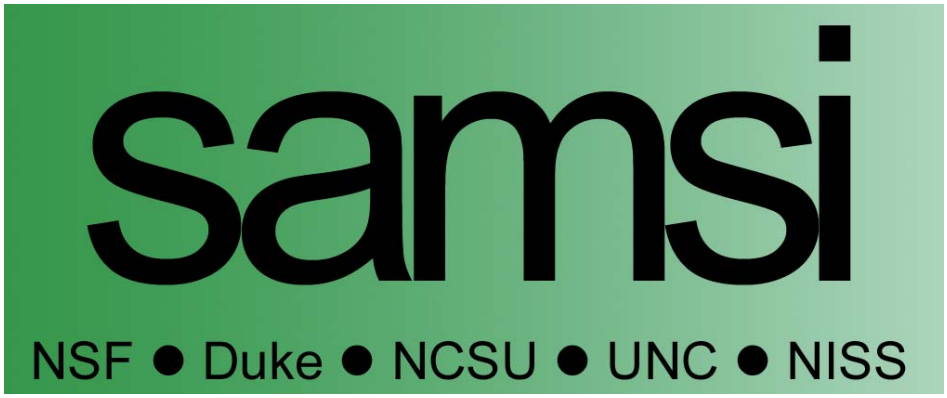


Annual Scientific Report

2004-2005

May 1, 2005

SAMSI Annual Scientific Report for 2004-2005

This report is a version of the SAMSI Annual Report to the National Science Foundation, with sensitive financial data and personal information removed. It covers the period of SAMSI activities from July 1, 2004 – June 30, 2005. Past and future activities of SAMSI are also discussed.

0. Executive Summary

The Executive Summary contains

A. Outline of SAMSI Activities and Initiatives for Year 3 and the Future
B. Financial Summary
C. Directorate's Summary of Challenges and Responses
D. Synopsis of Research, Human Resource Development and Education
E. Evaluation and Third Year Review
F. Evaluation by the SAMSI Governing Board.

A. Outline of Activities and Initiatives

1. Third Year Programs and Activities

Regular Programs

- Genomes to Global Health: the Computational Biology of Infectious Disease (Fall 2004-Spring 2005)
  - Opening Tutorials and Workshop (9/18/04-9/22/04)
  - Mid-Program Focused Workshop (1/31/05-2/1/05)
  - Transition Workshop and Symposium (5/22/05-5/24/05)
- Latent Variable Models in the Social Sciences (Fall 2004-Spring 2005)
  - Tutorials and Opening workshop (9/11/04-9/15/04)
  - Symposium on Causality (3/29/05)
  - Workshop on Latent Variable Models and Survey Data (joint with NPCDS, 5/4/05-5/6/05)
  - Transition Workshop (5/19/05)
  - Closing Workshop (11/11/05-11/12/05)
- Data Assimilation for Geophysical Systems (Spring 2005)
  - Tutorials and Opening Workshop (1/23/05-1/26/05)
  - Issues, Challenges & Interdisciplinary Perspectives (joint at IPAM, 2/22/05-2/26/05)
  - Mini-Workshop on Lagrangian Ocean Data Assimilation (4/12/05)
  - Mini-Workshop on Bridging Statistical Approaches and Sequential Data Assimilation (5/12/05)
  - Summer School - "Fusing Models with Data: From Theory to Practice to Theory" (joint at NCAR, 6/13/05-6/17/05)

Education and Outreach

- Industrial Mathematical and Statistical Modeling Workshop for Graduate Students (7/26/04-8/3/04)
- Two-Day Workshop for Undergraduates (2/18/05-2/19/05)
- PREP Workshop (joint with MAA, 5/25/05-5/28/05)
- Undergraduate Interdisciplinary Workshop (5/30/05-6/3/05)
- Industrial Mathematical and Statistical Modeling Workshop for Graduate Students (7/25/05-8/2/05)
- Mathematical and Experimental Modeling Course on NC-REN TV (Fall, 2004)
- Graduate Courses at SAMSI
    - Computational Immunology and Immunogenomics, Fall 2004
    - An Overview of Latent Variable Models in the Social Sciences, Fall 2004
    - The Biophysics of Cell Signaling, Spring 2005
    - Mathematical Modeling of Infectious Diseases, Spring 2005
    - Data Assimilation Methods for the Ocean and Atmosphere, Spring 2005
    - Computational and Statistical Methods for Inverse Problems, Spring 2005

Distinguished Lecture Series

- Bette Korber, "Diversity considerations in HIV Vaccine Design" (11/30/04)
- Eugenia Kalnay, "Data assimilation and ensemble forecasting: two problems with the same solution?" (1/25/05)
- Alan Perelson, "Modeling viral infections" (1/31/05)
- James Robins, "Optimal sequential decisions and causal inference" (3/29/05)

Planning, Hot Topic, Technology Transfer, and Closing Workshops

- Design and Analysis of Computer Experiments for Complex Systems (joint with NPCDS, 7/13/04-7/17/04)
- Closing Workshop: Multiscale Model Development and Control Design (9/27/04-9/28/04)
- Workshop on Data Mining Methodology and Applications (joint with NPCDS, 11/28/04-11/30/04)
- Planning Meeting on Complex Data Structures (joint with NPCDS, 4/9/05-4/14/05)
- Stochastic Modeling for Financial Mathematics (joint with CRM, 6/1/05-6/5/05)
- Random Graphs and Stochastic Computation (6/13/05-6/14/05)
- Data Mining Technology Transfer Workshop (6/20/05-6/24/05)

2. Fourth Year Program Schedule

Regular Programs

- Financial Mathematics, Statistics and Econometrics (Fall 2005)
    - Opening Tutorials and Workshop (9/18/05-9/21/05)
    - Mid-Program Focused Workshop (TBA)
    - Transition Workshop and Symposium (2/26/06-2/28/06)
- National Defense and Homeland Security (Fall 2005-Spring 2006)
    - Tutorials and Opening workshop (9/11/05-9/14/05)

    - Mid-Program Focused Workshop (TBA)
    - Transition Workshop and Symposium (5/15/06-5/16/06)
- Astrostatistics (Spring 2006)
    - Planning Workshop (7/14/05-7/15/05)
    - Tutorials (1/16/06-1/22/06)
    - Opening Workshop (1/23/06-1/25/06)
    - Closing Workshop (at Penn State, 6/11/06-6/14/06)

Education and Outreach

- Two 2-Day Workshops for Undergraduates will be held during the academic year
- A PREP Workshop (joint with MAA) will be held in late May 2006
- An Undergraduate Interdisciplinary Workshop will be held in early June 2006
- The Industrial Mathematical and Statistical Modeling Workshop for Graduate Students will be held in late July 2006
- Graduate Courses at SAMSI
    - E. Gheysels, Fall 2005
    - R. Sircar, Fall 2005
    - NDHS topics course, Fall 2005
    - Astrostatistics topics course, Spring 2006

Tentative Programs for 2006-2007

- High Dimensional Inference and Random Matrices (Fall 2006)
- Development, Assessment and Utilization of Complex Computer Models (Fall 2006, Spring 2007)
    - Subprogram in Engineering Models
    - Subprogram in Biomedical Models
    - Subprogram in Ecological Models

3. Developments and Initiatives

Third-Year Developments

- Technology-transfer workshops are being tried as a mechanism for dissemination of research from SAMSI programs.
- The website redesign was completed.
- An online postdoctoral application system was instituted, enabling much faster response by the directorate and remote program leaders.
- PREP workshops (joint with the MAA) were instituted as a way to reach teaching faculty at four-year colleges about the SAMSI vision.
- Another staff person was added, in part to deal with the more intensive evaluation schemes that were approved, and in part to systematize program development activities.
- Consideration of expansion of space, through an addition to the NISS building, was undertaken.

- Additional collaborations with other institutes were instituted, to enhance the overall impact of mathematics and statistics; these initiatives included
    - activities with the National Center for Atmospheric Research, relating to the Data Assimilation program, including joint postdoctoral appointments and a planned joint summer graduate educational program;
    - a joint workshop with the Centre de Recherches Mathématiques in financial mathematics;
    - a joint workshop with the Institute of Pure and Applied Mathematics in data assimilation;
    - a variety of coordinated activities with the Canadian National Program on Complex Data Structures, including
        - a planning workshop for the SAMSI program on the Design and Analysis of Computer Experiments for Complex Systems
        - a co-sponsored Data Mining workshop in Toronto, serving as a significant outlet for dissemination of DMML program results;
        - a co-sponsored meeting on social sciences and complex surveys, in concert with the SAMSI social sciences program
        - joint working groups were formed and active in latent variable modeling and complex surveys.

Planned Fourth-Year Developments

- Databases will be updated:
    - An internal financial database (not dependent on Partner universities) will be installed.
    - Sophisticated participant and scheduling databases (from IMA) will be installed.
    - Data-reporting to NSF will be standardized among institutes.
- Continuing efforts will be made to optimize the operation of workshops; in particular, directorate liaisons will play roles as workshop facilitators where warranted.
- Recognition of the centrality and success of working groups means that future one-semester programs will be extended (before or after the program semester).
- National accessibility to working groups will be greatly enhanced through initiatives discussed in part E.
- Summer schools are being instituted (this summer in Data Assimilation, and next summer in Computer Modeling).
- A SAMSI Science Advisory Board will be created to provide input into the initiation and development of interdisciplinary SAMSI programs.
- We will formalize (and advertise) the policy that *any* graduate student can apply to participate in a SAMSI program. The details of this program are given in E.
- An NAC member will become liaison to the E&O committee, allowing for more extensive national input into this crucial activity.
- Changes in cooperation among DMS institutes.
    - Yearly directors meeting
    - Broadening participation, including diversity:
        - Best practices will be exchanged, including diversity databases.

        - At activities involving broadening, an institute will provide time for other institutes to report relevant opportunities.
        - Diversity events, such as CAARMS and the Blackwell-Tapia conference, will be coordinated.
    - A common institutes webpage is being created.
    - Related scientific programs will be discussed, with overlap reduced, and joint workshops and activities instituted when warranted.

C. Directorate's Summary of Challenges and Responses

SAMSI has been successful in achieving its goals: the scientific programs have been of high caliber, and have lead to significant new and ongoing research collaborations between, statistics, applied mathematics and disciplinary sciences; there has been significant human resource development, through the postdoctoral and graduate programs and through involvement of senior researchers in new interdisciplinary areas; and many students across the country have been shown the SAMSI vision through educational outreach programs and courses. We feel that these successes are amply demonstrated throughout the report, and will here confine discussion to the challenges that arose in Year 3 and the Directorate's response to these challenges. Additional issues were raised during the recent Third-Year Review of SAMSI; these issues and our planned response to them are outlined in Part E of the Executive Summary.

Program Initiation: Most of the programs conducted during the first two years had been part of the initial SAMSI grant proposal, and hence had local individuals as leaders or co-leaders. During the third year, programs had a roughly 50-50 mix of local and non-local leaders. The key to this success was being able to bring the outside leaders (e.g., Byron Goldstein, Kayo Ide and Leonard Smith) to SAMSI for a significant part of the programs.

In Year 4, we expect programs to be roughly 75% driven by non-local scientists. The Financial Mathematics program has a mix of local and non-local leaders; but the National Defense Program and Astrostatistics Program have primarily non-local leadership. The Year 5 programs are likely to be entirely driven by outside scientists: High Dimensional Inference and Random Matrices is being lead by Iain Johnstone and Craig Tracy, while Development, Assessment and Utilization of Complex Computer Models is being lead by Thomas Santner and other (still under negotiation) outside scientists. Of course, these programs will still have significant participation of local scientists; indeed, one of the major strengths of SAMSI is its ability to draw to its programs stellar local talent in applied mathematics, statistics and disciplinary sciences.

Few programs come 'ready made' with leaders attached. Rather, it is a process of working with key individuals over time, to craft a program of key scientific interest and which they are willing to lead. We are finding that exploratory workshops are very helpful in the process of program creation, and that collaborating with other organizations (currently NISS, NCAR, IPAM, CRM and NPCDS) is also valuable in this process.

Program Development: As mentioned above, the initial SASMI programs were directed by individuals who had been heavily involved in the creation of SAMSI, and so the process of developing the programs (recruiting participants and postdocs, forming working groups, planning workshops, etc.) was well-understood by the program leaders. The upcoming programs, however, are being lead by individuals with no or minimal previous connection to SAMSI, and they do not know how SAMSI 'works.' Documents were prepared outlining the process of program and workshop development, but we came to realize that these are not sufficient guidance. We are thus instituting other mechanisms to provide more hands-on guidance, including having directorate and NAC liaisons on each program committee and instituting regular meetings or conference calls between the

program leaders and directorate. Also, another staff person was hired, part of whose responsibility is ensuring that program development keeps to schedule.

Program Operation and Evaluation: We are continually adapting program operations to reflect our experiences in running programs. Year 2 of operations involved considerable experimentation as to types of sessions within workshops, with the focus being on developing sessions designed to maximize interaction and discussion. Year 3 involved solidifying these changes, but also an increased realization that the primary purpose of the opening workshop in a program is finalization of the working groups – and their research agendas – which will be at the heart of the research activity throughout the year. We thus formalized this activity to a considerably greater extent.

During this year we had the first successful instances of working groups maintaining connection to researchers who could not reside at SAMSI. This success made it clear that we need to make sure that such outside participation in working groups is the rule, rather than the exception, and that enabling technologies and policies need to be developed.

All workshop participants are asked to submit an evaluation of their experience, and postdocs have been involved in an extensive process of evaluation of their experience. We ensure that responses from postdocs are obtained, but have not had a large response rate from workshop participants. Additionally, feedback from other program participants has been sought, but the response has been irregular – good in some programs and bad in others. As part of the SAMSI Evaluation Plan, obtaining such evaluations is becoming more institutionalized. Indeed, the new staff member mentioned above is playing a key role in regularizing the evaluations. Follow-up evaluations are reported in Appendix A.

Human Resources: We continue to focus on the postdoctoral program, enhancing both its collegiality (with specific postdoctoral program events) and its administration (with an online application/review system). This year we seem to have crossed a threshold in terms of visibility, with the applicant pool (84 candidates) radically increased from previous years. Furthermore, we have succeeded in hiring the top applicants; of the eight offers made to the top candidates, 5 accepted and 2 others are very likely to accept.

Diversity is a never-ending challenge. Carlos Castillo-Chavez became co-leader of the National Advisory Council, and one of his major roles will be to provide additional insight as to how best to tackle this challenge. As reported in Part E below, we seem to be performing well in regards to diversity, but continual improvement is needed.

SAMSI Graduate Fellows have the rare opportunity to be immersed in a combined statistics/mathematics/disciplinary interface, and the educational experience is of great benefit to them. We are looking at ways to enable this opportunity to also be available to graduate students from across the nation.

Inter-Institute Collaboration: The synopsis of events and plans in Part A above make it clear that we have found collaboration with our sister institutes to be of enormous benefit to our programs. The intensity and extent of such collaborations will only be increasing in the future.

D. Synopsis of Developments in Research, Human Resource Development, and Education

In later parts of the report, the extensive developments in research and education that have occurred under SAMSI research programs are discussed in detail. To give a flavor of these developments, we highlight some of their findings here.

1. Research

a) DATA MINING AND MACHINE LEARNING: The DMML Program of 2003-2004 significantly impacts the following scientific and societal issues.

A new method was developed for determining the key binding features of compounds to a protein, a method capable of handling *multiple binding modes* and compounds that bind differently to different proteins. This method has the potential to revolutionize the drug discovery process. A patent application has been filed.

Analysis of General Motors demand sensing data was undertaken. The scale and complexity of this data – more than 2.5 million vehicles and more than 1200 option codes in more than 500,000 combinations – had deterred GM itself from undertaking analyses. The analyses led to important insights into how one can deal with complex and largely non-quantitative data sets.

The first rigorous statistical justification for the *relevance vector machine* was given, together with a hierarchical description. The main result is that the extended prior structure will make it possible to obtain a unique solution.

Research on robustness in data mining included new ideas on the surprising success of over-completeness. In particular, Bayesian methods were applied to shed considerable light on this success, and to suggest significantly improved procedures. This exemplifies the ability of SAMSI to provide theoretical and methodological insight into areas previously dominated by attention only to algorithms.

The solution of the standard support vector machine (SVM) utilizes all the input covariates without discrimination. A typical high-dimensional low-sample-size dataset may contain many noisy and redundant variables. Without selecting the important variables, the standard SVM may give misleading classification results. A new approach to feature selection in SVM was developed and compares very favorably with other SVD methods, using a metabolomic dataset from Metabolon.

b) MULTISCALE MODEL DEVELOPMENT AND CONTROL DESIGN: The highlights of the activities of the three major research thrusts of the program are as follows:

Paradigms for Bridging Scales: We discussed fundamental energy mechanisms for piezoceramic and magnetic materials employed in present and projected high performance applications, using a magnetic prototype. Both deterministic and stochastic models were developed which incorporate multiple scales ranging from microscopic to macroscopic. Considerable progress was made on the goal of developing data-driven modeling frameworks for advanced materials which characterize hysteresis and nonlinear dynamics in a manner which facilitates control design.

Control Design: We discussed a number of nonlinear control issues pertaining to systems exhibiting the nonlinear, hysteretic behavior characteristic of present smart

materials. We numerically implemented a nonlinear hybrid method comprised of feedforward and feedback loops and investigated the real-time implementation of this design. A considerable start was also made on development of stochastic models and control designs that incorporate concepts related to wavelet analysis and stochastic Kalman filtering, with the eventual goal of providing robust control designs suitable for real-time implementation.

Model Development for Ionic Polymers: This group investigated the development of models for the ionic polymer Nafion which is presently being investigated for applications ranging from artificial muscle design to chemical sensing. Statistical mechanics (Boltzmann) principles were used to relate the identified densities to energy and macroscopic stiffness attributes of the compounds. Additionally, we investigated the degree to which the techniques developed in this program can be extended to biomedical applications including the modeling and treatment of cartilage disorders. Several research proposals are under development as follow-ups to this program.

c) LATENT VARIABLE MODELING IN THE SOCIAL SCIENCES: Working together, a sociologist, a business professor, and statistician from one of the working groups have developed an improved way to predict categorical outcomes, such as political affiliation, disease status or even potential terrorist activities, from independent variables that are measured with error.

In complex survey working group, a statistician and sociologist have joined together gauge the degree of measurement error in the US government's unemployment figures. The goal is to assess their accuracy and to develop better unemployment estimates. The economic (stock markets, in particular) and political impact of unemployment estimates attest to the importance of this work.

Structural Equation Modeling (SEM) is ubiquitous in the social sciences, and the methodology that is currently used for selection from among contending SEM's has recently come under considerable criticism in the social science literatures. While this problem seems highly technical, the implications of making model selection a science rather than an art touch the entire spectrum of the social sciences, since the "wrong" model can have negative consequences ranging from removing an effective drug from the market to policy that does not work.

In this working group involving numerous social scientists and statisticians, a natural thrust was to seek to apply more modern statistical model selection techniques to selection from among SEM's. One of the most commonly used criteria for selection from among SEM's is Bayes Information Criterion (BIC) which was designed as an asymptotic model selection technique for a simple setting. The statisticians in the group observed that this was known to be highly inaccurate for complicated settings such as SEM's and a joint effort has created a breakthrough – what seems to be a much more accurate version of BIC useable for complex models such as SEM's.

The Social Networks working group is studying multiple problems lying at the applied mathematics-statistics interface, including numerical solution of stochastic differential equations used to model agent characteristics and relationships in dynamic social networks; use of perturbation methods to incorporate stochastic fluctuations in dynamic social networks in which "edges" are defined by threshholding numerical pairwise associations; use of sensitivity analysis for SDEs to classify parameters of

stochastic models in terms of their qualitative effects (for instance, for low but non-zero "noise" levels, the behavior of the system is richer than when there is no noise, but as the level of noise increases, eventually all structure disappears); and nonlinear optimization of affinities in stochastic social networks.

d) COMPUTATIONAL MODELING OF INFECTIOUS DISEASE: One of the exciting research developments is the steps taken towards creation of *An Automated Reasoning System for Immunobiology*. The amount of information in the scientific literature about the immune system is so vast that the human mind can utilize only a tiny fraction of it. Assimilation of even this tiny fraction requires tremendous effort: the information is distributed over many disparate sources and the terminology is imprecisely defined and inconsistently used. Current methods for accessing and processing information pose a critical barrier to progress in the study of human health and disease. The solution is a comprehensive knowledge base (KB) equipped with a standard ontology and a powerful reasoning platform that supports efficient and uniform access to knowledge, *in silico* generation and testing of hypotheses, and inferences and analysis in general.

To achieve this goal, it is necessary to develop i) a rigorously defined and consistently applied vocabulary, ii) efficient methods for extracting and cleansing knowledge from the scientific literature, iii) knowledge models that permit nontrivial yet computationally tractable reasoning, iii) tools for a wide range of inference, analysis, and simulation tasks, iv) user and application interfaces for accessing the knowledge and accompanying tools. The computer science community has made significant strides in areas such as artificial intelligence, information retrieval, and database systems. A lot of progress remains to be made, however, and application of these techniques to biology is still in its infancy. Furthermore, the complex dynamics and the inherent variability of biological systems provide unique challenges, as does the rapid rate at which knowledge in biology evolves. To assess the feasibility and potential of an immune system knowledge base with automatic reasoning support, and to provide a foundation of tools and techniques on which to build such a system, we have begun the development of a knowledge base of signal transduction pathways for the pattern recognition receptors (PRR) of the innate immune system, together with a standard ontology and reasoning platform that supports numerous inference and analysis functionalities.

e) DATA ASSIMILATION: Two key areas, which are focusing our efforts and energies, have emerged. First, the promise of assimilating Lagrangian and quasi-Lagrangian data is of enormous importance in ocean state estimation. This program has allowed us to consider a number of different approaches to this all-important problem, including stochastic DE methods, smoothing and the use of indistinguishable states. The Lagrangian DA working group is spawning a number of group efforts, involving a variety of combinations of core group members, postdocs and students, to implement these approaches.

Secondly, a very active group interested in model inadequacy is working to combine statistical and dynamical systems inspired methods. The merging of these perspectives promises a more powerful technique that will provide a systematic and effective approach to ensemble forecasts, the key to effective weather and climate prediction.

2. Human Resource Development

SAMSI's impact on human resources is fully discussed in sections I.B and I.C, with impact on diversity highlighted in section I.H. The individual program reports also contain significant insight into human resource development. Here we give several illustrations indicating the unique impact that SAMSI has in these areas.

Research plans and paths. Not many people have left SAMSI doing exactly the same kind of research as when they arrived. For instance of the 22 SAMSI graduate fellows in years 1 and 2, 12 (55%) have completed or are preparing dissertations arising from their SAMSI experiences. A few specific examples of those who became engaged in completely different lines of research include:

- An INVERSE program postdoc with a degree in applied mathematics now at the University of Montana, who is actively engaging in statistical collaborations.
- A STOCOM program postdoc with a degree in statistics, with no previous experience in bioinformatics, who has co-authored papers and participated in proposals dealing with bioinformatics.
- An INVERSE postdoc with a degree in applied mathematics, who is now a tenure-track faculty member in statistics at Texas A&M University.
- An INTERNET program postdoc with a degree in applied probability (point processes), who is co-author of papers and pursuing further research and proposals on analysis of network traffic data.
- A MULTISCALE program postdoc, with a PhD in Operations Research having a focus on statistics, collaborated with students, postdocs and faculty in applied mathematics, mechanical engineering, chemical engineering, and material science on multiscale characterization techniques for advanced materials using a synergistic deterministic/statistical framework. She is presently a staff scientist at SAS and plays an integral role in investigations focused on the extension of these techniques to biological materials such as cartilage.
- A DMML-SAMSI University Fellow, who pursued and received a grant from the Canadian NSERC "based on the proposals for the research I ended up starting at SAMSI."

Here are some comments in the final reports from 2004 of the postdoctoral fellows, and some comments from the reports submitted by long-term visitors. More details can be found in sections I.B. and Appendix A.

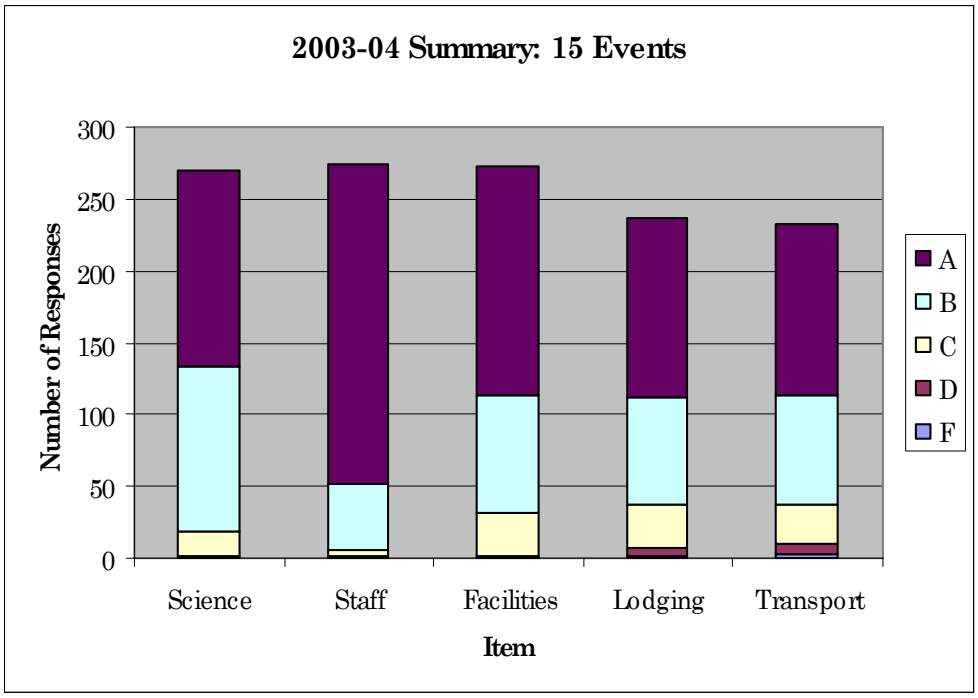
- DMML Postdoc: "Coming to SAMSI has turned out to be one of the best decisions of my academic life […] My SAMSI experience has been richer, more fulfilling and more rewarding than anything I could have had in a typical university setting."
- INT Postdoc: "Though I prefer probabilistic modeling to statistical modeling, talking to the practitioners is very important in understanding the real model."
- MULTI Postdoc: Before I came to SAMSI, I was a research scientist at a lab supported primarily by the military. While I was at this lab I did not receive any type of mentoring. I often felt rushed to complete a project and move on to the

next one. At SAMSI I have had more time to slow down, learn, and give a project my undivided attention. […] Learning to "speak the language" and communicate thoughts and ideas to researchers in other fields will be an invaluable skill …."

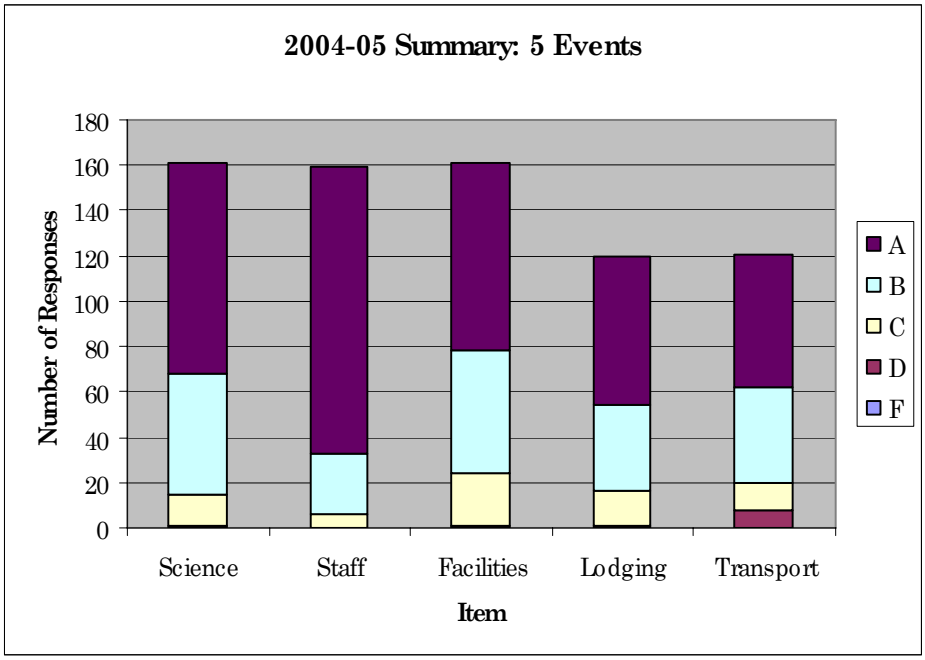
- DMML SAMSI-University Fellow: "This is the best intellectual/ research environment I've ever had the chance to enjoy. It seemed to me that everyone is keen to bring in the future, teach their fellow travelers and learn in turn, and enjoy the development of new ideas"
- DMML Faculty Fellow: "The main impact of the SAMSI data mining year is that is has gotten me back into the academic game"
- DMML Industrial Participant: "The support and advice we have gotten from SAMSI [have] allowed us to enter into new areas of exploration that we would not been able to enter without them."
- MULTI Long-Term Visitor: "This programme has been very valuable for developing a new research emphasis on environmental systems modeling."

There are, of course, numerous participants in SAMSI workshops other than the long-term participants. Details about their evaluation of the workshops are given in Appendix G. Here are summary graphs indicating their satisfaction with the workshops.

Workshop Evaluation Summary: 2003-04



Workshop Evaluation Summary: 2004-05



3. Education

The successes of the SAMSI courses and education and outreach programs are fully discussed in section I.E.4 and in the program reports. The new initiatives that have been undertaken in this area are detailed below.

- In an effort to reach college teachers (typically at primarily teaching institutions) with the SAMSI message, SAMSI joined with the Mathematical Association of America (MAA) in co-sponsoring one of their Professional Enhancement Programs (PREP, see http://www.maa.org/prep/) held May 26-29, 2004 at the University of Louisiana at Lafayette. Participants were college teachers chosen from a national applicant pool. A team from SAMSI will again lead this initiative on May 25-27, 2005.
- Technology transfer courses are being introduced (starting with one in June, 2005 on Data Mining) to enhance the dissemination of SAMSI research.
- Summer schools are being introduced (the first one on Data Assimilation in June 2005, joint with NCAR), to enable intensive exposure of a broad range of graduate students to research in a SAMSI program area.

E. Evaluation and Third Year Review

In addition to the usual evaluation process we undertake at SAMSI, there was a third-year review of SAMSI by the NSF from March 16-18, 2005. Here we highlight some of the main issues raised during the site visit, from the perspective of evaluation.

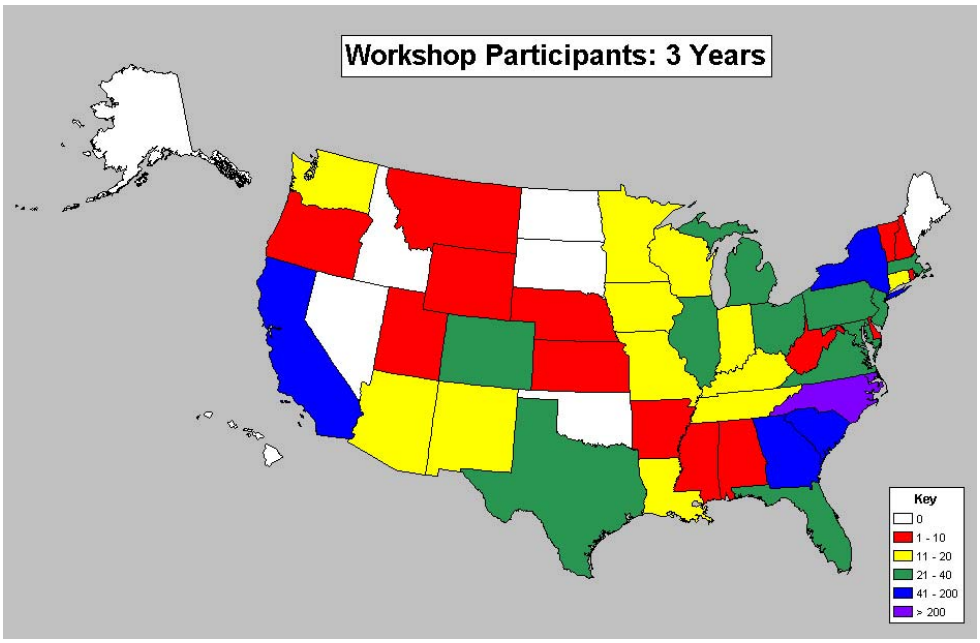
Issue 1. Since national development of the human resource base is central to assessing the effectiveness of SAMSI, three key questions are:

- Does SAMSI have a national presence?
- Is SAMSI reaching underrepresented minorities?
- Is SAMSI reaching a range of institutions, broadening the DMS research impact?

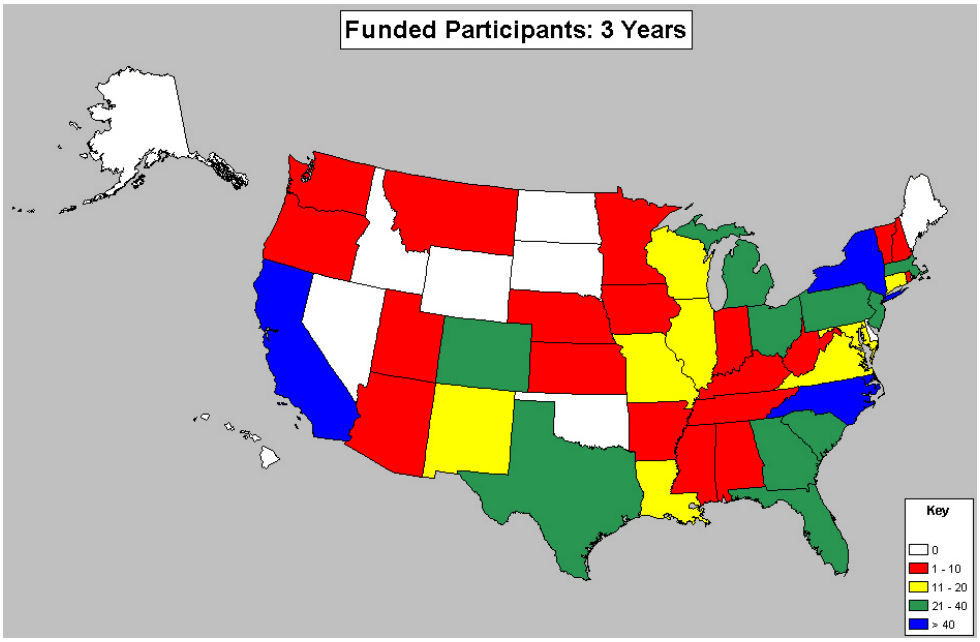
In terms of the national role of SAMSI, note first that the NSF funding to SAMSI goes almost exclusively to support non-local researchers; the efforts of local researchers are supported by the Partner institutions of SAMSI. This ensures that the focus of SAMSI is primarily national.

