

**32nd Symposium on the Interface:
Computing Science and Statistics**

**Modeling the Earth's Systems:
Physical to Infrastructural**

**April 5-8, 2000
Monteleone Hotel
New Orleans, Louisiana**

SPONSORED BY

The Interface Foundation of North America

HOSTED BY

Los Alamos National Laboratory
Neptune and Company, Inc.
RAND

CO-SPONSORS

U.S. Office of Naval Research
ASA Section on Statistical Computing
ASA Section on Statistical Graphics
SAS Institute, Inc.
MathSoft, Inc.
Bureau of Labor Statistics

COOPERATING ORGANIZATIONS

ASA CSNA ENAR IASC INFORMS IMS SIAM WNAR

INTERFACE 2000

SPONSOR

The Interface Foundation of North America is a nonprofit educational corporation founded in 1987 to sponsor the symposium and publish the proceedings. The IFNA is also a co-publisher of the *Journal of Computational and Graphical Statistics*.

BUSINESS OFFICE

Interface Foundation of North America
P. O. Box 7640
Fairfax Station, VA 22039-7460
(703) 993-4635
interface@galaxy.gmu.edu

PROCEEDINGS EDITORS

Yvonne M. Martinez
Los Alamos National Laboratory
yxm@lanl.gov

Edward J. Wegman
George Mason University
ewegman@gmu.edu

CONFERENCE CHAIRS

Sallie Keller-McNulty
Statistical Sciences Group
Los Alamos National Laboratory
P. O. Box 1663
Group TSA-1, Mail Stop F600
Los Alamos, NM 87545
(505) 665-3957
sallie@lanl.gov

Vicki Lancaster
Neptune and Company, Inc.
657 Magnolia Wood Avenue
Baton Rouge, Louisiana 70808-6053
(225) 766-7259
viancast@neptuneandco.com

Sally C. Morton
RAND
1700 Main Street
P. O. Box 2138
Santa Monica, CA 90407-2138
(310) 393-0411, x7360
Sally_Morton@rand.org

PROGRAM COMMITTEE

Paul Black
Neptune and Company, Inc.
pblack@neptuneandco.com

Andreas Buja
AT&T
andreas@research.att.com

Lorraine Denby
Bell Labs—Lucent Technologies
ld@research.bell-labs.com

Cathryn Dippo
Bureau of Labor Statistics
dippoc@ore.psb.bls.gov

Luis A. Escobar
Louisiana State University
karr@niss.org

Dorothy Merritts
Franklin and Marshall College
D_Merritts@acad.FandM.edu

Leslie M. Moore
Los Alamos National Laboratory
lmoore@lanl.gov

Barry Moser
Louisiana State University
bmoser@lsu.edu

Doug Nychka
National Center for Atmospheric Research
nychka@cgd.ucar.edu

Bonnie Ray
New Jersey Institute of Technology
borayx@m.njit.edu

William Shannon
Washington University School of Medicine
shannon@osler.wustl.edu

Nancy Spruill
Office of the Under Secretary of Defense
spruilln@acq.osd.mil

Edward J. Wegman
George Mason University
ewegman@gmu.edu

GENERAL INFORMATION

CONFERENCE FACILITIES

Interface 2000 is being held in the Monteleone Hotel conference facilities (see map inside back cover). The Monteleone Hotel is located in New Orleans beautiful French Quarter.

EVENING MIXER

Interface 2000 will be kicked-off with a mixer Wednesday evening at 8:00-10:00 pm in the River View Room atop the Monteleone Hotel overlooking the Mississippi River and the French Quarter. Please take this opportunity to renew old acquaintances and meet new friends. Refreshments will be provided.

EXHIBIT HALL

The exhibit hall will be located in the Queen Anne Room and will be open from 9:00 a.m.-5:00 p.m. Thursday and Friday. The following companies and organizations will exhibit:

- Books and Books
- Springer-Verlag
- American Statistical Association (ASA)
- ASA Section on Statistical Graphics
- ASA-SIAM Series on Statistics and Applied Probability
- Caucus for Women in Statistics
- Classification Society of North America
- Institute for Operations Research and the Management Sciences
- Institute of Mathematical Statistics
- Interface Foundation of North America
- International Association for Statistical Computing:
 - A Section of the International Statistical Institute
- Los Alamos National Laboratory
- National Institute of Statistical Sciences
- National Science Foundation
- Neptune and Company, Inc.
- Probability and Statistics Program, Office of Naval Research
- RAND

Please visit and support this year's exhibitors.

TECHNICAL SESSIONS

Room locations and times for all Invited and Contributed Technical Sessions are listed in the program.

INVITED POSTER SESSIONS

Several of the Invited Technical Sessions have companion Invited Posters. These posters will be displayed in the Queen Anne Room during the entire day of the corresponding Invited Technical Session. The poster presenters will be available for discussion at their posters during the morning and afternoon breaks. The poster abstracts are listed with their companion Invited Technical Sessions in the program.

BREAKS

Breaks on Thursday and Friday will be in the Queen Anne Room where exhibits and Invited Posters are displayed. The breaks are scheduled for 9:45-10:30 a.m. and 3:15-4:00 p.m. These have been intentionally scheduled longer than normal conference breaks to provide adequate opportunity to view the Invited Posters and foster discussion and collaborations.

BANQUET

The Banquet will be Thursday evening from 7:00–10:30 p.m. in the La Nouvelle Orleans Room. Since the banquet is included in the registration fee this is a great opportunity to sample the renowned cuisine of the Monteleone Hotel. The evening will include a banquet speaker, David J. Hand the Professor of Statistics at Imperial College, and an award presentation for the best paper in the first ten volumes of *Statistics and Computing*.

DINING

Apart from the banquet, meals will be on your own. New Orleans is known for its unique cuisine so this is an opportunity for you to try out some of the area restaurants. A restaurant guide put together by the conference chairs will be available at the registration table. The hotel concierge can help with lunch and dining reservations.

INTERNET ACCESS

There are telephone access lines in the hotel rooms to accommodate computers or fax machines. The Monteleone Business Center provides computers, fax machines, secretarial services and many other special services you may require.

MESSAGES

A message board will be on display beside the registration table.

SPEAKER PRACTICE ROOM

A speaker practice room with an overhead projector and screen will be available on Thursday and Friday from 7:00 a.m.–5:00 p.m. in the Bienville Room.

KEYNOTE ADDRESS

The Interface 2000 Keynote Address will be Thursday, 8:00 a.m. The Keynote speaker is Grace Wahba, John Bascom Professor of Statistics and Professor of Biostatistics at the University of Wisconsin, Madison.

PROCEEDINGS INFORMATION FOR PRESENTERS

Computing Science and Statistics, Volume 33, will be produced as a CD-ROM rather than a paper version. This planned production of the Interface 2000 proceedings will have several beneficial effects for Interface 2000 authors. First of all, the page restrictions formerly in place for the paper versions will be raised. Both invited and contributed authors are allowed to submit twenty pages of manuscript including text and graphics. This is an increase from 6 pages for invited speakers and 4 pages for contributed speakers. In addition, authors are encouraged to submit their PowerPoint presentations, animations (mpegs, avis), and software source code. It is anticipated that there will be enough space for all materials to be included.

The format for papers will remain as in previous years, a two-column format with printed size 8.5 by 11 inches. A LaTeX template is available on the Interface website (<http://www.galaxy.gmu.edu/stats/IFNA.html>). Papers will be produced in an Adobe pdf format on the CD. Papers can be submitted in a variety of formats including MS Word files, LaTeX files, TeX files, EXP files, postscript files, and Adobe pdf files. This first year, we also will accept paper versions for authors without the capability to provide electronic versions. Please note that graphics included in pdf files are compressed and may not be as legible as desired. This seems particularly true for pdf files produced from LaTeX and TeX files. Please verify that the graphics are legible before submitting pdf files. Authors are encouraged to use full-page graphics for this reason and are encouraged to submit graphics as separate files (gif files for line drawings, jpg files for continuous tone images, eps files for either).

As this will be Interface's first experience producing a professional CD volume, authors are encouraged to be flexible and to work with the publisher. It is planned to have both an Acrobat and an HTML interface for the CD. Papers must be submitted to Interface by July 15, 2000. Email submissions are preferred. Send electronic files to interface@galaxy.gmu.edu.

SPECIAL EXCURSION

Interface 2000 will offer an excursion into the Garden District of New Orleans on Friday evening beginning at 6:45 p.m. The cost of the excursion is \$20.00 and will include bus transportation to a two-hour walking tour of the Garden District and Lafayette No. 1 Cemetery. The Garden District is located uptown and is bounded by St. Charles and Magazine Streets. It is one of the most beautiful areas in the city. The Lafayette No. 1 Cemetery in the Garden District is bounded by Washington, Prytania, and Coliseum Streets. New Orleans is famous for her cemeteries, and this is a rare opportunity to view this miniature city of above-ground tombs at night. The tour will end at 9:00 p.m. at the Lafayette No. 1 Cemetery across the street from the famous Commander's Palace Restaurant. Bus transportation will be provided back to the French Quarter.

Participants can choose to stay in the Garden District and eat at the Commander's Palace (call 504-899-8221 for reservations). This is a great opportunity to eat at one of the best restaurants outside the French Quarter, especially one recognized for their outstanding desserts. For those wishing to make reservations, do so *now*; this is a very popular restaurant. **Participants who choose to eat at Commander's Palace are responsible for their own transportation back to the hotel.** The streetcar is located on St. Charles Street just two blocks from the final destination of the tour and can be taken back to the French Quarter for \$1.00. Taxi transportation back to the hotel can also be arranged.

Reservations for the Friday night excursion and can be made only during the conference at the registration table. Space is limited to 50 people so sign up early!

WEDNESDAY, APRIL 5, 2000**8:00 A.M.–NOON****SHORT COURSES**

REGISTRATION FOR INTERFACE 2000	7:00 A.M.–6:00 P.M. AND 7:30–9:00 P.M.	ROOM: QUEEN ANNE MEZZANINE
REGISTRATION FOR SHORT COURSES	7:00 A.M.–1:30 P.M.	ROOM: QUEEN ANNE MEZZANINE

SHORT COURSE I	8:00 A.M.–NOON	ROOM: LA NOUVELLE ORLEANS EAST
-----------------------	-----------------------	---------------------------------------

BUILDING AND FITTING RANDOM EFFECTS MODELS

William Cleveland, Lorraine Denby, Chuanhai Liu
 Statistics Research Department
 Bell Labs, Murray Hill, New Jersey

The use of random effects models in practice, often in the form of Bayesian hierarchical models, is growing rapidly because of major developments in computational methods for these models. In this short course we present models and building methods for data with random location and scale effects. Data visualization methods play a fundamental role in all phases of this model building: data exploration, model identification, and model checking. From the model building stage we move to Bayesian models for the data because, as a practical matter, the location and scale distributions fit readily into a hierarchical Bayesian framework. We describe computational methods for fitting these models.

LUNCH	12:00–1:30 P.M.
--------------	------------------------

SHORT COURSE II	1:30–5:30 P.M.	ROOM: LA NOUVELLE ORLEANS EAST
------------------------	-----------------------	---------------------------------------

**AN INTRODUCTION TO MODEL BUILDING WITH REPRODUCING KERNEL HILBERT
 WITH APPLICATIONS IN BIOSTATISTICS AND ATMOSPHERIC SCIENCES**

Grace Wahba
 Department of Statistics
 University of Wisconsin
 Madison, Wisconsin

We assume no knowledge of reproducing kernel Hilbert spaces, but review some basic concepts, with a view towards demonstrating how this setting allows the building of interesting statistical models, which allow the simultaneous analysis of heterogeneous, scattered observations, and other information. Methods appropriate for very large data sets will be discussed.

EVENING MIXER	8:00–10:00 P.M.	ROOM: RIVER VIEW
----------------------	------------------------	-------------------------

THURSDAY, APRIL 6, 2000

8:00–9:45 A.M.

KEYNOTE ADDRESS

REGISTRATION FOR INTERFACE 2000	7:00 A.M.–NOON AND 1:00–5:00 P.M.	ROOM: QUEEN ANNE MEZZANINE
SPEAKER PRACTICE	7:00 A.M.–5:00 P.M.	ROOM: BIENVILLE

KEYNOTE ADDRESS	8:00–9:45 A.M.	ROOM: LA NOUVELLE ORLEANS EAST AND WEST
-----------------	----------------	---

WELCOME TO INTERFACE	Vicki Lancaster, Neptune and Company, Inc.
WELCOME TO NEW ORLEANS	Sallie Keller-McNulty, Los Alamos National Laboratory
PREVIEW OF INTERFACE 2001	Arnold F. Goodman, University of California, Irvine
KEYNOTE INTRODUCTION	Sallie Keller-McNulty, Los Alamos National Laboratory

**COMBINING OBSERVATIONS WITH MODELS: PENALIZED LIKELIHOOD AND
RELATED METHODS IN NUMERICAL WEATHER PREDICTION**

Grace Wahba
University of Wisconsin

We will look at variational data assimilation as practiced by atmospheric scientists, with the eyes of a statistician. Recent operational numerical weather prediction models operate on what might be considered a very grand penalized likelihood point of view: A variational problem is set up and solved to obtain the evolving state of the atmosphere, given heterogeneous observations in time and space, a numerical model embodying the nonlinear equations of motion of the atmosphere, and various physical constraints and prior physical and historical information. The idea is to obtain a sequence of state vectors, which are “close” to the observations, close to a trajectory satisfying the equations of motion, and simultaneously respects the other information available. The state vector may be as big as 10^7 , and the observation vector 10^5 or 10^6 , leading to some interesting implementation questions. Interesting nonstandard statistical issues abound.

BIOGRAPHY



Grace Wahba is the John Bascom Professor of Statistics and Professor of Biostatistics at the University of Wisconsin, Madison. She is a Fellow of the Institute of Mathematical Statistics, the American Statistical Association, and the American Association for the Advancement of Science, and was recently elected to the American Academy of Arts and Sciences. She received the first Emanuel and Carol Parzen Prize for Statistical Innovation, the COPSS Elizabeth Scott Award, and the International Meetings on Statistical Climatology Achievement Award. Her research involves multivariate function estimation and model building with heterogeneous sources of information with applications in numerical weather prediction, climate, biostatistical model building and risk factor estimation, and supervised machine learning. She is most proud of her many and talented former students.

REFRESHMENT AND POSTER SESSION	9:45–10:30 A.M.	ROOM: QUEEN ANNE
--------------------------------	-----------------	------------------

THURSDAY, APRIL 6, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSIONS

CONTRIBUTED SESSION: Wavelets, Splines, State-Space and Adaptive Models**CHAIR:** Hyunjoong Kim, Worcester Polytechnic Institute**1:30–3:15 P.M.****ROOM: IBERVILLE**

- | | | | |
|-----------|--|-----------|---|
| 1:30 P.M. | <i>NORM Thresholding Method in Wavelet Regression</i>
Dongfeng Wu
University of Texas | 2:15 P.M. | <i>Partially Adaptive Bandwidth Used in Prediction and Local Regression</i>
Janis Grabis
Riga Technical University |
| 1:45 P.M. | <i>Data-Driven Optimal Denoising and Recovery of Derivatives Noisy Signals Using Multiwavelets</i>
Nathaniel Tymes and Sam Efromovich
University of New Mexico | 2:30 P.M. | <i>Self-Modeling Regression with Random Effects Using Penalized Regression Splines</i>
Naomi S. Altman and Julio C. Villarreal
Cornell University |
| 2:00 P.M. | <i>Adaptive Splines and Genetic Algorithms for Optimal Low-Dimensional Statistical Modeling</i>
Jennifer I. Pittman
Pennsylvania State University | 2:45 P.M. | <i>Estimation of Nonlinear State-Space Models in the Presence of Censored Observations</i>
Craig Johns
Colorado University and NCAR
Robert H. Shumway
University of California, Davis |

CONTRIBUTED SESSION: Applications of Exploring, Modeling and Presenting Large Datasets**CHAIR:** Derek Stanford, MathSoft, Inc.**1:30–3:15 P.M.****ROOM: LA NOUVELLE ORLEANS EAST**

- | | | | |
|-----------|---|-----------|---|
| 1:30 P.M. | <i>Predictive Statistical Models for Detecting Anomalies and Congestion in IP Based Networks</i>
Elisa M. Santos
Telcordia Technologies | 2:15 P.M. | <i>Data Mining on Time Series: An Illustration Using Fast Food Restaurant Franchise Data</i>
Lon-Mu Liu, S. Bhattacharyya,
S. L. Sclove, R. Chen, and W. J. Lattyak
University of Illinois, Chicago and
Scientific Computing Associates Corp. |
| 1:45 P.M. | <i>A Hierarchical Mixture Model for WWW-Usage</i>
Dee Denteneer
Philips Research | 2:30 P.M. | <i>Redesigning Tables and Graphics for Federal Statistical Agencies</i>
Daniel B. Carr
George Mason University |
| 2:00 P.M. | <i>Tracking Timing Patterns for Millions of Customers in Real-Time</i>
Jose C. Pinheiro, Diane Lambert,
and Don X. Sun
Bell Labs–Lucent Technologies | 2:45 P.M. | <i>Remote Medical Evaluation and Diagnostics: A Testbed for Hypertensive Patient Monitoring</i>
John C. Dumer, Timothy P. Hanratty,
and Barry A. Bodt
U.S. Army Research Laboratory
H. Mitchell Perry
and Sharon E. Carmody
Veterans Administration |

REFRESHMENT AND POSTER SESSION**3:15–4:00 P.M.****ROOM: QUEEN ANNE**

THURSDAY, APRIL 6, 2000

4:00–5:45 P.M.

INVITED SESSIONS

INVITED SESSION: CSNA Sponsored Session: Applications of Clustering and Classification to Large Datasets

ORGANIZER/CHAIR: William Shannon, Washington University School of Medicine

4:00–5:45 P.M.

ROOM: LA NOUVELLE ORLEANS EAST

4:00 P.M. *Inner-Loop Statistics in Automated Scientific Discovery from Massive Datasets*
Andrew Moore
Carnegie Mellon University and
Schenley Park Research, Inc.

4:45 P.M. *Current Approaches to Gene Chip Data Analysis*
Daniel Weaver
Genomica Corporation

5:30 P.M. **DISCUSSANT:** William Shannon

POSTERS*:

Preliminary Studies on Combining Wavelet and Cluster Analysis for Gene Chip Data
William Shannon
Washington University School of Medicine

The UC Irvine Knowledge Discovery in Databases Archive
Stephen D. Bay, Dennis Kibler, Michael J. Pazzani,
and Padhraic Smyth
University of California, Irvine

INVITED SESSION: Best of the *Journal of Computational and Graphical Statistics*: New Developments in EM

ORGANIZER: Andreas Buja, AT&T

CHAIR: Bonnie Ray, New Jersey Institute of Technology

4:00–5:45 P.M.

ROOM: IBERVILLE

4:00 P.M. *An Interval Analysis Approach to the EM Algorithm*
Kevin Wright
Pioneer Hi-Bred International
William J. Kennedy
Iowa State University

4:45 P.M. *Fitting Mixed-Effects Models Using Efficient EM-Type Algorithms*
David van Dyk
Harvard University

INVITED SESSION: Characterizing Large Complex Natural Systems and Beyond

ORGANIZER/CHAIR: Lorraine Denby, Bell Labs–Lucent Technologies

4:00–5:45 P.M.

ROOM: LA NOUVELLE ORLEANS WEST

4:00 P.M. *Statistics and Models for Complex Systems in Engineering and Biology*
John Doyle
Caltech University

5:00 P.M. **Discussants:** Edward J. Wegman,
George Mason University
and Chris Barrett, LANL

THURSDAY, APRIL 6, 2000

6:00–10:30 P.M.

RECEPTION AND BANQUET

RECEPTION

6:00–7:00 P.M.

