both types of uncertainty and demonstrates their use on two simple fire models.

In particular, consider an outcome of a computational model $\widetilde{y}$ based on a vector x of p input quantities. There is also a set of n physical measurements, $y_1, ..., y_n$, with uncertainties $u(y_i)$, and an associated set of $x_1 ..., x_n$. Assume that $Y_i | m_i, u(y_i) \sim \mathrm{N}(m_i, u^2(y_i))$, where m_i denotes the value of the physical quantity for the ith input. Further let $M_i = \widetilde{y}_i + D(x_i)$, where $D(x_i)$ represents the bias of the prediction at the x_i input. The M_i is a random function of x_i, a set of M_is is a single outcome from a multivariate Gaussian distribution. Specifically, for any $x_1, ..., x_k$, the $D(x_1), ..., D(x_k)$ has a multivariate normal prior distribution such that: $\mathbb{E}[D(x_i)] = 0$, $\mathbb{V}[D(x_i)] = \sigma^2$, $\mathrm{Cov}[D(x_i), D(x_{i'})] = \sigma^2 r(x_i - x_{i'})$, and $r(x_i - x_{i'}) = \exp\{-\sum_{j=1}^{p} |w_j(x_{ij} - x_{i'j})|\}$. This prior model is one particular form of a Gaussian Random Function (GRF) model. To fully specify the prior, the parameters σ^2 and w_j need to be given prior distributions. In the example, these were non-informative in the case of σ^2, and somewhat informative in the case of the weights. Bayesian updating produces posterior distributions for the original set of M_i conditional on the observations.

The main objective is to produce a probability distribution for M_{new}, the physical quantity corresponding to a set of inputs x_{new}, for which there is *no matching physical measurement*. This is accomplished by specifying that $M_{\mathrm{new}} = \widetilde{y}_{\mathrm{new}} + D(x_{\mathrm{new}})$, where $\widetilde{y}_{\mathrm{new}}$ is the prediction, and $D(x_{\mathrm{new}})$ follows the above GRF. Markov Chain Monte Carlo is used to produce samples from the posterior distribution.

Example: Fire modeling codes — MQH and Beyler methods

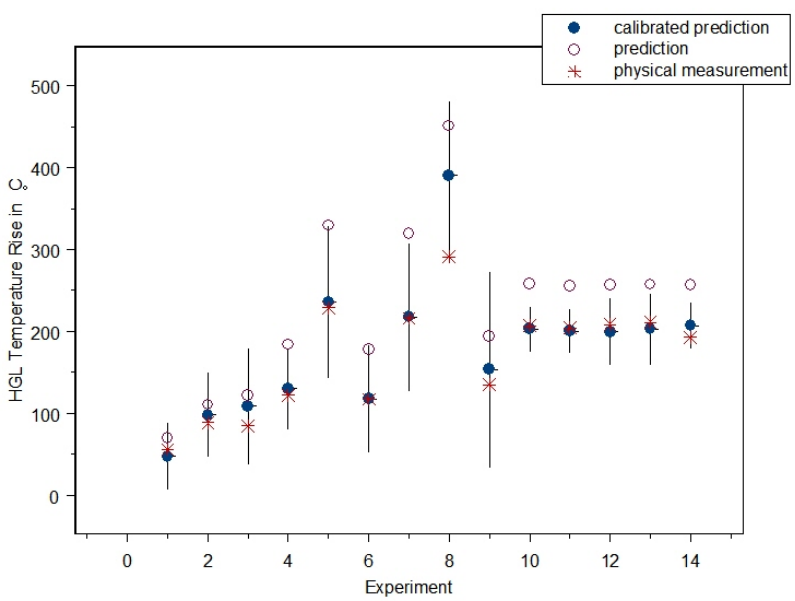
Predicting the outcome of a fire in an enclosure is often needed for the evaluation of safety of a planned building. In particular, the temperature estimate is necessary to evaluate danger to various features of the structure such as electrical cables. There are two widely accepted simple models for the calculation of the hot gas layer (HGL) temperature. The most frequently used model for the change in the HGL in a compartment with a single opening is due to McCaffrey,

Quintiere, and Harkleroad, the MQH method. For the case when the compartment is closed, the model due to Beyler is used.

In this work, these two computational models were applied to input values from fourteen fire experiments which produced the physical measurements of HGL temperature. The first, conducted inside the Technical Research Center of Finland Fire Test Hall, were intended for use in evaluating model predictions of fires in turbine halls of nuclear power plants. The hall has dimensions of 19 m high by 27 m long by 14 m wide. The second set of experiments was performed at NIST, where the compartment with one door had dimensions of 21.7 m by 7.1 m by 3.8 m high, designed to represent a compartment which may contain power and control cables.

The GRF model was applied to the two computational codes. Some of the experiments were appropriate for the MQH method and some for the Beylor method. In order to *mimic prediction* of an HGL temperature rise for a *"new"* input vector, a leave-one-out strategy was employed. The results show that the calibration aligns the predictions more closely with the actual physical measurements. Further, the size of the uncertainty of the predictions clearly shows the effect of the make-up of the training sample. For "new" experiments similar to the training sample the uncertainty is quite low. For "virtual" experiments performed under conditions "farther away" from those tested, the uncertainty is much larger.

The figure summarizes the results. It shows the 95 % uncertainty intervals (the vertical lines), the calibrated predictions (filled circles, which are the posterior means of the M_{new}), the physical measurements (crosses), and the non-calibrated predictions (empty circles). In all cases but one (experiment 8), which is an outlier in that the Beyler method greatly overestimates, the physical measurement falls within the 95% uncertainty interval. The advantage of close neighbors is apparent in the results for the final five experiments.



53 Model based Uncertainty Analysis in Interlaboratory Studies

AUTHOR Blaza Toman
COLLABORATORS Antonio Possolo (Statistical Engineering Division, ITL, NIST)

Introduction

Data from interlaboratory comparisons produce information about a particular measurand. In key comparisons, the primary goal is to produce measures of equivalence of the participating laboratories, both unilaterally with respect to a reference value, and bilaterally with respect to each other. The relationship between the *measurements* and the *measurand* is best accomplished using a *statistical model*.

The laboratories provide measurements $x_1, \ldots, x_n$, and their standard uncertainties $u_1, \ldots, u_n$. Each lab's measurement summarizes replicated measurements, each of which may be a combination of indications and measured values, and often other information. The standard uncertainties are computed according to the *Guide to the Expression of Uncertainty in Measurement* (GUM) . The analysis produces a reference value (x_{ref}) (RV or KCRV) and unilateral, $x_i - x_{\text{ref}}$, and bilateral $x_i - x_j$ degrees of equivalence (DoE).

Plots may suggest that the measurements are inconsistent, that is, that some of the laboratories' measurements are too far away from the reference value with respect to the laboratory uncertainties. Methods based on a chi-square statistic have been proposed to determine whether the laboratories belong to a consistent set. If not, then a consistent subset is chosen and used to compute the KCRV. This practice has many serious shortcomings, being based largely on faulty assumptions.

We propose alternative models for the analysis of key comparison data without excluding any measurements. Exclusion should be done only for cause, without which even the most discrepant measurement cannot logically be ruled out as erroneous.

Laboratory Effects Model

Under the Laboratory Effects Model (LEM), measurements are modeled as outcomes of independent Gaussian random variables with means $\theta_i = \mu + \beta_i$, where μ is the measurand, and variances given by $u_1^2, \ldots, u_n^2$. Two alternative models which differ in the definition and interpretation of the β_i exist.

In some cases, the number of degrees of freedom associated with the standard uncertainties is available. In these circumstances one must take into account the fact that the $\{u_i^2\}$ are only estimates of the true variances. One particular way to achieve this is to treat $\nu_i u_i^2/\sigma_i^2$ like an outcome of a chi-squared random variable with ν_i degrees of freedom.