The following two graphs specifically address the first question above, showing that workshop participation is highly geographically diverse nationally, both in terms of participation and in terms of funding. The graphs present the cumulative totals over the three years of SAMSI operation.

Geographical Distribution of Workshop Participants (2002-05)



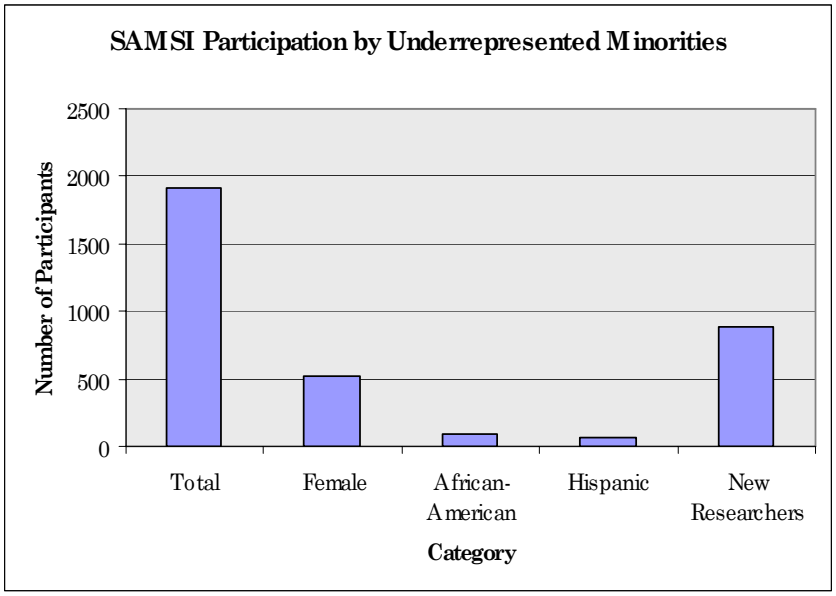
Distribution of Funded Workshop Participants (2002-05)



Also of interest is the geographical distribution of participants internationally. This is described in the following table. We note that foreign participation in SAMSI is less than at other DMS mathematical sciences institutes. This is, in large part, due to the focus of SAMSI on statistics and applied mathematics, areas of the mathematical sciences that are globally more heavily focused in North America than are many other areas of the mathematical sciences. Also, the SAMSI policy is to bring in the leading scientists, regardless of nationality, but to otherwise focus on domestic participants.

Year	US Citizen or Permanent Resident	Foreign National Residing in US	Foreign National not Residing in US	TOTAL
2002-03	209	87	36	332
2003-04	220	90	29	339
2004-05 (to date)	105	39	9	153
TOTAL	534	216	74	824
Percent	65%	26%	9%	

That SAMSI is reaching underrepresented groups is indicted by the following table, which again gives the cumulative experience from 2002-05. The percentages for the past year mirror the percentages in this table almost exactly, and so are not reported. While these are above average percentages for the mathematical sciences, we will be continually working to raise them.



The final key issue concerning participant distribution is whether SAMSI is reaching institutions that are not already heavily funded by DMS. One of the major roles of institutes is to help develop the national research base by enabling individuals who are not at institutions already heavily supported by DMS to engage in research. The SAMSI record in this regard is excellent:

- Participants from 137 US academic institutions were recipients of SAMSI funds, including 46 of the top 50 recipients of DMS funds.
- 50% of SAMSI funds go to the DMS top 50 institutions (while 85% of DMS funds go to the DMS top 50).
- 34% of SAMSI funds go to the DMS 51-200 institutions (while 15% of DMS funds go to the DMS 51-200).
- 16% of SAMSI funds go to 46 institutions not in the DMS top 200

Issue 2. How does SAMSI engage the national and international community in addition to its engagement through the National Advisory Council?

SAMSI uses multiple strategies to reach and engage the national community, including the following:

- The web contains complete descriptions of SAMSI programs and workshops and SAMSI visiting opportunities, and solicits proposals for future programs.
- The directorate gives roughly 50 talks around the world each year in which the opportunities at SAMSI are discussed.

- Annual one-page articles are sent to the newsletters or bulletins of AMS, ASA, IMS, and SIAM (and occasionally elsewhere, e.g. IEEE).
- Full or half page advertisements are placed in the newsletters of ASA, IMS, and SIAM for the postdoctoral program. Advertisements are also placed in regards to visiting opportunities.
- One summary poster of our yearly programs and a poster about each of the three major programs are mailed, for display on bulletin boards, to all math and stat/biostat departments (and many other places).
- There is regular announcement of all upcoming activities to the organizations in the affiliates program.
- Each major event (including E&O events), are announced by mass e-mails to relevant departments and organizations nationally (and internationally). For interdisciplinary events, considerable effort is also spent in trying to identify relevant departments in the other discipline to be notified. Available discipline-based individual mailing lists are also utilized when appropriate.
- At the annual Joint Statistical Meetings and SIAM meetings, at least one SAMSI specific session is held, to not only discuss SAMSI research, but also to highlight the next year's (and future) programs and strongly encourage participation and proposal of new programs. At the last JSM for instance, there was one SAMSI invited session about this, and two other presentations about SAMSI (one to the affiliates and one to the Chairs of statistics departments).
- At the JSM, there is a SAMSI reception one evening (joint with NISS, and paid for by NISS) and, at the Joint Mathematics Meeting, SAMSI participates with the other institutes in a joint reception/information session.
- Programs are strongly encouraged (and many do) submit scientific sessions of SAMSI research (identified as such) at a wide range of scientific meetings.
- There are a number of miscellaneous chances to advertise SAMSI that are taken, such as an entry last year in the Encyclopedia of Statistical Sciences.

Issue 3. How does SAMSI ensure that there is sufficient input from disciplinary scientists in planning interdisciplinary SAMSI programs?

The SAMSI Directorate, the National Advisory Council, the Local Development Committee, the Chairs Committee and the Governing Board comprise roughly 40 individuals who are heavily connected with other disciplines and able to place program planners in contact with relevant disciplinary scientists. Of course, those planning interdisciplinary programs also typically have many such contacts themselves. Proposed individuals in other disciplines are routinely placed on scientific and organizing committees of SAMSI programs and workshops, and routinely are major participants therein. As but one example, in the upcoming Astrostatistics program 3 astronomers and one physicist are currently planning long-term visits to the program.

In addition to the current strategies, we will be taking the following steps to ensure sufficient input from scientists in other disciplines:

- The conference calls involving program organizing committees, the directorate, and the NAC liaison will also include disciplinary scientists when relevant.

- A SAMSI Science Advisory Board will be created (with input from all the individuals mentioned above, especially the NAC). SSAB members will provide input into the initiation and development of SAMSI programs, through conference calls to relevant planning meetings, especially the annual meeting of the National Advisory Council with the Directorate.

Issue 4. How can the involvement of non-local graduate students be increased?

Graduate students from non-Partner universities already heavily participate in workshops: indeed, in the graduate student/new researcher category that we recorded in our database, there were 210 Partner university participants in workshops but 268 participants from non-Partner universities. There have been eight non-Partner university graduate students that visited SAMSI for a semester, but the availability of this opportunity is not widely known. We will formalize (and advertise) the policy that *any* graduate student can apply to participate in a SAMSI program. The details of this program are as follows:

- Visiting graduate students will be called SAMSI National Graduate Fellows, to enhance the prestige of such an appointment.
- Critical to such appointments, mentoring must be arranged. When a student is visiting with their current advisor, mentoring is automatically available. Otherwise, mentors will be sought from program leaders or other participating scientists. (If no mentor can be found, an appointment will not be made; we feel it is highly detrimental to the student to visit without solid mentoring.)
- SAMSI does not pay tuition for any graduate student, but support for visiting students can be provided, as well as health insurance.
- Appointments for the students as 'visiting scholars' will be arranged at Partner universities, to allow the students access to facilities at a university.

Issue 5. How can SAMSI Working Groups be made more accessible to non-local researchers?

Working Groups are the heart of SAMSI programs, meeting weekly or bi-weekly during a program to advance their research agenda. The core of most working groups is formed by the long-term visitors to SAMSI, the postdocs, and interested local participants. Because of the enormous success of the working group concept, it is clearly desirable to enable participation by individuals who cannot be present over a long period.

This has already happened for certain working groups – e.g., two of the working groups in the LVSS program have had significant year-long involvement by national and international participants. SAMSI is committed to enhancing this opportunity by the following mechanisms:

- Technology will be made available to all working groups whereby external participation is made easier: in addition to the current extensive use of working group websites for posting of material and talks, cameras and quality teleconferencing equipment will be available at SAMSI to allow any external participant to be fully involved in the meeting with only a telephone and a computer.

- The nature and interests of working groups will be publicized immediately after the opening workshop of a program, so that individuals who could not attend the opening workshop are able to become involved.
- We will consider the possibility that working groups can primarily be based elsewhere, with the individuals residing at SAMSI participating in these working groups by the technological means mentioned above. Committed working group leaders at the other location would be crucial to the success of such a working group.

Issue 6. Is selection of scientific themes for programs and working groups systematic and based on broad input?

The choice of scientific programs arises from an involved process that can be described in four phases:

- *Phase 1:* Ideas come in from a wide variety of sources, including national (and international) individuals and the NAC. The directorate very actively solicits such input. Ideas are given an initial 'credibility screening' by the directorate. Ideas in the past that have failed this screening tend to be ideas that are far too specialized for a SAMSI program.
- *Phase 2:* Ideas that pass the initial screening either proceed to preparation of a formal pre-proposal, or are embedded in other programs – often on a trial basis. Examples of the latter include Causality, Granular flow, and Nanotechnology of Soft Materials, that were all embedded in other programs as workshops, to assess their potential for development as a full program. (For various reasons, it was deemed that the time was not right for pursuing these as full programs.)
- *Phase 3:* Pre-proposals are brought before the NAC (and the local committees) to assess their potential as SAMSI programs. Those that are approved for development by the NAC enter a developmental stage, leading to a Full Proposal. A Full Proposal must be formally accepted by the directorate and NAC, and includes many of the specifics of the program.
- *Phase 4:* Details of the approved program are planned, including workshops and the research topics for working groups. While there is very considerable discussion of working group research topics in advance – among the long-term program participants and national leaders – the final working groups (and their research topics) are not set until the end of the Opening Workshop of a program. Indeed, the primary purpose of the opening workshop is to obtain input from the assembled national leaders as to the most promising research foci of the working groups; a variety of break-out and discussion sessions at the workshop are oriented towards this result.

F. Evaluation by the SAMSI Governing Board

(Bruce Carney, Vijay Nair, John Simon, Daniel Solomon – chair)

The Governing Board provides broad oversight for the Institute's administration, finances, and evaluation, and for relationships among the partnering institutions. As part of the annual evaluation, the Governing Board has elected to address four broad questions. That evaluation follows:

1) What is the synthesis of applied mathematics and statistics enabling?

The synthesis of applied mathematics, statistics and the disciplinary sciences is a central tenet of the SAMSI mission. There are notable examples of this synthesis in specific SAMSI programs, but the extent varies substantially across the full portfolio.

Such synthesis this past year has been perhaps most evident in the program on Data Assimilation for Geophysical Systems (see Section E.3 of the full report). The program brought mathematical experts in dynamical systems and statisticians expert in the general art of data assimilation to work with meteorologists and oceanographers. Having both quantitative groups involved with the disciplinary scientists is leading to a much greater clarification of the key issues and barriers to research.

The Social Networks working group associated with the program on Latent Variable Models in the Social Sciences is studying multiple problems lying at the applied mathematics-statistics interface, including numerical solution of stochastic differential equations (SDEs) used to model agent characteristics and relationships in dynamic social networks; use of perturbation methods to incorporate stochastic fluctuations in such networks; use of sensitivity analysis for SDEs to classify parameters of stochastic models in terms of their qualitative effects; and nonlinear optimization of affinities in stochastic social networks.

The program in Network Modeling for the Internet provided a venue that brought together statisticians and computer scientists from academe and industry that led to an ongoing, fruitful collaboration on data collection and network tomography research. As a side note, the partnership with the National Institute of Statistical Sciences provided software and hardware for the project.

In addition to the scientific outcomes of these programs, we see evidence of SAMSI's strong impact upon early career researchers.

Emily Lada was a postdoctoral fellow associated with the program in Multiscale Model Development and Control Design who came to SAMSI with a PhD in Operations Research having a focus on statistics. Through SAMSI she collaborated with students, other postdoctorals and faculty in applied mathematics, mechanical engineering, chemical engineering, and materials science on multiscale characterization techniques for advanced materials using a synergistic deterministic/statistical framework. She is

presently a staff scientist at SAS Institute and plays an integral role in investigations focused on the extension of these techniques to biological materials such as cartilage.

We see graduate students who, following involvement in the Multiscale program and the Data Assimilation program are undertaking thesis research combining advanced tools from both statistics and applied mathematics.

We also see examples of the integration being promulgated still earlier along the career development path. For example, in connection with the program in Data Mining and Machine Learning, SAMSI developed a workshop for undergraduates that moves from the introductory level to current research frontiers. The workshop was given twice, to audiences that spanned mathematics, statistics and computer science.

2) Is the impact of SAMSI on science and human resources growing?

Section D of the Executive Summary highlights some of the developments in science and education that have occurred through SAMSI's programs.

Examples of new science include the development (in connection with the Data Mining and Machine Learning program), in which "a new method was developed for determining the key binding features of compounds to a protein, a method capable of handling multiple binding modes and compounds that bind differently to different proteins. This method has the potential to revolutionize the drug discovery process. A patent application has been filed."

In connection with the Data Assimilation for Geophysical Systems program, "a very active group interested in model inadequacy is working to combine statistical and dynamical systems inspired methods. The merging of these perspectives promises a more powerful technique that will provide a systematic and effective approach to ensemble forecasts, the key to effective weather and climate prediction."

The lists of refereed publications associated with SAMSI programs (see Section I.G. of the full report) provide another measure of evidence of impact on the mathematical and disciplinary sciences.

SAMSI continues its strong commitment to the development of human resources in the mathematical sciences. Its impacts are discussed in Sections I.B, I.C and I.H (which highlights diversity) of the full report. Indeed some of the effects on early career researchers are described in item 1) above.

Participation rates of women and people of color in SAMSI activities compare favorably to those of similar organizations in the mathematical sciences but deserve continued attention. New researchers comprise a significant fraction of SAMSI participants, and a significant fraction of SAMSI funds go to participants from institutions that are not among the top recipients of other NSF-DMS funds.

SAMSI effects are proving to be powerful and persistent; for example influencing the research paths of those who have participated in its programs. Notably, over half of the SAMSI graduate fellows of the first two years have pursued dissertations arising out of their SAMSI experiences. In a specific example, the program on Network Modeling for the Internet brought together a student from a small graduate program and a faculty member from a major department. That connection ultimately led to the hiring of the student as Assistant Professor in the major program. Other examples of such positive impacts on senior as well as junior scientists, including statements from individual participants, are cited in Section D of the Executive Summary and in the body of the full report.

Educational outreach continues to be a prominent feature of SAMSI activity. In an effort to reach college teachers (typically at primarily teaching institutions) with the SAMSI message of the importance of combining mathematics and statistics, SAMSI joined with the Mathematical Association of America (MAA) in co-sponsoring in 2004, one of their Professional Enhancement Programs (PREP, see http://www.maa.org/prep/) titled "Mathematics Meets Biology: Epidemics, Data Fitting, and Chaos." It was held at the University of Louisiana at Lafayette. Participants were college teachers chosen from a national applicant pool. A team from SAMSI will again lead this initiative on May 25-27, 2005.

In addition, technology transfer courses are being introduced (starting with one in June, 2005 on Data Mining) to enhance the dissemination of SAMSI research.

3) Is the national recognition and respect for SAMSI growing?

After only three years of existence, SAMSI has become a noteworthy element of the national infrastructure of the mathematical sciences. This is documented throughout the full report and highlighted in the Executive Summary.

While most of SAMSI's early programs were part of the original grant proposal and so featured local leadership, programs now being planned (see Section II of the full report) show a preponderance of leadership by scientists from outside SAMSI's partner institutions, with about 75% non-local leadership in Year 4 and entirely non-local leadership anticipated in Year 5. That leading scientists from outside the local area are eager to lead SAMSI programs is a strong indicator of SAMSI's national stature.

The geographic distribution of participants is summarized in Section E of the Executive Summary, while the detailed participant lists for concluded programs provide ample evidence of the national and international draw of SAMSI activities. The solid participation rate of members of local institutions strengthens SAMSI programs and proves an attraction to high quality participants from across the nation and the world. Other evidence of SAMSI's reach is in the offers of partnerships with other organizations including the National Center for Atmospheric Research, the Centre de Recherches

Mathématiques in Montreal, the (Canadian) National Program on Complex Data Structures, as well as NSF's own Institute for Pure and Applied Mathematics.

SAMSI continues its focus on the postdoctoral program, enhancing both its collegiality (with specific postdoctoral program events) and its administration (with an online application/review system). This year, SAMSI visibility crossed a threshold, with the applicant pool (84 candidates) radically increased from previous years. Furthermore, SAMSI has succeeded in hiring the top applicants; of the eight offers made to the top candidates, 5 accepted and 2 others are very likely to accept.

The Directorate continues to seek ways to ensure national engagement in the implementation and shaping of SAMSI programs.

The SAMSI commitment to facilitating broad participation is evidenced in its plan to provide technology to working groups that enables distant members to participate fully in meetings while requiring only a telephone and computer. This has already proven popular and successful for working groups of the Latent Variable Models in the Social Sciences program. Working groups are a unique and powerful feature of SAMSI programs, and their remote accessibility should prove a strong attraction to broad geographic participation.

Section E of the Executive Summary includes a discussion of strategies for increasing the involvement of graduate students who are pursuing degrees at universities outside the local area. We agree that local mentoring is critical to a successful graduate experience at SAMSI, and such mentoring is automatic for graduate students from the partner universities. Of course even those students diffuse the SAMSI experience when they eventually take employment elsewhere.

Finally, to further enhance input from disciplinary scientists across the nation into the selection and development of its programs, SAMSI will create a Science Advisory Board. This will supplement the disciplinary connections already in place in the Directorate, Local Development Committee, Chairs Committee, Governing Board and Program leaders.

4) Is the Directorate meeting the needs of an evolving SAMSI?

The directorate model continues to serve SAMSI very well, and transitions in the directorate have gone smoothly. Among the strengths of the model is that there are clear divisions of responsibility among the members of the directorate, and the incumbents have excellent working relationships. In addition, a four person directorate whose members have diverse backgrounds and contacts has proven to provide a rich array of perspectives and a broad ability to recognize and respond to the evolution of the disciplines. The relatively short timescale on which programs move from idea to implementation enables SAMSI to respond quickly to such change.

The directorate has also been responsive to suggestions for improvement, e.g. in the integration of technology to enhance working group communication as noted earlier.

Members of the directorate are each active and visible nationally, for example giving some 50 talks around the world each year that describe SAMSI opportunities. They are also active in other ways to disseminate information on SAMSI programs and to build relationships with other professional organizations.

The Governing Board Chair and the SAMSI Director have a biweekly telephone conference at which administrative and personnel matters are regularly discussed and issues addressed where they have arisen. There is also excellent cooperation among the partner universities and NISS to ensure that obligations are met and that SAMSI continues to flourish.

Table of Contents

I. Annual Progress Report

The previous annual progress report was complete in all details only through April, 2004. Hence, we also report activities in Year 2 programs that occurred subsequently and were not itemized in the report. These Year 2 programs were *Data Mining and Machine Learning, Network Modeling for the Internet*, and *Multiscale Model Development and Control Design;* their final reports are in Appendices B, C, and D, respectively.

A. Program Personnel

1. Program and Activity Organizers

Program Organizers

Program	Name	Affiliation	Field
Data Mining and Machine Learning	David Banks (Co-Chair)	Duke	Statistics
	Mary Ellen Bock	Purdue, NAC	Statistics
2003-2004 SAMSI Program	Jerome Friedman	Stanford	Statistics
	Alan F. Karr (Directorate Liaison)	NISS	Statistics
	David Madigan	Rutgers	CS and Stat
	William DuMouchel	AT&T	Statistics
	Warren Sarle	SAS Institute	CS and Stat
Network Modeling for the Internet	Kevin Jeffay	UNC	Computer Science
	James Landwehr	Avaya Labs	Statistics
2003-2004 SAMSI Program	John Lehoczky	Carnegie Mellon, NAC	Probability
	J. S. Marron (Co-Chair)	UNC	Statistics
	Ruth Williams (Co-Chair)	UC San Diego	Probability
	Walter Willinger	AT&T	Computer Science
	Donald Towsley	Massachusetts	Computer Science
Multiscale Model Development	M. Gregory Forest	UNC	Applied Math
and Control Design	Doina Cioranescu	U Pierre & Marie Curie	Applied Math
	Alan Gelfand (Co-Chair)	Duke	Statistics
2004 SAMSI Program	David Schaeffer	Duke	Mathematics
	Murti Salapaka	Iowa State	Applied Math
	Ralph Smith (Co-Chair)	NCSU	Applied Math
	Christopher Wikle	Missouri	Statistics
	Margaret Wright	NYU, NAC	CS and Applied Math
Education & Outreach Program	H.T. Banks (Chair)	NCSU	Applied Math
	Johnny Houston	Elizabeth City State	Math and CS
	Rachel Levy	NCSU	Mathematics
	J. Blair Lyttle	Enloe HS, Raleigh	Statistics
	Negash Medhin	Clark Atlanta	Mathematics
	Daniel Teague	NC Sch Math & Sci	Mathematics
	Wei Feng	UNC-Wilmington	Math and Stat

Genomes to Global Health	Roy Anderson	Imperial College	Disease Epidemiology
Computational Biology of	Rustom Antia	Emory University	Biology
Infectious Disease	Carl Bergstrom	U of Washington	Biology
2004-05 SAMSI Program	Arturo Casadevall	Einstein College of Med	Microbiology & Immunology
	Carlos Castillo-Chavez	Arizona State U	Math and Stat
	Lindsay Cowell	Duke University	Biostats & Bioinformatics
	Sunetra Gupta	Oxford University	Mathematical Biology
	Tom Kepler (Chair)	Duke University	Biostats & Bioinformatics
	Denise Kirschner	U of Michigan	Microbiology & Immunology
	Jun Liu	Harvard University	Biostatistics
	Alan Perelson	Los Alamos Nat Lab	
	Man-Wah Tan	Stanford University	Genetics
Latent Variable Models in the	Kenneth Bollen (Chair)	UNC	Sociology
Ssocial Sciences	James Heckman	U of Chicago	Economics
	Alan Karr (Directorate Liaison)	NISS	
2004-05 SAMSI Program	Susan Murphy	U of Michigan	Social Research
Data Assimilation for Geophysical	Jeffrey Anderson	NCAR	Data Assimilation Initiative
Sciences	Mark Berliner	Ohio State U	Statistics
	Andrew Bennett	Oregon State U	Atmospheric & Ocean Sci
2005 SAMSI Program	Craig Bishop	Navy Research Lab	Marine Meterology
	Montserrat Fuentes	NCSU	Statistics
	Sujit Ghosh	NCSU	Statistics
	Kayo Ide	UCLA	Atmospheric & Ocean Sci
	Christopher Jones (Chair)	UNC	Mathematics
	Eugenia Kalnay	U of Maryland	Meterology
	Susan Lozier	Duke U	Earth & Ocean Sciences
	Authur Mariano	U of Miami	Marine & Atmospheric Sci
	Ian McKeague	Florida State U	Biostatistics
	Robert Miller	Oregon State U	Atmospheric & Ocean Sci
	Douglas Nychka	NCAR	Geophysical Statistics
	Juan Restrepo	U of Arizona	Mathematics
	Leonard Smith	Oxford U	Industrial & Applied Math
	Chris Synder	NCAR	Data Assimilation Initiative
	Istvan Szunyogh	U of Maryland	Physical Science & Tech
	Olivier Talagrand	Ecole Normale Superier	Meterology & Oceanography
	Keith Thompson	Dalhousie U	Math & Stat
	Zoltan Toth	NCEP	Environmental Modeling
	Francisco Werner	UNC	Marine Sciences
	Carl Wunsch	MIT	Earth, Atmos, Planetary Sci

Activity Organizers

Activity	**Name**
Data Mining Closing Workshop -- *May 17-18, 2004*	David Banks
PREP Workshop on Mathematics Meets Biology: Epidemics, Data Fitting and Chaos -- *May 26-29, 2004*	H.T. Banks and Azmy Ackleh
SAMSI-CRSC Undergraduate Workshop -- *May 31-June 4, 2004*	H.T. Banks
Internet Program Closing Workshop -- *June 25-26, 2004*	J.S. Marron
NPCDS/SAMSI Workshop on the Design and Analysis of Computer Experiments for Complex Systems -- *July 13-17, 2004*	Jim Berger, Derek Bingham, Randy Sitter, Jamie Stafford, Will Welch
SAMSI-CRSC Industrial Mathematical & Statistical Modeling Workshop for Graduates -- *July 26-August 3, 2004*	H.T. Banks, Pierre Gremaud, Negash Medhin, Hien Tran, Ralph Smith
Latent Variables in the Social Sciences Opening Workshop -- *September 11-14, 2004*	Ken Bollen
Computational Biology Opening Workshop -- *September 19-22, 2004*	Tom Kepler, Lindsey Cowell
Multiscale Program Closing Workshop -- *September 27-28, 2004*	Ralph Smith, Alan Gelfand
NPCDS/SAMSI Workshop on Data Mining Methodology and Applications -- *October 28-30, 2004*	Hugh Chipman, Antonio Ciampi, Michael Vainder
Data Assimilation Program Opening Workshop -- *January 23-26, 2005*	Chris Jones, Kayo Ide
CompBio Midterm Workshop on Mathematical Modeling of Infectious Diseases -- *January 31-February 1, 2005*	Alun Lloyd
Undergraduate Two-Day Workshop -- *February 18-19, 2005*	H.T. Banks

2. Program Core Participants and Targeted Experts

For each of the major programs, the following tables present the key participants for the programs. The participants are categorized and coded as follows:

D – *Distinguished Lecturer* for the program.
F – *Faculty Release Person,* defined as an individual from a partner university of SAMSI who is accorded release time for participation in the SAMSI program; the cost-sharing value of this release time is indicated.
FA – *Faculty Associate,* defined as an individual from a partner university of SAMSI who leads a working group in a SAMSI program but receives no financial support.
G – *Graduate Student,* receiving a research assistantship, in the indicated amount, from SAMSI
N – *New Researcher,* receiving the indicated support (salary and fringe benefits) from SAMSI
P – *Postdoctoral Fellow,* receiving the indicated support (salary and fringe benefits) from SAMSI
PA – *Postdoctoral Associates,* receiving the indicted support (salary and fringe benefits or reimbursement of expenses) from SAMSI
T – *Targeted Expert*, an individual with particular expertise that is felt to be needed for progress in key elements of program research. Such individuals are brought in for shorter intervals of time, for transference of expertise to the program participants.
U – *University Fellow*, a key program participant, visiting for a semester or year, whose primary support is via indicated cost-sharing from a partner university.
V – *Core Visitor*, an individual from outside the Triangle who plays a major role in the program activities, by either a lengthy visit to the program or repeated visits involving ongoing program research.
Grey – is used to indicate funds that are provided by partner university cost sharing.

Note: For visitors who have yet to visit SAMSI or who are still at SAMSI, dollar amount in the tables below are the expense allotment for the visitor.

I. Latent Variables for Social Sciences

Name	Gender	Affiliation	Department	Status
Bayarri, Maria Jesus	F	U of Valencia	Statistics & OR	V
Bauer, Dan	M	U of North Carolina	Psychology	FA
Biemer, Paul	M	U of North Carolina & RTI	Odum Institute	FA
Bollen, Ken	M	U of North Carolina	Sociology	F

Clarke, Bertrand	M	UBC/SAMSI	Statistics	U
Dunson, David	M	NIEHS	Biostatistics	FA
Edwards, Lloyd	M	U of North Carolina	Biostatistics	F
Ghosal, Subhashis	M	North Carolina State U	Statistics	F
Glymour, Clark	M	Carnegie Mellon U	Philosophy	T
Gu, Jiezhun	F	North Carolina State U	Statistics	G
Hipp, John	M	U of North Carolina	Sociology	G
Kamata, Akihito	M	Florida State U	Educational Psychology	N
Kinney, Saki	F	Duke U	Statistics	G
Lin, Xiaodong	M	U of Cincinnatti	Mathematics & Statistics	V
Medhin, Negash	M	North Carolina State U	Mathematics	F
Miyazaki, Yasuo	M	Virginia Tech	Education	T
Nguyen, Hoan	F	SAMSI		P
Palomo, Jesus	M	SAMSI		P
Rabe-Hesketh, Sophia	F	U of CA Berkeley	Education	T
Robins, James	M	Harvard U	Biostatistics	DL
Samuels, Jr., Johnny	M	NCSU & SAMSI		G
Skrondal, Anders	M	London School of Economics	Statistics	T
Visser, Ingmar	M	U of Amsterdam	Psychology	PA
Winship, Chris	M	Harvard U	Sociology	T
Zavisca, Jane	F	SAMSI		P

II. Genomes to the Global Health: Computational Biology of Infectious Diseases

Name	Gender	Affiliation	Department	Status
Castillo-Chavez, Carlos	M	Arizona State U	Mathetmatics & Statistics	V
Chakraborty, Arup	M	U of CA Berkeley		T
Cliburn Chan	M	Duke U	Bioinformatics & Computational Biology	N
Cooke, Ben	M	Duke U	Mathematics	G
Cowell, Lindsay	F	Duke U	Biostatistics & Bioinformatics	F
Datta, Sujay	M	Northern Michigan U	Mathematics, Statistics & Computer Science	V
Ellwein, Laura	F	North Carolina State U	Mathematics	G
Elston, Timothy	M	U of North Carolina	Mathematics	F
Fitch, Walter	M	U of California - Irvine	Eco-Evo	T
Fueyo, Joanna	F	IBM	Information Based Medicine	T
Goldstein, Byron	M	Los Alamos National Laboratory	Theoretical Biology & Biophysics	V
Greenwood, Priscilla	F	Arizona State U	Mathematics & Statistics	V
Ickstadt, Katja	F	U of Dortmund	Statistics	V
Kepler, Thomas	M	Duke U	Bioinformatics & Computational Biology	F
Korber, Bette	F	Los Alamos National Laboratory		DL
Miller, Mark	M	Washington U St. Louis	Pathology	T
Lloyd, Alun	M	North Carolina State U	Mathematics	F

Nobel, Andrew	M	U of North Carolina	Statistics	FA
Park, Soyoun	F	U of North Carolina	Statistics	G
Perelson, Alan	M	Los Alamos National Laboratory		DL
Ray, Surajit	M	SAMSI		P
Rodriguez, Abel	M	Duke U	Statistics	G
Root, Morgan	M	North Carolina State U	Mathematics	G
Schmidler, Scott	M	Duke U	Statistics	F
Strohmaier, Karl	M	North Carolina State U	Computer Science	G
Wang, Xiao	M	U of North Carolina	Statistics & Operations Research	G

III. Data Assimilation for Geophysical Sciences

Name	**Gender**	**Affiliation**	**Department**	**Status**
Apte, Amit	M	SAMSI & U of North Carolina	Mathematics	PA
Broecker, Jochen	M	U of Denmark		V
Budhiraja, Amarjit	M	U of North Carolina	Statistics and Operations Research	F
Choudhury, Roy	M	U of Central Florida	Mathematics	V
Clarke, Liam	M	London School of Economics		V
DelSole, Tim	M	George Mason U	Ocean, Land & Atmosphere Studies	T
Dijkstra, Henk	M	Colorado State U	Meterology	T
Foster, Steven	M	U of North Carolina	Mathematics	G
Ghosh, Sujit	M	North Carolina State U	Statistics	F

Hansen, Jim	M	Massachusetss Institute of Technology	Ocean Sciences	V
Herbei, Radu	M	Florida State U	Statistics	T
Ide, Kayo	F	U of California Los Angeles	Atmospheric and Oceanic Sciences	U
Jones, Chris	M	U of North Carolina	Mathematics	F
Judd, Kevin	M	U of Western Australia	Mathematics	V
Kalnay, Eugenia	F	U of Maryland		DL
Khare, Shree	M	SAMSI		P
Kim, Sangil	M	U of Arizona	Mathematics	T
Kyung, Minjung	F	North Carolina State U	Statistics	G
Lawson, Greg	M	Massachusetss Institute of Technology	Ocean Sciences	V
Liu, Liyan	F	U of North Carolina	Mathematics	G
Lozier, Susan	F	Duke U	Earth and Ocean Sciences	F
Mason, Simon	M	Columbia U	IRI	V
McKeague, Ian	M	Columbia U	Biostatistics	T
Mezic, Igor	M	U of CA Santa Barbara	Mechanical Enginnering	T
Mich, Nicole	F	Duke U & SAMSI		G
Mullen, Steve	M	U of Arizona	Atmospheric Sciences	T
Pensky, Marianna	F	SAMSI		U
Restrepo, Juan	M	U of Arizona	Mathematics	V
Reynolds, Carolyn	F	Naval Research Laboratory		V
Smith, Leonard	M	London School of Economics		U

Stephens, Monica	F	Spelman College	Mathematics	N
Szunyogh, Istvan	M	U of Maryland	Meterology	T

3. Summary of Activity Participants *

		Underrepresented Groups		
Activity	**# Participants**	**# Female**	**# African-American**	**# Hispanic**
Data Mining Closing Workshop -- *May 17-18, 2004*	40	8	2	2
PREP Workshop on Mathematics Meets Biology: Epidemics, Data Fitting and Chaos -- *May 26-29, 2004*	30	9	0	1
SAMSI-CRSC Undergraduate Workshop -- *May 31-June 4, 2004*	18	8	5	3
Internet Program Closing Workshop -- *June 25-26, 2004*	25	4	0	2
NPCDS/SAMSI Workshop on the Design and Analysis of Computer Experiments for Complex Systems -- *July 13-17, 2004*	38	3	0	0
SAMSI-CRSC Industrial Mathematical & Statistical Modeling Workshop for Graduates -- *July 26-August 3, 2004*	30	13	2	2
Latent Variables in the Social Sciences Opening Workshop -- *September 11-14, 2004*	124	45	3	1
Computational Biology Opening Workshop -- *September 19-22, 2004*	93	26	2	2
Multiscale Program Closing Workshop -- *September 27-28, 2004*	23	9	0	0
NPCDS/SAMSI Workshop on Data Mining Methodology and Applications -- *October 28-30, 2004*	120	17	1	0
Data Assimilation Program Opening Workshop -- *January 23-26, 2005*	87	26	2	4

CompBio Midterm Workshop on Mathematical Modeling of Infectious Diseases -- *January 31-February 1, 2005*	27	11	0	4
Undergraduate Two-Day Workshop -- *February 18-19, 2005*	38	24	2	1
Latent Variables in the Social Sciences Symposium on Casuality -- *March 29, 2005*	No Formal Registration			
Data Assimilation Mini-Workshop on Lagrangian Data -- *April 12, 2005*	No Formal Registration			
Latent Variables in the Social Sciences GLAMM Seminars -- *April 13-15, 2005*	No Formal Registration			
CompBio Closing Workshop	scheduled May 22-24, 2005: will be reported next year			
SAMSI/CRSC Interdisciplinary Workshop for Undergraduate	scheduled May 30-June 3, 2005: will be reported next year			

* Participant lists for workshops are given in Appendix A

B. Postdoctoral Fellows

This section starts with a brief synopsis of the activities of each postdoctoral fellow and associate, with further details in the following sections. An overview of SAMSI activities and strategies for effective mentoring of Postdocs is given in Section B.1. The mentors for the postdoctoral fellows are given in Section B.2. The midyear activity reports, written by the postdocs that were at SAMSI in the Fall of 2004, appear in Section B.3. Annual activity reports from the postdocs appear in Section B.4. The results from the SAMSI postdoc questionnaire, which is aimed at assessing the quality of the SAMSI Postdoc experience, and at directly soliciting information for improvement, appear in Section B.5. Final reports, which track the post-SAMSI experiences of the 2003-2004 postdocs are in Section B.6. The SAMSI Postdoctoral Fellows, for 2004-2005, with a brief synopsis (details available in Sections B.3, B.4 and B.5 below) of their activities were:

Emily Lada, (at SAMSI for the Spring and Fall of 2004) participated in the Multiscale Model Development and Control Design Program working groups on Control Design and Paradigms for Bridging Scales, and in the program Latent (Hidden) Variable Models in the Social Sciences. She developed a simulation model of nafion, and automated steady state simulation output analysis.