ROOM: OUTSIDE LA NOUVELLE ORLEANS EAST & WEST

BANQUET**INTRODUCTION:** Sally C. Morton, RAND**AWARD PRESENTATION:** Best Paper in the First Ten Volumes of *Statistics and Computing***PRESENTED BY:** David J. Hand, Editor; and Scott Delman, Kluwer Academic Publishers**BANQUET SPEAKER:** David J. Hand, Imperial College, London, England

7:00–10:30 P.M.

ROOM: LA NOUVELLE ORLEANS EAST AND WEST

MEASURING THE EARTHDavid J. Hand
Imperial College

The computer has made statistics one of the most exciting professions. By necessity statisticians are in at the kill when the structures start to emerge from the morass of numbers. But structures worth knowing about are only discovered if good quality data are available. This is ensured by effective measurement procedures. Measurement technology goes in hand with substantive theory, and statistical theory and method go hand in hand with both.

BIOGRAPHY

David J. Hand is a Professor of Statistics in the Department of Mathematics at Imperial College in London. Previously, he was a Professor of Statistics and Head of the Department of Statistics at the Open University. He is the founding editor and continuing editor-in-chief of the journal *Statistics and Computing*, and is the former editor of *Journal of the Royal Statistical Society, Series C*. Professor Hand has published over 100 research papers and fifteen books, including "Construction and Assessment of Classification Rules," "Practical Longitudinal Data Analysis," and "Statistics in Finance." His research interests include data mining, classification methods, and the interface between statistics and computing. He has consulted broadly, including in the areas of medicine, psychology, and finance.

FRIDAY, APRIL 7, 2000

8:00–9:45 A.M.

INVITED SESSIONS

REGISTRATION FOR INTERFACE 2000	7:00 A.M.–NOON	ROOM: QUEEN ANNE MEZZANINE
SPEAKER PRACTICE	7:00 A.M.–5:00 P.M.	ROOM: BIENVILLE

INVITED SESSION: A Tutorial on Inverse Theory
ORGANIZER/CHAIR: Vicki Lancaster, Neptune and Company, Inc.
8:00–9:45 A.M.

ROOM: IBERVILLE

8:00 A.M. *A Tutorial on Statistical Inverse Theory*
 Luis Tenorio
 Colorado School of Mines

Generalizing Wiener-Levinson Deconvolution
 Luis Tenorio and Alberto Villarreal
 Colorado School of Mines

POSTERS*:

Velocity Estimation in Exploration Geophysics, a Bootstrap Approach
 Alberto Villarreal, Colorado School of Mines

Estimating the Influence of Random Noise on Measured Travel Times
 Alben Mateeva, Colorado School of Mines

INVITED SESSION: Defining, Measuring, and Analyzing Quality of Care: Statistical and Computational Challenges
ORGANIZER/CHAIR: Sally C. Morton, RAND

8:00–9:45 A.M.

ROOM: LA NOUVELLE ORLEANS EAST

8:00 A.M. *The HIV Performance Measurement Perspective from NYC, Framing the Clinical Context*
 Bruce Agins
 New York State Department of Health

9:10 A.M. *Simulation in Models of Health Care Quality*
 Karl Heiner, SUNY at New Paltz

8:20 A.M. *Measuring and Improving Quality in Managed Care: Some Statistical and Computing Issues*
 Randall K. Spoeri
 HIP Health Plans, New York

POSTERS*:

Casemix Adjustment of the National CAHPS Benchmarking Data 1.0
 Marc N. Elliott, RAND; Richard Swartz, Rice U;
 John Adams, RAND; Ron D. Hays, UCLA

8:45 A.M. *Building Aggregate Health Care Quality Scales*
 John Adams, RAND

Imputing Treatment Differences in Meta-Analyses with Missing Data
 I. Elaine Allen, Babson College
 Ingram Olkin, Stanford University

INVITED SESSION: The Use of Modeling and Statistics in Defense Analysis

ORGANIZER/CHAIR: Nancy Spruill, Office of the Under Secretary of Defense (Acquisition and Technology)

8:00–9:45 A.M.

ROOM: LA NOUVELLE ORLEANS WEST

8:00 A.M. *Using Advanced Modeling and Simulation Technologies and Techniques in the Analysis of Defense*
 Colonel William Crain
 Defense Modeling and Simulation Office

9:00 A.M. *Use of Decision Support Simulation Tool in Army Resource Decisions*
 Craig E. College, Program Analysis and Evaluation Office, Army

8:20 A.M. *The Joint Warfare System (JWARS): A Tool to Improve Combat Simulation for the Information Age Military*
 Lieutenant Colonel Daniel Maxwell
 JWARS, Office of the Secretary of Defense

9:20 A.M. **DISCUSSANT:** Nancy Spruill

POSTERS*:

Modeling Support Infrastructure for the Expeditionary Aerospace Force
 Lionel Galway, RAND

8:40 A.M. *Improving Inventory Performance by "Rightsizing" Inventory Reorder Points*
 Ron Fricker, RAND

Information and Knowledge Organization to Support Modeling and Statistics in Defense
 Yvonne M. Martinez, LANL

REFRESHMENT AND POSTER SESSION	9:45–10:30 A.M.	ROOM: QUEEN ANNE
--------------------------------	-----------------	------------------

*POSTER SESSIONS WILL BE UP ALL DAY IN THE QUEEN ANNE ROOM.

FRIDAY, APRIL 7, 2000

10:30 A.M.–12:15 P.M.

INVITED SESSIONS

INVITED SESSION: Enterprise Modeling: Supply Chain Design to Statistical Performance Analysis**ORGANIZERS:** Bonnie Ray, New Jersey Institute of Technology and Leslie M. Moore, LANL**CHAIR:** Bonnie Ray, New Jersey Institute of Technology**10:30 A.M.–12:15 P.M.****ROOM: IBERVILLE**

10:30 A.M. *Systems Thinking in Supply Chain Management*
Charu Chandra, University of Michigan

11:30 A.M. *Forecasting Methods for Supply Chain Management*
Bonnie Ray, NJ Institute of Technology

11:00 A.M. *Planning Experiments with Computer Models of Complex Phenomena*
Leslie M. Moore, LANL
Bonnie Ray, NJ Institute of Technology
Dennis R. Powell, LANL

NOON **DISCUSSANT:** Max Morris
Iowa State University

INVITED SESSION: Using Statistical Modeling to Identify Perturbations in Earth System Processes: Examples from Landscape Evolution and Tectonics**ORGANIZER/CHAIR:** Dorothy Merritts, Franklin and Marshall College**10:30 A.M.–12:15 P.M.****ROOM: LA NOUVELLE ORLEANS EAST**

10:30 A.M. *River Network Scaling Laws: Deviations and Fluctuations*
Peter Dodds and Daniel Rothman
MIT

11:40 A.M. *Deviations in Slope-Area Relations that Indicate Geologically Recent Crustal Deformation*
Tim C. Hesterberg, MathSoft, Inc.
Dorothy Merritts, Franklin and Marshall College

11:05 A.M. *Quantitative Testing of Landform Evolution Models*
Garry Willgoose and Greg Hancock
University of Newcastle, Australia

POSTER*:
Conjugate Gradient Methods for Large-Scale Sparse Regression with Applications to Seismic Deformation Estimation
Derek Stanford, MathSoft, Inc.

INVITED SESSION: Statistics in Precision Agriculture**ORGANIZER/CHAIR:** Barry Moser, Louisiana State University**10:30 A.M.–12:15 P.M.****ROOM: LA NOUVELLE WEST**

10:30 P.M. *Hypothesis Tests in the Presence of Spatial Correlation*
Robert G. Downer, LSU

11:30 P.M. *Use of Spatial Statistics to Design Sampling Plans for Monitoring Rust in Coffee Trees*
Raúl E. Macchiavelli
and Rocío del P. Rodríguez
University of Puerto Rico

11:00 P.M. *Statistical Issues in the Analysis of Remotely Sensed Data as Pertains to Precision Agriculture*
Patrick D. Gerard, David Evans,
and Michael Cox
Mississippi State University

NOON **DISCUSSANT:** Barry Moser

POSTERS*:
Effect of Number of Seed Bulked when Using the Multiple-Seed Procedure for Self-Pollinated Crops
James Beaver and Raúl E. Macchiavelli
University of Puerto Rico

Geostatistical Analysis of Spatial Nutrient Data in a Precision Agriculture Experiment
Bradley Tiffée and Robert G. Downer, LSU

LUNCH**12:15–1:30 P.M.**

*POSTER SESSIONS WILL BE UP ALL DAY IN THE QUEEN ANNE ROOM.

FRIDAY, APRIL 7, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSIONS

CONTRIBUTED SESSION: Exploration and Visualization in High Dimensions**SPONSORED BY THE CAUCUS FOR WOMEN IN STATISTICS****CHAIR:** Matthias Schonlau, RAND**1:30–3:15 P.M.****ROOM: LA NOUVELLE ORLEANS EAST**

- | | | | |
|-----------|--|-----------|---|
| 1:30 P.M. | <i>Exploration and Estimation of North American Climatological Data</i>
James A. Shine and Paul F. Krause
U.S. Army Topographic Engineering Center | 2:15 P.M. | <i>Graphical Techniques for the Exploration of Functional Data</i>
E. Neely Atkinson
MD Anderson Cancer Center |
| 1:45 P.M. | <i>Visualizing Abandoned Hazardous Waste Sites in the United States</i>
Carolyn K. Offutt
U.S. Environmental Protection Agency | 2:30 P.M. | <i>Authenticating Vulnerability Measurements</i>
Edward J. Wegman
George Mason University |
| 2:00 P.M. | <i>High-Dimensional Visualization Using Continuous Conditioning</i>
William C. Wojciechowski and David W. Scott
Rice University | 2:45 P.M. | <i>2D Classification Trees</i>
Hyunjoong Kim
Worcester Polytechnic Institute
Wei-Yin Loh
University of Wisconsin, Madison |

CONTRIBUTED SESSION: Multivariate Modeling: Issues and Applications**CHAIR:** Jerome Reiter, Williams College**1:30–3:15 P.M.****ROOM: LA NOUVELLE ORLEANS WEST**

- | | | | |
|-----------|--|-----------|--|
| 1:30 P.M. | <i>Statistical Analysis of Rhizosphere Microbial Communities</i>
Jayson D. Wilbur, Cindy H. Nakatsu, Sylvie M. Brouder, and R. W. Doerge
Purdue University | 2:15 P.M. | <i>CCA: Canonical Correlation or Correspondence Analysis? Which is Better for Analysis and Interpretation of Multivariate Data?</i>
A. Dale Magoun
University of Louisiana at Monroe
Linda Peyman
U.S.A.E./WES |
| 1:45 P.M. | <i>A Study of Faculty Equity Salary Using Derived Data</i>
Trong Wu
Southern Illinois University, Edwardsville | 2:30 P.M. | <i>Penalized Score Equations and Penalized GEE</i>
Wenjiang J. Fu
Michigan State University |
| 2:00 P.M. | <i>Use of Latent Variable Models in Air Quality Monitoring</i>
William F. Christensen and Stephan R. Sain
Southern Methodist University | 2:45 P.M. | <i>Assessing Deformation in Glaciers</i>
S. Huzurbazar
University of Wyoming |
| | | 3:00 P.M. | <i>Some Statistical Measures on the National, Distribution Center and Dealer Demands Along the Supply Chain</i>
Nick T. Thomopoulos
Illinois Institute of Technology
Wayne E. Bancroft
Motorola Corporation
Nick Z. Malham
Forecasting & Inventory Consultants, Inc. |

FRIDAY, APRIL 7, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSIONS

CONTRIBUTED SESSION: Innovations in Model Diagnostics and Fitting Algorithms**CHAIR:** David van Dyk, Harvard University**1:30–3:15 P.M.****ROOM: IBERVILLE**

- | | | | |
|-----------|--|-----------|---|
| 1:30 P.M. | <i>Multiple Outlier Detection</i>
David W. Scott
Rice University | 2:15 P.M. | <i>Using the Response Variable in Principal Components Regression</i>
Roy E. Welsch
MIT |
| 1:45 P.M. | <i>Parameter Selection for Constrained Solutions to Ill-Posed Problems</i>
Bert W. Rust
National Institute of Standards and Technology | 2:30 P.M. | <i>Case Studies of Normal Diagnostics in Regression Using Recovered Errors</i>
Donald E. Ramirez
University of Virginia
Donald R. Jensen
Virginia Polytechnic Institute |
| 2:00 P.M. | <i>Optimal Algorithms for Unimodal Regression</i>
Quentin F. Stout and Janis Hardwick
University of Michigan | 2:45 P.M. | <i>Tree-Based Models for Fitting Stratified Linear Regression Models</i>
William Shannon
Washington University School of Medicine, St. Louis
Maciej Faifer and Cezary Janikow
University of Missouri, St. Louis |

CONTRIBUTED SESSION: Time Series and Proportional Hazards**CHAIR:** Jennifer I. Pittman, Pennsylvania State University**1:30–3:15 P.M.****ROOM: CABILDO**

- | | | | |
|-----------|--|---|--|
| 1:30 P.M. | <i>Inference About the Change-Points in a Sequence of Random Vectors</i>
A. K. Gupta
Bowling Green State University
J. Chen
University of Missouri, Kansas City | Rice University
Xuemei Wang
MD Anderson Cancer Center | |
| 1:45 P.M. | <i>Detecting Change in Variance for Unequally Spaced Time Series</i>
Tze-San Lee and N. Hou
Western Illinois University | 2:30 P.M. | <i>The Introduction of Local Spread as a Measure of Non-Stationarity</i>
Robert A. Hedges and Bruce W. Suter
Air Force Research Laboratory |
| 2:00 P.M. | <i>Dynamic Modelling of Spectral Density Power Rhythms in All-Night Electroencephalograph (EEG) Recordings</i>
Matthew R. Marler, J. Christian Gillin, Arlene Schlosser, and Hanspeter Landolt
University of California, San Diego | 2:45 P.M. | <i>Multivariate Time Series Analysis in Principal Component Space</i>
Joseph N. Ladalla
University of Illinois, Springfield |
| 2:15 P.M. | <i>Importance Bootstrap Resampling for Proportional Hazards Regression</i>
Kim-Anh Do
MD Anderson Cancer Center
Bradley M. Broom | 3:00 P.M. | <i>Sequential Testing of Proportional Hazards Models</i>
Victor D. Zurkowski
University of Toronto |

FRIDAY, APRIL 7, 2000

4:00–5:45 P.M.

INVITED SESSIONS

INVITED SESSION: The Utility of Bayesian Decision Analysis for Environmental Problems**ORGANIZER/CHAIR:** Paul Black, Neptune and Co., Inc.

4:00–5:45 P.M.

ROOM: LA NOUVELLE ORLEANS EAST

4:00 P.M. *Scenario and Parametric Uncertainty in GESAMAC: A Methodological Study in Nuclear Waste Disposal Risk Assessment*
David Draper
University of Bath, England

4:50 P.M. *Bayesian Assessment of Uncertainty and Variability in Deterministic Environmental Exposure Models*
Samantha Bates and Adrian Raftery
University of Washington

4:25 P.M. *A Probability Network for Water Quality Modeling and Decision Support*
Kenneth H. Reckhow and
Mark E. Borsuk
Duke University

5:15 P.M. *Environmental Modeling and Bayesian Analysis for Assessing Human Health Impacts from Radioactive Contamination*
Tom Stockton and Paul Black
Neptune and Company, Inc.

INVITED SESSION: IASC Sponsored Session: Applications to Earth Systems**ORGANIZER/CHAIR:** Edward J. Wegman, George Mason University

4:00–5:45 P.M.

ROOM: LA NOUVELLE ORLEANS WEST

4:00 P.M. *Using Smoothing to Reconstruct the Holocene Temperature in Lapland*
Lasse Holmström
Rolf Nevanlinna Institute

4:35 P.M. *A Computational Geometry Approach for Peeling and Outlier Detection*
Giancarlo Ragozini
Universita di Napoli Federico II

5:10 P.M. *Applications of Deepest Regression*
Mia Hubert, Peter J. Rousseeuw,
and Stefan Van Aelst
Universitaire Instelling Antwerpen

EXCURSION

6:45–9:00 P.M.

SATURDAY, APRIL 8, 2000

8:00–9:45 A.M.

CONTRIBUTED SESSIONS

REGISTRATION FOR INTERFACE 2000

7:30–8:30 A.M.

ROOM: QUEEN ANNE MEZZANINE

CONTRIBUTED SESSION: Uncertainty Quantification in Complex Models**CHAIR:** Todd L. Graves, Los Alamos National Laboratory

8:00–9:45 A.M.

ROOM: LA NOUVELLE ORLEANS EAST

8:00 A.M. *Quantifying the Effects of Noise on Biogeochemical Models*
Barbara Bailey
University of Illinois, Urbana-Champaign
Scott Doney
NCAR

8:45 A.M. *Statistical Quantification of Prediction Error Associated with Computational Predictions*
Robert G. Easterling and Marcey Abate
Sandia National Laboratories

8:15 A.M. *The Selection of the Optimal Structure of the Earth's Model for Forecasting the Main Physical Parameters*
Alexander Dmitrievich Gorobets
Sevastopol State Technical University

9:00 A.M. *The Role of Statistical Methods in Atmospheric Model Intercomparison Projects*
Christiane Jablonowski
University of Michigan

8:30 A.M. *Cost-Effective Uncertainty Analysis*
Daniela Stoevska-Kojouharov
Monmouth University

9:15 A.M. *Confidence Bands for Nonparametric Curve Estimates*
David J. Cummins
Eli Lilly & Company
Doug Nychka
National Center for Atmospheric Research

CONTRIBUTED SESSION: Statistical Tests, Estimation and Stability**CHAIR:** Imola K. Fodor, Lawrence Livermore National Laboratory

8:00–9:45 A.M.

ROOM: LA NOUVELLE ORLEANS WEST

8:00 A.M. *An Adjusted, Asymmetric Two Sample t -Test*
Sandy D. Balkin
Ernst & Young LLP
Colin Mallows
AT&T Labs–Research

8:45 A.M. *A Test for Symmetry about a Known Median Based on a Runs Statistic*
Alex Leonardo Rojas Peña
University of Puerto Rico
Jimmy A. Corzo Salamanca
National University of Colombia

8:15 A.M. *The Wilson-Hilferty Transform is Locally Saddlepoint*
George R. Terrell
Virginia Polytechnic Institute

9:00 A.M. *Mining Evolutionary Data for Multidimensional Scaling of Gene Measurements*
Rida Moustafa and Edward J. Wegman
George Mason University

8:30 A.M. *On Numerical Stability of MGF and CF*
Jinhyo Kim
Seoul National University

9:15 A.M. *Efficient Nonparametric Estimation of a Distribution Function*
Reza Modarres
George Washington University

9:30 A.M. *Large One-Sample and Two-Sample Tests for Average Hazard Rates*
John J. Hsieh
University of Toronto

REFRESHMENT

9:45–10:30 A.M.

ROOM: OUTSIDE LA NOUVELLE ORLEANS EAST & WEST

SATURDAY, APRIL 8, 2000

10:30 A.M.–12:15 P.M.

INVITED SESSIONS

INVITED SESSION: Statistics and Information Technology**ORGANIZER/CHAIR:** Alan Karr, National Institute of Statistical Sciences**10:30 A.M.–12:15 P.M.****ROOM: ROOM: LA NOUVELLE ORLEANS EAST**10:30 A.M. *How Should We Publish Data Analyses in the Web Age?*

Todd L. Graves, LANL

11:05 A.M. *Geographic Aggregation Procedures for Data Disclosure Limitation*

Ashish Sanil

National Institute of Statistical Sciences

11:40 A.M. *Detecting Defection: Mining Massive Online Data to Model ISP Customer Churn*

Nandini Raghavan

AT&T Labs–Research

INVITED SESSION: Statistical and Computational Methods for Survival and Reliability Data**ORGANIZER/CHAIR:** Luis A. Escobar, Louisiana State University**10:30 A.M.–12:15 P.M.****ROOM: ROOM: LA NOUVELLE ORLEANS WEST**10:30 A.M. *A Case Study in Competing Risk Reliability Analysis Using JMP Software*

Bradley Jones

SAS Institute, Inc.