The Fixed Effects Model assumes that the β_i are systematic laboratory biases which are expected to re-occur in repeated similar experiments. Without independent information about μ there are no unique estimates for the parameters without a constraint on the β_i. Under $\sum_{i=1}^n \beta_i = 0$, the measurand is estimated (via least squares or maximum likelihood) as $\bar{x}$, and the β_i are estimated by $x_i - \bar{x}$, the unilateral DoE with respect to the arithmetic average. The

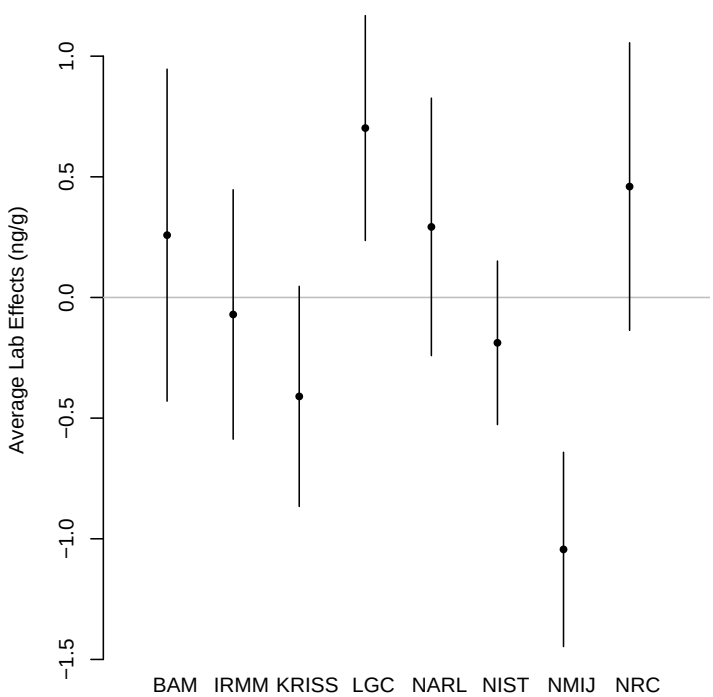
Fixed Effects Model attributes differences among measurements from various laboratories to differences among their means, the laboratory uncertainties are not increased.

The second type of LEM is the Random Effects Model. The β_i are random biases which are *not* expected to have the *same* value on repeated similar experiments, but instead are modeled as outcomes of a Gaussian random variable with zero mean and variance σ_β^2. Measurements from all laboratories have the same mean μ, but the variances are inflated to $u_i^2+\sigma_\beta^2$. Thus, apparent differences among laboratory measurements are explained by an increase in the laboratory uncertainties.

The Fixed Effects Model is used when the laboratory biases are likely to re-occur in similar sizes, across similar experiments, the Random Effects Model is used when the laboratory biases are due to some common underlying cause acting in a *varying nature*.

Experiments with multiple measurands afford a unique opportunity for model selection. The following model makes it possible to determine whether apparent differences between laboratories remain constant. The measurements of n laboratories of p measurands, $\{x_{ij} : i = 1,\ldots,n; j = 1,\ldots,p\}$, are outcomes of independent Gaussian random variables with means $\theta_{ij} = \mu_j + \beta_{ij}$, where μ_j are the measurands, and β_{ij} are the laboratory effects, and variances u_{ij}^2. The β_{ij} are Gaussian random variables with a mean λ_i which represents average laboratory bias. These are constrained as $\sum_{i=1}^n \lambda_i = 0$.

The figure shows the estimated λ_i and their confidence intervals for data from Key comparison CCQM-K25, on the determination of five different polychlorinated biphenyl (PCB) congeners in sediment. Two of the laboratories show significant biases that persist across experiments.



Average Laboratory Biases. Estimated $\{\widehat{\lambda}_i\}$ (dots), and corresponding, approximate 95% confidence intervals of the form $\{\widehat{\lambda}_i \pm 2u(\widehat{\lambda}_i)\}$. This suggests that, on average, and across all PCBs, NMIJ appears to have been biased low, and LGC biased high.

54 Dominic Vecchia

Biography

Dominic Vecchia received a B.S. degree in mathematics and a Ph.D. in mathematical statistics from Colorado State University, Ft. Collins, in 1972 and 1989, respectively. From 1972 to 1978 he was a graduate teaching and research assistant in the Department of Statistics, and a statistical consultant with the Natural Resources Ecology Laboratory.

Dominic joined the Statistical Engineering Division in 1978 as a mathematical statistician and was the Manager of the division's Boulder group from 1984 to 2002. His interests include research and applications of linear models, experiment design, stochastic modeling, permutation methods, robust statistics, statistics for calibration and measurement assurance, and functional data analysis.

Awards

James L. Madison Memorial Award, Colorado State University 1975.
Department of Commerce Bronze Medal, 1986.

Selected Publications

An empirical model for the warm-up drift of a commercial harmonic phase standard, *IEEE Trans. Instrum. Meas.*, 2007.

Higher order cumulants and Tchebyshev-Markov bounds for P-values in distribution-free matched pairs tests, *J. Stat. Plan. Infer.*, 2003.

Robust regression applied to optical-fiber dimensional quality control, *Technometrics*, 1997.

Optimum design of serial measurement trees, *J. Stat. Plan. Infer.*, 1995.

Exact moments of the quartic assignment statistic with an application to multiple regression, *Commun. Stat. — Theor. M.*, 1991.

Minimum cost inspection intervals for a two–state process, *J. Qual. Technol.*, 1990.

Calibration with randomly changing standard curves, *Technometrics*, 1989.

Problems with interval estimation when data are adjusted via calibration, *J. Qual. Technol.*, 1988.

Precision calibration of phasemeters, *IEEE Trans. Instrum. Meas.*, 1987.

Critical current measurements on a NbTi superconducting wire standard reference material, *Adv. Cryog. Eng. — Mater.*, 1984.

55 Pulse Shape Discrimination for Fast Neutron Spectroscopy

AUTHOR Dominic Vecchia
COLLABORATORS Kevin Coakley (Statistical Engineering Division, ITL, NIST), Jeffrey Nico and Brian Fisher (Ionizing Radiation Division, PL, NIST)

Introduction

The precise characterization of fast neutrons in terms of their energy and fluence yields quantitative information about neutron production that has critically important applications in national security, health physics, cosmology and basic science. Of foremost concern to the security community is the efficient detection of contraband fissile materials. An emitted neutron signal provides a crucial signature in detecting these materials. Current detectors are often expensive, bulky and prone to high rates of false positives.

We are developing a detector that uses a liquid scintillator loaded with ^{6}Li for the efficient measurement of fast neutrons and their energy. Such a detector has scientific applications, including quantification of neutron energy spectra in underground science facilities, as well as national security applications. Fast neutrons deposit energy in the scintillator by proton recoil scattering. If a neutron loses enough energy, it can be captured by lithium and produce a burst of scintillation photons.

In a prototype detector, we measured pulses in separate experiments with a ^{137}Cs gamma ray source and a ^{252}Cf neutron source. Using these data, we are developing methods to identify the sequence of proton recoil and delayed neutron capture events. Because background gamma rays in the ^{252}Cf data produce scintillation light, we must reject them with high probability. Discrimination is possible because scintillation light emission time probability density functions for electronic recoil events and nuclear recoil events are different.