Hoan Nguyen (at SAMSI for the full 2004-2005 year) participated in the Latent (Hidden) Variable Models in the Social Sciences program, and in the program Genomes to Global Health: Computational Biology of Infectious Disease.

Jesus Palomo (at SAMSI for the full 2004-2005 year) participated in the Latent (Hidden) Variable Models in the Social Sciences program.

Surajit Ray (at SAMSI for the full 2004-2005 year) participated in the Latent (Hidden) Variable Models in the Social Sciences program and the program on Genomes to Global Health: Computational Biology of Infectious Disease Data.

Jane Zavisca (at SAMSI for the full 2004-2005 year) participated in the Latent (Hidden) Variable Models in the Social Sciences program.

Shree Khare (at SAMSI for the Spring of 2005) participated in the program on Data Assimilation for Geophysical Systems.

Amit Apte (at SAMSI for the Spring of 2005) participated in the program on Data Assimilation for Geophysical Systems.

1. Overview of Postdoc Activities and Mentoring Strategies

The SAMSI Postdoctoral Fellowship experience has continued to include opportunities for collaboration in the SAMSI spirit of bringing together Statisticians and Applied Mathematicians. These opportunities came during the SAMSI Workshops, during the working groups that met weekly at SAMSI, from the SAMSI courses, and from informal discussions and contacts.

The enhancement of contacts between SAMSI and NISS Postdoctoral Fellows, particularly those participating in different programs, as well as their contact with the SAMSI Directorate, has been very successful. Monthly lunches cover various aspects of academic folklore, i.e. "things every academic should know", of a type that too often doesn't arise in other conversations. A typical format was that the Directors took turns offering their (sometimes rather different) views and experiences, with frequent questions by the postdocs. Topics covered in this context included:

- The job search process, application (what should and should not be included), the selection process, interviews (good and poor strategies), the job offer system.
- The grant process, how to write proposals, how they are reviewed, a comparison of different scientific cultures.
- The publication process, writing papers, how the review process works, writing reviews, editorial decisions, choice of journals, cross-cultural differences.
- The academic promotion process, ranks, tenure, the review system.

The postdoc research presentation part of the former pizza lunch was turned into a biweekly Postdoc & Graduate Student Seminar Series. Graduate students were included as both audience members and speakers. A challenge that arose early on was that because of the rather diverse research projects underway, it was not always easy for people from very different research areas to stay interested. To address this issue, we adapted the format of "practice job interview talks". Because successful job interview talks are able to interest both experts and non-experts, this seemed like an ideal way to both come up with broadly accessible talks, and also to allow the speakers to practice this important skill. The experience was further enhanced by limiting each talk to 50 minutes, which left 10 minutes for discussion of both technical matters, and also presentation. We found that non-experts tended to give very helpful pointers about presentational points.

Effective mentoring of postdoctoral fellows continues to be a central SAMSI goal. A mechanism for ensuring that each postdoctoral fellow had at least two people with whom they could personally discuss any concerns that might come up, was "double coverage" of mentoring assignments. This has been done by assigning both a "scientific mentor" (usually the senior scientist most connected with the research) and an "administrative mentor" (a member of the Directorate, different from the scientific mentor), to each postdoctoral fellow. The mentoring assignments for 2003-2004 are given in Section B.2.

To assess performance of SAMSI in terms of the overall postdoctoral experience, a Postdoctoral Questionnaire was used in March 2004. This was an updated version of last year's questionnaire. The questions and answers from each postdoctoral fellow can be found in Section B.5 below. The single clearest impression from these is that, overall,

the postdoctoral fellows were very happy with their SAMSI experience, and feel that it has given substantial value added to their careers.

As another means of assessing the quality of the SAMSI experience, the Scientific Mentors were asked to comment on each of the Postdoctoral Fellows. These reports are in Section B.6 below. Again the overall impression is very positive. It is clear that the Postdoctoral Fellows have made very important contributions to SAMSI.

In summary, the SAMSI Postdoctoral Program has been generally very successful. The postdoctoral fellows have been making well appreciated contributions to their programs, and been gaining valuable career skills for themselves.

2. Postdoctoral Fellow Mentoring Assignments

Name	Scientific Mentor	Administrative Mentor
Emily Lada	Ralph Smith	H. T. Banks
Amit Apte	Kayo Ide	Y. Truong
Shree Kahre	Chris Jones	A. Karr
Jesus Palomo	David Dunson	J. Berger
Hoan Nguyen	H. T. Banks	A. Karr
Surajit Ray	Tom Kepler	Y. Truong
Jane Zavisca	Ken Bollen	H. T. Banks

3. Postdoctoral Fellows and Associates Mid-Year Activity Reports

These reports were written by each postdoctoral fellow or associate, in December 2004. Reports do not appear for postdocs who joined SAMSI in 2005.

Jane Zavisca

I am participating in four working groups in the Latent Variables in the Social Sciences Program. All groups have begun working on papers for publication.

Categorical Observed Variables: I am co-authoring a paper with a graduate student in Statistics, Saki Kinney, on Bayesian modeling for regression-type models with covariates measured with error and limited dependent variables.

Complex Surveys (Latent Class Analysis Subsection): I am collaborating with Paul Biemer of RTI, and statisticians at the Census Bureau and Bureau of Labor Statistics on an account of rotation group bias in survey responses and the consequences for measurement error in calculating official unemployment rate.

Model Uncertainty: I am collaborating with a number of other SAMSI affiliates on a project on improved approximations to the Bayes factor (BIC-type statistics) for structural equation models.

Multilevel and Structural Equation Models: I presented preliminary results from a study I conducted last year, with one of my PhD advisors (Mike Hout), on changes in life satisfaction in Russia's turbulent economic climate. Group members advised me on multilevel and SEM approaches that would be appropriate for modeling nonlinear trends

in this data. An earlier version of this paper is currently under review at the *American Journal of Sociology*.

In addition to working groups, I presented a paper on latent class analysis of cultural clustering in the United States at the opening workshop for the LVSS program. I have completed a draft of this paper with co-authors from Berkeley, and will continue to work on this in the coming year.

Finally, I have been involved in teaching for the LVSS program. I helped Ken Bollen teach an introductory level class for graduate students on latent variable modeling in Fall 2004. I also recently presented introductory material on statistical modeling in the social sciences at SAMSI's undergraduate workshop. I developed materials using examples from my research on unemployment and on life satisfaction. The workshop was very successful, and I am pleased to have been asked by Dr. Banks to present an extended version of the life satisfaction ("happiness") example at SAMSI's weeklong workshop for undergraduates in June.

Hoan Nguyen

I started at SAMSI in August in the latent variable program. I attended the tutorial and the kickoff workshop of the program and have been involved in the social network working group and the longitudinal working group. For the longitudinal working group, I attend the meetings and maintain the website. In the social network working group, on top of attending the meetings and maintaining the website, I am actively working on modeling the social networks with stochastic differential equations. In the fall of 2004, I audited the latent variable class offered by Ken Bollen.

As for the infectious diseases program, I attended the tutorial/kickoff workshop in the September and the workshop in January/February. I am currently involved in the mathematical modeling of the immune system working group. For that working group, I am attending the meetings and maintaining the website of that working group. I am also auditing the infectious disease class by Alun Lloyd this semester.

Surajit Ray

Along with the development of high throughput genomic data (sequence, DNA array, and proteomics), we have seen a tremendous interest in developing statistical methodology to tackle this kind of high dimensional data. But in the important area of vaccine design, new information and statistical techniques have not yet been utilized. After attending the opening workshop on Computational Biology of Infectious Diseases Program at SAMSI (Sept 2004) and during my discussion with immunologists and vaccine designers over the past few months, it became clear to me that efficient use of existing statistical methodology and development of new statistical tools appropriate for the complex immune system would greatly improve vaccine design. Traditional vaccines are constructed from live attenuated (e.g. MMR) or killed-whole pathogens (e.g. Flu). But, these vaccines have two major drawbacks. First, the degree and longevity of the protective response induced by the dead pathogen can vary among individuals, and in some cases may not even lead to life-long protection. Secondly, the attenuated form of virus itself may revert to the wild-type - strong form - of the virus and cause disease. Also, for certain individuals having weakened immune system the attenuated form itself can cause disease and even, death. Moreover, cultivating the attenuated form is a time

consuming process, so, in case of the emergence of a new pathogen or the evolution of a unexpected mutant of a known pathogen, we may not have enough time to prepare the vaccine.

As a consequence, there is a tremendous interest in developing “peptide-epitope-based" and “DNA-based" vaccines. My research interest is in “peptide-epitope-based" vaccines, where instead of the whole pathogen, we find very short sequences of protein synthesized by the pathogen, which can potentially mimic the process of the whole pathogen, albeit with a concentrated impact and induce the desired immunity. This will provide a general template for designing new vaccines and test the effectiveness of existing vaccines. Moreover, to gain the largest impact of a particular vaccine we may have to target a genetically homogeneous section of the population, and develop different vaccines for each such group. The statistical challenges in this area are enormous. First of all, traditional classification tools have limited application in choosing the right epitopes for the vaccine, as we have very few experimentally verified epitopes which can induce immunity. The other challenge is to efficiently use the related proteomic and genetic information for predicting the right epitopes to be used.

Current Projects:

Prediction of binding affinity of epitopes: This is one of the projects in our working group on Mathematical Genomics for Vaccine Design(MAGVAD).We propose to develop prediction algorithm for a new peptides based on the small set of known binders to MHC molecules and structural and functional similarities with the binders. Moreover, understanding the gene-expression of dendritic cells, we can greatly improve the potential of the epitope-based vaccines triggering proper immune responses. Until now, we have compared existing methods and we are in the process of developing models based on structural binding groves of the MHC molecules. Finally, we want to verify our model based prediction methods on a test data, which is being generated at Duke under Dr. Georgia D. Tomaras and others.

Modeling Proteosomal Activity: This is the second project in our working group on Mathematical Genomics for Vaccine Design(MAGVAD). Before the peptides bind to the MHC molecules it goes through a selection process, which is a process where the long string of proteins gets chopped into small fragments (the peptides). These methods are based on heuristics or neural networks based algorithm which does not provide a clear idea of the actual data generation biological process. Our goal is to include structural information and develop protein specific models of determining the C and N termini in the proteosomal chops.

Model Selection and assessment of Model Uncertainty: In this working group under the social science program we are working on a better approximation of the Bayes factor for model selection which. The most commonly used approximation in the social sciences literature is the BIC (Bayes Information Criterion) which has many short comings especially in the Structural Equations model. First, we are designing model selection criterion where we have low sample size and huge number parameters. Moreover, our proposed methods will be designed to work in the case when the number of parameters

grows with the sample size. Especially, we are designing methods to incorporate effective sample size per parameter as a refinement of our information criterion.

4. Postdoctoral Fellows and Associates Year End Activity Reports

These reports were written by each postdoctoral fellow or associate, in March – April 2005.

Emily Lada

I joined SAMSI in January 2004 as part of the Multiscale Model Development and Control Design program. I participated in the following working groups: Control Design, Paradigms for Bridging Scales, and Ionic Polymer Modeling. Since September 2004, I have also been attending the working group Multilevel and Structural Equation Modeling as part of the Latent Variables program.

I participated in the development of a multiscale modeling approach for the prediction of material stiffness of the ionic polymer Nafion. Traditional rotational isomeric state theory was applied in combination with a Monte Carlo methodology to develop a simulation model of the formation of Nafion polymer chains on a nanoscopic level from which a large number of end-to-end chain lengths were generated. The probability density function of end-to-end distances was then estimated and used as an input parameter to enhance existing energetics-based macroscale models of ionic polymer behavior. Several methods for estimating the probability density function were compared, including estimation using Johnson distributions, Bézier distributions, and cubic splines. Collaborators on this project are Ralph Smith (NCSU), Jessica Matthews (NCSU), LisaWeiland (VA Tech) and Don Leo (VA Tech). Currently we have two papers related to this project under review and a third paper in progress. We are also in the process of extending this methodology to biological applications.

I am also continuing research related tomydissertation involving the development of methods for steady-state simulation output analysis. In particular, I have been focused on refining the wavelet-based spectral method for steady-state output analysis proposed in my dissertation. I have also started to develop a new method for output analysis involving a spaced batch means approach. Collaborators on this project are James Wilson (NCSU) and Natalie Steiger (University of Maine).

During the summer of 2004, I participated in a project to formalize the insurance negotiation process by modeling the adjustment process (as given by the offer/demand ratio differences) as a realization of a non-homogeneous Poisson process. A key component of the modeling process was obtaining a precise estimate of the underlying rate function. Our preliminary work, using a "naïve" aggregated estimate of the process rate based on a particular data set, seems to indicate the general feasibility of our approach but also the existence of possibly many factors affecting the form of non-homogeneity of the process rate function at different stages of the negotiation. In future research we would like to first explore and then compare some different ways of explicitly relating the process rate to claim covariates, such as fractures, wage loss, investigation, and suspicion of fraud (all in the case of auto injuries) of the claim

settlement. Collaborators on this project are Greg Rempala (University of Louisville) and Richard Derrig (OPAL Consulting).

Presentations:

- Multiscale analysis of high-performance materials. 2004 Joint Statistical Meeting, Toronto, Canada, August 2004.
- Monte Carlo simulation of a solvated ionic polymer with cluster morphology. SAMSI Multiscale Closing Workshop, Research Triangle Park, North Carolina, September 2004.
- Performance of a wavelet-based spectral method for steady-state simulation analysis. Invited presentation, 2004 national meeting of the Institute for Operations Research and the Management Sciences, Denver, Colorado, October 2004.
- Monte Carlo simulation of a solvated ionic polymer with cluster morphology. SAMSI Postdoctoral and Graduate Fellow Seminar, Research Triangle Park, North Carolina, October 2004.\
- Performance of a wavelet-based method for steady-state simulation analysis. 2004 Winter Simulation Conference, Washington, D.C., December 2004.

Papers:

- Matthews, J.L., E.K. Lada, L.M.Weiland, R.C. Smith, and D.J. Leo. 2004. Monte Carlo simulation of a solvated ionic polymer with cluster morphology. *Smart Materials and Structures*, in review.
- Weiland, L.M., E.K. Lada, R.C. Smith, and D.J. Leo. 2004. Application of rotational isomeric state theory to ionic polymer stiffness predictions. *Journal of Materials Research*, in review.
- Lada, E.K. and J.R.Wilson. 2004. A wavelet-based spectral procedure for steady-state simulation analysis. *European Journal of Operational Research*, to appear.
- Lada, E.K., J.R.Wilson, N.M. Steiger, and J.A. Joines. 2004. Performance of a wavelet-based spectral procedure for steady-state simulation analysis. *INFORMS Journal on Computing*, to appear.
- Steiger, N.M., E.K. Lada, J.R. Wilson, and J.A. Joines. 2005. ASAP3: A batch means procedure for steady-state simulation output analysis. *ACM Transactions on Modeling and Computer Simulation*, to appear January 2005.
- Lada, E.K., N.M. Steiger, J.R. Wilson, and J.A. Joines. Performance evaluation of a wavelet-based spectral method for steady-state simulation analysis. *Proceedings of the 2004 Winter Simulation Conference*, ed. R.G. Ingalls, M.D. Rossetti, J.S. Smith, and B.A. Peters.
- Steiger, N.M., E.K. Lada, J.R.Wilson, J.A. Joines, C. Alexopoulos, and D. Goldsman. Steady-state simulation analysis using ASAP3. *Proceedings of the 2004Winter Simulation Conference*, ed. R.G. Ingalls, M.D. Rossetti, J.S. Smith, and B.A. Peters.

Activities and Awards:

- In May 2004 Iwas awarded first place in the Institute of Industrial Engineers Pritsker Doctoral Dissertation Competition in recognition of outstanding doctoral dissertation research in industrial engineering.

- I assisted in the organization of the SAMSI workshop for undergraduate students held in June 2004. During the workshop, I gave a presentation on the SAMSI multiscale program, as well as presentations on standard least squares techniques and statistical analysis of the vibrating beam data.
- I participated as a mentor in the SAMSI workshop for graduate students in July 2004.
- I refereed journal articles for *Simulation Modelling Practice and Theory* and *INFORMS Journal on Computing*.
- I served as a session chair at the 2004 Winter Simulation Conference in Washington, D.C.

Jane Zavisca

Primary Working Groups:

A. *Complex Surveys (Latent Class Analysis Subsection):* I am collaborating with Paul Biemer of RTI, and statisticians at the Census Bureau and Bureau of Labor Statistics on an account of rotation group bias in survey responses.
B. *Model Uncertainty:* I am collaborating with a number of other SAMSI affiliates on a project on improved approximations to the Bayes factor (BIC-type statistics), and on a paper on stochastic search algorithms for structural equation models with random effects.

Secondary Working Groups:

C. *Categorical Observed Variables*: I have attended meetings and given group members feedback on working papers on models with covariates measured with error and limited dependent variables. I have also begun planning a paper with a graduate student, Saki Kinney, on Bayesian approaches to the same types of problems.
D. *Multilevel and Structural Equation Models*: I have been regularly attending meetings, and have been advised by group members on how I could apply multilevel and SEM approaches in my substantive research as a sociologist.

In addition to working groups, I presented a paper on latent class analysis of cultural clustering in the United States at the opening workshop for the LVSS program. I have completed a draft of this paper with co-authors from Berkeley, and will continue to work on this in the coming year.

Finally, I have been involved in teaching for the LVSS program. I helped Ken Bollen teach an introductory level class for graduate students on latent variable modeling in Fall 2004. I also presented introductory material on statistical modeling in the social sciences at SAMSI's undergraduate workshop. I developed materials using examples from my research on unemployment and on life satisfaction. The workshop was very successful, and I will present an extended version of the life satisfaction ("happiness") example at SAMSI's weeklong workshop for undergraduates in June.

<u>Complex Surveys (Latent Class Analysis Subsection)</u>

I am collaborating with Paul Biemer of RTI, and statisticians at the Census Bureau and Bureau of Labor Statistics, on an account of rotation group bias in survey

responses, applied to the estimation of unemployment in the Current Population Survey (CPS). Although the unemployment example has been the primary focus of our research, we have also read papers and discussed ongoing work by group members on related topics. For example, we have discussed modeling measurement error when the latent variable of interest is thought to be continuous rather than categorical, with applications to quality of response in surveys of adolescent drug and alcohol use.

Working Paper: "Latent Class Analysis of Rotation Group Bias: The Case of Unemployment." *My Role:* I have specified a list of hypotheses and done preliminary data analysis for a new data set on this topic that we received from the Census Bureau in February. I have also been investigating a new generation of software that could resolve some problems with model convergence with sparse data.

Long-Term Research Agenda: Professor Biemer has encouraged me to continue collaborating with him after I leave SAMSI. Our current work focuses on the cross-sectional re-interview sample; in the future we will apply Markov LCA models to panel data (for which re-interviews are not available).

The methods I have been learning with Dr. Biemer could also be fruitfully applied to a broader set of social science questions. While our focus has been on measurement error, confirmatory LCA modeling could also prove useful in some of my substantive research. For example, when I was a graduate student at Berkeley I served as a research assistant on a book chapter on cultural fragmentation and polarization in the United States. In that work I appled a simple exploratory LCA. We are now planning to do a follow-up paper for a peer-reviewed journal, and I will redo the analysis using more sophisticated latent variable methods.

Current Projects:

<u>Model Uncertainty:</u>

This working group is considering a broad range of issues in model selection and assessment of model fit, with a particular emphasis on Bayesian methods. My primary role has been to help the Bayesian statisticians in the group understand the types of applications, models, and methods that are common in social sciences. I have also focused on assessing in what contexts Bayesian methods could be appealing to social scientists, and whether and how they would realistically be applied, given that most social scientists are more familiar and more comfortable with classical (frequentist) statistical methods and software. I am a co-author on the following two papers:

Working Paper: Bayes Factors in Structural Equation Models (SEMs): Schwarz's BIC and Other Approximations. *My Contribution:* I have helped write some of the introduction motivating the paper, and have also run small-scale simulations and prepared results showing that our new approximations perform better for relatively small samples. I will take the lead on doing a more expansive simulation study with a broader range of simulated data.

Working Paper: "Bayesian Model Selection and Averaging in Structural Equation Models." *My Contribution:* I have developed an empirical example based on some of Ken Bollen's previous work on latent variable models for income trends. I have described the model space and written up a set of decision rules for the range of models that the stochastic search should test. I will run a standard frequentist analysis on this data, as

well as the other examples for the paper, to provide a baseline for comparing the fully Bayesian approach with standard procedures.

I

Future Research: I would like to continue work on Bayes Factor approximations, particularly those which would be relatively easy for the practitioning social scientists to apply. As a first step, I would like to extend the research on ABF's in the SEM context to a broader range of models. Participating in this group has gotten me very interested in issues of model selection and fit, and the relationship between theoretically- and stochastically-driven model design and selection approaches. I expect this will be a core area of methodological research for me in the long term.

List of Publications and Work in Progress:

- "Latent Class Analysis of Rotation Group Bias: The Case of Unemployment" (with Paul Biemer and Bac Tran) American Sociological Association Section on Methodology 2005 Annual Meeting, April 2005 (Chapel Hill).
- "Bayes Factors in Structural Equation Models (SEMs): Schwarz's BIC and Other Approximations." (with Ken Bollen and Surajit Ray) American Sociological Association Section on Methodology 2005 Annual Meeting, April 2005 (Chapel Hill)
- "Bayesian Model Selection and Averaging in Structural Equation Models." (with David Dunson and Jesus Palomo). Target journal: *Psychometrik*
- "Does Money Buy Happiness in Unhappy Russia?" (with Mike Hout). Under review at *American Journal of Sociology*

Surajit Ray

Along with the development of high throughput genomic data (sequence, DNA array, and proteomics), we have seen a tremendous interest in developing statistical methodology to tackle this kind of high dimensional data. But in the important area of vaccine design, new information and statistical techniques have not yet been utilized. After attending the opening workshop on Computational Biology of Infectious Diseases Program at SAMSI (Sept 2004) and during my discussion with immunologists and vaccine designers over the past few months, it became clear to me that efficient use of existing statistical methodology and development of new statistical tools appropriate for the complex immune system would greatly improve vaccine design. Traditional vaccines are constructed from live attenuated (e.g. MMR) or killed-whole pathogens (e.g. Flu). But, these vaccines have two major drawbacks. First, the degree and longevity of the protective response induced by the dead pathogen can vary among individuals, and in some cases may not even lead to life-long protection. Secondly, the attenuated form of virus itself may revert to the wild-type - strong form - of the virus and cause disease. Also, for certain individuals having weakened immune system the attenuated form itself can cause disease and even, death. Moreover, cultivating the attenuated form is a time consuming process, so, in case of the emergence of a new pathogen or the evolution of a unexpected mutant of a known pathogen, we may not have enough time to prepare the vaccine.

As a consequence, there is a tremendous interest in developing "peptide-epitope-based" and "DNA-based" vaccines. My research interest is in "peptide-epitope-based"

vaccines, where instead of the whole pathogen, we find very short sequences of protein synthesized by the pathogen, which can potentially mimic the process of the whole pathogen, albeit with a concentrated impact and induce the desired immunity. This will provide a general template for designing new vaccines and test the effectiveness of existing vaccines. Moreover, to gain the largest impact of a particular vaccine we may have to target a genetically homogeneous section of the population, and develop different vaccines for each such group. The statistical challenges in this area are enormous. First of all, traditional classification tools have limited application in choosing the right epitopes for the vaccine, as we have very few experimentally verified epitopes which can induce immunity. The other challenge is to efficiently use the related proteomic and genetic information for predicting the right epitopes to be used.

Current Projects:

Prediction of binding affinity of epitopes: This is one of the projects in our working group on Mathematical Genomics for Vaccine Design(MAGVAD).We propose to develop prediction algorithm for a new peptides based on the small set of known binders to MHC molecules and structural and functional similarities with the binders. Moreover, understanding the gene-expression of dendritic cells, we can greatly improve the potential of the epitope-based vaccines triggering proper immune responses. Until now, we have compared existing methods and we are in the process of developing models based on structural binding groves of the MHC molecules. Finally, we want to verify our model based prediction methods on a test data, which is being generated at Duke under Dr. Georgia D. Tomaras and others. Currently we are exploring the performance of existing classification methods e.g. Support Vector Machine, Boosting and Bagging, Random forest and fine tuning them with biological and chemical properties of the protein.

Modeling Proteosomal Activity: This is the second project in our working group on Mathematical Genomics for Vaccine Design (MAGVAD). Before the peptides bind to the MHC molecules it goes through a selection process, which is a process where the long string of proteins gets chopped into small fragments (the peptides). These methods are based on heuristics or neural networks based algorithm which does not provide a clear idea of the actual data generation biological process. Our goal is to include structural information and develop protein specific models of determining the C and N termini in the proteosomal chops.

Model Selection and assessment of Model Uncertainty: In this working group under the social science program we are working on a better approximation of the Bayes factor for model selection which. The most commonly used approximation in the social sciences literature is the BIC (Bayes Information Criterion) which has many short comings especially in the Structural Equations model. First, we are designing model selection criterion where we have low sample size and huge number parameters. Moreover, our proposed methods will be designed to work in the case when the number of parameters grows with the sample size. Especially, we are designing methods to incorporate effective sample size per parameter as a refinement of our information criterion.

Here I describe the paper titled “Bayes Factors in Structural Equation Models (SEMs): Schwarz’s BIC and Other Approximations” which is joint work with Dr Ken

Bollen and Dr. Jane Zavisca. Model fit and comparisons are subjects of much debate in the Structural Equation Models (SEMs) literature. Researchers typically apply likelihood ratio tests and numerous fit indices to assess the adequacy of a model's fit. Schwarz's (1978) Bayesian Information Criterion or BIC is one such measure of fit. The BIC measure is an approximation to the Bayes Factor. The Bayes Factor is $B_{12}=Pr(\mathbf{D}|H_1)/Pr(\mathbf{D}|H_2)$ where $Pr(\mathbf{D}|H_k)$ is the probability of the data (**D**) if hypothesis or model H_k is true. The BIC measure is the best known approximation to the Bayes Factor in SEM. However, the BIC is derived under simplifying assumptions that permit its calculation without explicit prior probabilities. Furthermore, the BIC derives from other approximations to make it simple to apply in SEMs. It is possible to develop other estimates of the Bayes Factor that make use of fewer approximations than the BIC and thus hold the potential to be more accurate. In this paper, we develop two such approximations, Approximate Bayes Factor 1 and 2, or ABF1 and ABF2. The paper provides the rationale for the BIC, ABF1, and ABF2, discusses their calculations using standard SEM software, illustrates and compares these measures for simulation examples, and finally discusses the evidence in favor of or against these approximations to the Bayes Factor. Our position is that the Bayes Factor could be a useful addition to the SEM literature, yet we need to evaluate the quality of measures that approximate it. We conclude with recommendations for the researcher.

Publications and Working Papers:

- Ray S. and Lindsay B.G. The Topography of Multivariate Normal Mixtures. (2005) (To Appear in the: *The Annals of Statistics)*
- With Kepler T., Cowell, L, Nobel, A., Schmeidler S., Classifying MHC-I binding epitopes based on amino acid proerties.
- With Marron, J. S., Feature selection of high dimensional low sample size data.
- With Bollen, K. and Zavisca, J, Approximation to Bayes factor for SEMs
- With Lindsay, B., Chen, S., Markatou M., Ray, S., Yang, K., Diffusion kernels and quadratic distances as building blocks for high dimensional inference.
- With Lindsay B.G. On using Quadratic Risk for high-dimensional Model Selection
- With Lindsay, B., Chen, S., Yang, K., Spectral Degrees of freedom and highdimensional smoothing.
- With Lindsay B.G. The Topography of Multivariate Normal Mixtures. II

Workshops and Professional Conferences:

1. International Conference on the Future of Statistical Theory, Practice and Education, December 29, 2004 - January 1, 2005, Indian School of Business, Hyderabad, India.

*Invited Talk***:** Selecting the Number of Components in a Finite Mixture: A Risk-based Approach

2. American Sociological Association Section on Methodology: 2005 Annual Meeting, Chapel Hill, April 22, 2005.

Invited Talk: Bayes Factors in Structural Equation Models (SEMs): Schwarz's BIC and Other Approximations.

3. Joint Annual Meeting of the Interface and the Classification Society of North America: Theme: Clustering and Classification Washington University School of Medicine St. Louis, Missouri June 8, 2005 - June 12, 2005
Invited talk: Selecting the Number of Components in a Finite Mixture: A Risk-Based Approach
Organizing/Chairing the Session on: Model-based clustering/classification in high-dimensional data.
4. Annual Meeting of the Statistical Society of Canada 12-15 June 2005 at the University of Saskatchewan
Invited talk: On using Quadratic Risk to Select High dimensional Mixture Model
5. WNAR/IMS annual meeting, Fairbanks, Alaska June 21-24, 2005
Invited Talk: Classification of MHC-I binding epitopes
Chairing session on: Challenges in Clustering/Classifying Data in Real Time

Jesus Palomo

I am actively involved in one working group that has since split in two. Both subgroups are focused on model selection in the structural equation models (SEM) framework. However, the first one focuses on Bayesian formulation of SEMs and Bayesian model selection, whereas the other focuses on generalizing the Bayesian Information Criterion (BIC). Besides these working groups, I have been attending some of the other groups, mainly the multilevel and structural equation modeling working group, lead by Dr. Bollen, and the categorical observed variables working group, lead by Dr. Bauer. I have also attended the course taught by Dr. Bollen ``An Overview of Latent Variable Models in the Social Sciences", and the Generalized Lineal Latent and Mixed Models (GLLAMM) course taught at NISS.

The activities I have developed so far are:

- Maintaining the web page of the model selection in the structural equation models working group: http://http://www.samsi.info/200405/socsci/workinggroup/lvmds/
- Collaborating in the logistics of the Kick-off and Mid-Term workshops.
- Searching for references on SEMs and applying the developed models and discussions in the working group to real sociological data (provided by Dr. Bollen) and biomedical data (provided by Dr. Dunson).
- Providing a Bayesian approach for SEMs. These novel methods have been written up, with Dr. Dunson and Dr. Bollen, as a book chapter that will be submitted for publication shortly. The abstract of the chapter is as follows: Structural equation models (SEMs) with latent variables are routinely used in social science research, and are of increasing importance in biomedical applications. Standard practice in implementing SEMs relies on frequentist methods. This chapter provides a simple and concise description of an alternative Bayesian approach. We provide a brief overview of the literature, describe a Bayesian specification of SEMs, and outline a Gibbs sampling strategy for model fitting. Bayesian inferences are illustrated through an industrialization and democratization case study from the literature. The Bayesian approach has some distinct advantages, due to the availability of samples from the joint posterior distribution of the model parameters and latent variables, that we highlight. These posterior samples provide important information not contained in the measurement and structural parameters. As is illustrated using the case study, this information can often provide valuable insight into structural relationships.