11:40 A.M. *Random Effects Survival Models for Familial Data*

Terry M. Therneau

Mayo Clinic

11:05 A.M. *Reliability Data Analysis Using S-Plus*

William Q. Meeker

Iowa State University

INTERFACE 2001

JUNE 13-16, 2001

ORANGE COUNTY, CALIFORNIA

Interface 2000 Abstracts

THURSDAY, APRIL 6, 2000 10:30 A.M.–12:15 P.M.

INVITED SESSION: Models for the Earth's Atmosphere and Ocean

ORGANIZER/CHAIR: Doug Nychka, National Center for Atmospheric Research

Air-Sea Interaction in the Labrador Sea: Deep Water Formation and Climate

Ralph Milliff

National Center for Atmospheric Research

The large-scale mean north-to-south overturning circulation in the Atlantic Ocean transports warm water poleward at the surface (e.g., the Gulf Stream and North Atlantic currents), and cold water equatorward at depth (e.g., the Deep Western Boundary Current system). This is an important branch of the so-called ocean conveyor-belt circulation that provides a conceptual model of the global ocean as a sink for atmospheric concentrations of greenhouse gases, e.g., carbon dioxide. The ocean deep convection process at high-latitudes is the energetic downward component of the conveyor-belt model. Ocean deep water is said to be formed in this process wherein oceanic parcels recently in contact with the atmosphere are sequestered at depth, and subsequently isolated from the surface over climate timescales, e.g., $O(1000 \text{ yrs})$. The Labrador Sea is one of a few locations in the world ocean where deep convection occurs. Ocean deep convection is driven there by vigorous air-sea exchanges of heat, momentum, moisture, and mechanical energy associated with the passages of intense storms in winter. An extensive observational program in oceanography and meteorology was implemented in the Labrador Sea region with a focus on the ocean deep convection process for the 1996-97 winter season. We will review the large observational dataset that is emerging from that campaign. In addition we will introduce the growing satellite observational datasets of sea-surface winds and sea-surface heights in the Labrador Sea region. This talk sets the stage for a description in the following presentation of Bayesian Hierarchical Model (BHM) approaches to geophysical problems, including an example of a BHM approach to ocean dynamics leading to ocean deep convection and deep water formation in the Labrador Sea.

Hierarchical, Space-Time Models: Physically Based Models for Combining Geophysical Data

L. Mark Berliner

Ohio State University

Phenomena studied in the geophysical sciences are high-dimensional, interrelated processes distributed in space and time. Modeling and prediction of such processes typically requires the combination of both scientific understanding, usually reflected in physical models, and observations. However, the physical models are often very complex and subject to a variety of uncertainties. Further, though very large to massive datasets are often available, they are typically composed of disparate types of observations, and almost paradoxically, cover a small portion of the processes of interest. The hierarchical Bayesian viewpoint is suggested to provide a framework for combining scientific reasoning and observational data, in a fashion that quantitatively accounts for our uncertainty.

I review a basic strategy for developing such hierarchical models indicate the relation between the Bayesian models and their analysis via Markov chain Monte Carlo and review some preliminary examples, motivated by Ralph Milliff's talk that precedes this one.

Computational Advances in Gibbs Sampling of Massive Spatio-Temporal Models

Timothy Hoar

National Center for Atmospheric Research

Christopher K. Wikle

University of Missouri, Columbia

The analysis of the prodigious amounts of data generated by orbiting platforms requires massive spatio-temporal models. Even for limited-domain cases, traditional covariance-based space-time statistical methods are generally not tractable. We explore programming methodologies for making hierarchical Bayesian simulations for such models (typically requiring a large number of iterations) a reality for very large datasets and spatio-temporal domains. The hierarchical model is fully described by Wikle et.al, JASA, in review. Our specific interest is the spatio-temporal prediction of the surface winds over the equatorial Pacific using remotely-sensed satellite wind observations and the output of a deterministic weather model.

Stochastic Parameterizations in General Circulation Models for the Atmosphere: Cloud Motion

Rachel Buchberger

National Center for Atmospheric Research and Colorado State University

Current models of the earth's atmosphere (General Circulation Models) rely on deterministic rules to decide the cloud amounts in a grid box. Clouds are an important feature of these models because they influence how incoming and reflected radiation interacts with the atmosphere. Unfortunately, such models do not capture the motion of the cloud field well. An alternative is to consider cloud amounts as a spatial and temporal process that evolves in a nonlinear and stochastic fashion over time. In this project, the form of this process is estimated from observational data and is found to be a good description of the measured cloud fields. The statistical models are Nongaussian and use a neural network form to represent the autoregressive relationship between current cloud amounts and those in the next time period.

THURSDAY, APRIL 6, 2000 10:30 A.M.–12:15 P.M.

INVITED SESSION: Critical Infrastructure Modeling

ORGANIZER/CHAIR: Sallie Keller-McNulty, Los Alamos National Laboratory

Simulation-Based Analysis of the Nation's Critical Infrastructure

J. Darrell Morgeson

Los Alamos National Laboratory

For the past 10 years, Los Alamos has pursued an aggressive simulation R&D program focused on the nation's critical infrastructure. The products of the program are described below together with their respective sponsors and funding. With internal discretionary funding, the Lab began to integrate these efforts into an overall "system-of-systems" simulation environment last year with a view toward addressing near, mid-, and long-term policy, investment, and operational issues for use by both public and private sectors. The scale and scope of these simulations extend the state-of-the-art in very large-scale complex systems simulations and computation (i.e., human based interactions on the order of 10^7 - 10^{12} interactions per second). The combined simulation and analysis system will effectively address such complex issues as: convergence, electrical grid restructuring energy security and reliability, interdependencies among critical infrastructure systems, environmental impacts, response to natural and manmade disasters, and others. The entire simulation system is designed to run on multiple hardware platforms including ASCI architectures. An S&T roadmap has been developed to show the evolution of this program over the next 10 years for both applied products and advances in the science that forms the foundation for these tools and their use. To date, over \$50M have been spent on these projects with another \$42M planned through 2004—the majority provided by other federal agencies.

Generation and Measurement of Large Dynamical Systems

Chris Barrett

Los Alamos National Laboratory

We introduce a new mathematical object, a Sequential Dynamical System (SDS), and explore the insights we obtain in relation to computer simulation of large, composed, dynamical systems, their measurement and interpretation, and a range of issues surrounding validity of a simulation as used in decision making and analysis of such systems. In large infrastructure design, analysis and policy questions, representation using computers of present and/or future systems is desirable in many respects. However, firm theoretical foundations for simulation and its appropriate use are lacking, which limits their credible use. In particular the following issues impose a need for rigorous foundations, 1. the short and long term dynamics of these systems are complicated and characterized by massive interaction among large numbers of subsystems, 2. a potentially endless data collection requirement warrants care in choice of approaches to represent the system, 3. the fact that knowledge of piecemeals usually exceeds understanding of the composed system, and 4. a broad range of issues surrounding any meaningful use of the term "validity" as it might apply to a computer simulation used as a model to which decision making processes refer. We will examine these issues from the perspective of our emerging theoretical foundations for simulations as well as applications in transport and communications infrastructure analysis.

THURSDAY, APRIL 6, 2000 10:30 A.M.–12:15 P.M.

INVITED SESSION: Information Technology and Federal Statistics

ORGANIZER/CHAIR: Cathryn Dippo, Bureau of Labor Statistics

Wrapping and Mediating Survey Data

Amarnath Gupta, Chaitanya Baru, Richard Marciano, and Ilya Zaslavsky
University of California, San Diego

The task of information mediation is to enable a user to query across a number of information sources as if they were a single integrated source. To accomplish this integration, the data from the sources are transformed to a common representation, through a process called wrapping. We show how such information integration can be done on survey data, when the data is presented in the DDI format and information integration is achieved in the MIX framework developed at UCSD. We point out how the role of statistical expert knowledge must be exploited to make the mediation meaningful for users. We also demonstrate how a survey analysis tool, called Sociology Workbench uses integrated survey information to perform its analysis.

Statistical Information Seeking and System Design

Carol Hert
Syracuse University

Statistical Websites have made it possible for a huge variety of users to access statistical data. Understanding how these users locate and use data, what expectations they have for use, and what they understand about data will enable website designers to provide information and tools that enable better access. This paper reports on a series of user investigations related to United States Federal Statistical Agency websites highlighting findings to date as well as providing a perspective on how to understand and support users via system design.

The Role of Ontologies in Statistical Information Seeking

Ed Hovy
USC Information Sciences Institute

Judith Klavans
Columbia University

The old saying “there’s nothing like more data” is only true if you can successfully access the data you need, not get lost in it. The proliferation of statistical databases in U.S. Government Agencies has not been accompanied by a single widely used access system. To try to facilitate terminology standardization and cross-database access and linkup, we are connecting databases to a large 100,000-node taxonomy of ‘concepts’ called SENSUS. SENSUS will be used as the basis of a multi-database query planning and access system. This paper describes SENSUS, methods of linking other terminology systems to it, and the work now being done with some statistical databases.

Statistical Information Seeking and System Design

Carol Hert
Syracuse University

This poster complements the presentation of the same title. It will graphically depict the connections among several streams of research all concerned with improving system design for statistical information seeking. The streams included are research about users (and non-users), users' interactions with statistical information seeking tools (including metadata and search engines), interface design, metric and tool development, and organizational impact research. The poster will indicate how these projects contribute to our understanding of user behavior and how to support it on web-based systems.

THURSDAY, APRIL 6, 2000

10:30 A.M.–12:15 P.M.

INVITED SESSIONS

The Role of Ontologies in Statistical Information Seeking

Ed Hovy
USC Information Sciences Institute

A poster accompanying the paper by Hovy and Klavans provides details about the SENSUS ontology and about the process of analysis and taxonomization required to integrate a database or collection of text with an ontology.

The Data Documentation Initiative: Current Status of an Attempt to Specify an XML DTD for Empirical Social Science Documentation

Peter Jofstis
University of Michigan

In early 1995, aided by a grant from the National Science Foundation, the Inter-university Consortium for Political and Social Research formed a committee, which now numbers approximately 20 stakeholders in the social science research and archiving process. The goal of the committee was to develop a specification for the documentation of empirical social science data collections. The project is known as the Data Documentation Initiative (DDI).

The group decided that the best format for the specification was an XML (Extensible Markup Language) based Document Type Definition (DTD). A 13-site beta-test of an initial DDI DTD was completed in mid-1999. Comments resulting from the beta-test were reviewed and many of the suggestions have been implemented. Version 1 of the DTD will be released in early 2000.

This poster session will present the goals of the DDI project. The structure of the Version 1 DTD will be presented and explained. The types of data collections that may be marked-up using the DTD will be discussed. Issues held for Version 2 will be presented. Finally, some thoughts on the role that other parts of the XML suite (XML Schema, XML Data, XLink, XPointer) will be presented.

THURSDAY, APRIL 6, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSION: Strategies for Investigating Geophysical and Other Complex Data**CHAIR:** Samantha Bates, University of Washington**Compressing Massive Geophysical Datasets Using Quantization**

Amy Braverman

Jet Propulsion Laboratory, California Institute of Technology

In this work we set forth a method for compressing massive geophysical datasets like those that will be obtained from NASA's Earth Observing System Terra satellite. We develop a statistical model for studying relationships between compressed and uncompressed data, and use it to evaluate compressors found by an iterative clustering method based on the ECVQ algorithm of Chou, Lookabaugh, and Gray (1989). The method arbitrates between error induced by compression and level of data reduction. Error explicitly includes a component that accounts for uncertainty due to multiple local minima of the ECVQ loss function. Dataset compressibility is identified as an important characteristic for setting parameters that determine the balance error and data reduction. We demonstrate this procedure using a well known dataset from our motivating field of application, Earth science.

Finding Bent-Double Radio Galaxies: A Case Study in Data Mining

I. K. Fodor, C. Baldwin, E. Cantó-Paz, C. Kamath, and N. Tang

Lawrence Livermore National Laboratory

Sapphire (<http://www.llnl.gov/casc/sapphire/>) is a project on large scale data mining and pattern recognition at the Center for Applied Scientific Computing at Lawrence Livermore National Laboratory. We are using state-of-the-art computational methods in order to help scientists extract useful information hidden in massive datasets. In one of our applications, we are collaborating with astronomers on the FIRST (Faint Images of the Radio Sky at Twenty-cm) survey to find radio galaxies with bent-double morphology.

We present a brief overview of Sapphire, then illustrate the challenges particular to finding bent-double galaxies. We derive features from the FIRST catalog and from raw images, then apply decision trees to classify the radio sources based on those features. Defining meaningful features poses a real difficulty, as bent-doubles vary considerably. The features we use must not only be scale, translation, and rotation invariant, but should also be robust to small changes in the data. We describe the features we use to discriminate bent-doubles from non-bent-doubles, and report on the sensitivity of our decision tree results to changes in the features.

Characterizing the Complexity of a High-Dimensional Classification Problem

Carey E. Priebe

Johns Hopkins University

David J. Marchette

Naval Surface Warfare Center

Classification of high-dimensional data is inherently difficult. We present an exploratory data analysis methodology for obtaining information about the high-dimensional decision boundary, and provide a nonlinear projection in which to perform classification. We focus on the two-class problem, although the methodology can be extended to the multiclass case. The idea is to characterize the support of one class as a collection of spheres covering the support, with each sphere centered at an observation in that class such that the radius is maximal without containing too many observations from the other class. A greedy algorithm for fitting the spheres is proposed. The spheres then provide a description of the support of the class, with information about the decision boundary implicit in the position, radii and adjacency of the spheres. Clustering the spheres by radius and projecting the data based on distances to the clusters yields a nonlinear projection to a lower-dimensional space in which classification can be performed. We illustrate the algorithm with pedagogical simulations and a chemical sensor data analysis application.

Multiresolution Stochastic Models for Object Recognition in Self-Similar Texture Images

Richard J. Barton, Jennifer Davidson, Lili Chen, and Fei Wan
Iowa State University

We consider the application of multiresolution stochastic modeling techniques to the analysis and synthesis of texture images. We adopt the approach of Crouse, et al., in which the wavelet coefficients of a texture image are modeled using a hidden Markov tree model (HMTM). The assumed tree structure arises naturally from a decomposition of the original image in terms of an orthonormal wavelet basis, and the Markov structure is imposed under the assumption that the wavelet coefficients decorrelate rapidly across both space and scale. One of the most common characteristics of image data that results in wavelet coefficients with this property is self-similarity. The existence of self-similarity in the image data allows us to reduce dramatically the number of parameters that must be estimated in order to accurately model the wavelet decomposition of the image using an HMTM. In addition, we extend the HMTM structure by modeling the relationship of the wavelet coefficients within a particular scale using a partially ordered Markov model (POMM). We show that POMMs fit naturally within the multiresolution HMTM paradigm, and that the multiresolution POMM structure leads to accurate and parsimonious models for texture data that are useful for texture segmentation, texture discrimination, and object recognition.

Predicting the Phase Transition in 3-Colorable Graphs

Hao Zhang
Washington State University

The phase transition in 3-colorable refers to the phenomenon that a graph abruptly becomes not 3-colorable when the connectivity of a large graph increases. Percentages of three colorable graphs were obtained by generating random graphs for each pair of different sizes and connectivity. We use these data to build a model and predict what the phase transition will occur. We will also address the model validation and distributions of estimators in the model.

On the Decomposition of Spatial Processes

Reinhard Furrer
Swiss Federal Institute of Technology

Let $\{Z(x):x \in D\}$ be a stochastic process in a domain $D \subset \mathbb{R}^d$, $d \geq 1$. To apply statistical procedures, it is often necessary to decompose the process into several parts. The most commonly used such decomposition is based on the separation according to different scales, a large-scale variation, a smooth small-scale variation, a microscale variation and a measurement error. Although this additive partitioning is of considerable utility it also has several drawbacks.

In this talk, we present an analysis based on state-space representation. Let $Z(x) = W(x) + \varepsilon(x)$, (observation equation) and $W(x) = \int_D k(x,s)W(s)ds + Y(x)$, $x \in D$, (state equation), where $k(x,s)$ is a sufficiently regular function, $Y(x)$ is a second-order stationary spatial process and $\varepsilon(x)$ is a zero-mean white-noise. The state at the point x is then a weighted mean of its neighborhood states plus a spatial process.

Other existing decompositions can be reconstructed by the new representation. The new model takes account of diverse shapes of trends and one does not have to decide whether the process is stationary or not. We will discuss estimates of the parameters of the covariogram of $Y(x)$ based on nonlinear optimization. And we discuss the efficiency of the proposed method and compare the results with those of other common models.

THURSDAY, APRIL 6, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSION: Computational and Estimation Issues in Modeling**CHAIR:** Barbara Bailey, University of Illinois at Urbana-Champaign**Bayesian Computations for Random Environment Models**Dhaifalla K. Al-Mutairi
Kuwait University

This paper deals with reliability data analysis from Bayesian perspective using Random Environment (RE) models. We review current literature on RE models and study statistical computational problems for these models that will arise in posterior and predictive analysis, test of hypothesis, and model selections. Computational methods to solve such problems are presented and we also give illustrative examples.

Borrowing Strength without Explicit Data Pooling: Estimating with External ConstraintsJerome Reiter
Williams College

When using regression models where units can be classified into distinct groups, similar parameters in each group can be estimated via explicit data pooling, such as in hierarchical models. Sometimes, however, external constraints prohibit explicit data pooling. For example, in the plan for the 2000 census that includes the Integrated Coverage Measurement survey, the Census Bureau avoids pooling data across states because the law may not allow data from one state to affect the population estimates in another state. Similar constraints may exist when auditing or comparing several groups' performances.

I present techniques that may be acceptable under such external constraints and yield more accurate estimates than those obtained by regressing separately in each group. These techniques utilize the information in multiple groups' parameter estimates to specify the model in each group, but ultimately estimate the parameters selected for each group's model using only that group's data. The techniques can be conceptualized as existing on a continuum ordered by how directly each relies on data pooling those techniques that look more like explicit data pooling are typically more accurate yet less likely to be acceptable. I investigate the techniques in a variety of simulation studies.

Do Blocks Make a Neighborhood? Approaches to Estimating Neighborhood Parameters Based on Localized ObservationsCarolyn A. Carroll
Stat Tech, Inc.

The talk will compare some approaches to parameter estimation based on "local" data readings. Data in some fields (e.g., environmental) can be viewed as "samples". But the samples may be poorly designed and contain some inherent but unknown bias. Finding methods of combining readings and making defensible, general statements across a larger area e.g., the neighborhood/community is difficult.

Semi-Parametric Nonlinear Mixed Effects ModelsYuedong Wang and Chunlei Ke
University of California, Santa Barbara

We present a class of semi-parametric nonlinear mixed effects models (SNMM) for repeated measures data. A SNMM assumes that the mean function depends on some parameters and nonparametric functions. These parameters provide interpretable data summary and these nonparametric functions provide the flexibility to allow data to decide some unknown/uncertain components. A second stage model with fixed and random effects are used to model the parameters. Smoothing splines are used to model the nonparametric functions. Covariate effects on parameters can be built into the second stage model and covariate effects on nonparametric functions can be constructed using smoothing spline ANOVA decompositions. SNMMs contain many existing models such as nonlinear mixed effects models and self-modeling nonlinear regression models as special cases. Therefore they can be used as diagnostic tools for many parametric and nonparametric models. Applications will be illustrated using real data sets.

Generalized Aspects of Fitting Generalized Nonparametric Mixed Effects Models

Peter Karcher and Yuedong Wang
University of California, Santa Barbara

Generalized Linear Mixed Effects Models (GLMM) provide useful tools for correlated and overdispersed non-Gaussian data. In this paper we consider Generalized Nonparametric Mixed Effects Models (GNMM) which relax the rigid linear assumption on the conditional predictor in a GLMM.