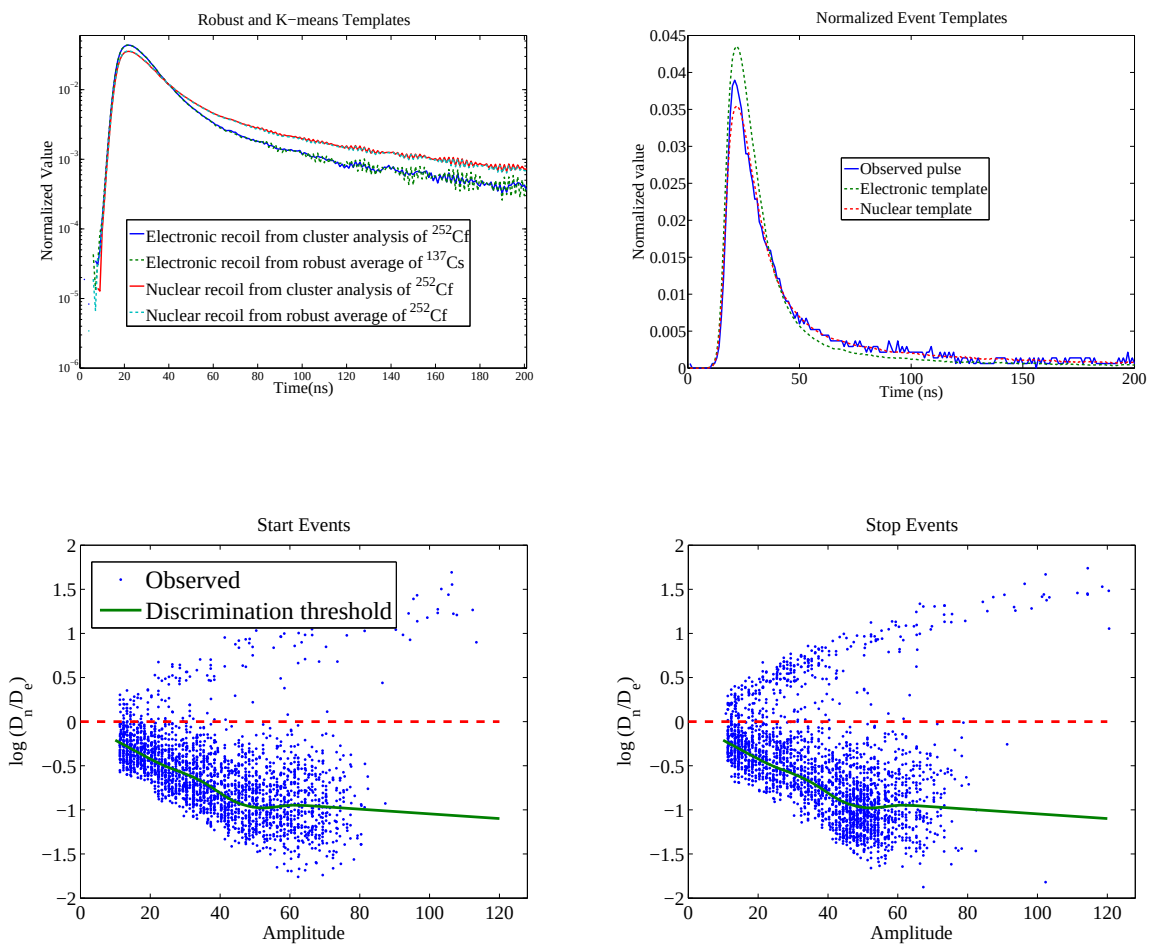
Pulse Shape Estimation

For event classification, we estimate the expected value of a background-corrected and normalized pulse (total energy summing to 1) for both nuclear and electronic recoil events. Pulses rapidly increase from a baseline to a maximum value and then slowly decay. We register baseline-corrected pulses based on the time at which each pulse first exceeds 30 % of its maximum value. Robust estimates of pulse templates for electronic and nuclear recoil events were computed as the normalized vector of 20 % time-indexed trimmed means of pulses from the appropriate calibration data. For comparison, we also estimated pulse templates by K-means cluster analysis where possible.

From the ^{137}Cs gamma ray data, we estimate an electronic recoil pulse by the normalized 20% trimmed mean of all registered pulses. A nuclear recoil pulse can be estimated similarly using data from the ^{252}Cf neutron source. Since the latter data is contaminated by gamma rays, we can estimate both nuclear and electronic templates with by a K-means cluster analysis. The upper left panel in Figure 1 shows that cluster analysis and robust signal averaging yield very similar templates. The upper right panel in the figure shows both estimated templates, and an observed pulse to be classified.

Event Classification

To classify unknown events, we compute the Matusita distances between each normalized pulse of interest, p_m, and both the electronic template $\hat{p}_e$ and nuclear template $\hat{p}_n$ as $D_e = \sum_i (\sqrt{p_m(i)} - \sqrt{\hat{p}_e(i)})^2$ and $D_n = \sum_i (\sqrt{p_m(i)} - \sqrt{\hat{p}_n(i)})^2$. We use $\log(D_n/D_e)$ to classify observed events. Examples of classified events from the ^{252}Cf neutron source are shown in the lower panels of Figure 1.



Pulse Templates & Event Classification. The upper left panel shows the close agreement of robust average and cluster analysis estimates of electronic and nuclear recoil. A pulse to be classified is shown with templates in the right panel. The distribution of $\log(D_n/D_e)$ for ^{252}Cf pulses in so-called Start and Stop time intervals are shown in the lower panels, along with discrimination thresholds corresponding to a nuclear recoil acceptance probability of about 0.5 for Start or Stop events.

56 Jack Wang

Biography

Jack Wang was born in Taipei, Taiwan. He received a Ph.D. degree in Statistics from Colorado State University, in 1978. Between 1979 and 1984 he worked for SPSS, Inc., in Chicago, where he developed, programmed, and documented statistical software for mainframe systems. From 1984 to 1988 he worked for the General Motors Research Laboratories in Warren, Michigan, where he collaborated with engineers and computer scientists on various automobile-related projects. He has been with NIST in Boulder since 1988.

Jack's research interests include statistical metrology and the application of statistical methods to physical sciences. Current research focuses on developing statistical methods for analyzing high-speed optoelectronic measurements.

Awards

James L. Madison Memorial Award, Colorado State University, 1977; Department of Commerce Bronze Medal, 1998; Fellow, American Statistical Association, 1998; Special Recognition Award, NCSL International, 2001; W. J. Youden Award, American Statistical Association, 2005.

Selected Publications

A robust algorithm for eye-diagram analysis, *Journal of Lightwave Technology*, 2009.

Propagation of uncertainties in measurements using generalized inference, *Metrologia*, 2005.

Models and confidence intervals for true values in interlaboratory trials, *JASA*, 2004.

Least-squares estimation of time-base distortion of sampling oscilloscopes, *IEEE Trans. Instrum. Meas.*, 1999.

Robust regression applied to optical fiber dimensional quality control, *Technometrics*, 1997.

Tolerance intervals for assessing individual bioequivalence, *Statistics in Medicine*, 1997.

Confidence limits for proportion of conformance, *Journal of Quality Technology*, 1996.

Tolerance intervals for the distribution of true values in the presence of measurement errors, *Technometrics*, 1994.

Optical fiber geometry by gray-scale analysis with robust regression, *Applied Optics*, 1992.

On the lower bound of confidence coefficients for a confidence interval on variance components, *Biometrics*, 1990.

57 Waveform Metrology

AUTHOR Chih-Ming (Jack) Wang
COLLABORATORS Paul Hale (Optoelectronics Division, EEEL, NIST), Dylan Williams (Electromagnetics Division, EEEL, NIST), Andrew Dienstfrey (Mathematical & Computational Sciences Division, ITL, NIST)

Introduction

NIST's waveform metrology project, aims to develop new techniques and measurement services at bandwidths currently unattainable for waveform verification, and to enable metrology for high-speed applications in Internet, wireless, remote sensing, and computing. It is a multidisciplinary collaboration between the Electronics and Electrical Engineering Laboratory and the Information Technology Laboratory. We are in the final year of this five-year project.