- Providing Bayesian model selection and averaging methods in SEMs. These have been written up, with Dr. Dunson and Dr. Zavisca, in a paper that will be submitted to Psychometrika by the end of the semester. The abstract of the paper is as follows: In fitting SEMs, there is typically uncertainty about various aspects of the model, such as whether to include certain cross loadings or structural parameters. Ideally, one could obtain inferences about the structural relations of interest, which are robust to misspecification in other aspect of the model. One approach would be to pre-specify a list of plausible models, fit all models in the list, and then base inferences on the final selected model. This approach does not account for uncertainty in the model selection process. This article instead considers a Bayesian model averaging approach. Each of the models in the list are assigned a prior probability, and conjugate priors are chosen for the included parameters. A stochastic search Gibbs sampling algorithm is developed for posterior computation. This approach, which can be used even when the list of models is very large, is contrasted to methods based on BIC approximations. The methods are illustrated using simulated and real data examples.
- Applying the proposed Bayesian approach to a common case study in the literature: ``Assessing Reliability and Stability in Panel Models" (1977) by Wheaton, B., Muthén, B., Alwin, D. and Summers, G.F. in Sociological Methodology. This case study has been reviewed several times from the frequentist viewpoint. Dr. Dunson and I have found that the Bayesian approach used in "Bayesian Estimation and Testing of Structural Equation Models" (1999) by Scheines, R., Hoijtink, H. and Boomsma, A. in Psychometrika, has several issues that can be critiqued and extended using our approach. We are currently working on a paper that will be submitted to Psychometrika by the end of the semester.
- Providing a semiparametric approach to SEMs where the linear assumptions have been relaxed. Dr. Dunson and I are currently working on this approach and we have written a draft with the following abstract: Structural equation models (SEMs) provide a general framework for modeling of multivariate data, particularly in settings where observable variables are designed to measure one or more latent variables. When implementing SEM analysis, it is typically assumed that the model structure is known, and that the latent variables have normal distributions. To relax these assumptions, this article proposes a semiparametric Bayesian approach. Categorical latent variables with an unknown number of classes are accommodated using Dirichlet process (DP) priors, while DP mixtures of normals allow continuous latent variables to have unknown distributions. Robustness to the assumed SEM structure is accommodated by choosing mixture priors, which allow uncertainty in the occurrence of certain links within the path diagram. A Gibbs sampling algorithm is developed for posterior computation. This methodology is illustrated using biomedical and social science examples.
- Developing a Bayesian statistics software for Objective Bayesian Analysis under the supervision of Dr. Berger. I am developing the software package 'OBayes' that allows for implementation of the procedures that are developed under the project entitled Objective Bayesian Analysis. It involves both software engineering and the developing of computational engines involving simulation and Markov Chain Monte Carlo procedures.
- Participating on the NISS project ``Statistical Framework for Evaluation of Complex Computer Models". My contributions as part of the team have been to develop methods for computer model validation, to write up the final report, and to implement the developed methodology as a software package called SAVE-2 along with its documentation.

The following presentations related with my activities at SAMSI occurred during this period:
- Computer Model Validation with Functional Output, Bormio (Italy), January 12-14 2005. Second IMS-ISBA Joint Meeting.
- SAVE-2, Simulator Analysis and Validation Engine, at NISS Affiliates and NISS/SAMSI University Affiliates 2005 Annual Meeting. March, 3rd 2005, Durham, NC (USA).
- Bayesian Structural Equation and Latent Variable Models, Varenna (Italy), June 2-4 2005, Bayesian Inference in Stochastic Processes (BISP4).
- Bayesian semiparametric structural equation models with latent variables. International Conference on Statistics, June 20-24, 2005. Hong Kong.

Regarding my private research, I have submitted three papers to the following journals: the American Statistical Association, the Operations Research, and the Applied Stochastic Models in Business and Industry. They are currently under revision and hopefully I will have an answer from the editor soon.

Additionally, I have written, with other co-authors, two book chapters: "Simulation in Industrial Statistics", in Applying Statistical Methods in Business and Industry", published by Wiley, 2005; and "On combining expertise in Dynamic Linear Models, in Multiple Participant Decision Making, J. Andrysek, M. Karny and J. Kracik eds. Advanced Knowledge International, 2004.

During this period I have been refereeing papers for the following journals: Bayesian Analysis Journal, Applied Stochastic Models in Business and Industry Journal, Journal of the American Statistical Association.

I have the following publications in progress:
- Uncertainty models in auctions processes, 2005. (with H. Hong, Duke University)
- Validation of Complex Computer Models with Multivariate Functional Outputs, 2005. (with J.O. Berger and G. Garcia-Donato).
- Validation of Complex Computer Models, 2005. (with J.O. Berger, M.J. Bayarri, J. Sacks, G. Garcia-Donato).
- Bayesian Semiparametric Structural Equation and Latent Variable Models, 2005. (with D. Dunson).

Finally, this year I have been awarded with the European Network for Business and Industrial Statistics (ENBIS) 2005 Young Statistician Award; and I have been nominated finalist of the Leonard J. Savage Dissertation Award. The winner will be announced in August 2005 at the Joint Statistical Meeting in Minnesota. Also, I have been awarded a grant to attend the Second IMS-ISBA Joint Meeting in Bormio (Italy).

Hoan Nguyen

Over the past year, I am mainly involved with the latent variable and infectious disease programs. On top of attending the workshops and the SAMSI class that are held for each program, I also participate in the working group. For the social network working group in the latent variable program, I am actively working on a model of the social networks using stochastic differential equations. Our group includes H.T. Banks, Alan

Karr, John Samuels and me. We believe that we have great preliminary results to prepare a paper and pursue the problem further. In fact, we are working on having a paper finished by early summer. Another excellent experience I have with this project is working with John Samuels, a graduate student from NCSU on this problem. I have gained more leadership and people skills from this experience.

As for the infectious disease program, I am attending mathematical modeling of the immune system leading by Alun Lloyd. Since we started meeting in January, we have been reading and discussing literatures on mathematical models of the immune system. With this experience, I have gained a tremendous knowledge about the immune system and the active research in the field. It is very beneficial with my current project at Center of Research in Scientific Computing (CRSC) working with oral viruses.

Since I am interacting with CRSC at NC State University since August, I spend 2-3 days a week at CRSC. I am happy with the arrangements there as well as the research project I am working on. At the moment, I am in training to be involved with the undergraduate workshop in the summer.

I am very satisfied with the mentoring at SAMSI; I have constant mentoring about research and advices on my career from Dr. H. T. Banks, my scientific mentor. When comparing my SAMSI experience to my government lab experience, I have more freedom on picking my topic research. In addition, I have more mentoring here at SAMSI in both scientific and career area. As if I was at a university setting, I probably would not be able to dedicate all my time to research since I might have to do some teaching.

While I have been at SAMSI, I presented at two conferences and will be presenting at another conference in the summer. I have finished two papers from my dissertation. The papers are at the final stage of editing and will be sent off to a journal by the end of the semester. In addition, H. T. Banks and I submitted a paper to Journal of Mathematical Analysis and Applications in February on Sensitivity of Dynamical Systems to Banach Space Parameters.

Overall, I really enjoy my experience at SAMSI on all levels. I feel more mature and well-rounded as a researcher. The staffs, post-docs, visitors, directors, and everyone else have made it a very friendly and exciting atmosphere to work at. I would have chosen to come to SAMSI all over again.

Amit Apte

The main focus of my activities has been data assimilation in ocean models and the activities of the “Low-dimensional behavior” group. A summary of the latter (to be presented on April 8, 2005 in the “NEDD special” seminar) is attached at the end of this report. As part of the Lagrangian data assimilation group, I have presented the work by Molcard et al. and compared it to the method by Ide et al. I am also working on the ocean model HYCOM (HYbrid Coordinate Ocean Model) with the eventual goal of implementing the Lagrangian data assimilation scheme of Ide et al. for the Gulf of Mexico using this model. Currently I have preliminary results from the model and am assessing them by trying to compare them with some benchmarks.

Summary of the Low-dimensional behavior group

1. Goals and obstacles

Inspired and intrigued by the work of Kalnay, Ott, et al., the goal of this group was to understand how very large dimensional systems exhibit behavior that can be effectively "captured" (or "explained") by low-dimensional "phenomena." Indeed, one of the goals was to really pin down or define what is meant by "lowdimensional behavior." In that sense, the group started off with an undefined or ill-defined or ill-understood term and hoped to make it well-defined or better understood! This is also the major obstacle we are trying to overcome. The hope was (and still is) that an understanding of various aspects of such low dimensional behavior could lead to a more efficient scheme for data assimilation. One aspect of our goal was to critically assess the previous work (e.g. by Ott et al. and by others).

2. Activities

The first few meetings focused on the dynamical systems motivated understanding of different ensemble generation schemes - singular value decomposition (SVD), bred vectors, Lyapunov vectors, etc. Juan started us off with an introduction (flavored by the exposure in Carl Wunsch's book) to SVD in a very general setting. Amit related that (using the articles by Legras and Vautard and by others) to the specific linearized dynamics used in ensemble generation. After having covered these basics, we discussed the relation between and shortcomings and advantages of these different schemes. A concurrent related discussion focused on the important (growing) directions and how these are captured by the ensembles. This naturally lead to the work on local ensemble Kalman filter, which is the focus for the past couple of weeks.

3. Some Conclusions and future directions

We managed to identify that there are (at least) three different "low-dimensional behaviors" which sometimes get mixed up.

- Phenomena that are local in the state space
- Phenomena that are local in the physical space (spatial localization, e.g. in local ensemble KF). This is simply a restricted class of the above.
- Low-dimensionality "in the covariance matrix." These are two different subcategories of this aspect:
  - Low rank (say k_D, where D is the dimension of the state space) of the "true" covariance matrix
  - Low rank (say k0) of the covariance matrix obtained from the ensemble of size N. If we know that k0 > k (of course, k0 < N), then at least there is a hope that the ensemble of size N will be "good enough" for capturing the growing directions. No result in this direction has been found yet and we are actively discussing this in current meetings. We are also trying to understand (or find a counterexample) that, even if k0 > k, why the true k growing directions are "contained" within the k0 growing directions indicated by the ensemble covariance matrix.

Shree P. Khare

1. Workshops/conferences/seminars attended

- SAMSI tutorials and kickoff workshop, January 23-26th, 2005, RTP
- IPAM workshop (issues, challenges and inter-disciplinary perspectives),

February 22-26th, 2005, UCLA
- Weekly lecture series on data assimilation at N.C. State University

2. Working groups report

As part of the geophysical data assimilation program, I have mainly been participating in the model inadequacy/empirical verification and low-dimensional behavior working groups. On occasion, I attend the Lagrangian assimilation and ocean data assimilation groups.

I have been actively involved in the model inadequacy/empirical verification group. This working group has focused on a subject fundamental importance to data assimilation. The ultimate goal of data assimilation is to achieve ensembles of state estimates which are perfectly accountable in a probabilistic sense. To define the problem from first principles, we started by defining the problem in both the perfect and imperfect model scenarios. Even if we assume that the dynamical model is perfect, comparing data assimilation algorithms to the perfect Bayesian solution (for ensembles) is a difficult task. In fact, we have discovered (through much debate and discussion) that there may not be a perfectly consistent way of comparing ensemble data assimilation algorithms. However, we have come up with a number of reasonable methods for comparing assimilation which are sensible for the types of systems and algorithms we are considering.

As a result of the discussion in this working group, I have become actively involved in a project related to comparison of various ensemble data assimilation schemes in both the perfect and imperfect model scenarios. The ultimate goal of this project is to study and understand the performance of various assimilation schemes in the context of model errors. One of the ensemble assimilation algorithms being tested claims to have a systematic way of handling model error and may perform better in the imperfect model scenario than current state of the art algorithms. The goal of this work is to write a short paper reporting these results. At this point, with more than two months left in the program, it is too premature to speculate where such a paper will be submitted. The goal now is to have enough research done to start writing the paper after the program is completed.

5. Postdoctoral Fellows Evaluations of SAMSI

In this section, the postdoctoral fellows respond to the questions in the SAMSI Questionnaire. This was designed with input from the postdoctoral fellows during NISS/SAMSI Postdoctoral Fellow Lunches. The questions asked were as follows:

1. Which SAMSI program have you been involved with, and at what level? (e.g. “been doing active research on …” or “just went to a tutorial/workshop to learn about that area which is new to me”).
2. Describe your interactions with other institutions, e.g. a Triangle university, NISS or CRSC.
3. What have been the high points of your SAMSI experience?
4. How could your SAMSI experience have been improved?

C. Graduate Student Participation

I. Data Mining and Machine Learning

Jeongyoun Ahn (Statistics, NCSU) played an important role in the feature selection cluster of the support vector machine working group. She always attended the weekly meetings, participated group discussions actively, and often brought up some fresh ideas to the cluster. Jeongyoun is a main author of two publications finished in the DMML program and has contributed substantially to both works. She has been involved in model formulations, program coding, and real data analysis. In the mid-year workshop, Jeongyoun gave a very nice presentation on the proposed work. Furthermore, she successfully established a research connection between our cluster and a medical research group led by Dr. C. M. Perou from UNC-Chapel Hill, and the DMML working group benefited from this collaboration in terms of receiving valuable microarray data sets and expertise suggestions from the biologists.

Jeongyoun Ahn has benefited from the DMML program in various aspects such as more independent thinking and a stronger ability for research and seeking collaborations. Compared with other graduate students, she shows an amazing academic maturity and is well-prepared for an academic position though she was only a third-year student. In the long run, I believe, the SAMSI provides a very stimulating atmosphere for all level researchers especially healthy for junior researchers, and great mentoring opportunities for senior graduates like Jeongyoun.

Atina Brooks (Statistics, North Carolina State) was a second-year master's student in the Department of Statistics at North Carolina State University when she started duties as a SAMSI graduate fellow. Her responsibilities included attending weekly administrative meetings, attending weekly technical meetings for the Bioinformatics Working Group, active involvement in research endeavors, technical presentation at the DMML Mid-year Workshop for the Bioinformatics Working Group, technical presentation at the DMML Closing Workshop, and assistance with the CSRC Undergraduate Workshop during summer 2004. Atina's research during the DMML year focused on using support vector machines to identify multiple mechanisms of activity among compounds being studied for the purpose of identifying leads for drug discovery. Her dissertation work continues to investigate the properties of support vector machines for modeling multiple mechanisms. Additionally, she has initiated research using other data-mining techniques, such as specialized recursive partitioning software, for building predictive models for drug discovery. Participation in the DMML year was very helpful to Atina and it has already had a long-term impact in terms of her dissertation topic.

Jen-hwa Chu (Statistics, Duke) was a second-year graduate student during the data mining year. He worked closely with Bertrand Clarke on large-p/small-n problems, and they have written two papers together that have been submitted as chapters for a monograph on data mining. Jen-hwa is planning to do his thesis in this area. Additionally, Jen-hwa worked on a problem in applied data mining related to how option patterns in GM automobiles are related to the time between manufacture and sale, and

gave a professional talk on this application at the Spring Research Conference in Gaithersburg, MD in 2004.

Leanna House (Statistics, Duke) was a fourth-year graduate student during the data mining year. She co-authored two papers with David Banks on the topic, both of which have appeared, and she was co-editor of a collected proceedings volume on the topic of data mining. She gave invited talks on her research at the Quality and Productivity Research Conference in 2004 and at the International Federation of Classification Societies meeting in 2004. Additionally, she co-authored a paper on data mining in metabolomics that appeared in a refereed proceedings volume, and is doing her doctoral research on the related topic of data mining in proteomics.

Balaji Krishnapuram (Electrical Engineering, Duke) was a fourth-year student in Electrical and Computer Engineering when the year began, and he completed his Ph.D. at the end of the year. He has written eight papers in data mining, mostly with people outside of SAMSI, and is now working for Siemens Medical Solutions. As part of the SAMSI data mining year he gave an invited talk at the meeting of the International Federation of Classification Societies and was the discussant for a special guest lecture by Vladimir Vapnik.

Fei Liu (Statistics, Duke) initially planned to work on GM machine down time data in the DMML program, but that data set was not available, so she switched to IBM data. Fei developed an algorithm for mining temporal patterns for system data based on minimum description length principle. The results were presented at DMML final workshop, Spring Research Conference (NIST, MD), and IBM T. J. Watson research center (Hawthorne, NY).

Peng Liu (Statistics, North Carolina State) was involved in the support vector machines (SVM) working group. He has worked on compactly supported SVMs (a paper listed below is currently under review) and on spatio-temporal data mining. He has also made several presentations to the working group about SVM-related topics. The DMML program has obviously influenced Peng's interest in statistics. He is currently working on his PhD thesis with Dr.\ Stefanski at NCSU on a topic related to data mining for pharmaceutical problems.

Katja Remlinger (Statistics, North Carolina State) was a third-year Ph.D. student in the Department of Statistics at North Carolina State University when the SAMSI DMML year began. Although she was not a SAMSI fellow, Katja was a major participant in the DMML year. She was a presenter at the DMML Opening Workshop and at a Special Contributed Session at the 2004 Joint Statistical Meetings that showcased the DMML program year. Katja also attended weekly technical meetings for the Bioinformatics Working Group. She also contributed to the group project that investigated the properties of leave-one-out-cross-validation. Participation in the DMML year was very helpful to Katja. She defended her dissertation immediately after the closing workshop.

Eric Vance (Statistics, Duke) was a second-year student graduate student during the SAMSI data mining year. He worked with Bertrand Clarke on generalizations of the LASSO and with Feng Liang and Merlise Clyde on overcompleteness. His thesis work will probably entail the application of data mining methods to social networks, using data from elephant herds.

Ke Zhang (Statistics, NCSU) was a first-year Ph.D. student in the Department of Statistics at North Carolina State University when the SAMSI DMML year began. He attended the DMML Opening Workshop and participated in weekly technical meetings for the Bioinformatics Working Group. He also made a presentation at the DMML Mid-year Workshop for the Bioinformatics Working Group. One-half of Ke's dissertation is focused on developing a particular data-mining tool (recursive partitioning) for identifying multiple mechanisms in a drug discovery set. The DMML program year was very helpful in providing stimulating discussions. Ke is expect to graduate during the 2005-2006 academic year.

II. NETWORK MODELING FOR THE INTERNET

Arka Ghosh, (Statistics & Operations Research, UNC), supported by SAMSI for the full 2003-04 year. Participated in all program workshops. Active member of Multifractional Brownian and Stable Motion, Semiexperiments – Formal Testing, SiZer and Wavelets, and Heavy Traffic working groups. Presented a Postdoc – Grad Student Seminar.

Felix Hernandez Campos, (Computer Science, UNC), supported by SAMSI for the full 2003-2004 year. Participated in all program workshops. Active member of Comparison of Hurst Parameter Estimators, Semiexperiments – Look and See, Testbeds – Lab Experiments working groups. Performed critical data base work, providing data for entire program.

Myung Hee Lee, (Statistics & Operations Research, UNC), supported by SAMSI for the Spring of 2004). Performed research on periodicities of flow arrival times within documents, and of round trip times. Started preliminary micro-array work, leading into Computational Biology Program next year.

Chuan Lin, (Operations Research, NCSU), not supported by SAMSI. Participated in Workshop on Heavy Traffic and Congestion Control. Active Member of Heavy Traffic working group.

Juhyun Park, (Statistics & Operations Research, UNC), supported by SAMSI for the 2003-04year. Participated in all program workshops. Active member of Changepoints and Extremes, Suite of Models, Comparison of Hurst Parameter Estimators, SiZer and Wavelets, and Heavy Traffic working groups, lead the Semiexperiments – Look and See working groups. Presented a Postdoc – Grad Student Seminar. Co-author on 2 papers in progress.

III. Multiscale Model Development and Control Design

Laura Ellwein (Mathematics, NCSU): Abnormal cerebral blood flow can be an indication of cerebral vascular disease. Laura's research is concerned with the simulation of cerebral blood flow during postural change from sitting to standing. A system of ordinary differential equations is used for a nine-compartment model representing the systemic arteries and veins in the upper extremities, lower extremities, brain, and heart. The system is solved with Matlab using steady-state parameters found in literature for initial values. Optimal control is used to adjust these parameters such that the model can be made to fit measured blood flow and pressure data. Physiologically based control mechanisms are used to describe how arterial and cerebral blood pressure drop during the posture change. Future analysis will include the addition of time delays to analyze the timing of the onset and the duration of the control. In addition, we plan to investigate the effect of replacing one or several compartments with a more detailed one-dimensional (with one spatial dimension) fluid dynamics model based on partial differential equations.

Whereas Laura's research focused on a biological application, it encompassed a number of multiscale and control facets. There was also inherent uncertainty which required analysis similar to that under investigation for materials. Laura additionally participated in the SAMSI Program on Computational Biology of Infectious Diseases in the Fall of 2004.

Jon Ernstberger (Mathematics, NCSU): Jon's research focused on the development of distributed parameter models (partial differential equation or PDE) for smart material uniforms operating in nonlinear and hysteretic regimes. In the first step of the model development, nonlinear constitutive relations were constructed by combining energy principles at the mesoscopic level with stochastic homogenization techniques to construct macroscopic models which characterize the hysteresis inherent to the materials. PDE based on these constitutive relations were constructed through conservation of force and momentum. The final step of the initial component of the investigation focused on the approximation of these PDE through Galerkin techniques to obtain finite-dimensional, vector-valued systems appropriate for simulation. The next step of this analysis will focus on the extension of these models and approximation techniques to more complex geometries (e.g., plates and shells) and alternative materials including magnetic compounds and shape memory alloys.

This research program fit very naturally within the context of the multiscale program since it encompassed aspects of multiscale model development and approximation for advanced material systems. The models developed in this investigation will employed for model-based control design for applications ranging from flow control to remote lens cleaning.

Joe Lucas (Statistics, Duke): Joe participated in the Multiscale Modeling Program at SAMSI. As part of this work, he attended the opening Workshop as well as the Tuesday evening class for which he Latexed Eric Vance's notes. These were being provided through Jon Ernstberger to Ralph Smith to provide a written document for the course. Also, he participated in the ``Paradigms for Multiscale Modeling" working group, for which he attended meetings and kept up with the reading.

Joe is studying the uses of partition of unity in simplifying local regression smoothing. Given the proper choice of weight functions, it can be shown that it is equivalent to splines. He wants to show that the technique can be arbitrarily close to loess as well (given a different choice of weight functions).

Joe contributed strongly to both the working group and class. While coming from a statistics background, he obtained a firm grasp of the deterministic energy relations used to construct nonlinear constitutive relations at the mesoscopic level. He is presently investigating the constructing of stochastic materials models based on inference principles.

Eric Vance (Statistics, Duke): Eric's involvement in the multiscale program consisted of being a part of the working group which is investigated paradigms for bridging scales. He attended the weekly working group meetings and read various articles concerning the items that were discussed. Related to this work is the SAMSI course on Multiscale Model Development in which he took notes for Joe Lucas to convert to Latex, and kept up with the readings and the homework assignments. Also, as part of his involvement with SAMSI, he continued research related to his work last semester in the {\bf Data Mining and Machine Learning} program. Eric's specific research includes developing a model of multiple latent factors in order to explain the covariance within metabolites for diseased and non-diseased individuals. This work spawned from the ``Large P small n" working group because the number of metabolites is larger than the number of individuals or data points observed.

Eric also contributed strongly to both the working group and class. While his research does not directly pertain to multiscale model development or control designs, it provided perspectives which may prove advantageous for constructing hierarchical models for advanced materials.

IV. Computational Biology of Infectious Disease

Ben Cooke (Mathematics, Duke) has participated in the weekly meetings of the working group Mathematical Genomics for Vaccine Design (MagVad). He has been involved in our detailed survey of the literature on mathematical, statistical, and biological aspects of epitope presentation and recognition by the immune system. Throughout the year, Ben has also done work on several aspects of the epitope binding problem. He has produced statistical and graphical analyses of publicly available databases of MHC binding peptide sequences. He has also obtained amino acid "feature sets" for a collaborative effort with SAMSI postdoc Surajit Ray on applying modern classification algorithms (random forests, CART with boosting, SVMs) to the prediction of MHC I peptide binding, which will result in a publication from the working group. Finally he has begun applying multiple structure alignment tools developed in the Schmidler group to examine X-ray crystallographic structures of known MHC binding peptides in complex with MHC I. Preliminary results have suggested new approaches to prediction of peptide/MHC binding by simulation. This will have an impact on Ben's dissertation work (with Schmidler), providing new directions for his work on quantitative peptide and protein simulation. Ben also helped run the audio/visual and computing at the opening program

workshop ``Genomes To Global Health: Computational Biology Of Infectious Disease".

Morgan Root (Mathematics, NCSU) has worked on the development of models for the spread of zoonotic infections, such as pneumonic plague, that are considered to pose significant bioterrorism threats. Root is in the early stages of his research career, so much of his time this year has been spent surveying both the biological and modeling literatures. He has developed deterministic and stochastic models for the spread of infection in a heterogeneous community. An example of a question of interest is whether the existence of groups that are underserved by the health system, such as minority communities or individuals of lower socio-economic status, can increase the likelihood or severity of disease outbreaks. Considering a two patch setting (representing the general population and a minority group), he has examined the time lag between outbreaks in the groups. If the underserved group is less closely monitored, then it is possible for the disease to become established there before it is noticed in the general population. The minority group can then become a source of infection for the community at large. Thus, any lag can have a major effect on the severity of an outbreak.

During the course of the year, we have started a collaboration with Jay Levine (NCSU College of Veterinary Medicine). Through Levine, Root will be funded by North Carolina's State Department of Health to model bioterrorism scenarios. Beyond this, Root and Lloyd are involved in the preliminary stages of a collaborative project between NCSU's Vet School and Purdue's Homeland Security Center that will develop scenario models examining the impact of and responses to bioterrorism attacks on the nation's agricultural and food supply industry.

Soyoun Park (Statistics/OR, UNC) and **Karl Strohmaier** (Computer Science, UNC) were each funded for one semester, and have been working with mentor Andrew Nobel on problems related to bi-clustering. Bi-clustering refers to a broad class of exploratory data analysis tools that are well suited to genetic, chemometric and microarray type data. The goal of bi-clustering is to identify scientifically interesting interactions between collections of samples (e.g. disease states) and sets of variables (e.g. genes). Bi-clusters correspond to submatrices of a given data matrix. Unlike traditional clustering, bi-clusters can overlap, and they need not cover all the entries of the data matrix.

Park has been working on connections between data mining and classification, with an eye towards producing more interpretable classification schemes for high dimensional data such as microarrays. Specifically, she has been looking at how bi-clustering algorithms can be applied to the classification of high dimensional data.

Strohmaier has been working on bi-clustering for noisy data. Existing bi-clustering algorithms do not account for the presence of noise in a direct way, a fact that may limit their applicability to gene expression and other data, where moderate or large amounts of noise are often present. Strohmaier and others in the group have been working on developing algorithms that will work when noise is present. In particular, he is currently working on the implementation of a heuristic, iterative bi-clustering algorithm. Once preliminary testing is complete, we expect to apply the algorithm to SNP and microarray data.

Abel Rodriguez (Statistics, Duke) has attended two of the program working groups: the Mathematical Genomics for Vaccine Design working group in the fall, and the Modeling

of the Immune System in the Spring. Abel also attended Alun Lloyd's course on "Mathematical Modeling of Infectious Diseases". In the MagVad group, Abel contributed to summary analyses of the first group of 86 epitopes we received, and particpated in the reading and presentation of papers on epitope and protein binding prediction. In the Immune System working group, Abel has participated in reading papers on two different lines: stochastic models for measles and spatial-temporal models for disease spread.

Miriam Nuno (Cornell), **Steven Tennenbaum** (Cornell), **Ariel Cintron** (Cornell), **Alicia Shim** (Arizona State Univ.), **Efrat Barzohar** (Arizona State Univ.), and **Yun Kang** (Arizona State Univ.) were visiting graduate students during Spring, 2005 in the Computational Biology Program. They participated extensively in the SAMSI activities including taking the courses offered by Alun Lloyd (every Monday evening) and Byron Goldstein (every Thursday evening) at SAMSI. In addition they contributed heavily to the success of the working group led by Alun Lloyd on immune response models, attending regularly and discussing papers for the group. At their request, Banks offered a special course on applied mathematical and statistical methods for inverse problems (a version of the course he and Davidian offered in the first year SAMSI program on Inverse Problems). This course was taught at SAMSI on Mondays and had 15 students.

These six students, who were recommended to SAMSI by Carlos Castillo-Chavez after his efforts with them in the Mathematical Biology Theoretical Institute (MBTI) at Cornell and ASU, were each in various stages of their research programs. Nuno finished and defended her Ph.D. thesis during the semester; Tennenbaum hopes to finish his in the Fall, 2005: Cintron should finish in one more year; Shim will finish in two more years; Barzohar and Kang just joined the program at ASU after completing MS degrees and are just getting started on their Ph.D research topics.

The semester went well for all students but especially so for Nuno, Cintron, and Shim. Nuno is headed for a postdoc at Harvard, while Cintron and Shim have developed a mentoring/working relationship with Banks and some of his postdocs and students and have expressed the desire to return to SAMSI for several weekly visits for collaboration during the coming year as their research progresses.

Banks agreed to act as local scientific mentor/personal advisor to the 6 students.This proved to be quite demanding, but necessary, since their advisor was able to visit SAMSI only periodically during the semester and the students needed extensive scientific assistance and counseling during the course of the semester. While this experiment in outreach and education was a successful one for the students and SAMSI, it reinforces the belief that SAMSI should pursue such efforts with caution in the future. Unless a student's advisor is also visiting for the semester or unless a local faculty member or SAMSI visitor agrees to closely mentor the student, such appointments are unwise.

V. Latent Variable Models in the Social Sciences

John Hipp (Sociology, UNC) played an active role in several of the working groups. He was a regular participant in the Multilevel and Structural Equation Model group. He regularly attended and participated in the group discussions. John also was a member of the Categorical Observed Variables in Latent Variable Models group. For this group, he

provided empirical and simulation examples that enabled us to explore several of the issues that emerged in our discussion. He also used these to demonstrate some of the new analytical results that we developed. In recognition of his role, John Hipp will be an author on the paper that Bollen, Thomas, and Wang are drafting, titled "Limited Dependent Variable Models with Covariates Measured with Error." John also participated in the social network working group. He is knowledgeable about social network data, theory, and examples and this knowledge was useful to the group. Finally, John Hipp helped with some of the logistics of making the working groups operate smoothly.

John Samuels (Mathematics, NCSU) has participated in the Social Networks working group that meets every Thursday at SAMSI. He has pursued specific research with Alan Karr, Hoan Nguyen (a SAMSI postdoctoral Fellow) and H.T. Banks. During the year they have developed a model for social dynamics (characteristics associated with a number of agents) of buddy/clique formation. The characteristics (on a continuous time, continuous value scale from -10 to +10) are assigned to each agent. A nonlinear model for degree of connectivity is coupled to a nonlinear stochastic differential equation for the evolution of the characteristics in the agent population. The resulting system of SDE is then solved by a classical fourth-order Runge-Kutta discretization procedure. Samuels has contributed substantially to the research and methodology and will be a co-author on a forthcoming publication this summer. He is also playing a major leadership role in the SAMSI Undergraduate Workshop to be held May 30-June 3.

Jen-hwa Chu (Statistics, Duke) has been involved in building agent-based models of social network dynamics. These models incorporate the latent-variable space approach described by Hoff, Raftery and Handcock (2002), as well as covariate information such as gender and memory of past relationships. The intent is to build rule-sets such that the dynamics of agent behavior mirror the dynamical models being from two other perspectives by other teams in the working group. The comparison of the models is based upon summary statistics from repeated runs, such as the mean and standard deviation of the number of persistent cliques, the first three moments of the in-degrees, the mean and standard deviation of the number of triad completions, and so forth.

Jiezhun Gu (Statistics, NCSU) has also been involved with the program. She has attended all of the weekly meetings, and has expressed interest in working on the asymptotic theory associated with model uncertainty. Subhashis Ghosal is also interested in working on this problem.

Satkartar Kinney (Statistics, Duke) has also made substantial progress on model uncertainty. She has focused on the problem of selecting fixed and random effects in logistic mixed effects models. A number of methods have been proposed in the literature for subset selection in regression, but little has been done for the challenging problem of selecting predictors that vary in their effects for different individuals and hence have associated variance components. Satkartar initially adapted an approach proposed by Chen and Dunson (2003) for the linear mixed model to the probit case. She then further modified this approach to logistic regression using a data augmentation scheme, with

slice sampling used in the implementation. Ongoing work focuses on parameter expansion methods, which lead to more efficient computation and improved priors for the variance components.

Eric Vance (Statistics, Duke) has been studying social network behavior in elephant herds using data collected by ethologists. He has fit the Hoff, Raftery, and Handcock (2002) latent variable version of the dyadic p* model and found that group dynamics change between the wet and dry season, and that genetic relatedness and the social hierarchy play a large role in elephant networks. The inferences are Bayesian, and use Markov chain Monte Carlo to find the posterior distributions of each of these effects. He is beginning to explore non-dyadic models that allow one to represent three-way interactions that might be described as jealousy.

Chien-Chung Wong (Mathematics, NCSU), working with Medhin, developed a model where each actor is endowed with a set of dynamic personal attributes, values, and preferences, and a set of statistical information on each of the other actors in the social group. If the social network consists of N actors we construct an NxN matrix of zeros and ones, called sociomatrix. If actor i is friendly toward actor j, then the ij-th entry of the matrix will be one, otherwise zero. The diagonal entries of the matrix are set to zero. If the ij-th entry of the matrix is 1, then we say there is a link from i to j. We have developed a model where the ij-th link depends on the maximum of a payoff of an appropriately constructed nonlinear programming problem involving the attributes and values of the actors in the social group. The model can be modified to handle social status as well as general network dependence structure. For example, the link from i to j may not be completely independent of the link from i to k, and/or from j to k.

The model also incorporates migration and preferred attributes. The approach developed captures the ideas of the well established P1 model introduced by Holland and Leinhart, and the more recent extension, P2 model, due to van Duijn, Snijders, and Zijlstra. In particular the model developed reflects reciprocity, and attributes and values of actors i and j play a role in determining whether or not there is linkage between these actors. In addition, a general network dependence is reflected in the model.

VI. Data Assimilation in Geophysical Systems

Kristen Foley (Statistics, NCSU) is currently enrolled in the DA course taught by Kayo Ide and will be working on a project on particle filtering methods. She presented a poster at both of the workshops. She is also planning on attending the week long summer school on Fusing Geophysical Models with Data at NCAR this June. She is involved with the working group on Estimation and Prediction. She is currently attending the weekly lecture series presented by Lenny Smith in the Statistics department at NC State.

She is currently working on her dissertation with her advisor Montserrat Fuentes on a data assimilation project in collaboration with a group of meteorologists and oceanographers at NC State. The goal is to improve the prediction of the onshore flooding caused by hurricane force winds and pressure along the coastal Carolinas and Georgia. The group she is working with has implemented some basic nudging techniques to combine observed data with the ocean model they use to predict storm surge associated

with hurricanes. They are now working, with the help of Foley and Fuentes, on implementing statistical data assimilation methods, beginning with the Ensemble Kalman Filter.

Nicole Mich (Geosciences, Duke) is involved with the Lagrangian Data Assimilation working group and is working with Susan Lozier on a study of the pathways that Labrador Sea water takes from the subpolar gyre to the subtropical gyre. The overall objective of her work is to understand climate pathways in the deep ocean. Her work is in support of an observational program that involves the deployment and tracking of Lagrangian floats (RAFOS floats) from the Labrador Sea to the subtropical basin. Her work sets the Eulerian framework for our understanding of the Lagrangian data that will first be available in July of 2005. To date, she has established a link between the Labrador Sea water pathways and the North Atlantic Oscillation, the dominant mode of decadal variability in the basin.