We use smoothing splines to model fixed effects. The random effects are general and may also contain stochastic processes corresponding to smoothing splines. We show how to construct smoothing spline ANOVA (SS ANOVA) decompositions for the predictor function. Components in a SS ANOVA decomposition have nice interpretations as main effects and interactions. We estimate all parameters and spline functions using stochastic approximation with Markov Chain Monte-Carlo.

Likelihood Based Tests for Over and Underdispersion Against General Alternatives

Gordon K. Smyth and Heather Podlich
University of Queensland

Variations from Poisson and binomial variation are a common concern when modelling count data. Tests for overdispersion are usually based on unrealistically specific alternatives, such as the negative binomial or beta-binomial distributions, or are not model based and therefore lack power. Convincing methods for detecting and modelling underdispersion are not generally available. We use extended Poisson process models, in which an arbitrary count distribution can be represented as the realization of a pure birth process. Under and overdispersion relative to the Poisson or binomial distributions can be represented in terms of the slope and curvature of the unobserved birth rate sequence. We give a new saddlepoint approximation for birth processes which is exact in the neighborhood of Poisson, negative binomial and binomial models. This allows us to compute score tests for the goodness of fit of standard models against very general alternatives.

THURSDAY, APRIL 6, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSION: Wavelets, Splines, State-Space and Adaptive Models**CHAIR:** Hyunjoong Kim, Worcester Polytechnic Institute**NORM Thresholding Method in Wavelet Regression**Dongfeng Wu
University of Texas

We present a new method called the NORM method for finding threshold values in wavelet regression. We use Wavethresh software in S-Plus to implement this method, and compare it with existing methods, such as Donoho & Johnstone's SureShrink, AdaptShrink and Nason's Cross-Validation, and with optimal thresholding. The goal is to minimize the average mean squared error. We use 3 different kind of noise: iid normal, iid t variable, and correlated noise, on 8 different test applications, including BLOCK, BUMP, HEAVISINE and DOPPLER. For iid normal noise, any method could be the best. The Cross-Validation method works best for independent long tailed t noise. In the case of correlated noise, the Norm method behaves best when the lag one correlation is large positive. The SureShrink or AdaptShrink method behaves best when the lag one correlation is large negative. We give a heuristic explanations of these behaviors. We also evaluate the accuracy of our method for estimation of noise level sigma.

Data-Driven Optimal Denoising and Recovery of Derivatives Noisy Signals Using MultiwaveletsNathaniel Tymes and Sam Efromovich
University of New Mexico

Multiwavelets are relative newcomers into the world of wavelets. Thus it has not been a surprise that the used methods of denoising are modified universal threshold procedures developed for uniwavelets. On the other hand, the specific of a multiwavelet discrete transform is that typical errors are not identically distributed and correlated whereas the theory of the universal thresholding is based on the assumption of identically distributed and independent normal errors. Thus we suggest an alternative denoising procedure based on Efromovich-Pinsker algorithm. We show that this procedure is asymptotically optimal over a wide class of spatially inhomogeneous functions. Moreover, together with a new "cristina" class of biorthogonal multiwavelets the procedure implies an optimal method for recovering the derivative of a noisy signal. The asymptotic results are supported by intensive Monte Carlo experiments.

Adaptive Splines and Genetic Algorithms for Optimal Low-Dimensional Statistical ModelingJennifer I. Pittman
Pennsylvania State University

Due in part to the increased availability of computational power, spatially adaptive smoothing methods involving regression splines have become a popular and rapidly developing class of nonparametric modeling techniques. Most existing algorithms for fitting adaptive splines are based on non-linear optimization and/or stepwise selection. Although computationally fast and spatially adaptive, stepwise knot selection is necessarily suboptimal while determining the best model over the space of adaptive knot splines is a very poorly behaved non-linear optimization problem. A possible alternative is to use more intensive numerical optimization techniques such as genetic algorithms to perform knot selection.

A spatially adaptive modeling technique referred to as adaptive genetic splines (AGS) is introduced which combines the optimization power of a genetic algorithm with the flexibility of polynomial splines. Preliminary simulation results comparing the performance of the genetic algorithm method to other current methods, such as HAS (Luo and Wahba 1997) and SUREshrink (Donoho and Johnstone 1995), will be discussed, as well as a current application of AGS in the engineering sciences. Topics for future research will also be mentioned.

Partially Adaptive Bandwidth Used in Prediction and Local Regression

Janis Grabis
Riga Technical University

A bandwidth parameter of the local regression model can be set either globally or locally. This paper considers partially adaptive bandwidth selection. The partially adaptive bandwidth is used to adjust the global bandwidth for particular data points. The adjustment takes place if a specified quality criterion of the given local model fails. The localized Akaike's Information Criterion acts as this quality measure. The global bandwidth is set either by cross-validation or arbitrary. In the first case the partially adaptive bandwidth is designated to improve the results of cross-validation. The second approach allows skipping of cross-validation. That provides substantial computational savings with small accuracy losses. The partially adaptive bandwidth attempts to encompass both the robustness of the global bandwidth and the flexibility of the local bandwidth. Performance of the partially adaptive bandwidth is evaluated by prediction of several empirical times series.

Self-Modeling Regression with Random Effects Using Penalized Regression Splines

Naomi S. Altman and Julio C. Villarreal
Cornell University

Self-modeling regression is a semi-parametric method for describing a family of similar curves. The overall shape of the curve is estimated nonparametrically, but differences among the curves are described through a parametric model. In designed experiments in which the response is a curve, modeling the parameters through a random effects model is often desirable.

In this paper, we describe a self-modeling regression model in which the nonparametric curve is defined by a penalized regression spline. Since the spline can be estimated as a linear random effects model, this allows us considerable simplification in both computation and inference. This simplicity can then be exploited to extend the model to generalized regressions.

Estimation of Nonlinear State-Space Models in the Presence of Censored Observations

Craig Johns
Colorado University and NCAR

Robert H. Shumway
University of California, Davis

State-space models involving time varying parameters are often used for describing broad classes of biological and physical phenomena. In some cases, measurement devices that produce data suitable for such models are hampered by an inability to measure beyond certain specified upper or lower detection limits. Traditional approaches to estimation for nonlinear state-space models use maximum likelihood procedures. These procedures depend on being able to compute conditional expectations via Kalman filtering and smoothing and are intractable under censoring or when using nonlinear models.

Carlin, Poulson and Stoffer (1992) develop a Markov Chain Monte Carlo (MCMC) estimation procedure for nonlinear state-space models. This MCMC method is extended to fit linear and non-linear state-space models when observations have been censored due to detection limits.

These MCMC estimation procedures are applied to filtering and parameter estimation for nonlinear state-space models of spatio-temporal data collected by a laser detector (lidar) measuring airborne particulate matter created by a moving point source.

THURSDAY, APRIL 6, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSION: Applications of Exploring, Modeling and Presenting Large Datasets**CHAIR:** Derek Stanford, MathSoft, Inc.**Predictive Statistical Models for Detecting Anomalies and Congestion in IP Based Networks**

Elisa M. Santos
Telcordia Technologies

Network performance monitoring has become a necessity to ensure quality of service. Currently, the tools available for network monitoring are mainly geared towards monitor to destination performance and cannot identify midway (node/link level) congestion or anomalies.

A methodology is presented to evaluate the performance of an IP based network with respect to the delay metric at individual node/link level using available methods of data collection. Delay data are collected with respect to each link and a statistical model is developed, taking into consideration potential trends and periodicities. The model is automatically upgraded to reflect normal changes of delay level and reflects the normal state of the network and, for this, is based on under control delay data. Congestion detection is based on model predictions and user-defined thresholds. The use of the model concentrates the detection effort on real signal, instead of chasing peaks/bursts that are mostly natural random variation. A formal statistical test is proposed for identifying congestion. To detect anomalies in a link it is necessary to model the delay under normal network conditions, regularly looking for abnormal behavior in the latest data, suggested by significant patterns in the residuals using appropriate statistics.

A Hierarchical Mixture Model for WWW-Usage

Dee Denteneer
Philips Research

Access networks (e.g. cable networks) are currently being standardised (e.g. DOCSIS, IEEE, DVB) and are the focus of extensive commercial activity.

Characteristic for such networks is a sequential procedure for data transfer from a station at the customer premises to a central node, which consists of two stages. First, a contention stage is carried out, in which a station requests a number of data slots (in contention with other stations). Second, a data transfer stage is carried out, in which the data is transferred in the data slots that have been reserved for this station.

This procedure implies that the performance of access networks is sensitive to both long-range dependencies and short-range dependencies in the traffic carried over the network. Hence it is mandatory that the traffic models used in performance analysis accurately reflect both types of dependencies.

We propose hierarchical mixture models for this purpose and develop an EM-algorithm to fit them. Relevance and use are demonstrated by applying the model to data on WWW-usage.

Tracking Timing Patterns for Millions of Customers in Real-Time

Jose C. Pinheiro, Diane Lambert, and Don X. Sun
Bell Labs–Lucent Technologies

Business applications such as credit card fraud detection and e-commerce require tracking the behavior of millions of customers who are making transactions. The behavior of each customer must be summarized separately and updated whenever the customer makes a transaction. Because storage space may be limited to a few hundred bytes per customer and computing time may be limited to milliseconds, the updating can depend only on the new transaction and the current summary for a customer. If a characteristic, such as the amount of a purchase is observed at random, then its distribution can be updated by exponentially weighted moving averaging. Timing variables, like day-of-week, however, are not observed at random, and standard sequential estimates of their distribution can be badly biased. To develop good estimates, we model timing variables by a Poisson process that has piecewise constant rates that evolve over time. This leads to a variant of exponentially weighted moving averaging that requires little storage space, is simple to compute, and is appropriate for timing variables. The new sequential estimator approximates the mean under a dynamic timing model and has good asymptotic properties. Simulations show that it also has good finite sample properties.

Data Mining on Time Series: An Illustration Using Fast Food Restaurant Franchise Data

Lon-Mu Liu, S. Bhattacharyya, S. L. Sclove, R. Chen, and W. J. Lattyak
University of Illinois, Chicago, Scientific Computing Associates Corp.

With the prevalent use of modern information technology, a large number of time series may be collected during normal business operations. We use a fast-food restaurant franchise as an example to illustrate how data mining can be applied to such time series, and help the franchise reap the benefits of such an effort. Time series data mining at both the store level and corporate level are discussed. Related data warehousing issues are also addressed. Box-Jenkins seasonal ARIMA models are employed to analyze and forecast the time series. Instead of a traditional manual approach of Box-Jenkins modeling, an automatic time series modeling procedure is employed to analyze a large number of highly periodic time series. In addition, an automatic outlier detection and adjustment procedure is used for both model estimation and forecasting. The improvement in forecast performance due to outlier adjustment is demonstrated. Adjustment of forecasts based on stored historical estimates of like-events is also discussed. To illustrate the feasibility and simplicity of the above automatic procedures for time series data mining, the SCA Statistical System is employed to perform the related analysis.

Redesigning Tables and Graphics for Federal Statistical Agencies

Daniel B. Carr
George Mason University

This talk describes the redesign of tables and graphs for communicating federal statistics summaries. The goal is to use perceptual and cognitive principles to make improvements while abiding within a given set of constraints. A set of constraints might be that the result to be a table, half-toning is limited to one or two colors, the table body elements may not highlight atypical values and the table must not take up much more space than in the previous publications. The talk focuses on a small set of examples from federal applications that are chosen to provide diversity. One example uses perceptual grouping and layering to better convey table header hierarchies and improve the readability of table body elements. Logical consideration even lead to improved wording for footnotes and suggest further modification. Another example shows conversion of a table to graphics and inclusion of metadata. Some changes simply make the result more attractive and provide a sense of value added. The examples illustrate the use of Splus™ and published statistical summaries from various agencies such as the Bureau of Labor Statistics and the Bureau of Transportation Statistics.

Remote Medical Evaluation and Diagnostics: A Testbed for Hypertensive Patient Monitoring

John C. Dumer, Timothy P. Hanratty, and Barry A. Bodt
U.S. Army Research Laboratory

H. Mitchell Perry and Sharon E. Carmody
Veterans Administration

Health care costs have surpassed \$2 billion per day in the United States. For government agencies faced with shrinking budgets, reduced medical staffs, and increased patient commitments this poses an especially challenging problem. One approach touted to allay this situation is to use remote medical diagnostics, involving the collection, monitoring, and analysis of patient data from remote locations (home) via a communication device. Toward this end, the U.S. Army Research Laboratory is developing a system for remote medical evaluation and diagnostics (RMED) that combines remote monitoring capabilities with intelligent decision support technology.

The first area identified for use with the RMED system is hypertension. A current effort with the Veteran's Administration through the St. Louis Hypertension Program employs prototype blood pressure cuffs in the field. This paper will demonstrate the prototype system for data collection, storage, and retrieval, and summarize the results of a small pilot study to assess the reliability of the measurements relative to standard measurement procedures.

THURSDAY, APRIL 6, 2000

4:00–5:45 P.M.

INVITED SESSION: CSNA Sponsored Session: Applications of Clustering and Classification to Large Datasets**ORGANIZER/CHAIR:** William Shannon, Washington University School of Medicine**Inner-Loop Statistics in Automated Scientific Discovery from Massive Datasets**

Andrew Moore

Carnegie Mellon University and Schenley Park Research, Inc.

Intensive statistical analysis of massive data sources ("data mining") has been embraced as one of the final areas with a need for massive computation beyond that available on a \$2000 computer or \$200 videogame. We begin this talk with two examples of software, instead of hardware, giving 1000-fold speedups over traditional implementations of statistical algorithms for prediction, density estimation, and clustering.

We then pause to examine directions in which these software solutions seemed blocked when faced with Physics, Biology and commercial scientific data discovery problems. The primary blocks were a curse of dimensionality and limitations on machine main memories. This is followed by four examples of new pieces of research that circumvent these barriers: lazy cached sufficient statistics, exact accelerated k-means, multiresolution ball-trees for very high dimensional real-valued data, and filament identifiers.

We then reveal the reason for our new-found respect for super-computation: when an algorithm you previously ran overnight executes in seconds, you find yourself wanting to run it ten thousand times. We show the impact of being able to run intensive statistics as an inner loop has had on our analysis of cosmology data (preliminary data from the Sloan Digital Sky Survey) and biotoxin identification, where desirable but hopelessly extravagant operations such as model selection, bootstrapping, backfitting, randomization and graphical model design now become somewhat non-hopeless.

Joint work with Andy Connolly (U Pitt Physics), Artur Dubrawski (Schenley Park Research), Geoff Gordon (Auton Lab), Paul Komarek (Auton Lab), Bob Nichol (CMU Physics), Dan Pelleg (Auton Lab) and Larry Wasserman (CMU Statistics).

Current Approaches to Gene Chip Data Analysis

Daniel Weaver

Genomica Corporation

Information from the Human Genome Project is allowing scientists to perform systematic experiments and gather data in unprecedented amounts. This talk will review the mathematical classification techniques being applied to gene expression data and will frame the scientific questions that such data can address. Current techniques being applied to such data range from simple average-linkage clustering to self-organizing maps. While these techniques are sufficient for the existing data volumes, they are unlikely to work efficiently on the large data sets that will be generated in the coming years. Some key scientific questions that will be raised include: What constitutes a statistically significant, diagnostic gene expression pattern? How are gene expression data used to functionally classify genes? How can large volumes of gene expression data be used to predict the underlying gene expression control mechanisms? No biological background is needed; the relevant biological concepts will be described.

Preliminary Studies on Combining Wavelet and Cluster Analysis for Gene Chip Data

William Shannon

Washington University School of Medicine

It is now recognized that examining patterns of gene expression (i.e., a gene's state of activity) in patients can assist health professionals in detecting, diagnosing, and treating human disease. Recently, a new technology named the nucleic acid array or 'Gene Chip' allows clinicians to measure these patterns and begin relating them to clinical diagnosis and outcomes.

In this poster we present an overview of the effort underway at Washington University School of Medicine in St. Louis to develop and apply microarray technology to basic and clinical research in surgery, molecular microbiology, genetics, and cancer. Attention is focused on the bioinformatic and biostatistical challenges we are faced with, and early efforts to bring this under control. In addition, we will discuss a novel application of wavelet transformations to gene chip data we are currently exploring.

THURSDAY, APRIL 6, 2000

4:00–5:45 P.M.

INVITED SESSIONS

The UC Irvine Knowledge Discovery in Databases Archive

Stephen D. Bay, Dennis Kibler, Michael J. Pazzani, and Padhraic Smyth
University of California, Irvine

Advances in data collection and storage have allowed organizations to obtain massive, complex and heterogeneous databases, which have stymied traditional methods of analysis. This has led to the development of new analytical tools that often combine techniques from a variety of fields such as statistics, computer science and mathematics to extract meaningful knowledge from the data. To support research in this area, UC Irvine has created the UCI Knowledge Discovery in Databases (KDD) Archive (<http://kdd.ics.uci.edu/>). This is a new online repository of large and complex databases, which encompass a wide variety of data types, analysis tasks and application areas. Our goal is to foster research in knowledge discovery by making these databases publicly available. The archive is supported by the Information and Data Management Program at the National Science Foundation, and is intended to expand the current UCI Machine Learning Database Repository to databases that are orders of magnitude larger and more complex.

THURSDAY, APRIL 6, 2000

4:00–5:45 P.M.

INVITED SESSION: Best of the *Journal of Computational and Graphical Statistics*: New Developments in EM**ORGANIZER/CHAIR:** Andreas Buja, AT&T**An Interval Analysis Approach to the EM Algorithm**

Kevin Wright
Pioneer Hi-Bred International

William J. Kennedy
Iowa State University

The EM algorithm is widely used in incomplete-data problems (and some complete-data problems) for parameter estimation. One limitation of the EM algorithm is that upon termination, it is not always near a global optimum. As reported by Wu, when several stationary points exist, convergence to a particular stationary point depends on the choice of starting point. Furthermore, convergence to a saddle point or local minimum is also possible. In the EM algorithm, although the loglikelihood is unknown, an interval containing the gradient of the EM q function can be computed at individual points using interval analysis methods. By using interval analysis to enclose the gradient of the EM q function (and, consequently, the loglikelihood), an algorithm is developed which is able to locate all stationary points of the loglikelihood within any designated region of the parameter space. The algorithm is applied to several examples. In one example involving the t distribution, the algorithm successfully locates (all) seven stationary points of the loglikelihood.

Fitting Mixed-Effects Models Using Efficient EM-Type Algorithms

David A. van Dyk
Harvard University

In recent years numerous advances in EM methodology have lead to algorithms, which can be very efficient when compared with both their EM predecessors and other numerical methods (e.g., algorithms based on Newton-Raphson). In this paper we focus on mixed-effects models and combine several of these new methods to develop a set of mode-finding algorithms, which are both fast and more reliable than standard algorithms such as `proc mixed` in SAS. We present efficient algorithms for Maximum Likelihood, Restricted Maximum Likelihood, and computing posterior modes. These algorithms are not only useful in their own right, but also illustrate how parameter expansion, conditional data augmentation, and ECME can be used in conjunction to form efficient algorithms. In particular, we illustrate a difficulty in using the typically very efficient parameter-expanded EM algorithm for posterior calculations, but show how algorithms based on conditional data augmentation can be used. We also present a result that extends Hobert and Casella's (JASA, 1996) result on the propriety of the posterior for the mixed-effects model under an improper prior, an important concern in Bayesian analysis. Finally, we show how similar methods applied to the Data Augmentation algorithm can lead to very efficient stochastic algorithms for posterior sampling.