We have accomplished many of the goals that we laid out at the beginning of the project. Specifically, we have developed a measurement service for waveform calibration that is traceable to fundamental physics through the NIST eletro-optic sampling system. The calibration includes the whole measured waveform along with a covariance matrix that describes the covariance structure of the sampled points in the waveform epoch. In addition, the calibration supports test equipments that operate in both time and frequency domains.

The Statistical Engineering Division has made significant contributions toward these efforts. Much of the joint work, such as the estimation, correction, and uncertainty analysis of timebase errors, has been published or is documented in previous years' SED Reports of Activities. Here, we summarize some of the recent accomplishments.

Calculation of Pulse Parameters and Their Uncertainties

The fundamental starting point for the analysis of all two-state waveforms is the determination of the low and high state levels. Once the state levels are estimated, pulse parameters such as amplitude, transition duration, overshoot, and undershoot can be calculated.

The current industry standard recommends methods for determining state levels and determining pulse parameters, but gives no guidance for propagation of uncertainty, particularly in the presence of systematic and/or correlated sources of error. Correlations are important because certain pulse parameters, such as transition duration and pulse duration, are invariant with respect to multiplicative error, which is highly correlated.

We proposed a new procedure for determining the pulse states that involves clustering the data and then using a robust location estimator to determine the state levels. This method allows the propagation of uncertainty from the covariance of a sampled waveform representation all the way to the calculation of pulse parameters. This work will be published in an upcoming issue of the IEEE Transactions on Instrumentation and Measurement.

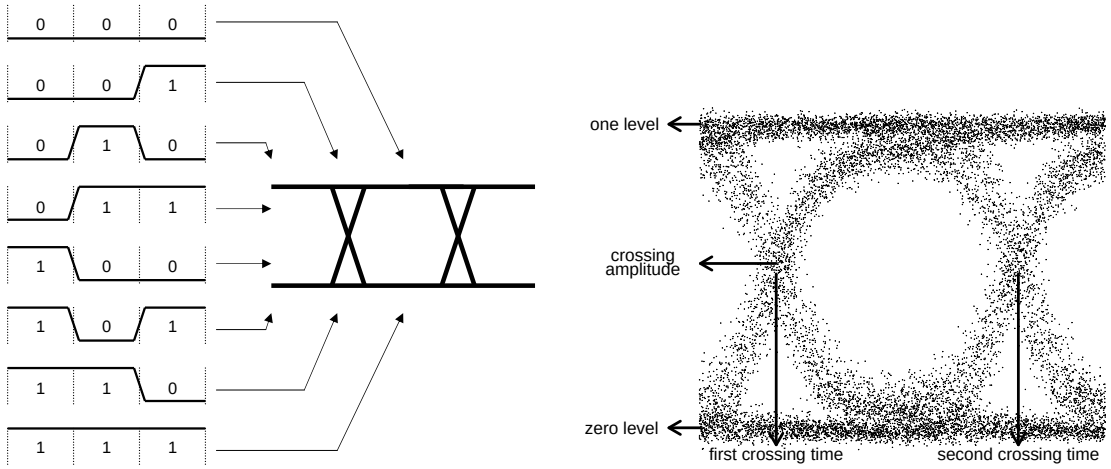
Analysis of Eye Diagrams

Eye-diagrams are multivalued displays used for assessing the quality of high-speed digital signals. They are usually constructed by applying a data waveform to the input of a sampling oscilloscope, and then overlapping all possible one-zero combinations on the instrument's display (see figure).

Eye diagram measurements have an enormous economic impact on the optical and electrical communications industries. With cost pressures driving manufacturers to create products that just meet specifications, the ability to make accurate and repeatable measurements for equipment testing is becoming more important. Conflicts may arise between component manufacturers and their customers when different test equipment leads to measurement inconsistencies. These discrepancies can be attributed to both software and hardware differences.

The focus of the current work is on an algorithmic method that can be implemented in software. The NIST algorithm makes use of a robust location estimator. In contrast to commonly used histogram techniques, this algorithm provides a repeatable solution that is insensitive to outliers and data distributions. The motivation for developing this algorithm was to create an independent, benchmark method that is both amenable to a thorough uncertainty analysis and can function as a comparison tool since no standardized industry algorithms currently exist.

Utilizing this technique, the researchers can calculate the fundamental parameters of an eye diagram, namely the one and zero levels, as well as the time and amplitude crossings (see figure). With these parameters determined, eye-mask alignment can be performed and various performance metrics can be derived, such as extinction ratio and root-mean-square jitter. The algorithm will be published in an upcoming issue of the IEEE Journal of Lightwave Technology.



The left panel shows how an eye diagram is constructed by overlapping all possible one-zero combinations on an instrument's display. The right panel shows the fundamental parameters of an eye diagram.

58 Fiducial Prediction Intervals

AUTHOR Chih-Ming (Jack) Wang
COLLABORATORS Hari Iyer (Department of Statistics, Colorado State University),
Jan Hannig (Department of Statistics and Operation Research,
The University of North Carolina at Chapel Hill)

Introduction

Prediction intervals, used in many practical applications, are statistical intervals that contain, with a specific probability, future realizations of a random variable. A method based on an extension of R. A. Fisher's fiducial argument can be used to construct prediction intervals.

Fiducial Prediction Quantities

Let $\boldsymbol{X}$ be a random vector with a distribution indexed by a parameter $\boldsymbol{\xi}$. Assume that $\boldsymbol{X}$ has a *structural representation* given by $\boldsymbol{X} = G(\boldsymbol{W},\ \boldsymbol{\xi})$, where G is a function and $\boldsymbol{W}$ is a random variable or vector with a completely known distribution free of $\boldsymbol{\xi}$. Let $H(\boldsymbol{x},\ \boldsymbol{w}) = \{\boldsymbol{\xi} : \boldsymbol{x} = G(\boldsymbol{w},\ \boldsymbol{\xi})\}$. The function H may be viewed as an inverse function of G. A fiducial distribution of $\boldsymbol{\xi}$ is then defined as a conditional distribution of $H(\boldsymbol{x},\ \boldsymbol{W}^*)$ given that $H(\boldsymbol{x},\ \boldsymbol{W}^*)$ is not empty. Here $\boldsymbol{x}$ is the observed value of $\boldsymbol{X}$ and $\boldsymbol{W}^*$ is an independent copy of $\boldsymbol{W}$. Denote a random variable having the same distribution as the fiducial distribution of $\boldsymbol{\xi}$ by $R_{\boldsymbol{\xi}}(\boldsymbol{x})$. We call this random variable a *fiducial quantity* for $\boldsymbol{\xi}$. A *fiducial prediction quantity* for the future random observation $\boldsymbol{Y}$ from the distribution of $\boldsymbol{X}$ is defined as

$$\widetilde{\boldsymbol{Y}} = G(\boldsymbol{W},\ R_{\boldsymbol{\xi}}(\boldsymbol{x})).$$

For problems where a close-form expression of $R_{\boldsymbol{\xi}}(\boldsymbol{x})$ does not exist, that is, $\widetilde{\boldsymbol{Y}}$ cannot be obtained using the above "plugging" method, simulation is employed to generate realizations of $\widetilde{\boldsymbol{Y}}$ based on realizations from the distribution of $R_{\boldsymbol{\xi}}(\boldsymbol{x})$. In both cases, the sampling distribution of $\widetilde{\boldsymbol{Y}}$ can be used to construct prediction intervals for observations from the distribution of $\boldsymbol{X}$. We use two examples to illustrate the procedure.