Stephen Foster (Mathematics, UNC) is a second year student and is involved in the program as a younger student with potential interest in this area. He is involved in various working groups and is an energetic participant in the SAMSI course. He is focusing on Lagrangian data assimilation and is working on a problem aimed at design smart float placement strategies. The idea is to have optimal float placement for assimilation of the resulting data. He is trying the ideas out on a model problem based on the shallow water equations with wind forcing.

Minjung Kyung (Statistics, NCSU). At the Tutorial and Opening Workshop of the DA program in January 2005, she took detailed notes of the lecture on "A Brief Introduction to Bayesian Statistics and Application to Data Assimilation" given by Chris Wikle. She has been the teaching assistant for the Data Assimilation course taught at SAMSI, in which she has been producing careful notes. Her course project is on Bayesian statistical modeling concerning spatial-temporal variability. She participates regularly in the working group on Estimation and Prediction.

Liyan Liu (Mathematics, UNC) is working with Jones. She has been developing Lagrangian data assimilation schemes for two-layer point vortex problems. She is deeply involved in the Lagrangian data assimilation working group. She presented a poster at both of the workshops.

D. Consulted Individuals

The individuals consulted for the broad selection of topics within programs and workshops were the members of two groups:

- The **Program Organizers**, listed in Section I.A.1.
- Members of the **Advisory Committees**, listed in Section I.J.

The specific topics that Program Working Groups chose to pursue were, in general, selected by the Working Group participants themselves, according to their combined interests. In almost all cases, however, a Program Leader headed each working group, so that specific research topics remained consistent with overall program goals. In Section I.E, the various Program Working Groups, and their members, are discussed.

E. Program Activities

1. Program on Latent Variable Models in the Social Sciences

1.1 Introduction, Motivation and Initial Ideas

The purpose of the 2004–05 SAMSI program on the Latent Variable Models in the Social Sciences (LVSS) was to address an area that:

- Offers multiple points of intersection between the social sciences and the statistical and applied mathematical sciences, in particular in order to more extensively introduce applied mathematical modeling in social sciences contexts;
- Is not customarily addressed at other DMS-funded mathematical sciences research institutes;
- Leverages across multiple research centers at the Research Triangle universities – including the Center for Demographic Studies and the Center for Decision Studies at Duke, and the Odom Institute and Carolina Population Center at University of North Carolina at Chapel Hill (UNC)—as well as at organizations such as North Carolina State University (NCSU), RTI International and the SAS Institute.

The organizing theme of the program was *latent variables*, which are widespread in the social sciences. Whether it is intelligence or socioeconomic status, many variables cannot be directly measured. Factor analysis, latent class analysis, structural equation models, error-in-variable models, and item response theory illustrate models that incorporate latent variables. This SAMSI program will take a broad look at latent variables and measurement error. Issues of causality, multilevel models, longitudinal data, and categorical variables in latent variable models are examples of the SAMSI topics for this program.

A great success of the program, which was evident from the Kickoff Workshop and the spin-off working groups, is that the program has brought together social scientists, statisticians, biostatisticians, and mathematicians who otherwise would not have gotten together. Members of these disciplines are no different than other disciplines in that their typical interactions are with others from the same discipline. When mixing occurs it is most likely for disciplines that are close in subject matter (e.g., psychology with sociology or statistics with biostatistics). Diffusion of knowledge across disciplines can be a slow process. Furthermore, the latent variable models developed in the social sciences are typically not known in statistics and biostatistics. However, statistics and biostatistics are much more interested in latent variable models today than they were a decade ago.

The program has provided—and through continuing collaborations will continue to provide—a more rapid sharing of knowledge across disciplines than would have

occurred without it. The program has clearly generated collaborations across these disciplinary lines that would have not otherwise have occurred.

1.2 Working Group Activities

1.2.1 Categorical Variables

Many categorical observed variables in the social sciences are imperfect measures of underlying latent variables. The working group on categorical observed variables was formed to study the statistical issues that emerge when categorical observed variables are part of a model with latent variables or with measurement error. Long-distance participation in this group has been particularly lively. Roland Thomas, a business school professor and statistician of Carleton University in Canada regularly calls in and has been active in developing the group's research agenda, and Liqung Wang, a statistician, has actively participated in email discussions. The group met weekly during the fall and spring semester.

The working group began by studying the relationship between two approaches to models with categorical outcomes: James Hardin et al's recent advances in models that correct for measurement error in nonlinear models within the GLM framework, and Ken Bollen's two-stage least squares approach to latent variable models.

After contrasting the differences and similarities in these approaches, we began to pursue specific research papers that grew out of this examination of the different approaches. The group is currently drafting a paper entitled "Limited dependent variable models with errors in covariates." James Hardin, Ray Carroll, and colleagues have examined an instrumental variable approach to generalized linear models that corrects for measurement error in covariates. In their work, published in the Journal of the American Statistical Association and elsewhere, they describe their technique as "approximately consistent" with the consistency related to the degree of measurement error in the covariates. Several members of the working group derived methods to permit consistent estimation of the coefficient parameter under the same conditions. The paper we are drafting focuses on the theoretical derivation and application to empirical examples. Coauthors include Professors Bollen, Thomas, and Wang, as well as John Hipp, a graduate student in the Department of Sociology at UNC. Future planned research may include a paper on Bayesian approaches to the problem of nonlinear models with measurement error; Jane Zavisca, a SAMSI postdoc, and Saki Kinney, a graduate student in Statistics at Duke, have begun a literature review to that end.

1.2.2 Complex Surveys

This working group is working on a range of problems for analysis of survey data. Since most of the groups' participants are long-distance, and their interests cluster into two categories, the group split early in the program into two subgroups with the following themes: Latent Class Analysis (LCA) of Measurement Error in Surveys, and Weighting and Estimation for Complex Sample Designs. Each subgroup meets approximately once a month via teleconference, and also corresponds via a listserv between meetings.

LCA Subgroup: This subgroup is developing latent class models for assessing measurement error in survey responses, with the goal of improving questionnaire design. Many questions in social surveys have categorical, multiple choice responses, making latent class analysis an appropriate technique for estimating measurement error when repeated measures are available. Group members include Paul Biemer, the group leader, Bac Tran of the US Census Bureau, Clyde Tucker of the Bureau of Labor Statistics, Brian Meekins of the Bureau of Labor Stastics, and Jane Zavisca, a postdoc at SAMSI.

The group's primary work to date has concentrated on the problem of rotation group bias in longitudinal surveys, using the empirical example of labor force classification in the Current Population Survey. The census bureau provided the group with access to specialized data that permits fitting a wide range of LCA models for measurement error. Jane Zavisca presented preliminary findings at the American Sociological Association Section on Methodology 2005 Annual Meeting. We plan to finish analysis and submit a paper for peer review this summer. Our current work focuses on the cross-sectional re-interview sample; in the future we will apply Markov LCA models to panel data (for which re-interviews are not available).

Although the unemployment example has been the primary focus of our research, we have also read papers and discussed ongoing work by group members on related topics. For example, we have discussed modeling measurement error when the latent variable of interest is thought to be continuous rather than categorical, with applications to quality of response in surveys of adolescent drug and alcohol use. All group members have broader research agendas for which this type of modeling is appropriate, and future collaboration beyond the SAMSI program is likely.

Weighting and Estimation Subgroup: The complex sampling working group (CSWG) is the other LVSS survey subgroup. The group originated with 35 members and has approximately 15 active members. Group membership is international with only a few members local to SAMSI. Therefore, most communication occurs via email and conference call. he group meets monthly via conference call to keep abreast of the group progress and happenings. Communication via the listserv is ongoing and includes sharing research, feedback on research, question and answers, and discussion. Several members are working on independent projects dealing with proper weighting in multi-level modeling. Therefore, the group is collaborating on a document intended to provide the most current information on weighting in a multilevel model.

Group members are meeting in person at a conference organized by members of the working group: Mary Thompson (Waterloo), Chris Skinner (Southampton), Paul Biemer (North Carolina), Jamie Stafford (Toronto), Milorad Kovacevic (Statistics Canada), Randy Sitter (Simon Fraser), David Bellhouse (Western Ontario), Roland Thomas (Carleton). Group members will be making presentations including the research on weighting in multi-level modeling. In addition, A. Skrondal and S. Rabe-Hesketh will provide a tutorial on Generalized latent variable modeling. Also, Mplus representatives are presenting on their survey data modeling project, which is the implementation of complex sample analysis capabilities in latent variable modeling software. Information on the conference can be found at the following website: http://www.crm.umontreal.ca/Latent05/index_e.html

1.2.3 Longitudinal Data

The participants in the working group can be grouped into three categories: 1. MetaMetrics; 2. Chris Kelly (Post-doc, UNC Institute of Aging); 3. Minimal participation by 2 or 3 others. Representatives from MetaMetrics (5-6 people) have made up the bulk of the group and consistently attended the meetings. Chris Kelly has attended many of the meetings but he and Lloyd Edwards have consistently met outside of SAMSI (at UNC) to address his needs.

Initially, an assessment was made regarding the needs and interests of the various participants in the group. As a result, since the inception of the program the working group has covered what amounts to a semester's worth of graduate material in applied longitudinal data analysis using the linear mixed model. Both MetaMetrics and Chris Kelly have made presentations to the group.

As a result of the LVSS Program, a long-term collaborative effort has been identified between MetaMetrics and me. MetaMetrics and I will be working together to address methodological and applied aspects of longitudinal data analysis of a very large dataset. The objective is to determine the profile of standardized reading scores across time and accurately assess predictors of the reading scores. The development of standardized reading scores, called Lexiles, is company proprietary information. In addition, since the dataset is large, we will research the use of cross-validation methods in longitudinal data analysis using the linear mixed model.

Chris Kelly and Lloyd Edwards are working together on a publication from his doctoral research regarding factors predicting the extent of nursing home regulation in the 50 states of the U.S. Kelly originally attempted to apply ordinary least squares and hierarchical linear modeling techniques to his longitudinal study. After publication rejection, Kelly and Edwards are working to apply appropriate linear mixed model techniques and are having great success. Through our individual meetings, Edwards has given Chris a crash course in using SAS Proc Mixed.

1.2.4 Model Uncertainty

The primary goal of the Model Uncertainty Working Group is to develop new statistical methods for accommodating model uncertainty in latent variable models motivated by social science applications. Because structural equation models (SEMs) with latent variables provide a broad and flexible framework for analysis of multivariate data, the working group chose to focus on uncertainty that arises in specifying an SEM as a starting point. The group began by reviewing the literature on model selection and averaging in SEMs. The BIC criteria, which was originally proposed as an approximation to the Bayes factor, was chosen as a good starting point for discussion. There are a series of articles in the literature by Adrian Raftery and others arguing in favor of the BIC.

Several members of the working group were interested in exploring improvements to the BIC, which could be calculated using the output produced by standard software for fitting SEMs. This work eventually lead to a paper, "Bayes factors in structural equation models (SEMs): Schwarz's BIC and other approximations", by

Kenneth A. Bollen, Surajit Ray, and Jane Zavisca. This paper was presented recently at the Methodology Conference of the American Sociological Association.

It was agreed by the working group that Bayesian methods provide an appealing framework for accommodating model uncertainty in SEMs, allowing for both model averaging and selection. The working group therefore reviewed the literature on Bayesian SEMs, which was found to be quite limited, motivating a summary article "Bayesian Structural Equation Modeling" by David Dunson, Jesus Palomo and Kenneth Bollen (to be included in the Handbook on Structural Equation Models, to be published by Elsevier).

In many cases, the number of plausible models a priori can be very large, and it can be difficult to fit all the models in the list. In addition, BIC and other approximations typically reply on normal approximations, which may not be appropriate for comparing SEMs having different numbers of latent variables and variance component structures. For this reason, the working group developed a stochastic search approach for Bayesian model averaging, carefully considering the important issue of prior choice. The paper *"Bayesian Model Selection and Averaging in Structural Equation Models,"* by David B. Dunson, Jesus Palomo, and Jane Zavisca, is complete except for the data example, and should be ready to submit within a month or so.

Ongoing work of interest to the group focuses on (1) defining an "effective sample size" for use in model selection criteria for latent class and time series models; (2) developing approaches for model selection in nonlinear variance component and random effects models; and (3) developing methods that relax parametric assumptions on latent variable distributions.

In summary, the model uncertainty working group has been very productive at generating new collaborations, which have already lead to real progress in this challenging and important area.

1.2.5 Multilevel Models

Over the past year, the working group explored a number of issues with respect to fitting multilevel models with latent variables (also known as hierarchical, random coefficients, or mixed-effects models). Our primary goal was to examine and/or develop models that could allow for random effects when either or both the criterion and predictors were constructs that could not be measured directly and without error. The general structure our group followed was to complete and discuss targeted readings and present illustrative analyses at the working group meetings.

In the fall, we began by considering multilevel models with random coefficients but without latent variables – that is, where the criterion and predictors are all assumed to be perfect (error-free) measures of the constructs they are said to represent. We examined both frequentist (likelihood-based) and Bayesian approaches to fitting these models. Using this as a base, we then extended our consideration to multilevel models with latent variables. We read work on the correspondence between multilevel models and structural equation models, including papers by Bauer (2003) and Skrondal and Rabe-Hesketh (2004). These readings allowed us to contextualize and compare two basic multilevel latent variable models originating in different literatures. Although useful, these models apply only under a limited set of conditions. Specifically, these models are linear in their

parameters, require continuous observed variables, and allow only random intercepts or means in addition to the latent variables.

In the spring, we examined ways to expand the model to overcome these limitations. One key extension was to allow for categorical observed variables, as these are quite frequent in social science research. We read and discussed further papers by Kamata (2001), Skrondal & Rabe-Hesketh (2004), Rijmen et al. (2003), and Thissen & Orlando (2001) on the correspondence between item-response theory models, binary factor analysis, and nonlinear mixed models. We were fortunate to have Anders Skrondal and Sophia Rabe-Hesketh come to SAMSI in April to present on their approach to fitting multilevel latent variable models with continuous or categorical observed variables. The second key extension that we examined with these models was the possibility to allow for both random intercepts and slopes; specifically, to have parameters of the measurement model (relating the latent variables to the observed variables) or structural model (relating the latent variables to one another) vary over units. The primary optimization difficulty in allowing both random intercepts and slopes is that one must integrate over more dimensions with a model that is nonlinear in the unknown parameters. A secondary difficulty is the need to have many observations per unit and many units to strongly identify and make possible the estimation of the model. We considered several possible ways of addressing these difficulties, with various strengths and weaknesses. These included the use of quadrature, Markov Chain Monte Carlo, an approximate method based on finite mixture models, and an approximate method that assumes the "random" effects are predetermined functions of other variables.

We are currently in the process of further exploring and writing up new results on several of these topics. One paper is nearing completion and I anticipate that at least one or two more papers will be completed and submitted for publication by the end of the summer.

1.2.6 Social Networks

This working group is the NSF-funded NISS project *Dynamics for Social Networks Processes: Comparing Statistical Models with Intelligent Agents*. The thrust of the project is to reconcile two methods for modeling change in social networks over time--*p** models and intelligent agent models. The latter family has received much attention from social scientists but little from mathematicians and almost none from statisticians, and so constitutes a promising and important opportunity for collaboration. The group pursued three complementary approaches to modeling dynamics of social networks:

- Using stochastic differential equations to describe evolution of participants characteristics and relationships ("edges": H.T. Banks, Karr, Nguyen, Samuels
- A mathematical programming formulation that maximizes "affinity:" Medhin, Hong.
- A discrete-time formulation with (initially) fixed relationships and participants that move with a "social space" of (possibly latent) characteristics: D. Banks, Chu.

Principal senior participants are David Banks (Statistics, Duke), H. T. Banks (Mathematics, NCSU), Kenneth Bollen (Sociology, UNC), Kathleen Carley (Computer Science, Carnegie Mellon), Alan Karr (NISS) and Negash Medhin (Mathematics, NCSU). SAMSI/CRSC postdoctoral Nguyen Hoan has also participated, as have graduate fellows John Hipp (Sociology, UNC) and Johnny Samuels (Mathematics, NCSU) and graduate associates and Jen-hwa Chu (Statistics, Duke), Chung-Chien Hong (Mathematics, NCSU) and Eric Vance (Statistics, Duke).

1.3 Personnel

1.3.1 Faculty Releases and Associates

Faculty releases for the program were Kenneth Bollen (Sociology, UNC), Lloyd Edwards (Biostatistics, UNC), Subhashsis Ghosal (Statistics, NCSU) and Negash Medhin (Mathematics, NCSU).

Dan Bauer (Sociology, UNC), Paul Biemer (RTI International) and David Dunson (NIEHS) led working groups as faculty associates. David Banks (Statistics, Duke), H. T. Banks (Mathematics, NCSU) and Alan Karr (NISS) participated in the social networks working group.

1.3.2 Research Visitors

Long-term research visitors to the LVSS program included:

- Maria Jesus Bayarri, University of Valencia
- Aki Kamata, Florida State University
- Katja Ickstadt, University of Dortmund
- Xiadong Lin, University of Cincinnati
- Ingmar Visser, University of Amsterdam

Targeted experts, here for shorter visits, included

- Yasuo Miyazaki, Virginia Tech
- Sophia Rabe-Hesketh, Berkeley
- Anders Skrondal, London School of Economics

1.3.3 Postdoctorals

Jane Zavisca has been active in most of the six working groups in the LVSS program. In the latent class subgroup of the survey working group, Jane has worked closely with Paul Biemer, a research statistician at the Odum Institute for Research in Social Science and RTI. She presented a coauthored a paper on their work on a latent class analysis of unemployment statistics at the spring meeting of the Methodology Section of the American Sociological Association in Chapel Hill, North Carolina. Jane also is working closely with the Model Uncertainty working group. Here, too, she was coauthor on a paper presented to the Methodology Section of the American Sociological Association.

The paper was on approximations to the Bayes factor in Structural Equation Models. Both the latent class unemployment and the Bayes factor paper will be revised and sent out for review for publication in refereed journals. Jane also has been an active participant in the Multilevel and Structural Equation Modeling working group and the Categorical Observed Variable and Latent Variables working group. In addition, to contributing to the discussions and preparing results for the group, Jane has helped with some of the web site maintenance for several of these groups.

She was also a major contributor to the February 18-19 Undergraduate Workshop, presenting material from the LVSS program.

Jesus Palomo and **Surajit Ray** were major contributors to the Model Uncertainty working group. There roles are described in Section B.

1.3.4 Graduate Fellows and Associates

Reports on the participation of the graduate fellows and associates, John Hipp, John Samuels, Jen-Hwa Chu, Satkartar Kinney, Jiezhan Gu and Eric Vance are given in part C.

1.4. Workshops

1.4.1 Opening Workshop

The Opening Workshop, held on September 11-14, 2004, was among the largest and most diverse to date at SAMSI. There were more than 100 attendees. The participants, program, and evaluations appear in Appendices E, F, and G.

1.4.2 Causality Mini-Workshop

A Mini-Workshop on Causality was held on March 29, 2005, with more than 80 attendees. The participants, program, and evaluations appear in Appendices E, F, and G.

1.5 Course

Kenneth Bollen (Faculty Fellow) and Jane Zavisca (postdoctoral) presented a seminar course entitled *An Overview of Latent Variable Models in the Social Sciences* during the fall semester of 2004-05. The course provided an overview of latent variable models that are common in the social sciences. It was designed to introduce students to a variety of statistical models that make use of latent or unobserved variables, including factor analysis, latent trait and latent class models, and structural equation models. The instructors and guest speakers lectured on the various models and take student questions, followed by class discussion and/or student presentations. Technical and substantive readings will further explain the models and provide examples of concrete applications.

In addition to attending lectures and reading background materials, students taking the course for credit applied one of the types of models covered in the course to a data set of their choice, and made a 15 minute presentation on their findings to the class.

2. Program on Genomes to Global Health: the Computational Biology of Infectious Disease

2.1 Introduction, Motivation and Initial Ideas

Infectious disease remains a major cause of suffering and mortality among people in the developing world, and a constant threat worldwide. The major killers are HIV, *Mycobacterium tuberculosis*, and the malaria parasites (eg *Plasmodium falciparum*)[1]. 45 million people worldwide are infected with HIV, and in 2003, three million people died from AIDS. One third of the world's six billion people are infected with latent tuberculosis (TB), and an estimated two million people die each year from the disease. Of the approximately 8.4 million new TB cases each year, 300,000 involve multi-drug resistant TB (MDR-TB), strains that are resistant to two of the four drugs used to cure TB. 79% of MDR-TB cases involve "super-strains" that are resistant to at least three of the drugs. Each year, the approximately 300 to 500 million malaria infections lead to over one million deaths. 75% of these deaths are in African children less than 5 years of age. 80% of malaria cases are not responsive to current treatments because of drug resistance.

The advent of genome science and the continuing rapid growth of computational resources together herald an opportunity for the mathematical and statistical sciences to play a key role in the elucidation of pathogenesis and immunity and in the development of the next generation of therapies and global strategies. We intend for this SAMSI program to spark interest in the study of infectious disease in the mathematics and statistics communities. The primary aims of this year of research and study are to identify those areas where mathematical/statistical innovation may have the greatest impact on the basic science and medicine of infectious disease, to progress materially toward major research efforts in these areas, to establish a greater sense of community among the researchers with skills and interests in these areas, and to contribute to the training of the next generation of mathematically-literate biomedical researchers, originating in both the biological and mathematical sciences.

Goals of the program:

1. Determine what the essential problems are in the study of infectious disease.

2. Identify those problems to which mathematicians, statisticians and computer scientists can make significant contributions.

3. Catalyze the flow of ideas between researchers who would normally not communicate but whose contribution is necessary for real progress in the study of infectious disease (eg statisticians, applied mathematicians, computer scientists, epidemiologists, immunologists, microbiologists, infectious disease experts, and physicians.

4. Forge new collaborations.

5. Generate interest and participation in the program through workshops, working groups, courses, seminars and visits from senior researchers.

2.2 Working Groups

The program has consisted of three working groups: mathematical genomics for vaccine design (fall and spring), cell communication (fall) and modeling the immune system (spring). The principle function of the working groups has been to identify relevant research questions and pursue at least one of them with the goal of publishing the results and writing a grant to fund continued work on the topic. The working groups will meet at least once per week to organize the research and ensure progress towards the stated goals.

A major concern of ours will be to integrate the views revealed from the disparate perspectives named above to explore novel, multiscale approaches to the fundamental problems.

2.2.1 Mathematical Genomics for Vaccine Design

The members of this working group were Thomas B. Kepler (DUMC, leader), Lindsay G. Cowell (DUMC), Joanna Fueyo (IBM), Andrew Nobel (UNC), Scott Schmidler (Duke), Cliburn Chan (DUMC), Jeff Frelinger (Immunology, UNC), Georgia Tomaras (Surgery, DUMC), Padraic Neville (SAS), Byron Goldstein (LANL), Surajit Ray (SAMSI), Ben Cooke (Duke), Karl Strohmaier (Duke), Abel Rodriguez (UNC), and Soyoun Park (UNC).

Two specific research foci are *Large Scale Antibody & T cell Epitope Discovery* and *High throughput identification of bacterial epitopes by T cell antigen discovery*. In each project, the experimental group is collecting large quantities of data on the molecular entities, known as peptide epitopes, that the immune system recognizes during an encounter with a pathogen.

These epitopes are short, averaging 9 amino acids in length, and bind to the molecules of the human leukocyte antigen complex. Cells that express these HLA molecules then interact with T lymphocytes in such a way that the MHC-epitope complexes interact with the T cell receptor. The nature of this molecular interaction, and the context in which it occurs, determines the nature of the T cell response, and consequently, of the immune response itself.

Some epitopes elicit a strong and protective immune response, while others are essentially inert. The magvad group is designing statistical methods to predict which epitopes will have the appropriate effects so that these can be used in the design of vaccines. This computational task is very important, and very difficult, since there are over 1000 distinct allelomorphs of HLA molecules, and on the order of potential epitopes in every pathogenic microorganism. The two experimental groups will be producing data in quantities sufficient to produce high-resolution predictive models.

2.2.2 Cell Signaling

The members of this working group are Tim Elston (UNC, leader),,Jeff Butterworth (Alien Skin Software), Rory Conolly (CIIT Centers for Health Research), Sujay Datta (Northern Michigan University, Department of Math and Computer Science), Chuanshu

Ji (UNC-CH, Department of Statistics and Operations Research), Julia Kimbell (CIIT Centers for Health Research), Delong Liu (CIIT/SAS), Kevin Thomas Morgan (Sanofi-Aventis), Abby Todd (UNC-CH, Math Department), Xiao Wang (UNC-CH, Department of Statistics and Operations Research), and Qiang Zhang (CIIT Centers for Health Research).

2.2.3 Modeling the Immune System

The members of this working group are Alun Lloyd (NCSU, leader), Brian Adams (NCSU), H.T. Banks (NCSU), Efrat Barzohar (ASU), Ariel Cintron (Cornell), Yun Kang (ASU), Grace Kepler (NCSU), Alun Lloyd (NCSU), Hoan Nguyen (SAMSI), Miriam Nuno (Cornell), Morgan Root (NCSU), John Samuels (NCSU), Alicia Shim (ASU) and Stephen Tennenbaum (Cornell).

Since many of the group's participants had little experience in this area, most of our semester's activities have taken the form of literature surveys. A new class of models, based around the so-called "immune program", have been of particular interest. A draft manuscript critiquing this class of models is currently in preparation.

The group has considered the ability of such models to describe acute infections. This is of great interest to Lloyd, since the development of such models will be a key component of a larger project that will study the impact of drug resistance on acute viral infections, such as the common cold.

2.3 Personnel

2.3.1 Faculty Releases

Faculty releases for the program were Tom Kepler (Biostatistics and Bioinformatics, Duke University Medical Center), Lindsay Cowell (Biostatistics and Bioinformatics, DUMC), Scott Schmidler (Statistics, Duke), Andrew Nobel (Statistics, UNC), Tim Elston (Applied Mathematics, UNC), and Alun Lloyd (Mathematics, NCSU)

2.3.2 Research Visitors

Long term research visitors to the LVSS program included:

- Sujay Datta, Department of Mathematics, Northern Michigan University, visited during the Fall semester.
- Byron Goldstein, Theoretical Biology and Biophysics, Los Alamos National Labs was the University Fellow for the program, visiting Spring semester.
- Cliburn Chan, was appointed the New Researcher Fellow for the program.
- Carlos Castillo-Chavez, Arizona State University
- Priscilla Greenwood, Arizona State University

Targeted experts who visited the program for shorter periods were

- William Scott, Duke Center for Human Genetics
- Mark Miller, University of California, Irvine

- Arup Chakraborty, University of California, Berkeley
- Jason Stout, Division of Infectious Disease, Duke University Medical Center
- Georgia Tomaras, Department of Surgery, Duke University Medical Center
- Bette Korber, Theoretical Biology and Biophysics, Los Alamos National Labs
- Alan Perelson, Theoretical Biology and Biophysics, Los Alamos National Labs

2.3.3 Postdoctorals

The contributions and activities of the postdoctoral fellow in the program, Surajit Ray, are documented in Section B above.

2.3.4 Graduate Fellows and Associates

Reports on the participation of the graduate fellows and associates, Karl Strohmaier, Ben Cooke, Morgan Root, Soyoun Park, Abel Rodriguez, Efrat Barzohar, Ariel Cintron, Yun Kang, Miriam Nuno, Alicia Shim, and Stephen Tennenbaum, are given in part C.

2.4. Workshops

2.4.1 Opening Workshop

The Opening Workshop, held on September 18-22, 2004, began with a day of tutorials. In the morning, there were three tutorials: statistics and data analysis (Tom Kepler, DUMC), mathematical modeling and and simulation (Tom Kepler, DUMC), and epidemiology (Alun Lloyd, NCSU). There were three additional tutorials in the afternoon: molecular evolution (Jeff Thorne, NCSU), microbiology (Denise Kirschner, UM), and immunology (Lindsay Cowell, DUMC). The tutorials were followed by three days of research talks and discussions. The first day of talks and discussion centered around the mathematical and statistical challenges in biostatistics and public health. The second and third days focused on the mathematical and statistical challenges in microbiology and immunology, respectively. A major goal of the opening workshop was to frame issues to be addressed by the programs working groups; the workshop concluded with a discussion of potential working-group topics and the formation of three groups. The participants, program, and evaluations appear in Appendices E, F, and G.

2.4.2 Mid-Program Focused Workshop

A midterm workshop on *Modeling Infectious Diseases* was held from Januaru 31 – Februaru 1, 2005. External participants included Alan Perelson (LANL), Carlos-Castillo Chavez (Arizona State Univ.), Priscilla Greenwood (ASU), Zhilan Feng (Purdue), Alison Galvani (Yale), Gerardo Chowell (LANL) and Christopher Kribs-Zaleta (Univ. Texas at Arlington). The main aim of the meeting was to discuss potential topics for the working group and to give the local students a better idea of the breadth of the field.

The workshop was notable for the wide range of topics and approaches that were

discussed. During the workshop, Alan Perelson delivered a SAMSI Distinguished Lecture, *Modeling Viral Dynamics,* discussing his important contributions to the field of virus dynamics. with more than 80 attendees. The participants, program, and evaluations appear in Appendices E, F, and G.

2.4.3 Transition Workshop

A workshop marking the formal end of the program will be held from May 22-24, 2005. The theme is *Collective Computational Biology for Infectious Disease*. The purpose of this three-day workshop is to explore novel approaches to the amelioration of infectious disease in the developing world through collective, open source and public efforts in computational biology and informatics. We will gather experts to help identify those scientific problems and approaches most susceptible to these methods, and the organization of public resources, the coordination of collective research efforts, and the dissemination of educational materials to address these critical problems. Our ultimate goal is to speed the development of drugs, vaccines and other therapeutic and prophylactic interventions where financial and market-based incentives are unlikely to lead to the desired results. The meeting will consist of two days of working sessions and a one-day public symposium.

Organizers: Tom Kepler, DUMC; Arti Rai, School of Law, Duke University (co-chairs) Stephen Maurer, School of Public Policy, Berkeley; Andrej Sali, UCSF; Lindsay Cowell, DUMC

Invited Participants

Martin, Bill	Epivax
Cook-Deegan, Bob	Duke
Krakauer, David	Santa Fe Institute
McKenzie, Ellis	Fogarty Inernational Center
Kissinger, Jessica	U Georgia
Hamilton, John	Duke, ID
Quackenbush, John	TIGR
Haldar, Kasturi	Northwestern
Woo, Katherine	Institute for One World Health
Connett, Maria	BIOS
Bond, Queta	BWF
Stallman, Richard	Free Software Foundation
Taylor, Terrie	Michigan State
Hide, Winston	South African National Bioinformatics Institute
Salij, Andrej	UCSF
Kalyanaraman., Chakrapani	UCSF
Chan, Cliburn	Duke
Burk, Dan	University of Minnesota Law School
Bonabeau, Eric	Icosystems
DeRisi, Joe	UCSF
Guy, Kip	UCSF
Sung, Nancy	BWF

Schmidler, Scott	Duke
Robbins, Zack	Duke
Sette, Alessandro	La Jolla Institute of Allergy and Immunology
Korber, Bette	LANL
Castillo-Chavez, Carlos	ASU
Wirth, Dyann	Harvard School of Public Health
Elston, Tim	UNC
Quill, Helen	NIAID
Giacomini, Kathy	UCSF
Lubkin, Sharon	NCSU
Jacobson, Matt	UCSF
Smith, Charlie	NCSU
McGovern, Victoria	BWF
Lloyd, Alun	NCSU
Fueyo, Joanna	IBM

2.5 Courses

Three courses were offered as part of the program. Each was listed at Duke, NCSU and UNC for graduate credit.

2.5.1 Computational Immunology and Immunogenomics

This course was taught in Fall, 2004 by Thomas B. Kepler, Duke Departments of Biostatistics & Bioinformatics and of Immunology and Lindsay G. Cowell, Duke Department of Biostatistics & Bioinformatics. Twenty graduate students from the three universities took the course for creditor audited it. Several postdoctoral fellows and researchers from local companies attended regularly. Guest lecturers joined on several occasions.

Brief description: This graduate course will integrate empirical and computational perspectives on immunology and host defense. Students are expected to have significant preparation in either biomedicine, or a quantitative science.

The ideal class will represent a diversity of backgrounds; individuals will be willing to contribute their own knowledge to the development of the topics as they learn those parts of the subject with which they are less familiar. The topics covered are intended to provide entrée into the use of computational methods for research and practice in immunology and infectious disease, from basic science to medical applications.

Topics:
Aug. 23, Lecture 1: Basic mathematical methods for computational immunology. This lecture will provide the necessary background on differential equations, probability and statistics and computation. It will be self-contained and much material will be provided for reference throughout the course.

Aug. 30, Lecture 2: Cells, molecules and tissues of the immune system. We will provide a self-contained introduction to the components, organization and activities of the vertebrate immune system, including cells, molecules, lymphoid tissues and phenomena such as self-tolerance and memory.

Sept. 6 and 13, Lectures 3 and 4: Introduction to Immunogenomics: disease susceptibility and genetic polymorphism, phylogenetics and molecular evolution, somatic diversification. This lecture will include discussion of statistical analyses of genetic polymorphisms as they relate to disease susceptibility and resistance to human disease, both infectious and non-infectious. We will explore SNP analysis, as well as classical genetics. We will investigate the comparative genomic structure of the antigen receptors, MHC molecules and cytokines across phyla. We will examine the mechanistic basis for somatic diversification, including V(D)J recombination, gene conversion and somatic hypermutation.

- William Scott

Sept. 20 and 27, Lectures 5 and 6: Pathways and signal transduction: gene expression, phosphorylation cascades, immunological synapses. We will explore the structure of signal transduction pathways, the biophysics of receptor signaling and formation of the immunological synapse. We will study gene expression networks, including the analysis of microarray data from lymphocytes and dendritic cells.

- Byron Goldstein
- Arup Chakraborty

*Oct. 4, Lecture 7: Intercellular communication: costimulation, cytokines, chemokines***.** We will discuss the modeling of various components of intercellular signaling, including antigen presentation, co-stimulation, inflammation, cytokines, chemokines and cytokine networks with specific examples drawn from dendritic cell – T cell and T cell – B cell interactions.

Oct. 18, Lecture 8: Lymphocyte homeostasis. This lecture will cover population dynamic models of the control of lymphocyte numbers and discussion of so-called homeostatic proliferation, lymphocyte lifetime, competition for antigen-receptor-specific resources as well as non-specific resources such as growth factors or space in lymphoid tissues.

*Oct. 25, Lecture 9: Motility and chemotaxis***.** Topics include the biophysics of neutrophil rolling, adhesion and extravasation. We will analyze video images from multi-photon microscopy and data from *in vitro* motility and chemotaxis assays.