THURSDAY, APRIL 6, 2000**4:00–5:45 P.M.****INVITED SESSIONS****THURSDAY, APRIL 7, 2000****4:00–4:45 P.M.****INVITED SESSION:** Characterizing Large Complex Natural Systems and Beyond**ORGANIZER/CHAIR:** Lorraine Denby, Bell Labs–Lucent Technologies**Statistics and Models for Complex Systems in Engineering and Biology**John Doyle
Caltech University

A great deal of attention has been given recently to describing features of complex systems in terms such as self-similarity, power laws, and entropy, phase transitions, criticality, fractals, chaos, and so on. This talk will focus on the fascinating statistical properties of web and internet traffic, and relate these to power law statistics in other domains, such as forest fires, power outages, natural and man-made disasters, and specie extinction. While it is now widely accepted that the commonly assumed Poisson traffic models poorly describe Internet traffic, it remains to be seen if these insights will lead to new approaches to network protocol design, which remains largely ad hoc. We critique the popular explanations from statistical physics and offer some novel explanations for the origins of power laws in terms of generalizations of source coding for data compression. The implications for future convergent, ubiquitous networking will also be discussed briefly as well as general issues of more rigorous approaches to analysis and robust design of complex multiscale systems in engineering and biology.

FRIDAY, APRIL 7, 2000

8:00–9:45 A.M.

INVITED SESSIONS

FRIDAY, APRIL 7, 2000

8:00–9:45 A.M.

INVITED SESSION: A Tutorial on Inverse Theory

ORGANIZER/CHAIR: Vicki Lancaster, Neptune and Company, Inc.

A Tutorial on Statistical Inverse Theory

Luis Tenorio
Colorado School of Mines

Ill-posed inverse problems arise when we try to recover information about a process from partial, indirect noisy observations. These problems are common in physical sciences like geophysics, astronomy and astrophysics where we have to rely on indirect measurements of processes in space or underneath the Earth's surface.

Through examples we will illustrate the questions that arise in ill-posed inverse problems and present some basic methods to determine a meaningful, stable solution. These methods include the Backus-Gilbert method, the method of regularization, singular value decomposition and wavelet estimation. We will also discuss methods to estimate the variability, and the bias of the inversion estimates, as well as minimax estimates of the mean square error.

Velocity Estimation in Exploration Geophysics, a Bootstrap Approach

Alberto Villarreal
Colorado School of Mines

The Seismic Reflection Experiment is one of the most important tools in geophysical exploration. This method consists of sending artificially generated seismic waves into the earth and recording them once they are reflected back to the surface by structural irregularities in the subsurface.

In a Seismic Reflection Experiment, the most important parameter of interest is the velocity at which waves travel through different media in the earth. The physical velocities in the subsurface covered by the seismic experiment define how the recorded data will look like, because seismic data consists of travel-time measurements (the time it takes for a wave to traverse the path source-reflector-receiver), and travel-time depends on the wave velocity in the earth. Therefore, we are interested in the inverse problem of estimating the medium velocities from data. These velocity estimates give important information about the structure and composition of the subsoil.

We use bootstrap resampling methods to improve and automate velocity estimation from seismic data. In the bootstrap approach, data samples are created by resampling the original seismic data. Next, an optimization procedure is used to obtain velocity estimates for each sample, and the variability of different velocity estimates is used to compute standard errors. This procedure is repeated iteratively with different trial velocities. The velocity estimate with the smallest error is selected. This is a computationally intensive method but can be efficiently implemented in parallel. Besides automating the velocity analysis, this method may be used to estimate errors of seismic velocities, which are essential for subsequent steps in the data processing sequence.

Generalizing Wiener-Levinson Deconvolution

Luis Tenorio and Alberto Villarreal
Colorado School of Mines

In reflection seismology a trace is modeled as a convolution of a seismic pulse with a reflectivity sequence that encodes information about the layering in the subsurface. To deconvolve the trace means to remove the blurring effect of the pulse to obtain an estimate of the reflectivity.

A usual assumption in deconvolution is that the reflectivity is a white random process. But, most of the time this assumption is not appropriate. We use Gaussian mixtures to generalize Wiener-Levinson deconvolution and obtain a procedure that is more robust to nonstationarities in the reflectivity and to correlation structure in noise.

Estimating the Influence of Random Noise on Measured Travel Times

Albena Mateeva
Colorado School of Mines

Tomography is used in seismic exploration for velocity-model building. Tomographic data consist of travel times of waves, excited near the earth's surface, which after having penetrated to a certain depth have been reflected back by some geological inhomogeneity. As any inversion procedure, tomography requires good knowledge of data uncertainties. Errors in measured travel times are introduced by a variety of factors but one of them is always present—random noise. This paper discusses its contribution to travel time uncertainty. Two objectives were set. First, to understand the interaction between signal and random noise that leads to uncertainty in travel time. Second, to estimate that uncertainty from seismic data. The latter is a hard statistical problem that seismic industry tends to ignore rather than tackle. A practical solution of it should be of great interest not only to geophysicists but also to anybody dealing with travel times of band-limited signals.

FRIDAY, APRIL 7, 2000

8:00–9:45 A.M.

INVITED SESSION: Defining, Measuring, and Analyzing Quality of Care: Statistical and Computational Challenges

ORGANIZER/CHAIR: Sally C. Morton, RAND

The HIV Performance Measurement Perspective from NYC, Framing the Clinical Context

Bruce Agins
New York State Department of Health

The NYSDOH AIDS Institute is responsible for systematic monitoring of the quality of medical care provided to HIV-infected individuals in NYS. Measurement of quality is based on indicators that are linked to optimal clinical outcomes, such as performance of PAP smears, PPD screening and use of antiretroviral therapy. Implemented in 1992, this initiative incorporates continuous quality improvement (CQI) techniques to stimulate health care providers to build and sustain quality within their organizations. Record abstraction at over 100 facilities is conducted annually, using standardized data collection forms. Through a new initiative, HIVQUAL, providers submit data directly using a software program developed in collaboration with HRSA. A subset of these records is reviewed to validate information. Results are presented as aggregated facility-specific data, comparatively and longitudinally to display historical trends. HIV performance data are also reported to the public. Accuracy and precision of data are paramount to the integrity and success of performance improvement efforts led by public health agencies. Impact at the state level ranges from individual agency actions to support for activities designed to improve care. This program represents an example of collaboration between a public health agency, clinicians and statisticians to incorporate sophisticated statistical techniques into routine CQI initiatives.

Measuring and Improving Quality in Managed Care: Some Statistical and Computing Issues

Randall K. Spoeri
HIP Health Plans, New York

With the rapid expansion of managed care, questions have been raised about the quality of care being delivered. In response to these concerns, quality management efforts have relied heavily on the measurement of performance. Associated with these measurement and improvement activities are various statistical and computing considerations. This talk will provide an overview of a number of these considerations, to include: measure definitions, data availability and quality, risk adjustment, analytical/interventional use of results, and predictive modeling. Thoughts about the future will conclude the talk.

Building Aggregate Health Care Quality Scales

John Adams
RAND

This talk will consider the emerging need to summarize information as multiple measures of health care quality proliferate, e.g., HEDIS, CAHPS and others. I will discuss various ways of building aggregate scales of health care quality and examine some of the technical issues that must be overcome to produce reliable aggregate information. These issues include case-mix adjustment, appropriate standard error calculations, developing weights for measures, and the communication of statistical uncertainty to the lay audience. The focus is on aggregate scales to profile HMO performance. Examples will be drawn from the development of the Combined Autos/UAW Reporting System (CARS), an HMO report card that is mailed to the employees of the big three auto makers during open enrollment.

Simulation in Models of Health Care Quality

Karl Heiner
SUNY at New Paltz

The AIDS Institute of the New York State Department of Health monitors the quality of care delivered by hospitals, community health centers and drug treatment centers to individuals infected with HIV. A medical peer review organization visits these facilities each year and applies a number of protocols reflecting the standard of care to random samples of medical records. Bayesian techniques are used to model aspects of the quality of care. For standards and indicators that have been employed for a number of years, conjugate analysis and simple dynamic models are useful when measuring facility specific performance. As standards and measures of quality of care evolve, or when comparisons among groups are required, more complicated models are indicated. In order to make inferences from the models, simulation methods are applied and trellis graphics are used to display overall and among facility trends.

Casemix Adjustment of the National CAHPS Benchmarking Data 1.0

Marc N. Elliott	Richard Swartz	John Adams	Ron D. Hays
RAND	Rice University	RAND	UCLA

Casemix adjustment of consumer ratings can provide more valid plan comparisons than unadjusted ratings by controlling for factors related to systematic response biases to questions about health care. Adjusted data are therefore potentially more appropriate for comparing the quality of care delivered. If members of a particular demographic group are less inclined than others to assign poor ratings to bad care, and members of this group are disproportionately enrolled in some plans, casemix adjustment for this systematic bias is useful when comparing assessments of different plans.

The CAHPS Implementation Handbook recommends adjusting for age and health status when comparing consumer assessments of health plans. Younger people and those in poorer health tend to report more problems and less positive evaluations of health care than do older people and those in better health. The current CAHPS approach uses a "health plan fixed effect" model to estimate the effects of casemix adjusters, which does not assume that all plans have equal true mean ratings (unlike simple casemix models). We also consider models that test whether casemix adjusters have different effects within different plans. We present graphical displays that compare the various casemix models considered.

Imputing Treatment Differences in Meta-Analyses with Missing Data

I. Elaine Allen	Ingram Olkin
Babson College	Stanford University

The problem of how to handle missing data occurs frequently in meta-analyses of clinical studies. Few cancer studies are randomized controlled trials and many include only one treatment arm. Often no comparative drug trials exist between competitive treatments in other therapeutic areas. Several techniques for imputing these differences will be described with examples from published meta-analyses and software imputation comparisons. These methods will include Bayesian hierarchical modeling to estimate missing random effects multilevel mixed models to estimate missing treatment differences and a proposed method to test the difference between active treatments when only placebo controlled studies exist.

FRIDAY, APRIL 7, 2000

8:00–9:45 A.M.

INVITED SESSION: The Use of Modeling and Statistics in Defense Analysis**ORGANIZER/CHAIR:** Nancy Spruill, Office of the Under Secretary of Defense (Acquisition and Technology)**Using Advanced Modeling and Simulation Technologies and Techniques in the Analysis of Defense**

Colonel William Crain
 Defense Modeling and Simulation Office

Modeling and Simulation (M&S) has become a very powerful and cost effective tool for analyzing the technical performance and warfighting utility of technologies being developed in Defense Science and Technology (S&T) programs as part of the DoD M&S strategy. An increased use of M&S-based experimentation is being seen in Defense laboratories, engineering centers, operational warfighting experiments (to address S&T impacts on force structure, doctrine, tactics, etc.), Advanced Concept Technology Demonstrations, and other aspects of the S&T community. This increase has built on many “success stories” associated with M&S-based analysis. To maximize the value of using advanced modeling and simulation technologies and techniques in the S&T community, it is important to understand the differences between models and simulations, as well as their appropriate use in experimentation. Additionally, it is important to become aware of the new technologies and applications being developed in the M&S community and their benefits to the acquisition, training and analysis communities.

The Joint Warfare System (JWARS): A Tool to Improve Combat Simulation for the Information Age Military

Lieutenant Colonel Daniel Maxwell
 JWARS, Office of the Secretary of Defense

The Joint Warfare System (JWARS) is a campaign-level simulation of military operations that is being developed under contract by the Office of the Secretary of Defense (OSD) for use by OSD, Joint Staff, Services, and Warfighting Commands. The motivation for JWARS is to provide insight into cause and effect relationships that influence the success or failure of military forces, and ultimately to support critical operational planning and multi-billion dollar resource allocation decisions. JWARS is a closed form simulation. It is mixed-mode, with both stochastic and deterministic components key uncertainties are reflected as stochastic events that potentially cause significantly different outcomes. JWARS also represents explicitly the effects that differences in information availability may have on operational success, as well as explicitly representing most critical combat and logistical systems. This highly resolved view of warfare provides a holistic view of combat operations that has been previously unachievable. This paper describes the JWARS design, modeling concepts, and sample model results, like they might be presented to a senior leader.

Improving Inventory Performance by “Rightsizing” Inventory Reorder Points

Ron Fricker
 RAND

This paper develops and applies a technique for “rightsizing” inventory levels. We demonstrate its power on a Marine Corps' inventory consisting of \$24 million of stock for 13,000 different types of items. We show that our “rightsized” inventory works significantly better than the existing inventory: We achieve equivalent performance at one-half to one-third the cost. Conversely, we demonstrate significant improvement in fill rates and other inventory performance measures for an inventory of the same cost. The computationally intensive method, based on the bootstrap, is only now becoming possible to apply with the advent of today's powerful desktop computers. It is an alternative to the standard approaches, which are often inappropriate and only applied out of computational convenience.

Use of Decision Support Simulation Tool in Army Resource Decisions

Craig E. College
Program Analysis and Evaluation Office, Army

Senior Army decision-makers constantly struggle to optimize the allocation of funds for readiness, modernization, and soldiers' quality of life. The analysis supporting resource decisions in the Army Program Objective Memorandum (POM) process comprises a large and complex series of tasks. There are time-delays between causes and effects, extensive "feedback" loops, and numerous critical qualitative and quantitative factors, which must be incorporated in the analysis. Many quick-reaction "what-if" scenarios must be investigated to detail the impacts of alternative decisions. In the face of reduced personnel and dollar resources, essentially manual execution of these activities is increasingly difficult. This situation generates a requirement for a suite of analytical models to assist in: 1) developing the POM 2) articulating the impact of resource decisions and 3) analyzing resource trade-offs. Army Program, Analysis, and Evaluation Directorate's (PAED) Decision Support System (DSS) is being developed and tested to enhance the Army POM process. DSS techniques and methodologies include rank-based hierarchical assessment, Quality Function Deployment (QFD), and System Dynamics Simulation. Among several functions, this DSS tool prioritizes and optimizes over 500 Management Decision Packages (MDEPs) which are the key programming structures supporting the Army's mission and Title X responsibilities. Risks associated with resultant funding levels in various sectors, e.g., readiness, modernization, and soldiers' quality of life are then made evident. Finally, this tool provides the capability to support a future executive-level display of resource options for real time decisions making.

Modeling Support Infrastructure for the Expeditionary Aerospace Force

Lionel Galway
RAND

In response to a "new world" of frequent deployments, the U.S. Air Force has developed a new operational concept, the expeditionary Aerospace Force, which replaces large overseas forces with units that can be deployed quickly from the U.S. Meeting demanding timelines for deployment will require a rethinking and potential restructuring of all support functions, such as munitions, fuel, maintenance, and supply. Because of the large uncertainties regarding proposed support alternatives in future scenarios, we argue that strategic support planning (i.e. decisions on overall structure, technology and process improvements, etc.) should be done with relatively simple and transparent models that have modest data requirements and run quickly to allow runs over many different scenarios. These models allow quick exploration of wide ranges of alternatives and help direct detailed analysis to promising options.

Information and Knowledge Organization to Support Modeling and Statistics in Defense

Yvonne M. Martinez
Los Alamos National Laboratory

The creation of knowledge bases to support decision-making has become an integral part of our Statistical Sciences Group's weapons reliability efforts. Designed properly, these knowledge bases become the critical infrastructure of expertise from which information (quantitative and qualitative) is extracted and modeled. An example of the Prototype Slapper Detonator Knowledge Base will be given. This is an electronic repository for capturing and preserving knowledge on a type of detonator—the Slapper. The Knowledge Base is a resource for the Department of Defense's (DoD's) decision making on the design, modeling, manufacturing, and reliability of the Slapper. Methods for developing the prototype originated with meetings of advisory experts and analysts on the contents, organization, and needed capabilities of the Knowledge Base. We refine the Knowledge Base through usability tests—observations of users as they search through the Knowledge Base and perform their decision-making tasks. The Statistical Sciences Group is further exploring the use of knowledge bases to integrate information, to perform statistical analysis and modeling, and to support collaborative updating of knowledge. This work will also be displayed through examples of PREDICT (**P**erformance and **R**eliability **E**valuation and **D**esign by **I**nformation **C**ombination and **T**racking) and RETAIN (**R**epository for **E**xpertise and **T**ools for **A**nalyzing **I**ntegrating **K**nowledge).

FRIDAY, APRIL 7, 2000

10:30 A.M.–12:15 P.M.

INVITED SESSION: Enterprise Modeling: Supply Chain Design to Statistical Performance Analysis**ORGANIZERS:** Bonnie Ray, New Jersey Institute of Technology and Leslie M. Moore, Los Alamos National Lab**CHAIR:** Bonnie Ray, New Jersey Institute of Technology**Systems Thinking in Supply Chain Management**

Charu Chandra
University of Michigan

The concept of supply chain is about managing co-coordinated information and material flows, plant operations, and logistics. It provides flexibility and agility in responding to consumer demand shifts with minimum cost overlays in resource utilization. The fundamental premise of this philosophy is synchronization among multiple autonomous entities represented in it. That is, improved co-ordination within and between various supply chain members. Co-ordination is achieved within the framework of commitments made by Members to each other. Members negotiate and compromise in a spirit of co-operation, in order to meet these commitments. Increased co-ordination can lead to reduction in lead times and costs, alignment of interdependent decision-making processes, and improvement in the overall performance of each Member, as well as the supply chain network (Group). Such an arrangement offers opportunities to design, model, and analyze problems with local perspective of a Member and global view of a Group. It also holds the potential of emergence of divergent supply chain network topologies, in order to satisfy dynamic market conditions. These unique configurations and associated problems require formulations in relation to a systems framework, recognizing their domain dependence within the domain independent environment of the supply chain. In this talk, we present a systems framework that utilizes an interdisciplinary approach to supply chain management with methods and techniques incorporated from production operations management, management science, industrial engineering /operations research, systems sciences, and artificial intelligence and computer science fields.

One of the important supply chain management problems is to transform incomplete information about the market and available production resources into co-coordinated plans for production and replenishment of goods and services in the network formed by co-operating entities. We illustrate the proposed framework to address this problem, for a textile supply chain.

Planning Experiments with Computer Models of Complex Phenomena

Leslie M. Moore
Los Alamos National Laboratory

Bonnie Ray
New Jersey Institute of Technology

Dennis R. Powell
Los Alamos National Laboratory

Computer codes are being developed to model many complex phenomena including a manufacturing supply chain. Use of computer simulation models leads to the consideration of statistical methods for gaining understanding from these models of underlying processes, perhaps to stimulate the development of science or to support decision making. We describe statistical methods for sensitivity and performance analysis of complex computer simulation experiments. Analysis of variance-based methods or regression tree analysis are useful for determining variables having substantive influence on the experimental results and to investigate the structure of underlying relationships between inputs and outputs. An approach to analysis leads to the need to design computer experiments from which estimates of the quantities of interest can be obtained with reasonable efficiency. Inputs to simulation codes may number in the tens to hundreds and information that allows focus on subsets of important inputs is invaluable. Some experiment design approaches based on fractional factorial design, or orthogonal arrays, are described.

Forecasting Methods for Supply Chain Management

Bonnie Ray
New Jersey Institute of Technology

Forecasting item demand across different segments of the manufacturing processes and across different time horizons is an integral part of supply-chain management. In this talk, we review time series methods that are commonly used for demand forecasting in inventory management applications and discuss some issues that arise when the methods are extended to the supply chain framework. An example of forecasting items in a textile supply chain is used to illustrate the discussion.

FRIDAY, APRIL 7, 2000

10:30 A.M.–12:15 P.M.

INVITED SESSION: Using Statistical Modeling to Identify Perturbations in Earth System Processes: Examples from Landscape Evolution and Tectonics

ORGANIZER/CHAIR: Dorothy Merritts, Franklin and Marshall College

River Network Scaling Laws: Deviations and Fluctuations

Peter Dodds and Daniel Rothman
MIT

The statistics and structure of river networks are commonly described by power laws. In practice, deviations in scaling are present making exact measurements difficult. The choice of parameter ranges used for regression analysis can markedly affect estimates of exponents. We show that many relationships possess several distinct scaling regimes linked by crossover regions. These scaling regimes, which may be present to varying degrees, pertain to the scale of linear, pre-network basins large length scales at which correlations in landscapes become negligible and outer length scales dictated by geology. We present evidence from real data for large-scale networks including the Mississippi, Amazon, Nile and Kansas River basins. We observe that improvements in topographic resolution would be unlikely to result in cleaner statistics that variations in measurements for small-scale basins are real and unavoidable and that strong deviations are indicative of geology being at work.