Normal Distribution

In this example, a one-sided fiducial prediction interval to contain at least p out of m future observations based on a random sample of sizes n from a normal distribution $N(\mu, \sigma^2)$ is derived. In this case, a fiducial quantity exists and is given by

$$R_{(\mu,\sigma^2)}(\overline{x}, s^2) = \left(\overline{x} - sZ^{(1)}/\sqrt{nV/(n-1)},\ (n-1)s^2/V\right),$$

where $\overline{x}$ and s are sample mean and standard deviation of the random sample; $Z^{(1)} \sim N(0,\ 1)$; and $V \sim \chi^2_{n-1}$. Let Y_i be the ith future observation. Then the data generating mechanism is given by

$$Y_i = \mu + \sigma Z_i^{(2)}, \quad i = 1, \ldots, m,$$

where $Z_i^{(2)}$ are iid $N(0, 1)$ random variables. Substituting (μ, σ^2) with $R_{(\mu,\sigma^2)}(\overline{x}, s^2)$ in the above equation, we obtain a fiducial prediction quantity for Y_i as

$$\widetilde{Y}_i = \overline{x} - sZ^{(1)}/\sqrt{nV/(n-1)} + sZ_i^{(2)}/\sqrt{V/(n-1)}.$$

Let $\widetilde{Y}_{(1)} < \ldots < \widetilde{Y}_{(m)}$. Then the probability that at least p out of m future observations will exceed L is equivalent to the probability that $\widetilde{Y}_{(m-p+1)} > L$. Thus, a one-sided $1-\alpha$ fiducial lower prediction limit is the α quantile of the distribution of $\widetilde{Y}_{(m-p+1)}$. Since the ordering of $(\widetilde{Y}_i - \overline{x})/s$ is identical to the ordering of $\widetilde{Y}_i$, this is equivalent to finding the α quantile of the distribution of the $(m-p+1)$-th order statistic of $Z_i^{(2)}/\sqrt{V/(n-1)} - Z^{(1)}/\sqrt{nV/(n-1)}$, which can be easily obtained by simulation. The quantiles so obtained are identical to the tabulated values in Fertig & Mann (*Technometrics*, 1977). Similarly, the two-sided symmetric fiducial prediction intervals are constructed using the $1-\alpha$ quantile of the distribution of the appropriate order statistic of the absolute value of $(\widetilde{Y}_i - \overline{x})/s$. These quantiles are identical to the tabulated values in Odeh (*Technometrics*, 1990).

Gamma Distribution

Let X_i, $i = 1, \ldots, n$, be a random sample from the gamma distribution Gamma(α, λ) with shape parameter α and scale parameter λ. Let $F_\alpha(\cdot)$ be the cumulative distribution function of Gamma$(\alpha, 1)$. Since $\lambda X_i \sim$ Gamma$(\alpha, 1)$, we have $U_i = F_\alpha(\lambda X_i) \sim$ uniform$(0, 1)$. We also write $\lambda X_i = F_\alpha^{-1}(U_i) \stackrel{\text{def}}{=} B(U_i, \alpha)$. Thus, we have the following structural equations

$$X_i = \frac{1}{\lambda} B(U_i, \alpha)$$

for the model. A fiducial distribution of (α, λ) can be obtained from these equations. In this example, a close-form fiducial quantity for (α, λ) is not available. Prediction intervals are obtained based on realizations from the fiducial distribution of (α, λ). For example, a lower prediction limit for at least p of m future observations from Gamma(α, λ) can be constructed as follows:

1. obtain a realization $(\widetilde{\alpha}, \widetilde{\lambda})$ from the fiducial distribution of (α, λ),
2. generate m independent uniform $(0, 1)$ random deviates $u_1, \ldots, u_m$,
3. calculate $y_i = B(u_i, \widetilde{\alpha})/\widetilde{\lambda}$, $i = 1, \ldots, m$,
4. calculate the $(m-p+1)$-th order statistic $y_{(m-p+1)}$ of $y_1, \ldots, y_m$,
5. repeat steps 1–4 a large number of times, say, M,
6. calculate the 95 percentile of $y_{(m-p+1)}^{(i)}$, $i = 1, \ldots, M$, as the 95 % lower prediction limit.

59 Grace Yang

Biography

Grace Yang was born in China, and moved to Taiwan in 1949. She graduated from National Taiwan University, and earned a Ph.D. in Statistics from the University of California at Berkeley, with Lucien Le Cam as advisor.

Grace is full professor in the Department of Mathematics at the University of Maryland, College Park. As a Faculty Appointee in NIST's Statistical Engineering Division since 1980, she has collaborated with NIST scientists on problems ranging from neutron lifetime measurement to sampling. Her research includes construction of stochastic models for the physical sciences, survival analysis and asymptotic theory in statistics.

She has performed editorial service for the Annals of Statistics, Journal of Statistical Planning and Inference, Mathematical Biosciences, and Statistics and Probability Letters. She has served on the Council of the Institute of Mathematical Statistics and the Bernoulli Society, and was a President of the International Chinese Statistical Association. During 2005–2008 she was a Program Director of Statistics at the National Science Foundation. She is a Fellow of the the Institute of Mathematical Statistics and an elected member of the International Statistical Institute.

Selected Publications

He, S., Yang, G. L., Fang, K.-T., and Widmann, J. F. (2005). Estimation of Poisson intensity in the presence of dead time, *J. of the American Stat. Assoc.*, 100(470): 669–679.

Le Cam, L. and Yang, G. L. (2000): *Asymptotics in Statistics*, 2nd ed., Springer.

Yang, G. L. and Coakley, K. (2000). Likelihood models for two stage neutron lifetime experiments, *Physical Review C*, 63, 014602 (16 pages).

Yang, G. L. and Chang, M. (1990) A stochastic model for the prevalence of Hepatitis A antibody, *Mathematical Biosciences* 98: 157–169.

Souders, T. M., Flach, D. R., Hagwood, C. and Yang, G. L. (1990) The effects of timing jitter in sampling systems, *IEEE, Trans. on Instrument. and Measurement* 39(1): 80–85.

Le Cam, L. and Yang, G. L. (1988) On the preservation of local asymptotic normality under information loss, *Annals of Statistics* 16(2): 483–520.

Chang, M. and Yang, G. L. (1987) Strong consistency of a nonparametric estimator of the survival function with doubly censored data, *Annals of Statistics*, 15(4): 1536–1547.

60 James Yen

Biography

James Yen has been a member of the Statistical Engineering Division since 1997. He works in statistical data analysis, and statistical computing in a wide variety of areas, including information technology and homeland security applications. He has long-standing collaborations with members of the Precision Engineering Division in ballistic indentification of firearms. In addition, he has worked with the Organic Chemical Metrology group in the certification of over two dozen food and supplement Standard Reference Materials.

James Yen earned a Ph.D in 1997 from the Stanford University Statistics Department; he wrote a dissertation titled *Robust estimation of effect sizes in meta-analysis*, under the direction of Professor Ingram Olkin.

Selected Publications

T. V. Vorburger, J. H. Yen, B. Bachrach, T. B. Renegar, J. J. Filliben, L. Ma, H. G. Rhee, A. Zheng, J. Song, M. A. Riley, C. D. Foreman, S. Ballou (2007), *Surface Topography Analysis for a Feasibility Assessment of a National Ballistics Imaging Database, NISTIR 7362* (submitted to *NIST Journal of Research*)

P. Volkovitsky, J. H. Yen, L. Cumberland (2007) Uncertainty in relative energy resolution measurements *Nuclear Instruments and Methods A*, 580 (3), 1497–1501.

R. D. Holbrook, J. H. Yen, T. J. Grizzard (2006), Characterizing Natural Organic Material from the Occoquan Watershed (Northern Virgina, US) using Fluorescence Spectroscopy and PARAFAC, *Science of the Total Environment*, 361, 249–266.

K. E. Sharpless, J. M. Betz, J. Brown Thomas, D. L. Duewer, K. Putzbach, C. Rimmer, L. C. Sander, M. M. Schantz, S. A. Wise, T. Yarita, J. H. Yen (2007), Preparation and Characterization of Standard Reference Material 3276 Carrot Extract in Oil, *Analytical and Bioanalytical Chemistry*, 389, 207–217.