- Mark Miller

Nov. 1, Lecture 10: Field-coupled agent-based models. We will introduce individual-based models of cellular reorganization of the immune system for studying phenomena such as germinal center development and granuloma formation.

Nov. 8 and 15, Lectures 11 and 12: Pathogenesis. We will use mathematical modeling to investigate the unfolding of events in the pathogenesis of two specific microorganisms, HIV and tuberculosis (choice of pathogen may change without notice).

- HIV: Georgia Tomaras
- TB: Jason Stout

Nov. 22, Lecture 13: Vaccine design. We will investigate computational approaches to vaccine design including the prediction of the binding affinities of MHC I and II molecules to specific peptides and the prediction of the efficiency with which epitope

precursors are processed. We will discuss the potential role of modeling for adjuvant and delivery system design.

- Bette Korber

2.5.2 Mathematical Modeling of Infectious Diseases

This course was taught in Spring, 2005 by Alun Lloyd (NCSU). Between 25 and 35 students and postdocs attended the course, and several local and visiting faculty attended one or more of the lectures. One long-term outcome of the course was the production of a detailed set (150 pages+) of lecture notes, which will in due course be expanded into a complete text book.

The course focused on the simplest biological situations, namely directly transmitted infectious diseases. Discussion of more involved settings, such as indirectly transmitted diseases (e.g. malaria and other vector-borne infections) and multi-strain infectious agents (e.g. HIV and influenza) will be given, but with reduced emphasis. The main emphasis will be on epidemiological dynamics, and the links to ecological (predator/prey) theory. The importance of evolutionary dynamics will be highlighted where appropriate.

2.5.3 The Biophysics of Cell Signaling

This course was taught by Byron Goldstein (SAMSI university fellow) in Spring, 2005. Approximately 10 graduate students took the course for credit or audited. Several postdoctoral fellows participated as well.

Description: Cells must constantly sense their environment and respond. The macromolecules (ligands) that cells detect, and the concentrations at which they detect them, are determined by the cell surface receptors they express. Once a particular ligand binds to a receptor this information must be transmitted across the cell membrane if the cell is to respond. When the information transfer is successful, a complex set of events is initiated (a signaling cascade) that culminates in the activation or suppression of one or more cellular responses. In many cases the binding of a ligand to a receptor is not sufficient to initiate signaling. Rather, for these ligands signaling is initiated by inducing receptors to aggregate with additional membrane proteins or among themselves. We will focus on receptor families that work in this way, including the growth factor receptors, the immune recognition receptors and the cytokine receptors. We will discuss models for ligand-receptor binding, ligand-induced receptor aggregation and ligand-initiated cell signaling. Two basic types of cell signaling models can be distinguished: simple cell signaling models which ignore the details of the signaling cascade, but give insight into how ligand-receptor binding properties affect signaling outcomes; and detailed models, which include specific molecular components and interactions beyond the ligand and receptor, and that are required to gain a mechanistic understanding of cell signaling cascades. Both types of models will be considered.

3 Program on Data Assimilation in Geophysical Systems

3.1 Program and its Objectives

The problem of assimilating data into a geophysical system, such as one related to the atmosphere or oceans, is both fundamental in that it aims at the estimation and prediction of an unknown true state, and challenging as it does not naturally afford a clean solution. It has two equally important elements: observations and computational models. Observations measured by instruments provide direct information of the true state. Such observations are heterogeneous, inhomogeneous in space, irregular in time, and subject to differing accuracies. In contrast, computational models use knowledge of the underlying physics and dynamics to provide a description of state evolution in time. Models are also far from perfect: due to model error, uncertainty in the initial conditions and computational limitations, model evolution cannot accurately generate the true state.

In its broadest sense, data assimilation (henceforth refereed as DA) is the subject that arises in all scientific areas that enjoy a profusion of data. By its very nature, DA is a complex interdisciplinary subject that involves statistics, applied mathematics, geophysics and engineering. Driven by operational demands for numerical weather prediction, however, the development of DA so far has been predominantly lead by the geophysical community. In order to further DA beyond the current state-of-art, the development of effective methods must now be viewed as one of the fundamental challenges in scientific prediction.

With this background, the two overarching goals of this SAMSI program for the Spring 2005 are:

1. to initiate and catalyze the collaboration of statisticians, mathematicians and scientists from the relevant domain science areas.

2. launch focused efforts aimed at understanding underlying issues and pioneering new approaches by providing a platform for interdisciplinary collaborations.

3.2 Core Group

A core group of researchers is gathering at SAMSI for the program in data assimilation. They include:

3.2.1 Visiting Researchers

- Kayo Ide (Atmospheric and Ocean Sciences, UCLA), UNC-SAMSI University Fellow, Spring 2005
- Leonard Smith (Time Series Institute, LSE and Pembroke College, Oxford) NCSU-SAMSI University Fellow, Spring 2005
- Monica Stephens (Mathematics, Spelman College)

- Roy Choudhury (Mathematics and Statistics, University of Central Florida)
- Marianna Pensky (Mathematics and Statistics, University of Central Florida)
- Juan Restrepo (Mathematics and Physics, University of Arizona)

3.2.2 Leaders from Triangle Universities

- Chris Jones (Mathematics, UNC)
- Susan Lozier (Geosciences, Duke)
- Sujit Ghosh (Statistics, NCSU)
- Amarjit Budirajah (Statistics, UNC)

3.3 Postdoctoral Fellows

- Shree Kare (PhD, Princeton, GFDL)
- Amit Apte (PhD, University of Texas at Austin, Physics)

3.3.1 Graduate Students

- Kristen Foley (Statistics, NCSU)
- Liyan Liu (Mathematics, UNC)
- Steven Foster (Mathematics, UNC)
- Nicole Mech (Geosciences, Duke)

3.3.2 Others

In addition, other researchers are involved in the activities including: Montse Fuentes (faculty, Statistics, NCSU), Jim Berger (Director, SAMSI), Lou Howard (emeritus faculty, MIT and FSU), Hayder Salman (postdoc, Math, UNC), Neil Martinssen-Burrell (postdoc, Math, UNC), Zhi Lin (grad, Math, UNC), Ke Xu (grad, Math, UNC).

3.4 Program Organization and Activities

The program goals are being achieved through carefully designed plans for interactive research activities that span the semester: including two major planned workshops, one-day mini-workshops with focussed themes, the initiation of collaborations through ongoing working groups, a weekly extended seminar bringing in an outside expert, and the training of young mathematical scientists through a full course and a (slightly less formal) seminar course.

A core group of researchers is gathering at SAMSI for the entire spring semester, 2005 and their activities and interaction will provide the main focal point of the program.

Part of the program is based on a collaborative effort with two other institutes: the Institute for Pure and Applied Mathematics (IPAM) at University of California, Los Angeles (UCLA) and the Institute for Mathematics Applied to Geosciences (IMAGe) at the National Center for Atmospheric Research (NCAR). In the following, we describe the key elements that are bringing the program to life.

The workshop series, with the first in January and the second in February, laid the foundation for the ongoing work of the program. They have informed the key participants about the state-of-the-art, and served to educate the students with an overview of advanced aspects of the subject. They further set the tone for the full program in its matching of expertise from various areas with a stake in data assimilation, both as consumers of the DA product and producers of techniques. These two main workshops have complementary purposes and together will provide a full exposé of the subject, its potential and its future. Details of these and other activities follow.

3.4.1 Tutorial and Kick-Off Workshop in Research Triangle Park

The program was officially inaugurated with the Tutorials on January 23rd and 24th, held at the Radisson Governors' Inn in RTP. These tutorials were the beginning part of the Kickoff Workshop, held Jan 24th through 26th.

The one-day tutorial covered the basics of DA statistical approaches, and filtering. This served to give a concise, but comprehensive, introduction to the participants unfamiliar with DA. These were given by Chris Wikle (Statistics, University of Missouri at Columbia) and Jeff Anderson (NCAR).

The Kick-Off workshop was organized to present the current state-of-art and the associated challenges in DA. The focus was on the needs for data assimilation on behalf of the operational community and the current techniques being used in practice. Among the speakers were users and developers of DA schemes from NRL, NCEP and NASA.

Covered were statistical methods and applications, ocean DA, atmospheric DA, new strategies, and current operational numerical weather prediction systems for the US. The aim of the workshop was to

formulate plans for a set of working groups that would meet at SAMSI through the semester. There was thus an emphasis on discussion after each presentation, the time length of which equalled that of the talk. The co-chairs of each session were asked to guide the discussion with an eye to formulating directions for future research and collaboration.

3.4.2 Distinguished Lecture

The workshop also showcased the Distinguished Lecture delivered by Eugenia Kalnay (University of Maryland-College Park) with the title: "Data Assimilation and Ensemble Forecasting: Two Problems with the Same Solution? "

3.4.3 IPAM workshop on "Mathematical Challenges in Geophysical Data Assimilation"

A major workshop was then held from February 22nd through 25th, 2005, at the Institute for Pure and Applied Mathematics (IPAM), UCLA. This was the first SAMSI-IPAM collaboration and proved to be highly successful.

While the kickoff workshop served to expose the needs of the DA community and the current DA strategies, the focus of the IPAM workshop was on engaging the mathematical community in the challenges presented by data assimilation. We therefore brought together mathematically-oriented DA practitioners with mathematicians in related fields including: information theory, applied probability, inverse problems, control theory, computational stochastic processes, dynamical systems, and computer science.

During the workshop, there were two discussion sessions with the specific aim of identifying outstanding problems in geophysical DA that require mathematical advances to help resolve.

3.4.4 Mini-Workshops

In addition to the two core workshops mentioned above, four mini (one-day) workshops are hosted to target the issues of hight interests among the SAMSI core group (section 3.2).

The first workshop focused on giving a statistical interpretation to the theory of indistinguishable states, which has emerged from a dynamical systems view of data assimilation. The aim is to produce a refined and effective technique by incorporating these two approaches. This was held on March 17th and will include Jim Hansen (MIT). The second mini-workshop will focus on Lagrangian data assimilation (currently planned for April 12th, 2005) during which we will host visits by Arthur Mariano (RSMAS, Miamai) and Keith Thompson (Mathematics, Dalhousie). The third will focus on bridging statistical approaches and sequential DA (date to be decided). The fourth will explore the impact and role of DA in the climate system (date to be decided).

3.4.5 NCAR Summer School

The SAMSI/IMAGe summer school will be held June 13-27, 2005 at NCAR. It will be run by statisticians and geophysicists who are leaders in the field of data assimilation. By bridging the gap between basic and applied research on ensemble data assimilation, the workshop will provide participants with an understanding of the most recent advances and the most critical unsolved problems in this rapidly growing field. Lectures and discussion will be supplemented by a series of computational explorations using the Data Assimilation Research Testbed facility at NCAR. By the end of the workshop, participants will be equipped with the tools to attack the problems posed in the lectures and to undertake research in data assimilation for a large variety of applications.

3.5 Ongoing Activities

3.5.1 Working Groups

At the end of the Kickoff Workshop, an outline of potential working groups was formulated. These groups are meeting once per week and will continue into Summer, 2005. The core group then met at SAMSI the following week and isolated five to six topics that would form the basis of working groups. These have different foci and expectations. Some consider fundamental issues, while others work on specific projects. These are:

1. Lagrangian Data
2. Ocean Data Assimilation
3. Model Error, Inadequacy and Empirical Verification
4. Low-Dimensional Behavior and Observing System Design
5. Estimation and Prediction

Each group is holding weekly meetings for about 1 to 1+1/2hrs. There is a weekly meeting of all participants to discuss the ongoing working group activities and organizational logistics.

3.5.2 Seminar Series “Nail’Em Down Day” (NEDD)

To support the efforts by the working groups, we have invited scientists from a variety of disciplines. these people will deliver a seminar (usually on Friday at 11am). This will be followed by a lunch and en extended discussion. The idea is to bring people from whom the core group can learn some specific idea or technique. The name of the seminar is intended to suggest that we will question the speaker until we collectively

understand the work they are presenting. In most cases, we ask the speakers to address a specific topic and/or give a certain emphasis in their presentation.

3.5.3 Courses

Two weekly educational events are being held in addition to the above activities.

The main SAMSI course (available as credit to students from the triangle universities) will be taught by Kayo Ide (UCLA and SAMSI-UNC University Fellow):

Data Assimilation Methods for the Ocean and Atmosphere

This course introduces the concepts of data assimilation derived in the context of estimation theory and covers a variety of methods for numerical weather prediction and ocean forecasting, such as optimal interpolation, Kalman-filtering and variational based methods. Advanced topics and the state-of-art data assimilation systems will also be discussed.

In addition, Leonard Smith (LSE, Oxford and SAMSI-NCSU University Fellow) will present a seminar course at NCState each week. The focus will be on model error, forecasting and a dynamical systems view of these issues.

3.6 Anticipated Outcomes

As outcomes, we anticipate the fruit of the working group activities to be represented by publications, student projects and the initiation of ongoing collaborations. The avowed goal is to bring to bear on data assimilation problems the combined expertise of applied mathematicians and statisticians and this will be manifest in these products.

This program has been conceived with the notion in mind that data assimilation is a subject that is at a crossroads. It stands to gain enormously from the increased input of applied mathematics, particularly in the form of dynamical systems, scientific computation and applied probability, and statistics, particularly in terms of filtering techniques and stochastic computing.

A specific outcome will be the formulation of future directions of research that will foster these connections and entice practicing mathematicians and statisticians into this rich and thriving area. This will be presented as a white paper. The white paper should help fertilize collaboration across the disciplines and serve as an agenda to the mathematical community engaged in the future advancement of DA.

We have been asked by the editorial board of the journal “Inverse Problems” to prepare a special volume outlining the state-of-the-art in data assimilation with a targeted audience being those slightly outside the subject but knowledgeable in a relevant area, such as inverse problems. We plan to create this volume based on papers solicited from key participants in the SAMSI program, particularly from the two main workshops.

4 Education and Outreach Program

The SAMSI Education and Outreach (E&O) program has become a truly national program during the third year of SAMSI operation. Three of our workshops (Outreach Days, the Undergraduate Interdisciplinary Workshop, the Industrial Mathematical and Statistical Modeling Workshop) have become so popular that we now must turn away qualified applicants for lack of capacity. In addition to a goal of major education and outreach to members of the scientific community at non-local (i.e., non-Research Triangle) and non-research oriented institutions, the program also focuses heavily on significant participation by SAMSI Graduate Fellows and Postdocs as a means to train young investigators in the important tasks and methodologies of outreach and education.

The SAMSI Education and Outreach program for 2004-2005 included the following activities:

Outreach days for undergraduates This Workshop (which has become an annual feature of SAMSI Outreach to undergrads) was held on February 18-19, 2005. Undergraduate students from a national applicant pool were invited to SAMSI for a day and half long program during which the members of the SAMSI Directorate gave an overview of SAMSI current and future programs, and the opportunities for participation were discussed. Technical presentations on three research programs were given. Lindsay Cowley (Duke U.) summarized Comp Bio Program activities, while Ken Bollen (UNC-CH) and Jane Zavisca (SAMSI postdoc) gave an afternoon tutorial and interactive session on the Latent Variables in the Social Sciences Program. A Friday evening pizza party allowed members of the Directorate and postdocs to meet with student participants and discuss career opportunities in statistics and applied mathematics. The Saturday morning presentation was led by Ralph Smith (NSCU) and Emily Lada (a former SAMSI postdoc now at SAS) and involved a presentation on the Multiscale Program that featured a number of lively demos and hands-on activities. There were 31 total participants including 20 females, 4 African-Americans, 1 Hispanic and 1 Native American.

Undergraduate Interdisciplinary Workshop: A one week workshop for 18 undergraduates (7 females, 5 African-Americans, 3 Hispanics) was held May 30 -June 4, 2004 and is scheduled to be repeated May 29 -June 3, 2005 with a substantially increased number of participants (27 applicants have been accepted to date). During this week, the students are exposed to an intensive program involving formulation of inverse problems, hands-on collection of experimental data, and mathematical and statistical analysis of data. The problem chosen (vibrations of a cantilevered beam mounted with sensors and actuators) is a paradigm for the theme of SAMSI (an interdisciplinary approach involving applied mathematics, statistics, and applications in domain science to solve complex practical problems) and embodies the principles to which we are trying to attract young mathematicians/statisticians. All pre-

sentations and tutorial sessions are organized and given by SAMSI Graduate Fellows and Postdocs under close supervision of the SAMSI Associate Director for E&O.

Industrial Mathematical and Statistical Modeling Workshop: This ten day annual workshop for Graduate Students was held July 25-August 3, 2004. This workshop involved 30 graduate students (13 Females, 2 African-Americans, 2 Hispanics) selected from a competitive national pool along with representatives from five industrial/government labs to work in teams on problems posed by the industrial representatives. Labs represented and topics in 2004 were: Lincoln Laboratories: Mobile sensing of aerosolized chemical and biological agents; Jet Propulsion Labs: Complexity of event-driven command sequences in parallel; Calabazas Creek Research, Inc.: Optimized design of electron guns for high power RF application; CIIT Centers for Health Research: Identifying respiratory parameters from plethysmography data; Air Force Research Lab/Kirtland AFB: The unifying of perspectives on attitude and shape control [for more details see http://www.ncsu.edu/crsc/reports/reports04.htm]. Again SAMSI Graduate Fellows and Postdocs were heavily involved as local mentors for the participants. A number of the projects (e.g., "Mobile sensing', "Complexity of event-driven", "Identifying respiratory parameters..") were excellent vehicles for illustration of the SAMSI focus on combining applied mathematical and statistical approaches to important scientific and engineering problems. A similar workshop (with new projects and presenters) is scheduled for July 25 - August 2, 2005.

Workshop on Mathematics Meets Biology: Epidemics, Data Fitting and Chaos: In an effort to reach college teachers (typically at primarily teaching institutions) with the SAMSI message, SAMSI joined with the Mathematical Association of America (MAA) in co-sponsoring one of their Professional Enhancement Programs (PREP, see http://www.maa.org/prep/) held May 26-29, 2004 at the University of Louisiana at Lafayette. Participants were college teachers chosen from a national applicant pool. The SAMSI Associate Director for E&O (H.T. Banks) along with several current SAMSI Graduate Fellows presented one full day of the program; their material included hands-on population modeling where combined deterministic and probabilistic aspects of the models are needed to describe experimental data featuring both inter- and intra- individual characteristics. There were 29 participants (10 females, 1 African-American, 1 Hispanic). SAMSI has been asked to participate and co-sponsor again in 2005; Banks will lead a team of SAMSI Graduate Associates and former Graduate Fellows in presenting one half (1 and days) of this years program scheduled for May 25-27, 2005. The planned topic is "Statistical and Mathematical Aspects of HIV Modeling".

Mathematical and Experimental Modeling Course on NC-REN TV: SAMSI again sponsored an innovative course on closed circuit TV (the NC-REN TV network) to multiple campuses in NC in the Fall, 2004 semester. This course (taught this year by Ralph Smith) combined (in the SAMSI spirit)

mathematical, statistical and experimental aspects of modeling of several physical and biological processes including thermal, mechanical and size-structured population systems. In addition to being given to a live audience on the NCSU campus, the course was live (and interactively) broadcast to classes at UNC-G. Students from the remote campuses came to the CRSC Lab at NCSU to carry out experiments with the NCSU students several times during the semester.

Diversity: See Section I.H for discussion of the efforts to achieve diversity.

Courses: See the program reviews in Section I.E for discussion of the SAMSI courses.

5 Planning, Hot Topics, and Technology Transfer Workshops

5.1 Design and Analysis of Computer Experiments for Complex Systems

This was a joint workshop with the Canadian National Program on Complex Data Structures (NPCDS) held July 13-17, 2004 in Banff, CA. It's purpose was to investigate the potential of a program in this area for both SAMSI and NPCDS. The following individuals were at the meeting and expressed considerable interest in a full program: M.J. Bayarri (U. Valencia), James Berger (SAMSI), Derek Bingham (Simon Fraser U.), Andrew Booker (Boeing), Hugh Chipman (Acadia U.), Dennis Cox (Rice), Michael Hamada (Los Alamos Nat. Labs), Dave Higdon (Los Alamos Nat. Labs), Leslie Moore (Los Alamos Nat. Labs), Nick Hengartner (LANL), William Notz (Ohio State U.), Shane Reese (Bringham Young U.), Tom Santer (Ohio State), Randy Sitter (Simon Fraser U.), Jerry Sacks (Duke emeritus), David Steinberg (Tel Aviv U.), and William Welch (U. British Columbia).

The excitement generated at the meeting led to the submission of the proposal discussed in II.B, which is on track to become a major SAMSI program.

5.2 Workshop on Data Mining Methodology and Applications

This workshop, joint with NPCDS and organized by Hugh Chipman of Acadia University, was held at the Fields Institute from November 28-30, 2004. SAMSI researchers participated in sessions designed to highlight the research and developments from the Data Mining and Machine Learning Program conducted by SAMSI during the previous year. Cosponsoring such events is a highly effective way to promulgate the research and insight as to future directions from a SAMSI program.

5.3 Planning Meeting on Complex Data Structures

This was joint with NPCDS, held April 9-14, 2005 at the BIRS conference center, organized by James Stafford of the University of Toronto. This was primarily a workshop for NPCDS to plan its major initiatives over the next few years. SAMSI was involved to see where synergistic activities could be developed, in relation to its own programs.

The most immediate development was the furthering of the discussion of the Program on Computer Modeling, since many of the key NPCDS individuals involved in their program on this topic were at the meeting. An additional opportunity arose, in this regard, when NPCDS also had a session on a new program on Modeling of Marine Ecological Systems, which will be led by Chris Fields and Joanna Mills of Dalhousie University. SAMSI was strongly considering having a subprogram on ecological modeling within the Computer Modeling program, and this would provide a good synergistic fit with the NPCDS effort. Also

of potential interest in this regard is a possible NPCDS program on computer modeling of forest fires.

More speculative opportunities were also discussed. In particular, both SAMSI and NPCDS have interests in pursuing programs in quantitative medical areas, and possible subjects and activities in this direction were discussed.

5.4 Random Graphs and Stochastic Computation

This workshop, to be held at SAMSI on June 13-14, 2005, has two goals:

- To provide a forum for cross-disciplinary communication on hot-topic research in random graphs, graph theory, related random matrix theory, applied statistical modelling with graphical structures and computational methods development related to graphical model analysis.
- To define specific areas for focus for a future SAMSI research program.

The workshop, being organized by Beatrix Jones (Massey Univ.) and Mike West (Duke Univ.), will bring together a small number of leading researchers from statistics, applied probability, applied mathematics and computer science, all of whom work on critical aspects of modelling, analysis and computation involving graphical models, graph theory, associated theories of random graphs and random matrices, and applications of statistical and mathematical models involving such structures in scientific and social scientific applications. Speakers will provide initial overview/orientation discussion aimed specifically at communication across the subdisciplinary boundaries, as well as highlights of core problems and research challenges from their own disciplinary perspectives. The invited oral presentations will all be intended to speak to the interest in communicating across the various flavours of mathematical sciences as opposed to focusing wholly and more traditionally on within-subdiscipline research advances. The oral presentations and discussions will be complemented by two poster sessions and a culminating ”open house” discussion that will aid in defining next steps towards a follow-on plan for a scientifically diverse but also goal-focused SAMSI research program proposal.

5.5 Data Mining Technology Transfer Workshop

The purpose of this workshop, to be held at SAMSI from June 20-24, 2005, is to present results from the 2003-04 SAMSI program on Data Mining and Machine Learning (DMML) to practicing applied mathematicians, statistical scientists and domain scientists from industry and government. David Banks, professor of the practice of statistics at Duke University and co-leader of the DMML program, is the organizer and will be the principal lecturer.

Necessarily, some background will be presented that covers material not arising form the DLLM program, but the focus will be on theory and methodology developed either during the SAMSI program or in subsequent collaborations catalyzed by the program. Topics will include classification, over-completeness and support vector machines. Computer laboratories each day will provide hands-on experience to participants.

The planned format for each day is

- 9:00-10:30 AM: Banks Lecture
- 11:00-12:30 PM: Banks Lecture
- 2:00-4:00: Computer Laboratory
- 4:00-5:30: Guest Lecture

6 Distinguished Lecture Series

This lecture series brought some of the worlds most prominent statistical and mathematical scientists to SAMSI. In addition to their very widely attended lectures, the distinguished visitors held highly useful discussions with SAMSI researchers. The list of distinguished lecturers for 2004-05, and the titles and abstracts of their talks, follow:

Speaker: *Bette Korber,* Fellow, Los Alamos National Laboratory

Date: November 30, 2004

Title: Diversity considerations in HIV Vaccine Design

Abstract: HIV-1 is a highly variable pathogen, and the variation is in part the direct the result of immune escape. Every person carries a population of distinct of the virus; generally viruses sampled from early in the infection are very similar, and within-patient diversity builds over time. HIV has in its repertoire many evolutionary strategies that increase the diversity, not only base substitution, but recombination, frequent insertion and deletions, and changes in patterns of glycosylation. Population diversity grows over time, evolving outward within sets of distinct lineages called subtypes. Inter-subtype recombination events are common, and such recombinants can found their own lineages. Part of the difficult in making an effective vaccine is overcoming this diversity. We are working at Los Alamos on the task of designing vaccine antigens that have increased potential to stimulate cross-reactive responses. Our strategies to date involves using artificial consensus sequences and maximum-likelihood derived ancestral sequences to find an approximation of a central position in antigenic space, and to derive an artificial antigen that has better cross-reactive

potential with circulating strains than any single natural strain. Our experimentalist collaborators at Duke and University of Alabama have tested these constructs, and the initial results are promising. We are now turning our attention to focusing on sequences derived from acute infection, as at least in the case of A and C subtypes, these seem to have shared distinctive properties and may have conserved features that make them a narrower target than the full spectrum of viruses isolated at any stage of progression. We are also developing strategies to define combinations of proteins that could provide maximum coverage of a population.

Speaker: *Eugenia Kalnay,* University Distinguished Professor at the University of Maryland

Date: January 25, 2005

Title: Data assimilation and ensemble forecasting: two problems with the same solution?

Abstract: Until 1991, operational numerical weather prediction models utilized a single control forecast representing the best estimate of the state of the atmosphere at the initial time. In 1992, operational NWP models began to utilize ensembles of forecasts from slightly perturbed initial conditions. Such ensemble forecasts provide human forecasters with a range of possible solutions, whose average is generally more accurate than the single deterministic forecast, and whose spread gives information about the forecast errors. It also provides a quantitative basis for probabilistic forecasting. The two essential problems in the design of an ensemble forecasting system are how to create effective initial perturbations, and how to handle model deficiencies, which make the ensemble forecast spread smaller than the forecast error. In this talk we present a brief historic review of ensemble forecasting and current methods to create perturbations. We point out that the promising approach of ensemble Square Root Kalman Filtering for data assimilation can solve, at the same time, the problems of obtaining optimal initial ensemble perturbations, and possibly estimating the impact of model errors. We also discuss the problem of coupled systems with instabilities that have very different time scales.

Speaker: *Alan Perelson,* Senior Fellow, Los Alamos National Laboratory

Date: January 31, 2005

Title: Modeling viral infections

Abstract: Viruses such as HIV and hepatitis B and C infect millions of people and lead to wide-spread disease and death throughout the world. Here I will show how mathematical and statistical analysis of data obtained from virally-infected people placed on antiviral treatment has helped unravel a set of mysteries about these virus and lead to improved therapies for them. A new field called viral dynamics has arisen with the aim of studying the kinetics of viral infection and treatment. For those new to this area, I will present a overview of work done to date emphasizing outstanding problems.

Speaker: *James Robins*, Mitchell L. and Robin LaFoley Dong Professor of Epidemiology, Harvard University

Date: March 29, 2005

Title: Optimal sequential decisions and causal inference

Abstract: How should one make optimal treatment or intervention decisions in the face of confounding by both high dimensional measured variables and by unmeasured variables. The problem is that 1) in the face of unmeasured confounding, Bayesian inference is the natural way to combine prior beliefs about the magnitude of confounding by unmeasured factors with the data but, 2) due to the curse of dimensionality, standard Bayesian methods fail when analyzing high dimensional data, because the standard Bayesian approaches are too ambitious and try to extract more information from the data than is feasible.

To make progress I propose a new type of Bayes-frequentist compromise that allows a valid Bayesian analysis of high dimensional data by reducing the data to a frequentist estimating function viewed as a stochastic process indexed by the causal parameterS of interest. I make connections between this approach and approaches based on causal discovery methods which assume the data distribution is faithful to some causal directed acyclic graph marginalized over some variables and conditioned on others.

F. Industrial and Governmental Participation

Government and industry participation in SAMSI programs and activities reflects broad interest in the SAMSI vision. The following summarizes participation during 2004-05.

Data Mining and Machine Learning Program: A variety of individuals participated in, or interacted significantly with, the program working groups. These included:

- *Abbott Labs:* Yvonne Martin, mid-year Bioinformatics Workshop visitor and participant in resulting proposal;
- *AT&T Labs Research:* William DuMouchel, Scientific Committee member and attendee at opening and closing workshops.
- *Bureau of Labor Statistics (BLS):* Testbed data, on which text mining was used to infer BLS occupational categories for Census long-form answers.
- *General Motors:* Testbed databases, the analysis of which is described in final report. More than 15 GM personnel participated at various stages, led by Clifford Hodges and Lynn Truss.
- *ICAGEN:* Michiel van Rhee, member of Bioinformatics Working Group.
- *Metabalon:* Chris Beecher, member of Bioinformatics and Theory & Methods Working Groups. Participant in subsequent NIH proposal.
- SAS Institute: Gerardo Hurtado, member of Bioinformatics and Support Vector Machines Working Groups; Warren Sarle, Scientific Committee Member

Network Modeling for the Internet: The Scientific Committee included 2 industrial members (AT&T Research and Avaya Labs). One of the theme problems of the Internet Tomography Workshop was a testbed developed by Avaya. The National Security Agency provided partial funding for the workshops on Internet Tomography and Sensor Networks. There many workshop participants from industry, including UAvaya Labs, CAIDA, Eurandom, IBM Research, Lucent Technologies Bell Labs, MCNC, NIST, Palo Alto Research Center, Sprint Labs, and the U.S. Navy,

Multiscale Model Development and Control Design: The interdisciplinary nature of the Multiscale Program made it highly amenable to industrial and government participation. During the Opening Workshop, there was significant participation, including invited presentations, by scientists from the Center for Disease Control (CDC), Lawrence Livermore National Lab, Los Alamos National Laboratory, NIEHS, and the US Army Research Office. The Workshop on Soft Matter Materials included participants from NIEHS and the US Army Research Office. Data collected by colleagues at NASA Langley Research Center was used to validate and motivate mechanisms incorporated in multiscale material models.

Computational Biology of Infectious Disease: The following people participated as speakers in workshops, participants in working groups or visitors. Joanna Fueyo, IBM Life Sciences; Padraic Neville, SAS; Anne De Groot, Epivax; Jeff Butterworth , Alien Skin Software; Rory Conolly , CIIT Centers for Health Research; Julia Kimbell, CIIT Centers for Health Research; Delong Liu, CIIT/SAS; Kevin Thomas Morgan, Sanofi-Aventis; and Qiang Zhang , CIIT Centers for Health Research.

Latent Variable Modeling in the Social Sciences: Government participation included David Dunson of NIEHS, who led a working group, and workshop attendees from NIH, Statistics Canada, the Census Bureau and the Bureau of Labor Statistics. Industrial participation included Paul Biemer of RTI International, who led a working group, multiple researchers from MetaMetrics – who participated throughout the program in several working groups – and workshop attendees from MetaMetrics, Pfizer, RTI, and Scientific Software International.

Data Assimilation in Geophysical Systems: Data assimilation is critical in the area of Numerical Weather Prediction that has great societal importance and impact. A number of Government Labs have focussed efforts in this area. We have made connections through bringing researchers from Government Labs to the workshops and to visit SAMSI for periods of time that have afforded a stronger connection with their work and the beginnings of collaborative efforts with our group. Craig Bishop (NRL), Ron Gelaro (NASA), Jeff Anderson (NCAR), Doug Nychka (NCAR) and Zoltan Toth (NCEP) all delivered lectures at the opening workshop. Jeff Anderson gave a tutorial. Carolyn Reynolds (NRL), Chris Snyder (NCAR) and Joe Tribbia (NCAR) delivered lectures at the second workshop, held at IPAM. In addition, Carolyn Reynolds has returned here for a week to deliver an in-depth lecture and consult further with the assembled group at SAMSI. A summer school on Fusing Geophysical Data and Models will be held at NCAR in June as part of the SAMSI DA program.

There was also strong industry/government/national laboratory participation through the NISS and NISS/SAMSI affiliates (see Section I.I.) and on the National Advisory Council, with members from Google, Microsoft and Los Alamos National Laboratory.

G. Publications and Technical Reports

I. Data Mining and Machine Learning

Publications and Technical Reports

- Banks, D.
  "A dimension reduction technique for local linear regression"
  Classification, Cluster Analysis, and Data Mining.
  Springer-Verlag, Berlin 2004
- Banks, D., House, L., Arabie, P., McMorris, F. R., and Gaul, W., eds.
  "*Classification, Clustering, and Data Mining"*
  Springer-Verlag, Heidelberg, 2004
- Banks, D., ed.
  "Data Mining: Principles and Methods"
  To be submitted to the ASA/SIAM Series on Statistics and Applied Probability, 2005. (This monograph will contain approximately ten individual papers arising from the program.)
- Banks, D. and L. House
  "Robust multidimensional scaling"
  Proceedings in Computational Statistics 2004, 251--260.
  Physica-Verlag, Berlin
- Banks, D. and F. Liang
  Review of "*The Elements of Statistical Learning"* by T. Hastie, R. Tibshirani, and J. Friedman.
  Journal of Classification, to appear.
- Banks, D., J. Woo, D. Burwen, P. Perucci, and R. Ball
  "Comparison of four methods of data mining in the vaccine adverse event reporting system."
  Phamacoepidemiology, to appear, 2005.
- Beecher, C., Cutler, A., House, L., Lin, X., Truong, Y., and Young, S.
  "Learning a complex metabolomic dataset using random forests and support vector machines."
  Submitted to SIGKDD 2004.
- Fokoué, E.
  "Parsimonious function representation and optimal predictive model selection"
  SAMSI technical report number 2004-19; Submitted to Canadian Journal of Statistics
- Fokoué, E.and B. Clarke
  "Optimal model list selection for prediction."
  SAMSI technical report number 2004-20
- Fokoué, E. (2004).
  "Sparsity through prevalence estimation."
  To be submitted to the J. Machine Learning Res.