Quantitative Testing of Landform Evolution Models

Garry Willgoose and Greg Hancock
University of Newcastle, Australia

Over the last 15 years a range of landform evolution models have been developed. These models are highly nonlinear and sensitive to small perturbations. It is not possible to carry our repeated combinatorial experiments to collect the data to test these models. Finally a key aspect of landforms are their spatial organisation in the form of drainage networks. This paper outlines a novel methodology the authors have developed to address these limitations using key indicator statistics (width function, cumulative area function, area-slope relationship) and the GLUE hypothesis testing methodology (Beven and Binley, 1992). Results from studies will be shown. Some unresolved statistical challenges will be outlined including

(1) accounting explicitly for spatial organisation and patterns of drainage for "random" networks. We cannot do repeated experiments in the field but we can do repeated experiments in the computer and see if the field data is likely to have come the computer population of generated landscapes.

(2) developing likelihood functions for simultaneous use of several test statistics. We must understand the correlation between test statistics. For instance, the hypsometric curve and area-slope relationship both characterise elevation properties but what model discrimination power does hypsometry provide over and above area-slope alone?

Deviations in Slope-Area Relations that Indicate Geologically Recent Crustal Deformation

Tim C. Hesterberg
MathSoft, Inc.

Dorothy Merritts
Franklin and Marshall College

In 1811-1812 three great (8.0+) earthquakes occurred near New Madrid, Missouri. We estimate coseismic deformation in this area using stream elevation data from topographic maps. Streams have a natural profile, the gradient of which depends on the resistance of underlying sediment and the volume of stream flow. If tectonic processes elevate the upstream end of a segment a different amount than the downstream end, the stream will attempt to return to its natural gradient by incising, aggrading, or altering its sinuosity. This adjustment takes time, so deviations from the natural gradient may indicate geologically recent deformation. We use penalized regression splines to estimate the natural stream profile and the deformation of the ground surface. Estimation of the natural profile and deformation is based on nonparametric regression of the form $y_2 - y_1 = f(x_2) - f(x_1)$, where the x 's are univariate or bivariate.

**Conjugate Gradient Methods for Large-Scale Sparse Regression,
with Applications to Seismic Deformation Estimation**

Derek Stanford
MathSoft, Inc.

To estimate seismic deformation, we need to fit a smooth surface to marked spatial data. Estimation of this surface can be characterized as a regression problem with smoothness constraints. We use a conjugate gradient method because the large size of the design matrix in this regression does not permit computation of an exact least squares solution. For example, a relatively coarse surface grid with 10^3 cells per side would lead to a design matrix with 10^6 columns an exact linear solution would require inversion of a matrix with dimensions 10^6 by 10^6 , which is not currently feasible. Our conjugate gradient approach allows us to avoid this difficulty by taking advantage of the sparse structure of the design matrix, and we have implemented this in software tailored to the use of topographic stream elevation data.

FRIDAY, APRIL 7, 2000

10:30 A.M.–12:15 P.M.

INVITED SESSION: Statistics in Precision Agriculture

ORGANIZER/CHAIR: Barry Moser, Louisiana State University

Hypothesis Tests in the Presence of Spatial Correlation

Robert G. Downer
Louisiana State University

Traditional experimental designs do not directly account for spatially dependent observations. However, this aspect of the data cannot be ignored in both design and analysis. The basic issues associated with spatially correlated observations in standard designs will be presented and some of the methods which have attempted to address the problem will be reviewed. The impact of these issues on hypothesis testing will be discussed. For fixed effects one-way analysis of variance, a permutation test is introduced.

Statistical Issues in the Analysis of Remotely Sensed Data as Pertains to Precision Agriculture

Patrick D. Gerard, David Evans, and Michael Cox
Mississippi State University

The potential environmental benefits of precision agriculture have been well documented. However, before these benefits can be fully realized, the economical feasibility of these methods must be addressed. One of the primary costs associated with precision farming involves the acquisition of information on quantities of interest, such as soil fertility and weed abundance, across a large area. Normally, obtaining the requisite information is both time and labor intensive. Agricultural scientists have turned to remote sensing to non-invasively obtain information, in the form of intensity of energy radiated from the earth, which may be related to the quantities of interest.

In this talk, we will briefly discuss some of the objectives of precision agriculture and how they relate to remote sensing. We will also discuss, in general terms, some statistical and data management issues that arise from the use of remote sensing in precision agriculture, with emphasis on new challenges as remotely sensed data moves from being multispectral in nature to hyperspectral.

Use of Spatial Statistics to Design Sampling Plans for Monitoring Rust in Coffee Trees

Raúl E. Macchiavelli and Rocío del P. Rodríguez
University of Puerto Rico

Rust is an important disease of coffee that decreases yield of coffee beans. For monitoring purposes, sampling is done in a two-step systematic plan: trees are sampled systematically (in a W pattern covering the field), and then leaves are randomly sampled from each selected tree. Since coffee in Puerto Rico is grown in areas with pronounced slopes, these plans require walking diagonally along slopes, which is not feasible for regular monitoring by farmers. In this work we compare by simulation different sampling plans in order to find one to be used for monitoring this disease. The disease incidence was evaluated for all trees in a coffee lot ($N=1269$) and two systematic patterns were studied: a W pattern every c trees ($c=2, \dots, 6$) and a pattern with parallel rows, where trees were selected along 4 equally spaced parallel rows, every c trees ($c=2, \dots, 7$). Simulations were carried out using a SAS macro, sampling 2, 5, 10, 30, and 40 leaves per tree every c trees. In order to generalize these results to other fields, the spatial distribution pattern was modeled using spatial statistics new data sets were generated changing the disease incidence and the spatial correlation and simulations were run using the approach described before. The results suggest that both systematic sampling patterns gave approximately the same standard errors (except in cases with large spatial correlation).

Effect of Number of Seed Bulked when Using the Multiple-Seed Procedure for Self-Pollinated Crops

James Beaver and Raúl E. Macchiavelli
University of Puerto Rico

Single-seed descent has been used by grain legume breeders to maintain genetic variability in populations of advanced generation lines. In this method, a single seed from each plant is chosen and planted, and the process is repeated several times (typically four). In order to reduce labor costs, breeders use a multiple seed procedure in which a single pod rather than a single seed is harvested from each plant and bulked. This paper studies the distribution of the proportion of the original plants represented after applying this method four times (advancing from the second to the sixth generations). Since the analytical solution to this problem is intractable (it involves a 4-fold convolution of a multivariate hypergeometric distribution), a simulation was run in SAS to estimate this probability distribution. Results show that the proportion of plants represented at least once is between .25 and .33 that an increase in size of the original population does not influence significantly the mean proportion but decreases its variability and that the number of seeds per pod affects both the mean and the standard deviation of the distribution.

Geostatistical Analysis of Spatial Nutrient Data in a Precision Agriculture Experiment

Bradley Tiffée and Robert G. Downer
Louisiana State University

Precision agriculture uses geographic information systems, computer technology and a global positioning system to map site-specific factors that affect production. The Dean Lee Research Station of the Louisiana Agricultural Experiment Station is using this technology to map soil nutrients and soybean yield. This may be achieved through spatial modeling using variograms and kriging. A sample variogram is estimated from sampled magnesium concentrations and kriging is used to predict values for the entire field. A comparison is made to prediction from a trend surface model and recommendations are discussed.

FRIDAY, APRIL 7, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSION: Exploration and Visualization in High Dimensions**SPONSORED BY THE CAUCUS FOR WOMEN IN STATISTICS****CHAIR:** Matthias Schonlau, RAND**Exploration and Estimation of North American Climatological Data**James A. Shine and Paul F. Krause
U.S. Army Topographic Engineering Center

The availability of spatial and temporal earth data is increasing for example, NASA's Earth Observing System (EOS) will soon be producing a terabyte of earth data per day. This data will permit more detailed exploration and analysis of earth systems than was possible in the past. One application of particular interest to the authors is the estimation of contour surfaces from point values and the visualization of these surfaces in map form. The authors explored a multivariate data set of approximately 6,000 points in North America and other global locations. Each point contains climatological and other information such as temperature, elevation and precipitation. Spatial correlation was modeled using a semivariogram and several estimation approaches were used to create estimated surfaces and resulting contours. Maps of these results and comparisons between different approaches will be presented.

Visualizing Abandoned Hazardous Waste Sites in the United StatesCarolyn K. Offutt
U.S. Environmental Protection Agency

The U.S. Environmental Protection Agency has identified over 35,000 abandoned hazardous waste sites across the country for investigation and possible remediation. Most of the sites do not require Federal action, but 1,400 sites are currently on the National Priorities List (NPL) for cleanup under the Superfund program.

The location and characteristics of the NPL sites form a large data set that has the capacity for analysis and subsequent visualization of the analyses. Characteristics of the sites include industrial activities, type of contamination, environmental media being contaminated (air, soil, surface water, ground water, sediment, structures, etc.), remediation planned or undertaken (excavation, in-situ bioremediation, soil vapor extraction, incineration, solidification, etc.), and many more characteristics. Spatial representation of this information allows further analysis of soil type, aquifers, endangered species habitats, and census data.

Public interest compels EPA to provide access to the data, as well as to the analytical tools. Much of the information is accessible on the Superfund Web site at: <http://www.epa.gov/superfund>.

This paper will demonstrate how this large data set is managed, updated, queried, and made accessible. In addition, the paper will discuss future plans for data accessibility and manipulation, as well as some of the data issues.

High-Dimensional Visualization Using Continuous ConditioningWilliam C. Wojciechowski and David W. Scott
Rice University

Many methods for visualizing hypervariate data have been employed. Among these are slicing, coplots (Cleveland, 1993), and dynamic coplots (Wojciechowski and Harner, 1995). Slicing and coplots display subsets of the data in static plots. The subsets are determined by the value of one or more conditioning variables. Dynamic coplots add animation and the displayed subset is updated in real-time as the conditioning values are modified. For all of these methods, the displayed points are determined by an indicator function. This produces a discontinuous transition as the value of the conditioning variable changes. The method we introduce uses a continuous weighting function based on a distance measure defined on the conditioning variable sub-space. The distance measure determines the color and transparency of a data point and produces a smooth visual effect during animation. The abstract notion of distance provides for many conditioning techniques. We present examples developed on Rice University's Immersadesk.

FRIDAY, APRIL 7, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSIONS

Graphical Techniques for the Exploration of Functional Data

E. Neely Atkinson
MD Anderson Cancer Center

This talk will demonstrate some tools for the graphical exploration of functional data, that is data in which some of the variables of interest may be considered as observations of an underlying smooth process. The functional data may be displayed or transformed in several ways and may be linked to a number of univariate and multivariate plots of covariates of interest. The methods are coded in LISP-STAT. The techniques will be illustrated on a data drawn from an ongoing study of the use of fluorescence spectroscopy to diagnosis cervical abnormalities.

Authenticating Vulnerability Measurements

Edward J. Wegman
George Mason University

The DoD is required by law to conduct "live-fire" tests on developing weapons systems in order to test their vulnerability. Because weapons systems are expensive, few actual live-fire tests are conducted. These are supplemented by simulations. Because weapons are complicated systems, the failure modes of the simulations usually do not correspond exactly to the live-fire tests. We use multidimensional clustering and visualization techniques to authenticate vulnerability measurements.

2D Classification Trees

Hyunjoong Kim
Worcester Polytechnic Institute

Wei-Yin Loh
University of Wisconsin,
Madison

Many classification tree algorithms summarize the data in each terminal node with univariate statistics such as the proportion misclassified. We propose a new classification tree algorithm which yields at each node a 2-dimensional plot of the data with superimposed linear discriminant boundaries. Our algorithm is distinct from "linear combination tree" algorithms that partition the data space with hyper-planes. Intermediate nodes are split using the same techniques as other univariate split algorithms. The difference lies in the model fitted to each terminal node: we fit a statistical model and summarize it using a 2-dimensional plot. The new algorithm thus employs visualization to further enhance our understanding of the data structure. It is shown that the new algorithm has better prediction power than many other classification tree algorithms. Examples using real data will be given for illustration.

FRIDAY, APRIL 7, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSION: Multivariate Modeling: Issues and Applications**CHAIR:** Jerome Reiter, Williams College**Statistical Analysis of Rhizosphere Microbial Communities**

Jayson D. Wilbur, Cindy H. Nakatsu, Sylvie M. Brouder, and R. W. Doerge
Purdue University

Rhizosphere soils of corn are used to determine if distinct microbial communities are associated with different root types, early plant growth stages, and history of soil used for planting. The exudates from the roots can promote microbial growth in the rhizosphere. However, very little is known about the diversity, composition and dynamics of this component of the terrestrial ecosystem. Corn plants were grown in disturbed and undisturbed soils with a 24 year history of growth as a monoculture crop or two crops grown in annual rotation. Both greenhouse and field experiments are presented. Characteristic profiles of the microbial communities were obtained by denaturing gradient gel electrophoresis (DGGE) of PCR amplified 16S rDNA from soil extracted DNA. The dominant rhizosphere bacterial populations, differed during plant growth and with soil treatment. Using various clustering algorithms, the microbial community DGGE fingerprints grouped according to agronomic treatment and within each agronomic treatment according to plant growth stage. Analysis of the community grown in the greenhouse showed less differentiation between growth stages and agronomic treatment but a distinct difference from the field community. The analysis of these data identified possible factors influencing the microbial ecology of the rhizosphere and aided in preliminary statistical modelling.

A Study of Faculty Equity Salary Using Derived Data

Trong Wu
Southern Illinois University, Edwardsville

Faculty's salary of a public institution is often depending upon the state revenue and governor budget. The faculty can be underpaid after several years of lower revenue and budget. Therefore, each university or college needs to study its faculty's salary intermittently to determine whether its faculty is underpaid or be paid sufficiently. This paper reports a study of faculty equity salary plan at a public institution in the state of Illinois. The plan selects a number of comparable institutions to be its salary peer institution then process the derived data of the salary information from these peer institutions. Together with the salary information for each discipline from a nationwide survey by the Oklahoma State University to be the target salary for the faculty of each discipline for the following year.

Use of Latent Variable Models in Air Quality Monitoring

William F. Christensen and Stephan R. Sain
Southern Methodist University

Latent variable analysis is a statistical approach for modeling the underlying structure in multivariate data in terms of a smaller number of latent variables or factors. In the environmental sciences, factor analytic techniques have been used to assess the number of pollution sources affecting the air quality at a monitoring site. Because air quality data often exhibit temporal and/or spatial dependence, we consider the importance for accounting for such correlation in estimating model parameters and making statistical inferences. Potential approaches for accounting for dependencies in the data are discussed, and the use of the block bootstrap as a tool for constructing appropriate inferential procedures is evaluated. An example using a set of air quality measurements is presented.

CCA: Canonical Correlation or Correspondence Analysis? Which is Better for Analysis and Interpretation of Multivariate Data?

A. Dale Magoun
University of Louisiana at Monroe

Linda Peyman
U.S.A.E./WES

Various researchers throughout the United States have recently studied patterns of bird distributions within riparian habitats. These studies focused on avian habitat use as observed by field biologists and physical data as measured from field sampling designs based on 0.25 ha plots. Multivariate studies, such as these, are extremely difficult to interpret due to the underlying interdependencies of the avian and the physical structures describing these habitats. Many researchers have used canonical correspondence analysis, factor analysis, and multivariate regression techniques as ways to explain the interdependencies found in data sets such as these. This paper presents some recent findings that relate satellite imagery data with avian habitat usage data and focuses on the use of multivariate techniques such as canonical correlation analysis and canonical correspondence analysis as tools for analysis and interpretation of such data. The paper further shows how each of these analysis techniques can be used to best interpret multivariate, environmental data.

Penalized Score Equations and Penalized GEE

Wenjiang J. Fu
Michigan State University

In a longitudinal study of environmental pollution, the levels of various pollutants are usually highly correlated. If these levels are predictors of a regression model for certain environmental concerns, the parameter estimator will have a large variance due to the collinearity. A penalty approach to deal with the collinearity is considered in the GEE model. The difficulty of this penalty approach due to the lack of joint likelihood in GEE models is overcome by a new approach to the penalty: the penalized score equations. A GEE model with the Lasso penalty will be demonstrated through an environmental study of air pollutants on asthma patients.

FRIDAY, APRIL 7, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSIONS

Assessing Deformation in Glaciers

S. Huzurbazar
University of Wyoming

A common method for obtaining data in glaciology is via borehole inclinometry. The data are then processed to study various aspects of the mechanics of glaciers, including construction of a three-dimensional deformation field. We consider a spatial data set collected over four time periods from the Worthington Glacier in Alaska. Problems with the data include measurement errors, censored observations as well as small sample sizes. As a first step, we construct confidence regions for the borehole trajectories and also discuss methods for dealing with the censored data and the modelling of the deformation field.

Some Statistical Measures on the National Distribution Center and Dealer Demands Along the Supply Chain

Nick T. Thomopoulos
Illinois Institute of Technology

Wayne E. Bancroft
Motorola Corporation

Nick Z. Malham
Forecasting & Inventory Consultants, Inc.

The monthly demands in three levels of the supply chain are measured using the coefficient of variation. The supply chain here includes the national, the distribution centers and the dealers. Two tables are presented. One table compares the national demands with distribution center demands, and another compares the national demands with the dealer demands.

FRIDAY, APRIL 7, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSION: Innovations in Model Diagnostics and Fitting Algorithms**CHAIR:** David van Dyk, Harvard University**Multiple Outlier Detection**

David W. Scott
Rice University

The detection of more than one or two simultaneous outliers in regression and density analysis remains a challenging practical problem. A priori knowledge of the fraction of bad data would greatly facilitate a solution. In this paper, we describe an iterative algorithm that attempts to simultaneously estimate the fraction of outliers, the scale of the good data, as well as robust parameter estimates.

Parameter Selection for Constrained Solutions to Ill-Posed Problems

Bert W. Rust
National Institute of Standards and Technology

Many physical measurements are modelled by linear integral equations expressing each measurement as the sum of an instrumental smearing of the desired function and a random measuring error. Discretizing the integrals gives an ill-conditioned linear regression model with a matrix whose columns are discrete response functions of the instrument. Linear least squares solutions give wildly oscillating, physically impossible estimates of the function being measured. Such estimates are often stabilized either by truncating the singular value decomposition of the response matrix or by introducing a regularization constraint on the solution vector. In the former case it is necessary to choose a "numerical rank" for the matrix, and in the latter case to choose the value of the Lagrange multiplier in the constrained minimization. This paper suggests methods for using the statistical properties of the measuring errors and the residuals in making those choices.

Optimal Algorithms for Unimodal Regression

Quentin F. Stout and Janis Hardwick
University of Michigan

We present algorithms for determining the best real-valued unimodal regression for a set of weighted univariate observations. This is a form of shape-constrained regression that is of use in several applications. If the n observations are given in order of the independent variable, and L_2 regression is desired, then our algorithm requires $\mathcal{O}(n)$ time. All previously published algorithms for this problem require quadratic time or worse. Our algorithm for L_∞ regression also requires $\mathcal{O}(n)$ time, while for L_1 regression $\mathcal{O}(n \log n)$ time is required. All published algorithms for this problem utilize multiple isotonic regressions. Our contribution is to organize these as a unified “scan” or “parallel prefix” calculation, eliminating redundancies inherent in earlier approaches.