R. Snelick, M. Indovina, J. H. Yen, A. Mink (2003), Multimodal Biometrics: Issues in Design and Testing, *The Fifth International Conference on Multimodal Interfaces* (Vancouver, Canada), November 2003.

61 Estimation of Detection Limits in Explosive Trace Detectors

AUTHOR James Yen
COLLABORATORS Mike Verkouteren, Jessica Coleman, Jennifer Verkouteren (Surface & Microanalysis Science Division, CSTL, NIST), Stefan Leigh, Andrew Rukhin, Alan Heckert (Statistical Engineering Division, ITL)

Introduction

Today's enhanced security environment has stimulated the development and widespread deployment of explosive trace detectors (ETD). The most likely siting for ETDs is in airports, but first responders and military personnel are among other users. Multiple technologies can be employed in ETDs, but ion mobility spectrometry (IMS) is prominent among them. They often employ swipe media, which are used to collect and hold residues from surfaces; an example is shown in the figure. Modern ETDs can detect extremely small quantities, often below the nanogram level, of explosives such as Pentaerythrite Tetranitrate (PETN) and Cyclotrimethylenetrinitramine (RDX). IMS-based devices can also be directed to detect drugs-of-abuse, including heroin, cocaine, and amphetamines, toxic industrial chemicals (TIC) such as hydrogen cyanide and phosgene, and chemical weapons such as sarin and sulfur mustard.

Limit of Detection

Currently, there is no consensus on the setting and testing of many critical performance standard for ETDs. Michael Verkouteren of NIST is preparing a proposal for standard limit of detection procedures to the ASTM E54.01 (Homeland Security Applications-CBRNE Sensors and Detectors) subcommittee, with sponsorship of the Office of Law Enforcement Standards (OLES, EEEL) and the Department of Homeland Security. NIST's efforts involving SED have centered on estimation methods for the limit of detection (LOD), in particular LOD95, which is *the lowest mass of an analyte deposited on a trap at which there is 95 % confidence that a single measurement in a particular ETD will have a true detection probability of at least 95 %, and a true non-detection probability of at least 95 % when measuring a process blank sample* (ASTM WK19817 — Proposed Standard Method for Determining Limits of Detection in Trace Contraband Detectors).

In practice, the ETD's response during field use is either a red light (presence detected) or green light (presence not detected). When there is a positive hit for a substance of interest, NIST researchers can extract the positive measurement value that induced the device's response. No such value is retrievable when the analyte is undetected. In this case, all one can ascertain is that the amount of analyte presence is below the threshold that prompts detection. In other words, the measurements are censored but the censoring value is unknown, and must be estimated as part of the process of determing the LOD. Matters are further complicated by the fact that the instruments undertake processing of the response, using proprietary algorithms, to dampen or remove background signals.

Experiments to determine LOD typically involve measurement of samples with controlled amounts of the analyte, with a sufficiently wide range of values that they straddle the LOD. One then

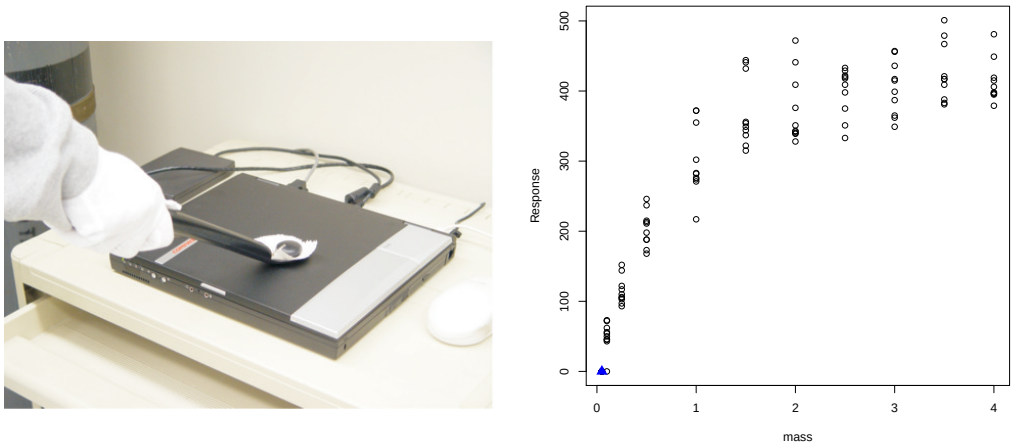
needs to decide how such experiments should be designed, including the choice of analyte concentration in the samples, and the choice of the number of replicates to include. A limiting factor to possible designs is that chemists find multiple samples of the same concentration much easier to produce and analyze than samples of different concentrations.

Modeling and Estimation

The figure shows a typical example of the response of a device for one substance of interest. The response levels off as concentration increases owing to sensor saturation, and it can vary considerably among samples with the same concentration of analyte. The observations with a non-detect are plotted at $y = 0$. The body of the data can be fit with a non-linear model, whose parameters can then be used to estimate the LOD95. Although such a global fit of the data is valuable to characterize the instrument, current LOD estimation efforts tend to focus on the smaller mass levels in the neighborhood of LOD95, thus insulating the LOD95 estimate from measurements at concentrations much large than the LOD95. Over such narrow ranges of concentrations, instrument response is more linear than over wider ranges, enabling simple approaches, for example, based on linear regression.

Andrew Rukhin has developed a model-based LOD estimation method involving the estimation of several parameters, including a threshold below which measurements do not register on the device. This method produces estimates of uncertainty and also confidence and tolerance bounds. An ASTM standard and associated website are currently in development for future use by ETD testers who need to determine LOD: the facilities provided will include the method just mentioned, a graphical technique, and a basic coverage factor method. As the standard and website evolve, the methods currently present may be modified or culled.

NIST's LOD contributions have been very well received by the interested ASTM community. In fact, the proposed LOD technology for ETDs may find itself being utilized for other applications as well, such as in environmental and nuclear monitoring.



The left panel shows a swipe media trap collecting surface residues for testing in an explosive trace detector. The right panel plots the response of one detector to varying amounts of a substance of interest. Actually there are ten non-detects at the lowest level (0.05) (blue triangle), and one non-detect at the next lowest level (0.1).

62 Motion Imagery Metrics

AUTHOR James Yen
COLLABORATORS Charles Fenimore, John Roberts, Hassan Sahibzada (Information Access Division, ITL, NIST), Darrell Young, Fred Petitti (National Geospatial-Intelligence Agency), Ivelisse Aviles, Dennis Leber (Statistical Engineering Division, ITL, NIST)

Introduction

Advanced technologies are making ever increasing amounts of image data, including motion imagery, available for intelligence analysis. Metrics are needed to quantify the interpretability of motion image data. The National Image Interpretability Rating Scales (NIIRS), developed in the 1970s, has long been used by the intelligence community to indicate the interpretability of still imagery. Recent efforts have tried to extend interpretability metrics to motion imagery.

The Statistical Engineering Division has had an ongoing collaboration on motion imagery research with members of ITL's Information Access Division and the National Geospatial-Intelligence Agency (NGA) and its contractors. Over the past several years, we have shown that motion quality metrics are feasible, and have examined how factors such as image resolution and camera motion affect the quality of motion imagery. In particular, we have ascertained that the interpretability of motion image data depends largely on image resolution. Current NIST work in this area focuses on two components: a format-frame rate study developed by ITL's John Roberts, Charles Fenimore, and Hassan Sahibzada, and a motion imagery criteria survey developed by NGA's Fred Petitti.

Format-frame rate study

Increased bandwidth has enabled increasing use of motion imagery, but questions remain on how best to utilize that bandwidth. NIST performed a comparative study of the relative interpretability of several motion imagery formats. The main comparison of interest is between two standard high definition (HD) formats, 1080p30 (1920×1080 progressive scan at 30 frames per second (Fps)) and 720p60 (1280×720 progressive scan at 60 Fps). In short, should the bandwith be used to carry more pixels or more frames per second?