- Hawkins, D. M., Wolfinger, R. D., Liu, L., and Young, S. S. (2003).
  "Exploring blood spectra for signs of ovarian cancer."
  Chance, 16, 19-23.
- House, L. and D. Banks
  "Cherry-picking as a robustness tool."
  Classification, Cluster Analysis, and Data Mining
  Springer-Verlag, Berlin, 2004
- House, L., and D. Banks
  "Robust multidimensional scaling."
  Submitted to Proceedings of COMPSTAT 2004.
- Jeske, D. and Regina Liu
  "Mining massive text data and developing tracking statistics."
  Classification, Cluster Analysis, and Data Mining, 495--510
  Springer-Verlag, Berlin 2004
- Karr, A. F. and A. P. Sanil
  "Analysis of demand sensing data."
  Report delivered to GM on December 21, 2004.
- Lin, X. and B. Clarke
  "Bayesian effective samples and parameter size."
  Submitted to Journal of the American Statistical Association; SAMSI technical report number 2004-21.
- Lin, X. and Y. Zhu
  "Degenerate expectation-maximization algorithm for local dimension reduction."
  Classification, Cluster Analysis, and Data Mining, 259—268
  Springer-Verlag, Berlin 2004
- Liu, L., Hawkins, D. M., Ghose, S., and Young, S. S.
  "Robust singular value decomposition analysis of microarray data"
  Proc. Nat. Acad. Sci. 2004 (100) 13167-13172
- Simmons, S. J., X. Lin, C. Beecher, Y. Truong, and S. S. Young
  "Active and passive learning to explore a complex metabolism data set."
  Classification, Cluster Analysis, and Data Mining, 447—457
  Springer-Verlag, Berlin 2004
- Young, S. S., and N. Ge
  "Design of diversity and focused combinatorial libraries in drug discovery."
  To appear in Current Opinions in Drug Design and Development, 2004
- Young, S. S., M. Wang and F. Gu
  "Design of diverse and focused combinatorial libraries using an alternating algorithm."
  J. Chem. Info. Comp. Sci. 2003 (43) 1916-1921.
- Zhang, H. H., J. Ahn, X. Lin and C. Park
  "Gene selection via learning with concave penalty."
  Submitted to Bioinfomatics, 2004
- Zhang, H. H., M. Genton and P. Liu
  "Compactly supported radial basis kernel."
  Under revision for Journal of Machine Learning Research, 2004

Reports in Preparation

- Fokoué, E., D. Sun and P. Goel, (2004).
  "A new hierarchical prior structure for the relevance vector machine."
- Young, S. S., Feng, J., and Sanil, A.
  "PharmID: Pharmacophore identification using Gibbs sampling."

II. NETWORK MODELING FOR THE INTERNET

Publications and Technical Reports

- Chang, H., A.Q. Fu, N.D. Le and J. Zidek
  "Designing Environmental Monitoring Networks for Measuring Extremes"
  SAMSI Technical Report 2005-4
- Dinwoodie, I. H., Mosteig, E., Gamunid, E.
  "Algebraic Equations for Blocking Probabilities in Asymmetric Networks"
  SAMSI Technical Report 2004-15
- Park, C., Godtliebsen, F., Taqqu, M., Stoev, S. and Marron, J. S.
  "Visualization and Inference Based on Wavelet Coefficients, SiZer and SiNos"
  SAMSI Technical Report No. 2004-10.
- Park, C., Hernandez-Campos, F., Marron, J. S., Rolls, D. and Smith, F. D.
  "Long-Range-Dependence in a Changing Internet Traffic Mix"
  SAMSI Technical Report No. 2004-9.
- Stoev, S., Taqqu, M., Park, C. and Marron, J. S.
  "LASS: a Tool for the Local Analysis of Self-Similarity"
  SAMSI Technical Report No. 2004-7.
- Stoev, S., Taqqu, M., Park, C. and Marron, J. S.
  "Strengths and Limitations of the Wavelet Spectrum Method in the Analysis of Internet Traffic"
  SAMSI Technical Report No. 2004-8.
- Xu, P., Devetsikiotis, M. and Michailidis, G.
  "Adaptive Scheduling using Online Measurements for Efficient Delivery of Quality of Service"
  SAMSI Technical Report No. 2004-12.
- Xu, P., Devetsikiotis, M. and Michailidis, G.
  "Online Scheduling for Resource Allocation of Differentiated Services: Optimal Settings and Sensitivity Analysis"
  SAMSI Technical Report No. 2004-13.
- Zhu, Z. and Taqqu, M. S.
  "Impact of the sampling rate on the estimation of the parameters of fractional Brownian motion"
  SAMSI Technical Report No. 2004-14.

Reports in Preparation

- Abry, P. and Pipiras, V.

"Wavelet-based synthesis of the Rosenblatt process"

- Hernandez-Campos, F., Le, L., Marron, J. S., Park, C., Park, J., Pipiras, V. Smith, F. D., Smith, R. L., Trovero, M. and Zhu, Z.
  "Long Range Dependence Analysis of Internet Traffic"
- Park, C., Hernandez-Campos, F., Marron, J. S. and Jeffay, K.
  "Thresholded Log-Log Correlation Analyses of TCP Characteristics"
- Park, C., Marron, J. S. and Rondonotti, V.
  "Dependent SiZer: Goodness of Fit Tests for Time Series Models"
- Park, C., Veitch, D., Shen, H., Hernandez-Campos, F., and Marron J. S.
  "Semi experiment analysis of the shifting knee wavelet spectrum"
- Park, J. and Park C.
  "Robust H estimation, automatic choice of parameters"
- Pipiras, V.
  "On the use and usefulness of wavelet-based simulation of fractional Brownian motion"
- Piparas, V. and Taqqu, M. S.
  "Identification of periodic and cyclic fractional stable motions"
- Piparas, V. and Taqqu, M. S.
  "Integral representations of periodic and cyclic fractional stable motions"
- Piparas, V. and Taqqu, M. S.
  "Semi-additive functionals and cocycles in the context of self-similarity"
- Rolls, D., Michailidis, G. and Hernandez Campos. F.
  "Queueing Analysis of Network Traffic"
- Smith, R. L., Taqqu, M., Shen, H., Park, J., Zhu, Z. and Park, C.
  "Change Points and Long Range Dependence"
- Zhu, Z., Shen, H., Park, C., Hernandez-Campos, F and Marron, J. S.
  "Shot noise model, start times, micro-bursts"

III. Multiscale Model Development and Control Design

- Hatch, A.G., R.C. Smith and T. De
  "Model Development and Control Design for High Speed Atomic Force Microscopy"
  Proceedings of the SPIE, Smart Structures and Materials 2004, Volume 5383, pp.457--468, 2004.
- Edmonds, Jr., B., J. Ernstberger, K. Ghosh, J. Malaugh, D. Nfodjo, W. Samyono, X. Xu, D. Dausch, S. Goodwin and R.C. Smith
  "Electrostatic Operation and Curvature Modeling for a MEMS Flexible Film Actuator"
  Proceedings of the SPIE, Smart Structures and Materials 2004, Volume 5383, pp.134--143, 2004; SAMSI Technical Report 2004-4.
- Raye, J.K. and R.C. Smith
  "A Temperature-Dependent Hysteresis Model for Relaxor Ferroelectric Compounds"
  Proceedings of the SPIE, Smart Structures and Materials 2004, Volume 5383, pp.1--10, 2004; SAMSI Technical Report 2004-5
- Ball, B.L. and R.C. Smith

"A Stress-Dependent Hysteresis Model for PZT-Based Transducers"
CRSC Technical Report CRSC-TR04-14; Proceedings of the SPIE, Smart Structures and Materials 2004, Volume 5383, pp.23--30, 2004; SAMSI Technical Report 2004-6

- Smith, R.C. and A.G. Hatch
"Parameter Estimation Techniques for a Polarization Hysteresis Model"
CRSC Technical Report CRSC-TR04-19; Proceedings of the SPIE, Smart Structures and Materials 2004, Volume 5383, pp.155--163, 2004.
- Hatch, A.G., R.C. Smith and T. De
"Experimental Implementation of a Model-Based Inverse Filter to Attenuate Hysteresis in an Atomic Force Microscope"
Proc 43rd IEEE Conf. Dec. and Control, Paradise Island, The Bahamas, pp.3062--3067, 2004.
- Smith, R.C., A.G. Hatch, B. Mukherjee and S. Liu
"A Homogenized Energy Model for Hysteresis in Ferroelectric Materials: General Density Formulation"
CRSC Technical Report CRSC-TR04-23; Journal of Intelligent Material Systems and Structures, to appear.
- Massad, J.E. and R.C. Smith
"A homogenized free energy model for hysteresis in thin-film shape memory alloys"
International Journal on the Science and Technology of Condensed Matter Films, submitted.
- Smith, R.C., S. Seelecke, M.J. Dapino and Z. Ounaies
"A Unified Framework for Modeling Hysteresis in Ferroic Materials"
Journal of the Mechanics and Physics of Solids, submitted.
- Nealis, J.M. and R.C. Smith
"Adaptive Parameter Estimation Techniques for Magnetic Transducers Operating in Hysteretic Regimes"
IEEE Transactions on Control Systems Technology, submitted.
- Smith, R.C. and A.G. Hatch
"Parameter Estimation Techniques for a Class of Nonlinear Hysteresis Models"
Inverse Problems, submitted.
- Matthews, J.L., E.K. Lada, L.M. Weiland, R.C. Smith and D.J. Leo
"Monte Carlo Simulation of a Solvated Ionic Polymer with Cluster Morphology"
SAMSI Technical Report 2005-1; Smart Materials and Structures, submitted.
- Banks, H.T., V.A. Bokil, D. Cioranescu, N.L. Gibson, G. Griso, and B. Miara
"Homogenization of Periodically Varying Coefficients in Electromagnetic Materials"
SAMSI Technical Report 2005-2.
- Weiland, L.M., E.K. Lada, R.C. Smith and D.J. Leo
"Application of Rotational Isomeric State Theory to Ionic Polymer Stiffness Predictions"
SAMSI Technical Report 2005-3; Journal of Materials Research, submitted.

IV. From Genomes to Global Health: the Computational Biology of Infectious Disease

The following papers have been submitted to peer-reviewed journals. Several others are in preparation.

- Kepler TB, He M, Tomfohr JK, Devlin BH, Sarzotti M, Markert ML (2005) *Statistical Analysis of Antigen Receptor Spectratype Data.* Bioinformatics, under review.
- He M, Tomfohr JK, Devlin BH, Sarzotti M, Markert ML, Kepler TB (2005). *SpA: Web-accessible Spectratype Analysis: data management, statistical analysis and visualization*. Bioinformatics, under review.
- Lu J, Tomfohr JK, Kepler TB (2005). Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, under review.
- Tomfohr JK, Lu J, Kepler TB (2005). *Pathway Level Analysis of Gene Expression using Singular Value Decomposition*, BMC Bioinformatics, under review.

V. Latent Variable Models in the Social Sciences

- Bollen, K., Ray, S. and Zavisca, J. (2005). *Bayes factors in structural equation models (SEMs):* Schwarz's BIC and other approximations. In preparation.
- Bollen, K., A., R. Thomas, L. Wang, and John Hipp. (2005). Limited *Dependent Variable Models with covariate measurement error: A consistent instrumental variable estimator.* In preparation.
- Dunson, D.B., Palomo, J. and Bollen, K. (2005). *Bayesian structural equation modeling.* Handbook on Structural Equation Models (ed. S.-Y. Lee). Elsevier, to be submitted.
- Dunson, D.B., Palomo, J. and Zavisca, J. (2005). *Bayesian model selection and averaging in structural equation models.* Psychometrika, to be submitted.
- Kinney, S. and Dunson, D.B. (2005). *Bayesian fixed and random effects selection for binary response models.* In preparation.
- Kamata, A. & Bauer, D.J. (2005). *A note on the relationship between factor analytic and item response theory models.* Psychometrika, to be submitted.
- Kelly, C.M., Leibig, P.S., Edwards, L.J. (2005). *Factors Predicting the Extent of Nursing Home Regulatory Activity in the 50 States.* In preparation.

VI. Data Assimilation for Geophysical Systems

The program only began 3 months ago; publications and technical reports will be reported in next year's annual report.

H. Efforts to Achieve Diversity

From the beginning, SAMSI has focused on achieving diversity. This begins with composition of advisory committees. On the National Advisory Committee, 3 of 11 are women and one member is Hispanic (a co-chair). On the Local Advisory Committee, 2 of 9 are women and one member is African-American. On the Education and Outreach Committee, there were and 4 women and 2 African-Americans on the 8 member committee.

Specific efforts, and successes, of each of the programs towards achieving diversity are indicated below.

Data Mining and Machine Learning: Women and new researchers were well-represented throughout the program, with the possible exception of the invited speakers at the Kickoff Workshop, only one of whom was female. However, four of the seven talks at the new researchers session were by women. Approximately one-half of visitors were women, and speakers at the January-February mid-term workshops included a number of new researchers (among them, graduate students and postdoctorals). Note, also, that minority participation at the two undergraduate 2-day workshops, which were based on DMML material, was very strong, as detailed below.

In terms of those present for most of the program, diversity was quite substantial. One of the two postsdoctoral fellows was black, as was one faculty fellow. There were also four women faculty fellows. There were four women graduate fellows among nine total fellows. There were seven women among the 21 research visitors.

Network Modeling for the Internet: A number of women have been major players at all phases of this program, as noted below. Furthermore, strenuous efforts have been taken to address diversity, as detailed below.

Multiscale Model Development and Control Design: Two significant goals of the program were to make it as widely accessible to young people as possible and to recruit a diverse range of participants. Both goals were addressed through aggressive solicitation by the program leaders and committee via personal and research contacts as well as formal symposia and presentations. For example, the majority of participants who attended the Opening Workshop were notified by either the organizers or committee. Of the 82 attendees on the second day of the workshop, 19 were women, 4 were African American and 1 was Hispanic. Similar demographics were observed during the remainder of the workshop as well as at the other workshop.

Computational Biology of Infectious Disease: Two of the three program leaders are women, and a very good representation of women was achieved in the graduate courses, workshops and working groups. Three Hispanic (and several non-Hispanic) graduate students visited for the spring semester from Carlos Castillo-Chavez's group. One African-American woman took the Kepler/Cowell course and worked with Drs. Kepler and Cowell on a SAMSI-related project during the fall semester.

Latent Variable Models in the Social Sciences: Because of both the nature of the social sciences and strong efforts by SAMSI, diversity in the program was high. In particular, one of the four faculty fellows is an African American; both the postdoctoral fellow and the postdoctoral associate participating in the program are female; one working group was led by an African American; two of the four graduate fellows are women; one of the two tutorial presenters and four presenters at the opening kickoff workshop were women.

Data Assimilation in Geophysical Systems: The core group at SAMSI has been put together with diversity in mind. Of the group of ten senior researchers (University Fellows, Faculty Fellows, New Researcher and Long Term Visitors), six are male and four are female. One of the female researchers is African-American. Among the students, women outnumber men by three to one. Diversity is only lacking at the rank of postdoc – both of these are male. The Distinguished Lecturer associated with this program is a female distinguished professor from Maryland, Eugenia Kalnay.

Education and Outreach Program: SAMSI continues to use its E&O program to enhance its diversity efforts by active recruitment of under represented participants. Special efforts are made to recruit from HBCU's for all programs. During the past year, we have begun special efforts in recruiting Hispanics. Participation breakdowns include:

- PREP (May, 2004): Out of 29 participants, 10 were female, 1 was African American, and 1 was Hispanic.
- Undergraduate Workshop (June, 2004): Out of 18 participants, 7 were female, 5 were African American, and 3 were Hispanic.
- Graduate Workshop (July, 2004): Out of 30 participants, 13 were female and 2 were African American, and 2 were Hispanic.
- Undergraduate Workshop (February, 2005): Out of 31 participants, 20 were female, 4 were African American, 1 was Hispanic, and 1 was Native American.

I. External Support and Affiliates

1. External Support

Kenan Foundation: provided $50,000 of supplementary support.

Internet Program: The workshops on Internet Tomography and on Sensor Networks were jointly sponsored by SAMSI and the National Security Agency. The National Security Agency provided $15,000 for this purpose. J. S. Marron, SAMSI, was P.I. on the grant and Deborah Estrin, University of California, Los Angeles and Robert Nowak, University of Wisconsin, were co-P.I.s.

Data Mining and Machine Learning Program: in-kind support from industry and government was provided as follows:

- More than 15 researchers and managers from General Motors took part in discussions associated with analysis of the vehicle sales data.
- The Bureau of Labor Statistics provided data and advice to support text mining to infer occupational categories for Census long-form answers.
- Michiel van Rhee from ICAGEN (Research Triangle Park, NC) participated in the bioinformatics working group.
- Chris Beecher from Metabolon (Research Triangle Park, NC) participated in the Bioinformatics Working Group throughout the year. The Metabolon relationship has generated one proposal to date; another is in progress.
- Gerardo Hurtado from the SAS Institute (Cary, NC) participated in the Bioinformatics Working Group throughout the year.

Computational Biology of Infectious Disease: The Institute for Genome Sciences and Policy (IGSP) is providing support ($10,000) for the final workshop, *Collective Computational Biology for Infectious Disease*.

Latent Variable Models in Social Sciences: Canadian NPCDS supported participation of its personnel, and is co-sponsoring the May 4-5 workshop on complex surveys. All government and industrial participants were supported by their employers, not by SAMSI.

Data Assimilation in Geophysical Systems: The SAMSI program has leveraged other supported efforts around the triangle universities. The ONR supported project at UNC on data assimilation is supporting one student and one postdoc, both of whom are very much involved in the working groups and the SAMSI course. One of the SAMSI postdocs, Amit Apte, has split support from SAMSI and this ONR grant. In addition, support of Juan Restrepo, who has visited for two months, has been shared with this grant. An NSF-CMG grant at UNC has supported one student and one postdoc who are both actively involved in the SAMSI program.

2. Affiliates Program

Background: The NISS Affiliates Program and NISS/SAMSI University Affiliates Program are the largest programs of their kind among the DMS-funded mathematical sciences research institutes.

As a benefit of membership, NISS Affiliates and NISS/SAMSI University Affiliates may receive reimbursement for expenses to attend SAMSI workshops as well as NISS events. Through meetings and other activities, the NISS Affiliates and NISS/SAMSI University Affiliates inform the development of SAMSI programs. To illustrate, the Latent Variable Models in the Social Sciences (LVSS) program for 2004-05, as well as DMML program for 2003-04 and the NDHS program for 2005-06, exists to a significant degree because of affiliate interest.

NISS Affiliates and NISS/SAMSI Affiliates are listed below:

- *Corporations:* Avaya Labs, Basking Ridge, NJ; General Motors, Detroit, MI; GlaxoSmithKline, Research Triangle Park, NC and Collegeville, PA; ICAGEN, Research Triangle Park, NC; Merck \& Company, West Point, PA; MetaMetrics, Durham, NC; Nuevolution, Copenhagen, Denmark; RTI International, Research Triangle Park, NC; SAS Institute, Cary, NC; SPSS, Chicago, IL; Telcordia Technologies, Morristown, NJ

- *Government Agencies and National Laboratories:* Bureau of Labor Statistics, Washington, DC; US Census Bureau, Washington, DC; Los Alamos National Laboratory, Los Alamos, NM; National Agricultural Statistics Service, Fairfax, VA; National Center for Education Statistics, Washington, DC; National Center for Health Statistics, Hyattsville, MD; National Institute of Standards and Technology, Gaithersburg, MD; National Security Agency, Ft. George G. Meade, MD

- *NISS/SAMSI University Affiliates:* UCLA, Department of Statistics and Statistical Consulting Center; Carnegie Mellon University, Department of Statistics; Duke University, Institute of Statistics and Decision Sciences and Department of Mathematics; Emory University, Department of Biostatistics; Florida State University, Department of Statistics; University of Florida, Departments of Biostatistics and Statistics; University of Georgia, Department of Statistics; University of Illinois Urbana-Champaign, Department of Statistics; University of Iowa, Department of Statistics; Iowa State University, Department of Statistics; Johns Hopkins University, Department of Applied Mathematics and Statistics; University of Michigan, Departments of Statistics and Biostatistics; University of Missouri-Columbia, Department of Statistics; North Carolina State University, Department of Mathematics; North Carolina State University, Department of Statistics; University of North Carolina at Chapel Hill, Department of Biostatistics; University of North Carolina at Chapel Hill, Department of Mathematics; University of North Carolina at Chapel Hill, Department of Statistics and Operations Research; Oakland University, Department of Mathematics and Statistics; Ohio State University, Department of Statistics; Pennsylvania State University, Department of Statistics; Rice University, Department of Statistics; Rutgers University,

Department of Statistics; Southern Methodist University, Statistical Science Department; Stanford University, Department of Statistics; Texas A&M University, Department of Statistics

Affiliate Participation: Expenditures from Affiliates Reimbursement Account expenditures to attend SAMSI events exceeded $20,000. Participation in the Latent Variable Models in the Social Sciences (LVSS) program and by corporate, government and national laboratory affiliates was especially deep. Examples include:

- Paul Biemer, a researcher from RTI International, led the Complex Surveys working group in the program.
- Multiple researchers from Metametrics participated throughout the year in the Categorical Variables working group.
- An "Affiliates Workshop on Total Survey Error," held in Washington, DC on March 17-18, 2005 was co-hosted by the Bureau of Labor Statistics, and drew more than 80 attendees from more than 20 affiliates, including *all* federal agency affiliates.

Plans for the Future: At the affiliates annual meeting, held at NISS/SAMSI on March 3-4, 2005, affiliates discussed ideas and proposals for future SAMSI programs, and were briefed on programs planned for 2005-06 and 2006-07. Although there was strong interest in all programs, support for National Defense and Homeland Security (NDHS) program in 2005-06 was especially widespread among corporate, government and national laboratory affiliates. In particular, the Centers from Disease Control and Prevention--through the National Centers for Health Statistics, Los Alamos National Laboratory and General Motors have committed to provide personnel and testbed databases for the NDHS program.

J. Advisory Committees

The four advisory/oversight committees of SAMSI are as follows:

- The Governing Board (GB), which oversees SAMSI's administration, finances, evaluation and partner organization relationships. The GB meets with the Directorate twice a year. The SAMSI Director also has a conference call with the GB Chair and/or GB every other week.
- The National Advisory Committee (NAC) consists of leading national scholars, and is the primary external input into program choice and development. The NAC met with the Directorate, at SAMSI, on November 12, 2004, to review the progress in the current programs and to consider the pre-proposals and proposals that had been submitted for programs in future years. In addition, there are frequent e-mails to the NAC asking for advice concerning developing or new programs. Finally, a member of the NAC serves as a Liaison with each of the Scientific Committees of the major SAMSI programs.
- The Local Development Committee (LDC) consists of leading local scholars, and has a crucial role to play in the involvement of local individuals in SAMSI programs, including the Faculty Release Fellows, the Graduate Associates, and the University Fellows. The LDC met with the Directorate on November 10, 2004.
- The Chairs Committee consists of the chairs of the following departments at the partner universities:
    - Duke: Biostatistics and Bioinformatics, Institute of Statistics and Decision Sciences, Mathematics
    - NCSU: Mathematics, Statistics
    - UNC: Biostatistics, Mathematics, Statistics and Operations Research

  Note that the Chairs are also ex officio members of the LDC. Meetings with the Chairs were held before (and during) the LDC meeting mentioned above.
- The Education and Outreach Committee provides guidance concerning new initiatives in education and outreach and provides contacts – around the nation – for advertising opportunities in education and outreach at SAMSI.

The membership of each of these committees during the past year is given in the table on the following page.

Committee	Name	Affiliation	Field	Term
Governing Board	Bruce Carney	UNC, Assoc. Dean	Astronomy	
	Vijay Nair	NISS Trustees Chair	Statistics	
	John Simon	Duke, Asst. Provost	Chemistry	
	Daniel Solomon	NCSU, Dean	Statistics	
National Advisory Committee	Mary Ellen Bock	Purdue	Statistics	2002-2007
	Peter Bickel (Co-Chair)	UC Berkeley	Statistics	2002-2005
	Lawrence Brown	Pennsylvania	Statistics	2002-2007
	Raymond Carroll	Texas A&M	Statistics	2005-2008
	Carlos Castillo-Chavez (Co-Chair)	Arizona State	Mathematics & Statistics	2003-2006
	John Lehoczky	Carnegie Mellon	Probability	2002-2005
	Sallie Keller-McNulty	LANL	Statistics	2002-2005
	Daryl Pregibon	Google, Inc	CS and Statistics	2003-2006
	G.W. Stewart	Maryland	CS and Mathematics	2003-2006
	Philippe Tondeur	Illinois	Mathematics	2002-2005
	Mary Wheeler	U. Texas	Math and Engineering	2005-2008
Local Development Committee	David Banks	Duke	Statistics	
	Lloyd Edwards	UNC	Biostatistics	
	Gregory Forest	UNC	Mathematics	
	Montserrat Fuentes	NCSU	Statistics	
	John Harer	Duke	Mathematics	
	Sharon Lubkin	NCSU	Mathematics	
	Richard Smith	UNC	Statistics	
	Butch Tsiatis	NCSU	Statistics	
	Mike West	Duke	Bioinfomatics & Stats	
Chairs Committee	Clarence (Ed) Davis	UNC	Biostatistics	
	Vidyadhar Kulkarni	UNC	Statistics	
	Jean-Pierre Fouque	NCSU	Mathematics	
	Richard Hain	Duke	Mathematics	
	Sastry Pantula	NCSU	Statistics	
	William Smith	UNC	Mathematics	
	Dalene Stangl	Duke	Statistics	
	Bill Wilkinson	Duke	Biostatistics	
Education and Outreach Committee	H.T. Banks (Chair, ex officio)	NCSU	Mathematics	
	Negash Begashaw	Benedict College	Mathematical Sciences	
	Carlos Castillo-Chavez (NAC liaison)	Arizona State	Mathematics & Statistics	
	Karen Chiswell	NCSU	Statistics	
	Cammey Cole	Meredith College	Mathematics and CS	
	Wei Feng	UNC-Wilmington	Math and Statistics	
	Marian Hukle	University of Kansas	Biological Sciences	
	Negash Medhin	NCSU	Mathematics	
	Masilamani Sambandham	Morehouse College	Mathematics	

II. Special Report: Program Plan

A. Programs for 2005-2006

1 Program on National Defense and Homeland Security (NDHS)

1.1 Program Foci

Currently, we expect one domain science focus and four more methodologically oriented foci for the program. The domain science focus is:

Biointelligence, which intersects planned development of a CDC Biointelligence Center. Such a center would need to analyze data from a variety of sources, which have differing characteristics in terms of temporal and spatial resolution, seasonal and regional variation, accuracy, completeness and complexity.

Among issues that intersect the theory/methodology foci listed below are benchmarking individual (multivariate) data series to establish mean level and trend, so that anomalous instances or changes can be identified. This requires identification and adjustment for seasonal, regional and other effects. There are also strong relationships to the real time inference and anomaly detection foci.

The theory/methodology foci are:

Real-time inference, also known as data streams. Clearly many of these problems present deep questions of estimation and control, and so naturally require collaboration of statistical and applied mathematical scientists, precisely in line with the SAMSI vision. These are also decision problems, leading naturally engagement of the decision sciences and operations research.

Anomaly detection, with particular attention to multivariate (possibly very high-dimensional) data, extremely rare events and false positives.

Data integration, with attendant problems of privacy, confidentiality and "new forms" of data such as images or biometric identifications.

Dynamics of massive databases, which is in part a fundamental issue of data quality. How long is it before an accumulating database (e.g., one containing facial images) becomes hopelessly contaminated. Is the contamination global or local? What strategies can retard or reverse the process?

There is also already interest in the program beyond the statistical sciences, applied mathematics and decision sciences communities. For example, Scientific Committee member Carey Priebe has already engaged the interest of the Information Security Institute at Johns Hopkins University.

1.2 Program Organization

1.2.1 Schedule

A full-year program is proposed, to run the entire 2005–06 academic year.

1.2.2 Leadership

Program Leaders Committee: This will consist of James Crowley (SIAM), Lawrence Cox (National Center for Health Statistics; co-chair), Alan F. Karr (NISS; directorate Liaison), Sallie Keller–McNulty (Los Alamos National Laboratory; NAC Liaison), Jon Kettenring (Drew University), and Nell Sedransk (National Institute of Standards and Technology; co-chair).

Scientific Committee: In addition to members of the Program Leaders Committee, this committee comprises Alicia Carriquiry (Iowa State University), Philip Hanlon (University of Michigan), Carey Priebe (Johns Hopkins University), Fred Roberts (DIMACS), and Stephen M. Robinson (University of Wisconsin).

1.2.3 Personnel

The following personnel are in place for the program:

Faculty Releases: David Dickey (Statistics, NCSU), Douglas Kelly (Statistics & OR, UNC), Mette Olufsen (Mathematics, NCSU), Bahjat Qaqish (Biostatistics, UNC)

Other Faculty Participants: David Banks (Statistics, Duke), Thomas Banks (Mathematics, NCSU),

Postdoctorals: Sava Dedui (Applied Math, RPI), Lisa Denogean (Statistics, Cornell), Francisco Vera (Statistics, South Carolina; split with NISS). At least two other NISS postdoctorals are expected to participate.

Research Visitors: James Lynch (Statistics, South Carolina), Sidney Resnick (OR, Cornell)

SAMSI–University Fellow: Gauri Datta (Statistics, Georgia)

Other arrangements, including New Researcher Fellows and graduate students, are in progress.

1.2.4 Testbed Databases

Commitments for testbed databases (and personnel to advise and assist SAMSI) are in place from:

CDC: Through the Biosense program; Henry Rolka is the contact.

GM: Debra Elkins is the contact.

In both cases, the exact nature of the data is yet to be determined.

	Sunday, September 11
PM	2-3 Tutorials; possible topics are biosurveillance and data streams
	Monday, September 12
AM	Two technical sessions (2 talks each)
PM	New researchers session
	Technical session
	Poster sales talks
Evening	Poster session and reception at NISS/SAMSI
	Tuesday, September 13
AM	Two Technical Sessions (2 talks each)
PM	Five-Minute Madness
	Breakout Meetings: Precursors of the working groups
	Reports from Breakout Groups
	Initial formation of working groups
	Wednesday, September 14
AM	Meetings of initial working groups

Table 1: Tentative Schedule for Opening Workshop

1.2.5 Workshops

The tentative schedule for the opening workshop, September 11–14, 2005, is in Table 1.

1.2.6 Leveraging

The program will benefit from strong interest among NISS affiliates, a group that includes the National Center for Health Statistics, the National Institute of Standards and Technology and the National Security Agency, all of which have participated in the predecessor efforts.

Ongoing research at NISS on data confidentiality, data integration and data quality will be able to inform specific aspects of the SAMSI research. Moreover, these projects and planned NISS initiatives on National Defense and Homeland Security (NDHS) provide excellent opportunities for second year support for SAMSI NDHS postdoctorals.

Finally, the program will leverage other SAMSI programs in 2005–06. In particular, both anomaly detection and (perhaps to a lesser extent) real time inference, are relevant to astrostatistics.

1.2.7 Course

While NDHS is not (yet, in any case) the subject of a “traditional” course, we plan a seminar course strongly focused on problems, problem owners and (potentially) testbed databases. The course will be offered in the fall semester.

2 Program: Financial Mathematics, Statistics, and Econometrics

Financial Mathematics, Statistics, and Econometrics is a rapidly expanding field comprising interdisciplinary and overlapping areas including Applied Mathematics, Probability and Stochastic Analysis, Economics and Finance, Financial Engineering, Econometrics and Statistics. Since the introduction of the geometric Brownian motion as a tool for modeling stock price evolution, and the discovery in the early seventies of the Black-Scholes formula for pricing options, there has been enormous progress scientifically, that has been matched by the growth of new markets trading new and more complex sources of risk. A tremendous amount of work has sought to understand and explain option prices observed in the markets, and to build tools to handle the associated uncertainties. Needless to say, derivatives markets currently represent an important percentage of trades and investments, especially with the seven-fold increase in the credit derivatives market since 1997. Moreover, in the last decade, four Nobel prizes in Economics were awarded for research in the fields of Financial Mathematics and Financial Econometrics.

The goal of the SAMSI program in Financial Mathematics and Econometrics is be to bring together these disciplines, and focus on the major challenges in the three essential tasks: **modeling, data analysis and computation**, in applications ranging from financial and energy derivatives to real options and defaultable securities.

- **Modeling.** In equity markets there is a profusion of models ranging from local volatility to stochastic volatility, multi-factors with and without jumps, based on Brownian motions or Lévy processes. The situation is similar in fixed income markets with short rates models, HJM or BGM models to name only a few, and intensity-based or structural default risk models for credit linked securities. There are also a variety of discrete time models such as the ARCH family for instance. Fundamental questions about performance, realism, implementation and analysis of these models will be addressed in the program, as well as links between physical measures and pricing measures through market prices of risks. Closely connected are the problems of hedging and portfolio optimization which will also be addressed.

- **Data.** The size of financial data can be considerable when looking at high frequency data for large numbers of stocks for instance. Data is essential in the modeling part in at least two ways: writing models which capture the main effects seen in the data (for instance "are jumps present?") and calibrating the models with an assessment of the stability of the estimated parameters over time. The program will bring together practitioners and researchers from Statistics, Econometrics and Applied Mathematics to present the state of the art and discuss issues of choosing, preparing and using financial data. The program will make sure that statistical software companies are involved.

- **Computation.** Once a model has been written and calibrated to data, it remains to compute quantities of interest. For instance, in option pricing, one has to compute expected values along the trajectories (time evolution) of multidimensional stochastic processes. These quantities are also often obtained as solutions of partial differential equations (or inequalities) with various boundary conditions. The program will address the question of choosing the most efficient computational method for classes of problems. In particular Monte Carlo methods and numerical methods will be discussed, keeping in mind that the computational difficulty has a feedback effect on the modeling and data calibrating parts.

Program Leadership: The Program Leaders Committee will consist of: **Marco Avellaneda** (NYU, Mathematics), **Jean-Pierre Fouque** (NC State, Mathematics), **Eric Ghysels** (UNC, Economics), **Ronnie Sircar** (Princeton, Operations Research & Financial Engineering), and **Ruey Tsay** (University of Chicago, Graduate School of Business Statistics department).

Program Committee: A group of internationally distinguished scholars have accepted to be members of this committee. They are: **Ole E. Barndorff-Nielsen** (Centre for Mathematical Physics and Stochastics, Aarhus, Denmark), **René Carmona** (Princeton University), **Darrell Duffie** (Stanford University), **Nicole ElKaroui** (Ecole Polytechnique, France) **Lars Hansen** (University of Chicago), **Robert Jarrow** (Cornell University), and **Thaleia Zariphopoulou** (University of Texas Austin, Mathematics).