Using the Response Variable in Principal Components Regression

Roy E. Welsch
MIT

A recent study by Frank and Friedman (1994) indicated that ridge regression (RR) performed well when compared to partial least-squares regression (PLS) and traditional principal components regression (PCR). Both RR and PLS make use of the regression response variable in determining which latent variables or principal components to use while PCR does not. This study did not examine modified principal components regression procedures (MPCR) which take explicit account of the response variable. Rawlings (1988), Krzanowski (1992), and Sharma and Welsch (1997) present a number of MPCR procedures. In this paper, we discuss these ideas and some new ones and compare the results to those obtained by Frank and Friedman.

Frank, I. E. and Friedman, J. H. (1994). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35:109-148.

Krzanowski, W. J. (1992). Ranking Principal Components to Reflect Group Structure. *Journal of Chemometrics*, 6:97-102.

Rawlings, J. O. (1988). *Applied Regression Analysis*. Wadsworth, Belmont, CA.

Sharma, V. and Welsch, R. E. (1997). Research on Variation Reduction Using Massive Data Streams. *Bulletin of the International Statistical Institute*, 51st Session, Istanbul 289-290.

Case Studies of Normal Diagnostics in Regression Using Recovered Errors

Donald E. Ramirez
University of Virginia

Donald R. Jensen
Virginia Polytechnic Institute

Diagnostics for normal errors in regression currently utilize ordinary residuals, despite the failure of assumptions validating their use. Case studies here show that such misuse may be critical even in samples of size exceeding currently accepted guidelines. A remedy is to employ recovered errors having the required properties.

Tree-Based Models for Fitting Stratified Linear Regression Models

William Shannon
Washington University School of Medicine, St. Louis

Maciej Faifer and Cezary Janikow
University of Missouri, St. Louis

This generalizes the methods developed in Shannon, Province and Rao (2000) to use recursive partitioning to identify subsets within which simple linear regression models are fit. This method is proposed as an alternative to multivariate regression modelling when the analyst is primarily concerned with the regression of an outcome onto a single predictor and needs to control for other covariates. Splitting rules and pruning methods are programmed in C and linked to 'RPART' to allow a full implementation of this method. Examples using data from the biomedical literature are presented.

FRIDAY, APRIL 7, 2000

1:30–3:15 P.M.

CONTRIBUTED SESSION: Time Series and Proportional Hazards**CHAIR:** Jennifer I. Pittman, Pennsylvania State University**Inference About the Change-Points in a Sequence of Random Vectors**A. K. Gupta
Bowling Green State UniversityJ. Chen
University of Missouri, Kansas City

In this talk, first I will review the change-point problems. Then the testing and estimation of multiple covariance change-points for a sequence of m -dimensional ($m > 1$) gaussian random vectors by using Schwarz information criterion (SIC) studied. The unbiased SIC is also obtained. The asymptotic null distribution of the test statistics is also derived. The result is applied to the weekly prices of two stocks ($m=2$), Exxon and General Dynamics from 1990 to 1991, and changes are successfully detected.

Detecting Change in Variance for Unequally Spaced Time SeriesTze-San Lee and N. Hou
Western Illinois University

To detect a change in the variance of unequally spaced time series, unobservable errors are modeled by the Ornstein-Uhlenbeck process. By applying the principle of likelihood ratio, a statistical test is proposed for detecting the change-point. The asbestos exposure data collected at Lackland Air Force Base from a team of five workers is used to illustrate the proposed test.

Dynamic Modelling of Spectral Density Power Rhythms in All-Night Electroencephalograph (EEG) RecordingsMatthew R. Marler, J. Christian Gillin, Arlene Schlosser, and Hanspeter Landolt
University of California, San Diego

When the digitized recordings of short EEG epochs (e.g. 4 seconds) are analysed by Fast Fourier transforms, the graphs of the spectral power in some frequency bands show a characteristic shape at night in a subset of people: peaks in power occur right after sleep onset and at approximately 90 minutes thereafter the heights of the peaks decline throughout the night the nadirs occur during episodes of rapid-eye-movement sleep (REM).

We investigate models in which: the REM sleep is the nonlinear output of a dynamic oscillator (Duffing, van der Pol) the power in some spectral bands (but not others) is a non-linear function of a compartment that accumulates an abstract "sleep propensity" a nonlinear output of the REM oscillator stifles discharge from the compartment that accumulates sleep propensity specific experimental manipulations (the "tryptophan-free" cocktail) affect some parameters of the model but not others. We display some data that are well described by the model, as well as some that are not well modelled. We assess the degree to which sets of parameter estimates differ between diagnosed groups (normal men versus age-matched depressed men.)

Importance Bootstrap Resampling for Proportional Hazards RegressionKim-Anh Do
MD Anderson Cancer CenterBradley M. Broom
Rice UniversityXuemei Wang
MD Anderson Cancer Center

The use of importance resampling methods to reduce the amount of resampling necessary for the construction of nonparametric bootstrap confidence intervals in the context of survival data with censored observations is investigated. Simulation results showed that relative mean-squared-error (MSE) efficiency gains, when compared to uniform resampling, increased significantly with sample size, was mildly associated with amount of censoring, but decreased slightly as the number of bootstrap resamples increased. The extra CPU time requirement for calculating importance resamples was negligible when compared to the large factor of MSE efficiency gains. The method is applied to a real data set of chronic lymphocytic leukemia, and an SPLUS program is presented for general use. Importance resampling should be used whenever bootstrap methodology is implemented in a survival framework.

The Introduction of Local Spread as a Measure of Non-Stationarity

Robert A. Hedges and Bruce W. Suter
Air Force Research Laboratory

Establishing measures for local stationarity is an open problem in the field of time-frequency analysis. One promising theoretical measure, known as the spread, provides a means for quantizing potential correlation between signal elements. However, it has been noted that the spread is not robust the finiteness of the measure is dependent on the smoothness of the signal covariance.

In this paper we undertake the issue of implementing such a measure for discrete signals and investigate its robustness. A more robust measure, the local spread, is introduced theoretically and implemented. Issues arising from the finite and discrete nature of the data are discussed. The technique is then applied to several examples and the robustness of the method is investigated.

Multivariate Time Series Analysis in Principal Component Space

Joseph N. Ladalla
University of Illinois, Springfield

Box-Jenkins (1976) have provided perhaps the most practically convenient methodology to analyze univariate time series data. Multivariate Time Series analysis can also be made as simple. This is made possible by transforming the given data to its principal components. These principal components are independent under assumption of normality, though they may be cross-correlated. We fit univariate ARIMA models to each principal component, obtain forecasts and combine the individual models into a multivariate model for the vector of principal components. This model may be converted into a multivariate ARIMA model for the original time series. The residual series of the original time series is subjected to all necessary diagnostic checks including the cross correlation function. Strong theory is developed, interesting examples are presented.

Sequential Testing of Proportional Hazards Models

Victor D. Zurkowski
University of Toronto

We consider the distributions of failure times in the context of processes in which instantaneous failure hazards depend on "risk status" covariate processes. In order to assess consistency of observed data with a proportional hazards model (Cox model), we look at the log-likelihood ratio process comparing the Cox model to a non-parametric hazard model. We state properties of the log-likelihood ratio process, and obtain its large sample distribution by means of Martingale analysis. We combine various residuals to produce a test that converges to a power one test as the sample size increases. We present examples of implementation of the test on real data.

FRIDAY, APRIL 7, 2000

4:00–5:45 P.M.

INVITED SESSION: The Utility of Bayesian Decision Analysis for Environmental Problems**ORGANIZER/CHAIR:** Paul Black, Neptune and Co., Inc.

**Scenario and Parametric Uncertainty in GESAMAC:
A Methodological Study in Nuclear Waste Disposal Risk Assessment**

David Draper
University of Bath in England

In this talk I will examine a Bayesian conceptual and computational framework for accounting for all sources of uncertainty in complex prediction problems, involving six ingredients: past data, future observables, and scenario, structural, parametric, and predictive uncertainty. I will apply this framework to nuclear waste disposal using a computer simulation environment—GTM-CHEM—which "deterministically" models the one-dimensional migration of radionuclides through the geosphere up to the biosphere. Focusing on scenario and parametric uncertainty, I will show that mean predicted maximum dose for humans on the earth's surface due to key radionuclides, and uncertainty bands around those predictions, are noticeably larger when scenario uncertainty is properly assessed and propagated. I will conclude by describing how Bayesian decision theory can take predictions such as these and turn them into recommendations for environmental action.

A Probability Network for Water Quality Modeling and Decision Support

Kenneth H. Reckhow and Mark E. Borsuk
Duke University

A probability network model is being developed and applied to the problem of eutrophication in the Neuse Estuary, USA. Also called a "Bayes net," this model consists of the variables of interest in the system and a set of assertions concerning the probabilistic relationships among the variables. The objective of the model is to provide a scientific assessment of the impact of nitrogen loading on estuarine algal blooms and fishkills model expressions are quantified using data analysis, mechanistic relationships, and/or expert judgment. Probabilistic predictions of model endpoints may then be made based on the entire set of conditional probabilities. Not only does this network structure provide an integrated approach to uncertainty analysis, but it also allows easy updating of prediction and inference when observations of model variables are made. This capability is particularly important when applied to a natural system in which additional monitoring is likely to occur concurrent with the modeling effort. The method is probabilistic in its approach, which facilitates a meaningful communication of uncertainty, and is consistent with the risk assessment paradigm. Model endpoints are chosen so that they are of vital interest to stakeholders and can be easily expressed for use in formal decision analysis.

Bayesian Assessment of Uncertainty and Variability in Deterministic Environmental Exposure Models

Samantha Bates and Adrian Raftery
University of Washington

In this paper we discuss Bayesian methods of analysis, which incorporate both prior knowledge of the distributions of the inputs to a deterministic model and any available data on the model inputs and outputs. These methods yield posterior distributions for the model output from which to find distributions for quantities of interest. The first method uses Monte Carlo simulation from the prior distributions for the inputs and resampling of these simulations with weights determined by the observed data under the sample importance-resampling scheme of Rubin. The second involves sampling from the posterior using MCMC methods.

We will present an application of the methods to modeling poly-chlorinated biphenyl (PCB) concentrations in various media at a Superfund site in New Bedford Harbor, MA. A deterministic model for PCB concentration in soil was developed by Cullen (1992). Expert opinion is reflected in the prior distributions for model inputs. Interest lies in developing a distribution for the PCB concentration in soil at this site, which accounts for uncertainty and variability in the model inputs and can be used in policy decisions.

**Environmental Modeling and Bayesian Analysis for
Assessing Human Health Impacts from Radioactive Contamination**

Tom Stockton and Paul Black
Neptune and Company, Inc.

EPA regulations and DOE orders require assessing the impact on human health of radioactive waste contamination over periods of up to ten thousand years. Towards this end complex environmental simulation models are used to assess “risk” to human health from migration of radioactive contamination. Typically there is very little data underlying these models and the data that is available is incorporated in input parameter distributions for Monte Carlo simulation. Expert judgment typically drives the level of model complexity chosen but model complexity choices are often not made within the context of the decision to be made. The utility and regulatory acceptability of a Bayesian approach to decision making regarding radioactive contamination is discussed within the context of radioactive contamination examples from Los Alamos National Laboratory, Hanford, and the Nevada Test Site. These examples highlight the desirability and difficulties of merging the cost of monitoring, the cost of the decision analysis, the cost and viability of clean up, and the probability of human health impacts within a rigorous decision framework.

FRIDAY, APRIL 7, 2000

4:00–5:45 P.M.

INVITED SESSION: IASC Sponsored Session: Applications to Earth Systems

ORGANIZER/CHAIR: Edward J. Wegman, George Mason University

Using Smoothing to Reconstruct the Holocene Temperature in Lapland

Lasse Holmström
Rolf Nevanlinna Institute

Small arctic and subarctic lakes are known to be sensitive to climatic variation. Changes in external conditions are continuously recorded in their sediments in the form of aquatic organisms. The abundance of such organisms can therefore be used to reconstruct past environmental conditions. We use data collected from the Finnish Lapland to reconstruct post ice-age temperatures. Nonparametric smoothing has not been used often in this context. We find smoothing a viable tool both in the actual reconstruction phase and in the subsequent time series smoothing, where the SiZer method is used.

A Computational Geometry Approach for Peeling and Outlier Detection

Giancarlo Ragozini
Universita di Napoli Federico II

From a geometrical point of view, outliers are those observations lying isolated on the periphery of data cloud. A large literature exists on the detection of multiple outliers in multivariate data sets. Most of recent proposals are based on some robust distance of each data point from a center. However, they are really effective only when the data scatter has a regular shape. The proposed method is based on the direct exploration of the data periphery, without considering any center or fixed shape, exploiting the geometrical properties of the sample convex hull. The first step of the proposed detection procedure consists of a new “weak” convex hull peeling, reducing the computational effort of the classical peeling procedures. In this step the set of candidate outliers is identified, evaluating gaps in the data scatter and proximities to its boundary region. In the second step, a block omission approach is performed, considering only some specific subsets among the candidate outliers, in order to reduce the combinatorial computational cost. The outlyingness of each subset is measured through a new index based on the variation of the convex hull volume when a subset is omitted.

FRIDAY, APRIL 7, 2000

4:00–5:45 P.M.

INVITED SESSIONS

Applications of Deepest Regression

Mia Hubert, Peter J. Rousseeuw, and Stefan Van Aelst
Universitaire Instelling Antwerpen

The deepest regression is a method for linear regression introduced by (Rousseeuw and Hubert 1999). It is the fit with maximal regression depth. We prove that this estimator is highly robust against outliers. We propose an approximate algorithm for fast computation of the deepest regression in higher dimensions, and apply it to several real data sets. From the distribution of the regression depth function we construct tests for the true unknown parameters in the linear regression model. We also propose a bootstrap method to construct regression confidence regions. For bivariate datasets we use the maximal depth to construct a test for linearity versus convexity/concavity. Finally, the deepest regression is applied to polynomial regression and to the Michaelis-Menten model.

SATURDAY, APRIL 8, 2000

8:00–9:45 A.M.

CONTRIBUTED SESSION: Uncertainty Quantification in Complex Models**CHAIR:** Todd L. Graves, Los Alamos National Laboratory**Quantifying the Effects of Noise on Biogeochemical Models**Barbara Bailey
University of Illinois, Urbana-ChampaignScott Doney
NCAR

The need to understand the effects of anthropogenic perturbations on the ocean carbon cycle has sparked a new interest in biogeochemical models in recent years. These models are now being coupled with physical ocean circulation models. For this coupling to generate realistic fluxes of nutrients and carbon, the biogeochemical equations need to exhibit realistic dynamics.

We propose an experimental design approach to quantifying the effects of different types of noise on biogeochemical system dynamics. The biogeochemical model we use in our investigation is a model of plankton dynamics and nitrogen cycling. It is a compartmental model (NPZ) consisting of a compartment for nitrogen (N), phytoplankton (P), and zooplankton (Z). The flows or intercompartmental exchanges are modeled as a nonlinear system of three first order differential equations.

The types of noise of interest are both independent and correlated over time. We propose to study the effects of noise on the state variables of the system and parameters. Because the noise is an integral part of the system's dynamics, a nonlinear time series approach is used to quantify the dynamics and predictability of the system. This involves fitting nonlinear models and estimating dynamical systems quantities of interest such as global and local Lyapunov exponents, along with measures of uncertainty for these estimates.

The Selection of the Optimal Structure of the Earth's Model for Forecasting the Main Physical ParametersAlexander Dmitrievich Gorobets
Sevastopol State Technical University

The Earth's model to be considered is a system of m nonlinear simultaneous equations $F(Y, X, A_i) = U$, where F is the true but unknown vector of models, Y is a m -vector of dependent (endogenous) variables, X is a k -vector of independent or controlled input (exogenous) variables, A_i is a vector of unknown parameters in i -th structural equation ($i=1..m$), and U is a vector of independent random variables with zero mean and variance-covariance matrix S . There is usually some prior information about the regions of possible values for variables: $Y \in W_1$ and $X \in W_2$, where W_1 and W_2 are sets of possible values of the vectors Y and X .

The problem is selection the optimal structure of simultaneous equation system which minimize the expected error of prediction of endogenous variables in some region of interest for making predictions. The loss function for each of the system of models depends on the prediction error of the vector of endogenous variables Y . The objective of research is to present the method of selection of the optimal structure of the functions $F(Y, X, A_i)$ and to investigate two criteria that assures one quality in the selection strategy, such as the average of the mean square error of prediction.

Cost-Effective Uncertainty Analysis

Daniela Stoevska-Kojouharov
Monmouth University

Uncertainty analysis is a tool for determining the relative influence of different inputs to a simulation model on the output(s) of interest. In many simulation models, at least some of the inputs are unknown parameters whose values must be estimated, thus inducing unwanted variation in the simulation output(s). A researcher who wishes to improve the simulation model by reducing the variation is not well served by current methods for uncertainty analysis, which rank the inputs based on their relative influence on the output(s), but do not account for any costs associated with improving the estimates. In our work we introduce an algorithm for obtaining the resource allocation that results, for a given level of expenses, in the greatest improvement of a simulation model. The method is called cost-effective uncertainty analysis. Cost-effective uncertainty analysis extends the concept of current methods for uncertainty analysis to include the fiscal cost of improving model precision. The method uses regression meta-modeling to study and make conclusions about the complicated simulation model. The information about the influence of the inputs is combined with information about the cost of obtaining additional observations on each input (as provided by the researcher) and the total amount of money available for improving the simulation model. An algorithm for best money-allocation is developed.

Statistical Quantification of Prediction Error Associated with Computational Predictions

Robert G. Easterling and Marcey Abate
Sandia National Laboratories

Confidence in computational predictions is enhanced if the potential 'error' in these predictions (the difference between the prediction and nature's outcome in the situation being modeled) can be credibly bounded. Determining such error-limits is a problem that has been solved for relatively simple mathematical models, not the complex, multi-physics codes used to predict, say, system responses in various environments. We develop a conceptual framework for solving the prediction-error-quantification problem, discuss several issues involved, and illustrate proposed methods through a damped spring-mass example and a contact-resistance experimental program. In general, this framework requires designing and conducting a suite of physical experiments and calculations (both ranging from phenomenological to integral levels), then analyzing the results to provide a basis for inferring the uncertainty of a model-prediction of system performance in a particular application, which may be in an environment or configuration that cannot be tested. Problems discussed include: the design and analysis of physical experiments for the purpose of quantifying uncertainty in computational predictions analysis methods for estimating prediction error at untested points in the parameter space merging prediction uncertainty results at single phenomenon or component levels to obtain system-level prediction uncertainty dimension reduction in order to make the problem tractable.

The Role of Statistical Methods in Atmospheric Model Intercomparison Projects

Christiane Jablonowski
University of Michigan

Since the early nineties, atmospheric model intercomparison projects have been initiated that reveal interesting agreements and disagreements among weather prediction and climate models. Especially the role of the dynamics (the so-called dynamical core of a general circulation model) has recently been discussed. This talk gives an overview of current results and presents ideas how to compare dynamical core experiments.

Special attention is given to easy-to-use statistical methods that help understand and explain model phenomena such as the influence of a varying resolution or diffusion parameter on the model's climate. The statistical framework provides insight into the significance of model variations. In particular, methods like the univariate Student-t and Fisher-F hypothesis test as well as a recurrence and empirical orthogonal function analysis (EOF) are shown. In addition, frequency and wavenumber analyses reveal characteristic model features.

The talk addresses the importance and use of these statistical analysis techniques using dynamical core examples of three different general circulation models. The models involved are two weather prediction models of the German Weather Service and the forecast model IFS of the European Center for Medium-Range Weather Forecasts. All models are different in design and numerics and therefore provide a suitable test bed.

A Test of Symmetry about a Known Median Based on a Runs Statistic

David J. Cummins
Eli Lilly & Company

Doug Nychka
National Center for Atmospheric Research

Curve estimates with no measure of their accuracy are not very useful. We address the problem of obtaining valid pointwise and simultaneous confidence bands for nonparametric curve estimates. Although recent work on confidence intervals show promising results, those intervals have coverage probability at the nominal level only when averaged across the design points and not uniformly at all design points. We present methods for obtaining bands which have more uniform coverage and at the same time are thinner than what are produced by established methods.