The ideal comparison data would be motion imagery of the same scenes shot in both formats simultaneously. While such data are not available, direct comparison data can be generated from existing video content. NIST possesses many motion-rich 1080p30 clips of action shot at the Marine Corps Marathon and at Pax River Air Station. 1080p30 content, when played at double speed, produces nominal 1080p60 content. This 1080p60 content can be downsampled temporally to produce 1080p30 content and downsampled spatially to produce 720p60 content. The unavoidable side effect is that the action in this *fastworld* of derived clips, appears to unfold at twice the normal speed. This creates difficulties in interpretation for some of the actions in the derived clips; however, other actions that originally were too slow, may now be fast enough for interpretation tasks. The value of *fastworld* comparisons will be its enabling direct comparisons between the two formats.

There were other parts of the study that varied only the frame rate. Also, some matching 720p30 and 853×480p60 clips were downsampled from original 720p60 content to produce yet another comparison of the tradeoff between pixels and frame rate.

In the study, image analysts were shown pairs of motion imagery clips of the same scene but in different formats. They were asked to rank both the relative confidence and relative ease engendered by the two clips while performing an image intelligence task associated with that clip. Previous work has shown that an analyst's self-estimate of task confidence is a good estimate of the actual success of that task being performed.

Preliminary results show that there is an overall marginal preference (in both confidence and ease of use) for the 1080p30 format over the 720p60 format. Tabulated for each question are the values of various factors such as the resolution of the clip. The factor with the strongest influence analysts preferring one video format is the speed of the target associated with the task. In general, if the target (the object that is the focus of the interpretability task) is moving very (relatively) fast, then there is more of a preference for the higher framerate 720p60 format against the 1080p30 format. However, taking target speed into account, if the target object is relatively very small, then that engenders a greater preference for the 1080p30 format's greater number of pixels. However, these factors explain only a modest amount of the variability present in the study data.

Motion imagery criteria survey

Fred Petitti of NGA and Raytheon, has developed a Motion Imagery Interpretability Rating Standard (MIIRS), which is a "NIIRS-like" scale for motion imagery. It contains a ranked list of intelligence tasks, known as criteria, for each of an assortment of different orders of battle, such as in the Naval or Air Force domains. An example of a criteria might be to visually track the movement of a convoy of vehicles. Such a scaled list should assist analysts in assigning interpretability levels to motion imagery clips.

A survey was conducted among a representative group of image analysts via a web interface. In the survey analysts produced relative comparisons and rankings of various pairs of tasks. Preliminary results show a good agreement with the projected ordering of the criteria.

These images from the Pax River airfield and the Marine Corps Marathon are captured from video clips shown to subjects in the NIST motion imagery format study.

63 Nien Fan Zhang

Biography

Nien Fan Zhang is a Mathematical Statistician in the Statistical Engineering Division at NIST. He received his M.S. and Ph.D. in Statistics from Virginia Polytechnic Institute and State University. His technical areas of research are statistical process control, time series analysis, and statistical metrology and uncertainty analysis.

At NIST, he has collaborated with scientists from MEL, EEEL CSTL on varous projects, for example on phase-sensitive scatterfield microscopy, and from ITL on fingerprint usability, and on studies of performance of electronic voting systems.

Awards

2004, 2003 and 1999 Commerce Department (DOC) Silver Medal Awards for Scientific or Engineering Achievement.

2000 Federal Laboratory Consortium (FLC) Award for Excellence in Technology Transfer.

Selected Publications

Zhang, N. F. (2008) Allan variance of time series models for measurement data, *Metrologia*, 45, 549–561.

Winkel, P. and Zhang, N. F. (2008), *Statistical process control in medicine — the perils of risk adjustment*, Encyclopedia of statistics in Quality and Reliability, Wiley.

Winkel, P. and Zhang, N. F. (2007) *Statistical Development of Quality in Medicine*, Wiley.

Zhang, N. F. (2006), Calculation of the uncertainty of the mean of autocorrelated measurements, *Metrologia* 43, S276–S281

Zhang, N. F. (2006), The batched moving averages of measurement data and their applications in data treatment, *Measurement*, 39, 864–875.

Zhang, N. F. (2006), The uncertainty associated with the weighted mean of measurement data, *Metrologia*, 43, 195–204.

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

Zhang, N. F. (1998) A statistical control chart for stationary process data, *Technometrics*, 40(1), 24–38.

Zhang, N. F. (1998), Estimating Process Capability Indices for Autocorrelated Data, *Journal of Applied Statistics*, 25(4), 559–574.

64 Allan Variance of Time Series Models for Measurement Data

AUTHOR Nien Fan Zhang

Introduction

In time and frequency metrology, the power spectral density has been proposed to measure frequency stability in the frequency domain. It has been found that random fluctuations in standards can be modeled by a power law spectral density of the form

$$f(\omega) = \sum_{\alpha=-2}^{2} h_\alpha \omega^\alpha$$

for small $\omega > 0$, where $f(\omega)$ is the value of the spectral density at the Fourier frequency ω and the h_α's are intensity coefficients. Among several noise types that are commonly encountered in practice and are consistent with this equation, there is a process called $1/f$ noise or flicker frequency noise, which has the property that $f(\omega) \sim 1/\omega$ when $\omega \to 0$. The variance of such processes is infinite, but the Allan variance still is meaningful. In addition to the spectral density, the Allan variance has been widely used in time and frequency metrology as a substitue for the classical variance to characterize the stability of clocks or frequency standards in the time domain because the concept is meaninful even in the presence of drift, that is, for non-stationary processes. In this paper, we review the properties of the Allan variance for a wide range of time series.

Stationary processes

Consider time series observed at equally spaced (time) points, modeled as discrete weakly stationary process $\{X(t), t = 1, 2, \dots\}$. Define the process $\{Y_n(T), T = 1, 2, \dots\}$, $n \geqslant 2$ of arithmetic means (*moving averages*) of n consecutive $X(t)$'s ($n > 1$).

$$Y_n(T) = \frac{X((T-1)n+1) + \cdots + X(Tn)}{n}$$

When $\{X(t)\}$ is a stationary and uncorrelated process (or a sequence of independent, identically distributed random variables), $\mathbb{V}[Y_n(T)] = \sigma_X^2/n$. This fact has been used in metrology to reduce the standard deviation or uncertainty of the $Y_n(T)$ or $\bar{X}$, using a large sample size n. When $\{X(t)\}$ is autocorrelated but stationary, the variance of $Y_n(T)$ is used to calculate the uncertainty of the mean of autocorrelated measurements. However, for some non-stationary processes the variance of $Y_n(T)$ may not decrease at the rate of $1/n$, , or it may not even decrease, with increasing n. This means that, in this case, continued averaging of repeated measurements does not reduce the uncertainty and improve the quality of the average of the measurements, as it does when the measurements are statistically independent. This concern was the main reason to use the Allan variance.

Allan Variance

The two-sample variance or Allan variance of $\{X(t)\}$ for the average size of $n \geq 2$ is defined as

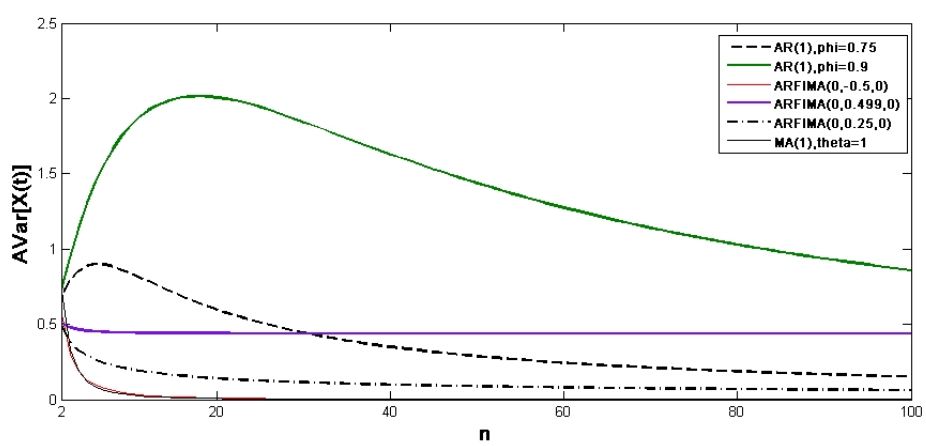
$$\mathrm{AVar}_n[X(t)] = \frac{E[Y_n(T) - Y_n(T-1)]^2}{2}$$

In geostatistics, the Allen variance is the semivariogram of the $\{Y_n(T)\}$ at lag 1.