We also present a new approach which avoids the usual practice of inflating pointwise intervals. We proceed by estimating a confidence set that contains the true function with probability $1-\alpha$ using an intersection of two balls in a Sobolev space. The upper and lower bands for the function are the boundary elements of this confidence set. Finding these boundary elements is an optimization problem which is solved using Nonlinear Programming. We provide an accurate approximation to this computer-intensive method that reduces the computation to be linear in the number of observations. This confidence set approach gains more uniform coverage, and unlike other methods, gives asymmetric bands that are not of a fixed width, adapting to the smoothness and shape of the estimated curve.

SATURDAY, APRIL 8, 2000

8:00–9:45 A.M.

CONTRIBUTED SESSION: Statistical Tests, Estimation and Stability

CHAIR: Imola K. Fodor, Lawrence Livermore National Laboratory

An Adjusted, Asymmetric Two Sample t-Test

Sandy D. Balkin
Ernst & Young LLP

Colin Mallows
AT&T Labs–Research

The Telecommunications Act of 1996 requires that Incumbent Local Exchange Carriers (ILECs) must provide, for a fair price, interconnection services to the customers of a Competitive Local Exchange Carrier (CLEC), these service being "...at least equal in quality to [those] provided by the local exchange carrier to itself...". To monitor the ILEC's performance, we need formal statistical tests of compliance. Inspection of data on several performance measures reveals severe positive skewness, violating the assumptions of the standard t-test. Also, since we want to detect not only shifts in mean but also increases in variance, use of the modified t-statistic of Brownie et al (Biometrics 1990) is indicated. Permutation testing would be preferable, but is unwieldy. In response to this need, we present adjustments (following the method of Johnson (JASA 1978)) to the standard and modified two-sample t-tests. We compare the resulting tests with permutation tests.

The Wilson-Hilferty Transform is Locally Saddlepoint

George R. Terrell
Virginia Polytechnic Institute

In 1931 Wilson and Hilferty discovered a quick, rough method for obtaining p-values for chi-squared statistics. Its usefulness declined with the advent of computers. Recently there has been interest in "saddlepoint" methods for approximate probability calculations. These are fairly general, and can therefore often be adapted to the ever more complicated test statistics that modern statisticians use. However, they do not as readily provide confidence intervals and simulated values as does a Wilson-Hilferty transform.

We will propose a generalized Wilson-Hilferty transform, and establish that it is locally a saddlepoint method. The method therefore combines traditional and modern virtues, and shows promise for difficult inference problems.

On Numerical Stability of MGF and CF

Jinhyo Kim
Seoul National University

Regarding the numerical computing of the moment generating function and the characteristic function, a series of publications appeared in the 1990's. (cf. McCullagh(1994), Waller(1995), Luceño(1997)). The system matrix in a discretized linear system, generated by the MGF, exhibits a serious backward error due to ill-conditioning. The inherently perturbed error in the MGF makes it hard to implement on a digital computer whereas the CF is not. Those phenomena arise because the real Vandermonde matrix associated with the MGF is extremely susceptible to numerical error whereas the complex Vandermonde matrix associated with the CF is not. This article explains those phenomena using algorithm stability, specifically backward stability. We discuss originality and properties of the inherently perturbed error in the MGF and the CF. More general arguments are given to show that the CF is superior to the MGF in terms of numerically stable behavior.

A Test of Symmetry about a Known Median Based on a Runs Statistic

Alex Leonardo Rojas Peña
University of Puerto Rico

Jimmy A. Corzo Salamanca
National University of Colombia

This paper presents a test, based on a runs statistic, for symmetry of a continuous distribution about a known median. A Monte Carlo's study, for twelve distributions from the Generalized lambda family (FDLG) (Ramberg and Schmeiser, 1974) which provides a wide range of asymmetric distributions, shows that the test is more powerful than tests proposed by McWilliams (1990) and Castillo (1993) when the distribution of which the data come is asymmetric and it possesses both tails, while it is less powerful than tests proposed by McWilliams (1990) when it possesses only one tail.

Mining Evolutionary Data for Multidimensional Scaling of Gene Measurements

Rida Moustafa and Edward J. Wegman
George Mason University

The performance of evolutionary algorithms are important issue of the optimization field. In this paper we use a mining tool to control parameter of the evolutionary algorithms applied to multidimensional scaling of gene measurements problem. Experimental results and comparison are used to demonstrate the feasibility of the performance in optimization process.

Efficient Nonparametric Estimation of a Distribution Function

Reza Modarres
George Washington University

We consider the efficient estimation of a distribution function F under several models and offer a unifying approach based on the nonparametric likelihood principle.

Under the symmetry model, we show that the nonparametric MLE of F coincides with the Schuster estimator. Under the auxiliary--sample model, we discuss an estimator based on the total law of probability. We show that this estimator coincides with the nonparametric MLE of F . Under the intersection of the two models, we present an efficient hybrid estimator. We show that the hybrid estimator is asymptotically normal and converges to the nonparametric MLE of F under the assumption of conditional symmetry. A Monte Carlo simulation assesses the efficiency of the proposed estimators.

Large One-Sample and Two-Sample Tests for Average Hazard Rates

John J. Hsieh
University of Toronto

Both one-sample and two-sample statistical inference on average hazard rates are frequently used in observational studies. One-sample test are usually employed to test departure of an observed average hazard rate in a study population from that of a general population or of some reference population, in order to detect excess hazard in the study population. Two-sample test are employed to compare hazard rates of two study groups for measuring effect and association and for comparing event intensities between populations. In observational studies, data on event counts and person-time of exposure are available, in terms of which test statistics are to be formed. The aim of this article is to develop asymptotic one-sample and two-sample test statistics, using likelihood methods and counting processes martingale techniques applied to the square root and logarithmic transformations of hazard rates. Various types of Wald's statistics, score statistics and likelihood ratio statistics are derived. These statistical inference procedures yield good statistical properties (such as consistency, asymptotic unbiasedness, efficiency and asymptotic normality). In addition, square root and logarithmic transformations of hazard rates improve normality approximations. We obtain, and compare the accuracy of, eight two-sample test statistics classified in four groups and five two-sample test statistics classified in three groups.

SATURDAY, APRIL 8, 2000

10:30 A.M.–12:15 P.M.

INVITED SESSIONS

SATURDAY, APRIL 8, 2000 10:30 A.M.–12:15 P.M.

INVITED SESSION: Statistics and Information Technology

ORGANIZER/CHAIR: Alan Karr, National Institute of Statistical Sciences

How Should We Publish Data Analyses in the Web Age?

Todd L. Graves

Los Alamos National Laboratory

Papers and journals about data analyses that are published online should be different from and more powerful than those in paper journals. Web papers can include interactive visualization and data analysis applets to allow the readers to perform exploratory analyses. Readers could also replay the authors' analyses, trying out minor modifications, or even importing their own software or data to see if different analyses or additional data change the authors' conclusions. Online journals could be reorganized so that all analyses of a particular data set or that bear on a particular real world problem could be reachable through the same web page. Readers could perform and submit their own analyses of these problems all within the web journal environment. This talk will include demonstrations of these concepts written in Java.

Geographic Aggregation Procedures for Data Disclosure Limitation

Ashish Sanil

National Institute of Statistical Sciences

Government agencies often report their data (gathered through sample surveys and censuses) in the form of statistical summaries by geographic units (e.g., by state, county, etc.). In many cases the public release of data on particular geographic units is considered too risky for preserving the confidentiality of the respondents. A possible strategy for such cases is to aggregate neighboring regions into larger units which satisfy the confidentiality requirements. Often, as in the case of an on-line query system, the computation of the aggregations needs to be automated, should be computationally efficient, and should produce meaningful aggregates. Procedures for carrying out such confidentiality-preserving geographic aggregation will be described, and illustrative examples will be presented.

Detecting Defection: Mining Massive Online Data to Model ISP Customer Churn

Nandini Raghavan

AT&T Labs–Research

From a statistician's point-of-view, the most striking (and daunting) aspect of the web phenomenon is extracting meaningful information from the tremendous volume of data available. In this talk I describe our efforts to develop evolving statistical profiles (signatures) of users of an internet service provider (ISP). ISPs are characterized as having large, fluid user populations and massive, dynamic data streams which record information at different granularities. I describe an application where we use these signatures to build formal statistical models of customer migration.

SATURDAY, APRIL 8, 2000 10:30 A.M.–12:15 P.M.

INVITED SESSION: Statistical and Computational Methods for Survival and Reliability Data

ORGANIZER/CHAIR: Luis A. Escobar, Louisiana State University

A Case Study in Competing Risk Reliability Analysis Using JMP Software

Bradley Jones
SAS Institute, Inc.

This presentation illustrates the use of JMP software to analyze a manufacturing oriented reliability problem with competing risks. The application is non-trivial and the data are real.

The demonstration shows the ability of fast modern computers to fit, diagnose, and refine such complex models interactively.

Reliability Data Analysis Using S-Plus

William Q. Meeker
Iowa State University

Censored and truncated data arises frequently in product reliability studies involving laboratory accelerated life tests, field tracking studies, and the analysis of warranty data. S-PLUS has powerful tools for analyzing such data. This talk will describe SLIDA, a collection of S-PLUS functions for Life Data Analysis that has been designed to extend and enhance the S-PLUS capabilities in this area. Some of these extensions include:

- Functions that link maximum likelihood estimation with probability plots to facilitate analysis steps ranging from model identification and diagnostics to sensitivity analysis and presentation of final results.
- Simulation tools for inference and test planning.
- Functions that allow the user to specify a likelihood function and easily do appropriate likelihood-based analyses for nonstandard models.
- Methods for recurrence data.
- A comprehensive set of example data sets and command scripts illustrating the methods.
- A graphical user interface for the core SLIDA functionality.

Random Effects Survival Models for Familial Data

Terry M. Therneau
Mayo Clinic

We are currently dealing with a large study investigating both genetic and environmental risk factors for breast cancer. The principle study data consists of the descendants and relatives of 426 incident cancer cases identified between 1940 and 1952. Full family pedigrees have been evaluated through 1992; currently there are data on 11848 women of which 5185 are marry-ins. Adjustment for familial genetic effects is an important part of the investigation of other risk factors such as oral contraceptive use, obesity, or smoking. The size of the data set can make this challenging, however. We have relied primarily on our S-plus implementation of random effects Cox and accelerated failure time models. The talk will focus on some of the computational and algorithmic challenges in implementing these, along with the results and benefits/problems of this type of model for such data. Liberal use will be made of examples from the above and selected other data sets.

Index

A

Abate, Marcey, 19, 58
 Adams, John, 13, 40
 Agins, Bruce, 13, 39
 Allen, I. Elaine, 13, 40
 Al-Mutairi, Dhaifalla K., 9, 29
 Altman, Naomi S., 10, 32
 Atkinson, E. Neely, 15, 48

B

Bailey, Barbara, 9, 19, 29, 57
 Baldwin, C., 9, 27
 Balkin, Sandy D., 19, 59
 Bancroft, Wayne E., 15, 50
 Barrett, Chris, 8, 11, 23
 Barry, Moser, 14
 Barton, Richard J., 9, 28
 Baru, Chaitanya, 8, 25
 Bates, Samantha, 9, 17, 27, 54
 Bay, Stephen D., 11, 36
 Beaver, James, 14, 46
 Beckman, Richard, 8, 24
 Berliner, L. Mark, 8, 22
 Bhattacharyya, S., 10, 34
 Bickel, Peter, 8, 24
 Black, Paul, 2, 17, 54, 55
 Bodt, Barry A., 10, 34
 Borsuk, Mark E., 17, 54
 Braverman, Amy, 9, 27
 Broom, Bradley M., 16, 52
 Brouder, Sylvie M., 15, 48
 Buchberger, Rachel, 8, 23
 Buja, Andreas, 2, 11, 36

C

Campbell, Katherine, 8, 24
 Cant?-Paz, E., 9, 27
 Carmody, Sharon E., 10, 34
 Carr, Daniel B., 10, 34
 Carroll, Carolyn A., 9, 29
 Chandra, Charu, 14, 43
 Chen, J., 16, 52
 Chen, Lili, 9, 28
 Chen, R., 10, 34
 Christensen, William F., 15, 49
 Cleveland, William, 6
 College, Craig E., 13, 42
 Cox, Michael, 14, 45
 Crain, William, 13, 41
 Cummins, David J., 19, 59

D

Davidson, Jennifer, 9, 28
 Denby, Lorraine, 2, 6, 11, 37
 Denteneer, Dee, 10, 33
 Dippo, Cathryn, 2, 8, 25
 Do, Kim-Anh, 16, 52
 Dodds, Peter, 14, 44
 Doerge, R. W., 15, 48
 Doney, Scott, 19, 57
 Downer, Robert G., 14, 45, 46
 Doyle, John, 11, 37
 Draper, David, 17, 54
 Dumer, John C., 10, 34

E

Easterling, Robert G., 19, 58
 Efromovich, Sam, 10, 31
 Elliott, Marc N., 13, 40
 Escobar, Luis A., 2, 20, 63
 Evans, David, 14, 45

F

Faifer, Maciej, 16, 51
 Fodor, I. K., 9, 19, 27, 59
 Fricker, Ron, 13, 41
 Fu, Wenjiang J., 15, 49
 Furrer, Reinhard, 9, 28

G

Galway, Lionel, 13, 42
 Gerard, Patrick D., 14, 45
 Gillin, J. Christian, 16, 52
 Goodman, Arnold F., 7
 Gorobets, Alexander Dmitrievich, 19, 57
 Grabis, Janis, 10, 32
 Graves, Todd L., 19, 20, 57, 62
 Gupta, A. K., 16, 52
 Gupta, Amarnath, 8, 25

H

Hancock, Greg, 14, 44
 Hand, David J., 3, 12
 Hanratty, Timothy P., 10, 34
 Hardwick, Janis, 16, 51
 Hays, Ron D., 13, 40
 Hedges, Robert A., 16, 53
 Heiner, Karl, 13, 40
 Hert, Carol, 8, 25
 Hesterberg, Tim C., 14, 44
 Hoar, Timothy, 8, 22
 Holmström, Lasse, 17, 55
 Hou, N., 16, 52
 Hovy, Ed, 8, 25, 26
 Hsieh, John J., 61
 Hubert, Mia, 17, 56
 Huzurbazar, S., 15, 50

J

Jablonowski, Christiane, 19, 58
 Janikow, Cezary, 16, 51
 Jensen, Donald R., 16, 51
 Joftis, Peter, 8, 26
 Johns, Craig, 10, 32
 Jones, Bradley, 20, 63

K

Kamath, C., 9, 27
 Karcher, Peter, 9, 30
 Karr, Alan, 20, 62
 Ke, Chunlei, 9, 29
 Keller-McNulty, Sallie, 2, 7, 8, 23
 Kennedy, William J., 11, 36
 Kibler, Dennis, 11, 36
 Kim, Hyunjoong, 10, 15, 31, 48
 Kim, Jinhyo, 19, 60
 Klavans, Judith, 8, 25
 Krause, Paul F., 15, 47
 Kwon, Jaimyoung, 8, 24

L

Ladalla, Joseph N., 16, 53
 Lambert, Diane, 10, 33
 Lancaster, Vicki, 2, 7, 13, 38
 Landolt, Hanspeter, 16, 52
 Lattyak, W. J., 10, 34
 Lee, Tze-San, 16, 52
 Liu, Chuanhai, 6
 Liu, Lon-Mu, 10, 34
 Loh, Wei-Yin, 15, 48

M

Macchiavelli, Raúl E., 14, 46
 Magoun, A. Dale, 15, 49
 Malham, Nick Z., 15, 50
 Mallows, Colin, 19, 59
 Marchette, David J., 8, 9, 24, 27
 Marciano, Richard, 8, 25
 Marler, Matthew R., 16, 52
 Martinez, Yvonne M., 2, 13, 42
 Mateeva, Alben, 13, 39
 Maxwell, Daniel, 13, 41
 Meeker, William Q., 20, 63
 Merritts, Dorothy, 2, 14, 44
 Milliff, Ralph, 8, 22
 Modarres, Reza, 19, 60
 Moore, Andrew, 11, 35
 Moore, Leslie M., 2, 14, 43
 Morgeson, J. Darrell, 8, 23
 Morris, Max, 14
 Morton, Sally C., 2, 12, 13, 39
 Moser, Barry, 2, 14, 45
 Moustafa, Rida, 19, 60

N

Nakatsu, Cindy H., 15, 48
 Nychka, Doug, 2, 8, 19, 22, 59

O

Offutt, Carolyn K., 15, 47
 Olkin, Ingram, 13, 40

P

Pazzani, Michael J., 11, 36
 Peña, Alex Leonardo Rojas, 19, 60
 Perry, H. Mitchell, 10, 34
 Peyman, Linda, 15, 49
 Pinheiro, Jose C., 10, 33
 Pittman, Jennifer I., 10, 16, 31, 52
 Podlich, Heather, 9, 30
 Powell, Dennis R., 14, 43
 Priebe, Carey E., 9, 27

R

Raftery, Adrian, 17, 54
 Raghavan, Nandini, 20, 62
 Ragozini, Giancarlo, 17, 55
 Ramirez, Donald E., 16, 51
 Ray, Bonnie, 2, 11, 14, 43
 Reckhow, Kenneth H., 17, 54
 Reiter, Jerome, 9, 15, 29, 48
 Rice, John, 8, 24
 Rodriguez, Rocío del P., 14, 46
 Rothman, Daniel, 14, 44
 Rousseeuw, Peter J., 17, 56
 Rust, Bert W., 16, 50

S

Sain, Stephan R., 15, 49
 Salamanca, Jimmy A. Corzo, 19, 60
 Sanil, Ashish, 20, 62
 Santos, Elisa M., 10, 33
 Schlosser, Arlene, 16, 52
 Schonlau, Matthias, 8, 15, 24, 47
 Sclove, S. L., 10, 34
 Scott, David W., 15, 16, 47, 50
 Shannon, William, 2, 11, 16, 35, 51
 Shine, James A., 15, 47
 Shumway, Robert H., 10, 32
 Smyth, Gordon K., 9, 30
 Smyth, Padhraic, 11, 36
 Solka, Jeffrey L., 8, 24
 Spoeri, Randall K., 13, 39
 Spruill, Nancy, 2, 13, 41
 Stanford, Derek, 10, 14, 33, 45
 Stockton, Tom, 17, 55
 Stoevska-Kojouharov, Daniela, 19, 58
 Stout, Quentin F., 16, 51
 Sun, Don X., 10, 33
 Suter, Bruce W., 16, 53
 Swartz, Richard, 13, 40

T

Tang, N., 9, 27
 Tenorio, Luis, 13, 38
 Terrell, George R., 19, 59
 Therneau, Terry M., 20, 63
 Thomopoulos, Nick T., 15, 50
 Tiffée, Bradley, 14, 46
 Tymes, Nathaniel, 10, 31

V

Van Aelst, Stefan, 17, 56
 van Dyk, David, 11, 16, 36, 50
 Villarreal, Alberto, 13, 38
 Villarreal, Julio C., 10, 32

W

Wahba, Grace, 3, 6, 7
 Wallet, Bradley, 8, 24
 Wan, Fei, 9, 28
 Wang, Xuemei, 16, 52
 Wang, Yuedong, 9, 29, 30
 Weaver, Daniel, 11, 35
 Wegman, Edward J., 2, 11, 15, 17, 19, 48, 55, 60
 Welsch, Roy E., 16, 51
 Wikle, Christopher K., 8, 22
 Wilbur, Jayson D., 15, 48
 Willgoose, Garry, 14, 44
 Wojciechowski, William C., 15, 47
 Wright, Kevin, 11, 36
 Wu, Dongfeng, 10, 31
 Wu, Trong, 15, 49

Z

Zaslavsky, Ilya, 8, 25
 Zhang, Hao, 9, 28
 Zurkowski, Victor D., 16, 53