We have shown that the variance of moving averages and the Allan variance of a stationary autoregressive movng average (ARMA) process and a stationary fractional differrence ARMA (ARFIMA) process are closely related. They decrease at the same rate when the size of the sample being averaged increases.

For a random walk process, which is nonstationary, or in general an ARMA$(0,1,1)$ process, the variance of its moving averages and the Allan variance will grow to infinity when the size of the averages increases. For a nonstationary fractional ARFIMA$(0,d,0)$ process as well as an ARFIMA$(1,d,0)$ or an ARFIMA$(0,d,1)$ process with $d \to 0.5$, which is a $1/f$ noise process, we have demonstrated that their Allan variances are stabilized at certain level when the size of the average is large enough while the variances of the moving averages will approach infinity. The property is expected to hold for a general ARFIMA$(p,0.5,q)$ process when the corresponding AR part is stationary and the MA part is invertible.

The following figure shows the behaviour of the Allan variance as a function of n, for AR(1) with $\phi_1 = 0.75$ and 0.9, ARFIMA$(0,0.499,0)$, ARFIMA$(0,0.25,0)$, ARFIMA$(0,-0.5,0)$, and MA(1) with $\theta_1 = 1$.



We conclude that the Allan variance is a measure of uncertainty similar to the variance of moving averages for measurements from stationary processes. For measurements from a nonstationary ARFIMA$(p,0.5,q)$ process the Allan variance is stabilized when the size of the average increases.

65 Obituary

John Mandel, 1914–2007

On the very day, May 10th, 1940, when the shadow of the evil scythe then sweeping across Europe first darkened the skies and souls of Belgium, Holland, and Luxembourg, John Mandel, then twenty five years old, his wife and best friend Ernestine, his parents and their extended family, fourteen people in all, chose to leave Antwerp, where he had been born on July 12th, 1914, and start an adventurous trek that would bring him to America, where his illustrious career, and his purposeful life, would end on October 1st, 2007, in Silver Spring, Maryland.

John Mandel had studied chemistry at the *Université Libre de Bruxelles* starting in 1933, earning a Master's degree in 1937, his inclination towards mathematics notwithstanding: what future would there have been for a Jewish mathematician in the Europe of the 1930s, friends and family had asked? During 1938–1940 he was research chemist at the *Societé Belge des Recherches et d'Études*, in Ghent.

After an anxious wait of many months in Marseille, the Mandels at long last obtained Venezuelan visas that would allow them to cross the Atlantic, which they sought to do departing from Lisboa, Portugal, owing to the shortage of ships then leaving southern France. Getting to their intended point of embarkation, however, involved first crossing the border into Spain, which all of the Mandels were allowed to do, except for John on account of his age that made him eligible for military service. However, his small stature and unusually youthful appearance, reinforced by children's clothes and shaved legs, made his second attempt at crossing the same border successful, some time later, posing as the youngest son of a friendly family whose eldest son then was fifteen years old. He came to America on board the *Excambion* (American Export Lines).

The sponsorship of a relative residing in New York allowed the Mandels to leave Ellis Island as immigrants into the United States, rather than bound for Venezuela. During 1941–1943 John Mandel worked for Foster D. Snell, Inc., as analytical and development chemist. His enduring fascination with mathematics motivated him to take courses in the night school of Brooklyn College, starting in 1943, then earning a two-year scholarship from Harold Hotelling to study statistics at Columbia University.

At Columbia during 1943 and 1944, John Mandel completed all the requirements for the Ph.D. in Mathematical Statistics but for the thesis. (His doctoral degree, from the University of Eindhoven, in the Netherlands, would come only much later, as a result of his sabbatical year of 1965.) During 1944–1947, he was a research chemist at the B. G. Corporation, which manufactured airplane spark plugs. For all the years thereafter, and until his retirement in 1990, John Mandel was a staff scientist at the National Bureau of Standards (NBS), in Washington, DC, which later became the National Institute of Standards and Technology (NIST).

Although never a member of the Statistical Engineering Division that Churchill Eisenhart founded at NBS in 1946, John Mandel remains the paradigmatic NIST statistician: deeply engaged in, and deriving inspiration from substantive problems in science and technology (including research supporting the production of plastics, leather, paper, and rubber tires).

Steadily he practiced most skilled data analysis and developed new ways to learn from data, and formulated innovative statistical models for the analysis of two-way tables and for the assessment of measurement uncertainty, especially in the context of interlaboratory studies. The American Statistical Association awarded him the W. J. Youden Award in Interlaboratory Testing twice: in 1988, together with Theodore W. Lashof, for "The Nature of Repeatability and Reproducibility," *Journal of Quality Technology*, vol. 19 (1987), pp. 29–36; and in 1996, for "Structure and Outliers in Interlaboratory Studies," *Journal of Testing and Evaluation*, vol. 25 (1995), pp. 364–369.

John Mandel's contributions to the execution of NIST's mission, by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life, include the development, jointly with J. Paul Cali (NBS), Larry Moore (NBS), and Donald S. Young (National Institutes of Health), of a method to determine the concentration of calcium in human blood serum with sufficiently low uncertainty to satisfy the requirements of clinical practice. For this and many other contributions of major significance, the U.S. Department of Commerce awarded him a Gold Medal in 1973.

Other public signs of recognition of his meritorious accomplishments as statistician include the Shewhart and Deming Medals, the Brumbaugh Award, and the Frank Wilcoxon Prize of the American Society for Quality Control (for best practical application paper of 1971, "A new analysis of variance model for non-additive data", *Technometrics* 13(1), 1–18), as well as election to the fellowship of several learned societies. In addition to many technical memoranda and reports, his publications include more than one hundred articles and three books: *The Statistical Analysis of Experimental Data* (1964), *Evaluation and Control of Measurements* (1991), and *The Analysis of Two-Way Layouts* (1994).

In a letter that he wrote to Cuthbert Daniel close to his retirement from NIST, John Mandel notes: "When I look back on the last 40 years, I find much to rejoice about, but also a lot of frustration. I sometimes wonder about statisticians. Do they live in closed clans, recognizing only members of the clan?" One of the examples he gives of this phenomenon relates to "Non-additivity in two-way analysis of variance" (*Journal of the American Statistical Association*, vol. 56 (1961), pp. 878–888), which he felt had been unduly neglected, especially when compared with alternative approaches to the same problem that he believed offered no practical advantage over his. In the same letter he also expresses the view that much too much gets published in the technical literature, yet fails to rise to the standard of practical usefulness that he adhered to steadfastly throughout his career.

John Mandel was always in great demand as a speaker and lecturer, even after his retirement from NIST. Standing behind the speaker's lectern, yet hardly rising above its height, he would often start by asking the audience not whether they could hear him, but whether they could *see* him: indeed, John, we can see your light shining brightly, now and for all the years to come.

— Antonio Possolo

NOTE: The author is much indebted to John Mandel's son and daughter, Paul Mandel and Judy Stapler, for sharing recollections and unpublished letters and notes of their father's; and should like to thank Joan Rosenblatt and Jim Filliben for their comments on an early draft.