In addition, the SAMSI partners Duke, NC State and UNC are already actively involved in the topics proposed in the program through their departments of Economy, Finance, Mathematics, Statistics and their Business and Management Schools. Key individuals include:

- At Duke: Pete Kyle (Finance), Jonathan Mattingly (Mathematics),
- At NC State: Paul Fackler (Agricultural and Resource Economics), Sastry Pantula (Statistics)
- At UNC: Chuanshu Ji (Statistics), Eric Renault (Economics)

Other individuals, including Tim Bollerslev, Bjorn Eraker, Ron Gallant and George Tauchen at Duke, Amarjit Budhiraja, Richard Smith and Harold Zhang at UNC, and Peter Bloomfield, David Dickey, Alastair Hall, Tao Pang, Denis Pelletier, and John Seater at NC State, will participate in the organization of the various activities in the Program.

Program Timing: The Program will take place during the **Fall semester of 2005**. In particular, the activities will consist of an opening workshop with tutorials, semester long courses in Finance, Mathematics and Statistics, one workshop in the middle of the semester on modeling, data and computation, and a closing conference.

The CRM workshop organized in June 2005 in Montreal by J.P. Fouque and R. Sircar will, in particular, gather together applied mathematicians, statisticians and econometricians to discuss the importance of the time scale content revealed in the long and short ends of market derivative term structures, from the perspectives of the three disciplines, and make sure that we understand each other. This would be very important in designing the tutorials before the opening conference. Every effort will be made to design a program with well defined topics around "Modeling, Data, Computation" with activities and meetings as a progression in these topics, which would makes this program unique.

Details are at:
http://www.crm.umontreal.ca/Finance05/index_e.html

We also plan to organize several working groups during the semester with the local participants, SAMSI visitors and postdocs, and graduate students. These subgroups will be involved in the organization of the workshops.

Program Participants: Besides the participants already mentioned, the program will attract PhD students from the three local institutions working on subjects in the Program scope. We will involve recent programs in Financial Mathematics, such as the ones at GeorgiaTech (Bob Kertz) and at Florida State (Alec Kercheval), by offering to their faculty and students unique opportunities of research collaborations. We also have contacted a number of people who have shown interest to participate as visitors. They include Yacine Ait-Sahalia (Princeton), Elena Andreou (Cyprus/Tilburg), Fabienne Comte (Paris), Rama Cont (Paris), Frank Diebold (U Penn), Ed George (Wharton), Vicki Henderson (Princeton), Eric Hillebrand (Louisiana State), Lojos Horwath (Utah), Piotr Kokoszha (Utah State), Knut Solna (UC Irvine), Jonathan Stroud (Wharton), Bas Werker (Tilburg), among many others.

The two **University Fellows**, Ronnie Sircar (Princeton) and Bas Werker (Tilburg), will give the tutorials at the opening workshop and they will co-teach the SAMSI courses "Advanced Topics in Financial Mathematics" and "Advanced Topics in Financial Econometrics" with J.P Fouque and E. Ghysels respectively. These courses will be of great benefit to newcomers to the field. This will include students in the newly created Professional Science Master's of Financial Mathematics at NC State. We will also organize short courses in order to facilitate the participation of students in comparable programs in other states.

The **Post-Doctoral Fellow**, Jesus Rodriguez (PhD 2005, Cornell), will be supported during 2006-07 with funds from a FRG NSF grant with PI, JP Fouque, at NC State.

The **Faculty Fellows** are: Peter Bloomfield (Statistics, NC State), Eric Ghysels (Economics, UNC), Jonathan Mattingly (mathematics, Duke), and Tao Pang (Mathematics, NC State).

The **Graduate Fellows** are: John Hyde (Duke), Arthur Sinko (UNC), Doug Vestal (NC State).

The **New Researcher Fellow** is Mingxin Xu (Mathematics, UNC-Charlotte).

Among the confirmed **Visitors**: Knut Solna (UC irvine), Elena Andreou (Greece), Rama Cont (France), Aytac Ilhan (Oxford), Gordan Zitkovic (UT Austin), Antje Berndt (CMU), Thaleia Zariphopoulou (UT Austin), René Carmona (Princeton). Other possible visitors are: Wendell Fleming (Brown), Jerome Stein (Brown), Phillip Protter (Cornell), Monique Jeanblanc (France), Tomasz Bielecki (Chicago), Dmitry Kramkov (CMU), Christian Houdré (GaTech), Sasha Stoikov (NYU), Roger Lee (Chicago), David Lando (in sabbatical at Princeton during 2005-06), Kay Giesecke (Stanford), Lin Xu (Morgan Stanley), Lisa Goldberg (Barra, San Francisco), Pierre Collin Dufresne (Berkeley), and Paul Glasserman (Columbia).

Every effort will be made to involve local financial institutions such as Wachovia and Bank of America, software developers (SAS), and energy companies such as Progress Energy and Duke Energy. These companies are potential sponsors for workshops in special areas of interest which will be part of the Program. Other non local companies will also be contacted including companies such Morgan Stanley and Bloomberg in New York, and Mark-It Partners as a source of financial data for the calibration and testing done during the program.

Program Outcomes:
1) The topic is de facto interdisciplinary and throughout the profession, excellence in the field

of financial mathematics, statistics and econometrics is only achieved when top scholars of the various subdisciplines work together to solve interesting research problems. The proposed program is unique with respect to its interdisciplinary content and it is designed to truly impact the field across the country by creating new connections.

2) The SAMSI program should create new collaborations and enforce existing ones across the various mathematics, statistics and economics departments in the US as well as the major business schools.

3) Objective measures should and will be used. These include impact and success of conferences organized; placement of doctoral students; impact of research made possible by the SAMSI program in terms of citations (though this will not be visible on the immediate horizon); quality of the scholars, postdocs and visitors we will be able to attract. We aim at giving the opportunity to a wide range of participants to start new collaborations and to benefit from the top people we will attract in the program.

4) Notes for the courses will be put together and a handbook in Financial Mathematics edited.

3 Program on Astrostatistics and Physics Statistics

Summary A vast range of statistical problems arise in modern astronomical and space sciences research, particularly due to the flood of data produced by space-based astronomical surveys at many wave-bands. A resurgence of interest in statistical and applied mathematical methods has emerged among space scientists as they seek insights into the physical phenomena underlying such complex data. Researchers at the frontiers confront problems for which the common approaches in astronomy either inadequately utilize known methods or require the development of new methods. In contrast with the biological and social sciences, the statistical needs of physical scientists have been neglected during recent decades.

To cope with the current and future needs of astronomy missions requires concerted efforts by cross-disciplinary collaborations involving astronomers, computer scientists, mathematicians and statisticians. SAMSI is an ideal place from which to broadcast these issues and involve the wider statistical and applied mathematical communities.

The semester–long Astrostatistics and Physics Statistics program at SAMSI will take place in Spring 2006. A vital ingredient of the program is to provide a single geographical location – a crossroads – where researchers at the interface between statistics, applied mathematics, astronomy, and particle physics can congregate and initiate lasting collaborations. The participation by graduate students and postdocs gives them a rare opportunity to develop skills needed for cross-disciplinary work.

Introduction: Astronomy at the beginning of the 21st century, and particularly research arising from robotic space-based observatories, finds itself with serious challenges in statistical treatments of data to achieve its astrophysical goals. Innumerable issues arise in the scientific interpretation of astronomical studies. Some issues involve sampling, multivariate and survival analysis, while others involve image and spatial analysis, signal processing or time series analysis. Nonlinear regression

is needed to model the spectra of astronomical objects in terms of continuum and line components deriving from the quantum mechanical properties of matter. Here are a few of the key questions:

> Is a collection of objects chosen for study an unbiased sample of the vast underlying population? When should a collection of objects be divided into two or more classes? What is the intrinsic relationship between two properties of a class, particularly in the presence of confounding variables such as redshift? How can we answer such questions in the presence of flux-limited samples and flux-dependent error bars? When is a blip in a spectrum or image a real signal rather than noise ([6])? How do we characterize blips embedded in larger structures? When is a signal variable rather than constant? How do we characterize the vast range of periodic, correlated and stochastic variations ranging from the Doppler wobble of normal stars due to invisible planets, X-ray manifestations of accretion onto black holes, and gamma-ray bursts from the exotic end-states of stellar evolution ([4])? How do we understand the 3-to-6-dimensional spatial point processes representing the location and motions of stars in the Galaxy or Galaxies in the Universe ([5])? How do we understand the structure of continuous entities like the cosmic microwave background or the interstellar medium?

Clearly, statistical and mathematical problems in astronomy today involve many more problems than can be addressed by any single method, any single field, or any single statistician or mathematician. We estimate that $\simeq 3,000$ distinct refereed studies each year require non-trivial statistical methodologies, 10% of which are principally involved with statistical methods ([3]).

The typical quality of statistical and mathematical methodology used to address such complex questions has not been very high among astronomers. Most astronomers are largely unaware of a host of important statistical and computational developments of the past half century: robust methods, bootstrap resampling, hidden Markov models, empirical Bayes and James-Stein estimation, survival analysis, semi-parametric methods, Bayesian decision theory, and much more. Some methods lie on the interface between mathematical statistics, mathematics, and computational methods: consider, for example, the EM Algorithm, Kalman filter and Monte Carlo Markov chain for likelihood calculations. The systemic need for improved communication between statistics and the physical sciences has been emphasized by Bradley Efron, eminent professor of statistics at Stanford, in his opening essay as incoming President of the American Statistical Association ([2]).

Other statistical and mathematical issues do not appear in research journals but rather arise deep inside the complex machinery of modern astronomical observatories. Many testing, monitoring, compressing, fitting and even intelligent decision-making operations are embedded in the operation, calibration and data reduction process of a contemporary astronomical satellite. With advances in high-speed radiation-hardened chips and high-data rate detectors, sophisticated data analysis operations often take place on-board. Telemetered data are then subject to pipeline processing which provide the basic input to hundreds of astronomical studies. Most of the codes are developed by engineers and scientists who have little formal training in statistics or applied mathematics.

Another level of astrostatistical challenge has recently emerged with the **Virtual Observatory (VO)**. Major efforts are underway around the world, including both NASA and the NSF, to federate huge, uniform, multivariate and image databases collected by specialized observatories ([1, 7]). Most VO efforts have focused on computational aspects of data access and mining from

distributed, heterogeneous databases. But after the scientists have collected the sub-datasets of interest, powerful statistical techniques should be brought to bear to help them make astrophysical inferences. We have begun this effort with the creation of a prototype **VOStat** Web service where scientists can interactively request a statistical analysis using software (located elsewhere) on data (located at other locations) and receive near-realtime answers. VOStat is being developed under a Focused Research Group funded by the NSF Division of Mathematical Sciences led by Babu. This is an example of synergy where collaboration between astronomers, statisticians and computer scientists resulted in an outcome that is substantially more powerful than the sum of individual efforts. This has been successfully demonstrated to the astronomical community at the recent AAS meeting in Denver in June 2004.

Program activities: As astrostatistics is a vast enterprise, a two-day *Planning Workshop* will be held July 14-15, 2005, to identify topics to focus on during the program period. The attendees would include those planning to spend significant time at SAMSI during the Program, and a few other key individuals.

The SAMSI Astrostatistics Program will begin with a set of *Tutorials* on *Bayesian Methods in Astrostatistics,* tentatively scheduled for January 16–22, 2006. These will complement tutorials happening in summer, 2005 at the Center for Astrostatistics at Penn State.

The *Opening Workshop* will be January 23-25, 2006, and will focus on the scientific agenda of the program. It will involve key leaders in astrostatistics, and will be open to the community.

Working Groups will meet throughout the semester, focusing on research areas suggested by the planning workshop and opening workshop. These research areas will almost certainly include some subset of the problems mentioned earlier.

The *Closing Workshop*, held during June 11-14, 2006, will be joint with the Center for Astrostatistics at Penn State, and indeed will coincide with the fourth conference on Statistical Challenges in Modern Astronomy. This will provide a perfect venue for disseminating the results from the SAMSI program. The workshop will also receive funding from the Astrostatistics Center and (hopefully) from NASA.

Program Leadership: The leadership committee consists of chair G. J. Babu (statistics, Penn State), Eric Feigelson (astronomy, Penn State), Donald Richards (statistics, Penn State), Alanna Connors (astronomy, Eureka Scientific), and Larry Wasserman (statistics, Carnegie-Mellon University).

A Scientific Committee will also be created from other scientists that will be consulted about the SAMSI program, including Francoise Genova (CDS, Director, Centre des Données Astronomiques de Strasbourg), George Djorgovski (Caltech, Prof. of Astronomy, former Sloan Fellow and NSF PYI, Chair, NVO Science Definition Team, cosmologist), Ian Johnstone (Stanford, 2002 IMS President, Associate Dean for Science, Fellow of American Academy of Arts and Science), Fionn Murtagh (Queen's, Professor of Computer Science, Fellow of British Computer Society, Editor-in-Chief, The Computer Journal), John Rice (Berkeley, Former Chair, Dept. of Statistics, UC Berkeley), David van Dyk (UC Irvine, Spectral analysis, image analysis), Lee Samuel Finn (Penn State, Gravitational Wave Physics), Ajit Kembhavi (IUUCA, Quasars, Elliptical Galaxies, Simulations), Thomas Loredo

(Cornell, Bayesian methods, extra-solar planet detection), Louis Lyons (Oxford, Particle Physics), Jean-Luc Starck (Saclay, Image analysis, wavelets), Vicent Martínez (Valencia, Statistics of Galaxy distribution), and Philip Stark (Berkeley, Helioseismology, optimization, Cosmology).

Long-term Visitors: Jogesh Babu will spend the spring Semester of 2006 at SAMSI, as a SAMSI University Fellow. David van Dyke will spend 3 months at SAMSI as the New Researcher Fellow for the program. Louis Lyons (physicist), Bill Jefferys (astronomer), Tom Loredo (astronomer), and Volker Dose (physicist) are likely long-term visitors. Discussions are under way with a number of other potential long-term visitors.

Faculty Fellows in the program are Merlise Clyde (Statistics, Duke), Arlie Petters (Mathematics, Duke) and Zhengyuan Zhu (Statistics/OR, UNC).

Graduate Fellows in the program are Nicholas Robbins (Mathematics, Duke) Floyd Bullard (Statistics, Duke), and Shenek Heyward (Statistics, NCSU).

Postdoctoral Fellows: Offers are currently out to two potential fellows.

Organizational Involvement: The Center for Astrostatistics at Penn State will be a partner in a number of the activities of the SAMSI program. In particular, funding for preparation of the the tutorial and educational materials discussed above has been secured from the NSF astronomy program, and this will thus provide an ideal feed-in to the SAMSI program.

Program Goals and Outcomes: Previous SCMA workshops have brought astronomers together with statisticians and applied mathematicians for brief periods, and instilled a sense that significant long-term collaboration of numerous individuals on each side is required. The SAMSI program is an ideal vehicle for providing the sustained interaction needed for such collaborations to 'take,' leading to greatly improved statistical and applied mathematical methodologies in astronomy, and providing a rich source of problems for development of new methodologies by statisticians and applied mathematicians. In addition, the educational and outreach goals of the program are significant, as discussed earlier.

References

[1] Brunner, R.J., Djorgovski, S.G. & Szalay, A.S. (eds.) (2001). *Virtual Observatories of the Future*. ASP.

[2] Efron, B. (2004) Statistics as a Unified Discipline. *Amstat News*, Issue # 319, 2–3.

[3] Feigelson, E. D. & Babu, G. J. (2004). Statistical Challenges in Modern Astronomy. In *PhyStat 2003: Statistical Problems in Particle Physics, Astrophysics, and Cosmology*, SLAC, in press.

[4] Maoz, D., Sternberg, A. & Leibowitz, E.M. (eds.). (1997). *Astronomical Time Series*. Kluwer.

[5] Martinez, V.J. & Saar, E. (2002). *Statistics of the Galaxy Distribution*. Chapman & Hall

[6] Starck, J.L. & Murtagh, F. (2002) *Astronomical Image and Data Analysis*. New York, Springer.

[7] Szalay, A. & Gray, J. (2001). The World-wide telescope. *Science*, **293**, 2037-1040.

B. Scientific Themes for Later Years

The programs listed below have not yet been formally approved, but all are well along in the development cycle and we are confident that they will be approved and implemented.

1 Development, Assessment and Utilization of Complex Computer Models

Background: This area had been proposed by the NAC as a natural program for SAMSI, since it inherently combines applied mathematics and statistics. Two recent workshops solidified this idea. A January 2004 meeting on *Opportunities at the Statistics-Operations Research Interface* highlighted the area as one needing a major interdisciplinary effort. A second relevant workshop was *The Design and Analysis of Computer Experiments for Complex Systems* held at BIRS in July, 2004. This was a joint workshop between SAMSI and the Canadian National Program on Complex Data Structures (NPCDS) and was, in part, a planning meeting to gauge interest in the area. Strong enthusiasm was also expressed there for a SAMSI program in the area, with significant co-participation by NPCDS.

1.1 Introduction

Mathematical models intended for computational simulation of complex real-world processes are a crucial ingredient in virtually every field of science, engineering, medicine, and business, and in everyday life as well. Cellular telephones attempt to meet a caller's needs by optimizing a network model that adapts to local data, and people threatened by hurricanes decide whether to stay or flee depending on the predictions of a continuously updated computational model.

Two related but independent phenomena have led to the near-ubiquity of models: the remarkable growth in computing power via Moore's law for raw speed and its analogue for data storage; and the matching gains in algorithmic speed and accuracy. Together, these factors have vastly increased the applicability and reliability of simulation—not only by drastically reducing simulation time, thus permitting solution of larger and larger problems, but also by allowing simulation of previously intractable problems.

The intellectual content of computational modeling comes from a variety of disciplines, including statistics and probability, applied mathematics, operations research, and computer science. However, most likely for historical reasons, models in certain application domains tend to be associated with a relatively small set of specific modeling techniques. For example, models of flow in physics rely on partial differential equations; statistics is the source for models in medicine (such as of life expectancy and drug side effects) that lack equations but for which data are plentiful; and transportation systems are represented with classical models from operations research. As a consequence of the narrow focus dictated by custom, modelers in some application areas may not be aware of the latest developments (or even of well-known features) in models from other domains.

Despite enormous recent progress in formulating and solving computational models, the issue of modeling in its broadest sense remains a grand challenge because there is so much more to be done. A new, unified paradigm is needed that integrates adaptive mathematical formulation, maximally effective use of observed data, and flexible solution methodologies that can cope with highly nonlinear, very large, specially structured problems.

1.1.1 Integrating experimental data

It is increasingly the case that models must make sense of massive amounts of physical data, often of high dimension. Furthermore, depending on the problem, it may be easy, difficult, or impossible to obtain high-quality experimental data that are well matched to the model being developed.

- In instances such as designing airplanes, drugs, and traffic systems, there is a coordinated interplay between an evolving sequence of models and collection of observed data. When further data are needed, they can be obtained more or less to order and on point.
- In other situations, such as modeling the consequences of individual cancer chemotherapy or the effects of "baby boomer" retirements on the budget of the United States, it might be possible in theory to obtain some experimental data, but in practice this would be impractical, inadequate, or too late to be helpful. Under these circumstances, although new data can be gathered, they will necessarily be inadequate as well as of possibly dubious relevance to the questions of interest.
- In still other situations, such as simulating nuclear weapons without testing them, the only experimental data are incomplete, gathered from systems that no longer exist, and subject to deep uncertainty about their quality.

When comparing the results of a model with observed data, it is crucial to know whether there is a guaranteed band of reliability within which changes in the data do not affect the quality of the solution. It is also important to quantify the potential effects on the model output of perturbations in the data, to provide a concrete bound on how much the solution may change. A further issue is related to experimental data of uneven quality; as noted above with respect to simulating nuclear weapons, the model should not be allowed to degrade its overall fidelity to match suspect data.

Physical data can be used in additional ways in developing models, in particular to identify and ameliorate bias. When, as is often the case, a computational model cannot capture all relevant features of a process that changes rapidly with time (for example, the behavior of the Internet), techniques like "parameterization" are needed to incorporate data from the time-varying elements automatically.

1.1.2 Coping with the varieties of uncertainty

Essentially all models of real-world processes contain uncertainties. A common form of uncertainty arises from unknown variations in the inputs or internal processes of the model. The effects on

the solution of small local perturbations in the problem data are understood for a relatively small class of optimization problems—e.g., for linear programming—but analysis is needed of the effects of larger perturbations, structured perturbations, and systematic (non-random) perturbation.

A danger of worst-case perturbation analysis is that it may lead to highly unrealistic conclusions, which may be either overly optimistic because large constants compromise bounds of apparently favorable order, or else overly pessimistic because the scenarios for which the worst case is achieved will never occur in practice. Hence methods are required that can characterize the typical effects of perturbation. The recent development of "smoothed analysis" to explain the complexity of various methods from operations research—notably, the simplex method for linear programming—offers hope for a realistic general approach.

Even at their best, some general techniques of accounting for uncertainty are impractical for non-linear models with huge input spaces or discrete variables. A less obvious issue is that some models contain numbers that must be treated as exact and unvarying—for example, the coefficients in a geometric formula—even when other quantities are subject to uncertainty.

A further source of uncertainty is associated with model parameters that must be fitted or "tuned" to available data. These parameters are usually determined by optimizing a measure of the quality of the model, and for the most part they do not represent anything physical or real, about which the modeler might have some intuition based on experience. Depending on how the model parameters are chosen, their mathematical properties are not necessarily stable with respect to perturbations in the problem data, nor is there any guarantee that they make sense in the overall framework of the model.

1.1.3 Validation

The fidelity of models to reality (often termed *model validation*) is central to their effectiveness in understanding and predicting real phenomena. However, techniques for validating complex models are limited and are often based on informal heuristics. Validation seems straightforward conceptually—data are collected that represent both the inputs and the outputs of the model, the model is run with those inputs, and the outputs are compared with the observed data corresponding to the outputs. In reality, complications abound: observed data may be expensive, scarce or noisy, the model may be so complex or time-consuming that only a few runs are possible, and uncertainty enters the process at every turn.

1.1.4 Model improvement

Statistics and optimization can help to improve a flawed model—"flawed" in the sense that it does not match reality well—by identifying and rank-ordering potential causes for the flaws and suggesting corrections. For example, if a model contains several components, statistical and optimization methods may be able to determine that a specific set of components (or perhaps one of those components) is most likely to be the major source of discrepancy from the real process. It may also be possible to determine that more attention is needed to an input variable, or that the model is

inadequate when an input variable lies in a certain region.

A major challenge is finding tractable diagnostic methods for the typically limited amount of available data. Further issues are related to the cost of model improvement—if a model is running in real time, it may be impractical to explore even a small part of the full space of possible errors.

1.1.5 Screening experiments

The purpose of many scientific investigations—and a major use of many computer models—is to identify which factors are most influential on the system response. Depending on the context, the initial set of candidate factors may range in size from a few to several hundred or more. When the computational model is expensive or time-consuming to run, correct determination of the most influential factors requires a careful design of experiments, a process that blends statistics and operations research in defining an "optimal" experiment design, creating algorithms to compute an optimal or near-optimal design, and analyzing the responses.

1.1.6 Model approximation; fitting in high dimensions

For complex simulations, statistical analysis may require thousands of model evaluations (for example, using Markov chain Monte Carlo, or "MCMC"). Thus the computational problem to be solved may be the composition of two highly nonlinear and potentially huge computational subproblems—i.e., an intractable problem. But a solution that is often almost as good, or at least good enough for many purposes, can sometimes be obtained through model approximations or "meta-models".

The ideas of "model reduction", "dimension reduction", "order reduction", "approximation algorithms", and "surrogate models" are currently active research topics in several areas of statistics, applied mathematics, computer science, and operations research. One idea is to develop an exact but much less complex (or much smaller) model that retains the most important features of the original; another approach is to compute an approximate solution of the original problem that retains a provable degree of closeness to the exact solution; and obviously a combination of these ideas is possible. In all cases, a mixture of rigor and application-specific knowledge is needed to deal with increasingly high-dimensional spaces and data.

1.1.7 Varying data types

Models are associated with input data, parameters, and output data, and it is no longer true in today's models that all of these are well-behaved real numbers. Tomorrow's models will certainly contain data with features that stretch (or exceed) today's capabilities, including:

- numbers that differ drastically in scale, e.g., by twelve orders of magnitude within a single model;

- data specified by algebraic functions, logical variables, or logical constraints;
- data chosen from a discrete set;
- data that are individual-specific, as is typical for current models of disease progression;
- spatial or image data.

Dealing with combinations of these data types will require new methodologies in statistics and operations research. To take the most basic example, mixed-integer optimization problems, which involve integer and continuous variables, are extremely common in practice, yet reliable techniques are available for solving them only when the problem is a linear program.

1.1.8 Getting inside the 'black box'

Often the computer model is viewed as a 'black box' from the statistical viewpoint, and a key question is how statistical analysis can be brought 'inside the box.' For instance, assessment of the uncertainty of outputs due to uncertainty in inputs can be approached either by simply simulating inputs and studying the variation in the corresponding outputs (the black box approach) or by formally processing the input distributions through the steps of the code for the computer model (perhaps using approximations such as local linearizations), arriving at a distribution for the output. SAMSI is ideally suited for exploring this issue.

1.2 Possible interdisciplinary subprograms

Study of computer models needs to take place in the context of actual computer models. But because of the inherent complexity of computer models, and the very different types of such models, it is proposed to have a SAMSI program with sub-programs, focusing on specific computer modeling scenarios. These sub-programs could take place in the same year, or span several years. This approach allows the depth of exploration of specific types of computer models that is needed to make real advances, while maintaining an overall 'SAMSI umbrella' that allows quick transfer of techniques developed in one sub-program to another. Some of the possible sub-programs that have been proposed are

- Engineering models
- Biomedical models
- Social network models
- Microsimulation models
- Combinatorial algorithms

1.3 Participants

Potential Program Leaders: The following have expressed possible interest in – or have been proposed for – leading the program or sub-programs: M.J. Bayarri (U. Valencia), Derek Bingham (Simon Fraser U.), Hugh Chipman (Acadia U.), Dennis Cox (Rice), John Dennis (Rice), Nick Hengartner (LANL), Max Morris (Iowa State), William Notz (Ohio State U.), Tony O'Hagan (Sheffield) Tom Santer (Ohio State), Randy Sitter (Simon Fraser U.), Jerry Sacks (Duke emeritus), David Steinberg (Tel Aviv U.), William Welch (U. British Columbia), Margaret Wright (NYU), and Jeff Wu (Georgia Tech).

Potential Participants: Others who expressed considerable interest in the program include Andrew Booker (Boeing), Michael Hamada (Los Alamos Nat. Labs), Dave Higdon (Los Alamos Nat. Labs), Leslie Moore (Los Alamos Nat. Labs), Shane Reese (Bringham Young U.), Henry Wynn (London School of Economics), Kenny Ye (SUNY Stony Brook).

Partner university involvement: There is serious interest in this area in virtually every one of the SAMSI affiliated partner university departments, and in affiliated centers such as CRSC.

Affiliates involvement:

- General Motors supported a significant NISS effort in this direction, and could be a good source of models and data sets.
- LANL would be heavily involved in the program. Indeed some activities (e.g. a proposed 2006 'summer school' in the area), would take place at LANL.
- Many university affiliates are heavy players in the area.

National lab involvement: In addition to LANL, mentioned above, Sandia has a large and active group in the area, which should be involved in the program.

International involvement: Eurandom, the European equivalent of SAMSI, has expressed interest in the area. Henry Wynn, an associate director of Eurandom, has a considerable background in the area.

1.4 Activities and timing

- In summer 2006, there would be a short course for students, postdocs and others.
- The opening workshop would be September, 2006.
- The three planned subprograms could be either a semester each or year-long, depending on interests and personnel.

2. Program on High Dimensional Inference and Random Matrices (Fall, 2006)

Massive data collection efforts across contemporary science, technology and commerce are yielding ever more high dimensional datasets. Often the dimensionality (number of variables p) is comparable to, or even larger than the number of cases (sample size n). Numerous methodologies, both old and newly proposed, are being adapted in specific applications to the tasks of dimension reduction and inference. There is a concomitant need to develop fundamental tools and theory to provide reliable understanding of the phenomena and properties thrown up by large-p data analysis. At the same time, recent advances in applied mathematics, and particularly within random matrix theory, offer new mathematical tools and approaches which can be brought to bear on the spectral aspects of high dimensional data reduction and inference.

We propose, therefore, to bring together a group of applied mathematicians (broadly interpreted) active in random matrix theory with theoretical statisticians (and probabilists) concerned with high dimensional inference particularly via eigen-structure methods. We hope also to engage methodologically oriented researchers from at least one application domain, climatology, in which large p data analysis has long played a major role (PCA, under the rubric of "empirical orthogonal functions").

Background: Initial planning began some time ago when Jim Berger approached Iain Johnstone about a possible program. Johnstone sought out Craig Tracy as a co-organizer, and they have worked together on planning, along with frequent and valuable advice from Peter Bickel.

A meeting of potential main organizers was held at the American Institute of Mathematics, Palo Alto, February 29 and March 1, 2004. (Thanks to Helen Moore and Brian Conrey of AIM for hosting!). Participants were: Iain Johnstone, Craig A. Tracy, Kenneth McLaughlin, Peter Bickel, Douglas Nychka, Hien T. Tran, James O. Berger, J. S. Marron. From the workshop emerged the core list of "likely research topics" and "potential participants" detailed below, with some subsequent additions.

Organizing Personnel

*Program Leadership***:** Iain M. Johnstone (Statistics, Stanford), Craig A. Tracy (Mathematics, U. C. Davis), Kenneth McLaughlin (Mathematics, U. Arizona), Peter Bickel (Statistics, UC Berkeley), Douglas Nychka (NCAR, Boulder).

Scientific Committee: Persi Diaconis (Stat & Math, Stanford), David Donoho (Stat, Stanford) , Neil O'Connell (Maths, Warwick). Other possibilities include Alan Edelman (Math, MIT), Estelle Basor (Math, Cal Poly SLO), Alice Guionnet, Sara van de Geer, Hien Tran, (Mathematics, North Carolina State).

Program Timing and Structure: Fall 2006 is proposed. Detailed structure of meetings, workshops and working groups remains to be developed. Certainly there would be an opening and closing meeting at SAMSI, and the possibility of a workshop linked with climatology and NCAR has been discussed with Doug Nychka. An orientational course

on random matrices and statistics at SAMSI would be a useful part of the program. This could be given e.g. by Johnstone and Tracy or possibly by Ofer Zeitouni.

Likely Research Areas: Major research areas that were delineated at the AIM planning workshop are listed, together with some important subtopics.

Extreme Sample Eigenvalues (Large *n, p* Setting) (Edge)
- PCA, CCA: distributions under alternative hypotheses...
- Integrable systems - long term project: beta = 1
- Connection to financial math program

Sample Eigenvectors
- Consistency and distribution
- Smoothing / filtering of lead estimated eigenvectors

Empirical Distribution Of Eigenvalues (Global)
- Marcenko-Pastur applications, contiguity, CLT for linear statistics

Approximations Of Empirical Data
- Compact energy decay - finite range approximations
- Invariant approximations
- Random Schrödinger Operators (ie random tri diagonal matrices)

Design Of Snapshots (PCA applied to numerical solution of PDEs)
- Design measure
- Deterministic errors (treated as stochastic?)
- PDE induced smoothing of Principal Orthogonal Decomposition (POD) empirical functions
- Pre-specified basis functions

Nonlinear / Topological Approaches To Dimensional Reduction
- Data sets; Dynamical systems
- Uncertainty assessment
- [NB related MSRI Workshop Dec 9-13 2004]

Bayesian Version Of Regularization In Large p Problems / Invariance Structure
- Coherence with frequentist uncertainty assessment
- Prior distributions on covariance structures: dispensing with Vandermondes
- Banded random matrices

Stochastic Evolution Of Random Matrices
- Dyson Brownian Motion model
- Wishart Processes

Statistical Issues In Empirical Orthogonal Functions (EOFs) In Climatology
- [Possible kick-off workshop at NCAR, some NCAR funding]

Other topics of potential interest include

Estimation of large covariance matrices: Bayesian approach - priors, sparsity models, frequentist properties,

Machine learning perspective: Connects to Bayesian regularization topic above. Kernel PCA, ICA; classification problems, models for large covariance structures. Possible names: John Shawe-Taylor (Southhampton), Michael Jordan (CS/Stat,

UC Berkeley), Nello Cristianini (Stat, UC Davis), Connections with FDA (Toulouse group: Dauxois, Pousse, Besse, Romain)

"Traditional'" multivariate distribution theory: Wishart distributions: Steen Andersson (Indiana), Mike Perlman (Seattle), Gerard Letac (Toulouse), Hèlène Massam.

Symbolic evaluation programs (MOPS, Ioana Dumitriu, UC Berkeley)

Potential Program Participants:

"Applied Math" side (broadly construed!): Harold Widom, UC Santa Cruz; Jean Bernard Zuber, CEA Saclay; Alan Edelman, MIT; Ioana Dumitriu, UC Berkeley; Peter Forrester, U. Melbourne; Sasha Soshnikov, UC Davis; Eric Rains, UC Davis; Peter Miller, U. Michigan; Edriss Titi, UC Irvine; Ioannis Kevrekidis, Princeton; Mark Adler, Brandeis; P. van Moerbeke, Boston U. and Louvain; Alexander Its, IUPUI Indianapolis; Percy Deift, Courant Inst.; Stephanos Venakides, Duke; Thomas Guhr, Lund; Estelle Basor, Cal Poly SLO; Kelly Wieand; Tom Spencer; Greg Anderson, U. Minnesota; Jinho Baik, U. Michigan; Momar Dieng, UC Davis

"Stat/Prob" side: Ofer Zeitouni, U. Minnesota; Persi Diaconis, Stanford; Dave Donoho, Stanford; Amir Dembo, Stanford; Steve Evans, UC Berkeley; Zidong Bai, Singapore; Neil O'Connell, Warwick; Steve Marron, UNC; Jack Silverstein, NC State; Kurt Johansson, RIT Stockholm; Bas Klein, UC Berkeley; Alice Guionnet, ENS Lyon; Jayanta Ghosh, ISI Calcutta/Purdue; Jon Wellner, U. Washington; Alexander Tsybakov, U. Paris VI; Vladimir Koltchinskii, U. New Mexico; Sara van de Geer, Leiden; Aad van de Vaart, Amsterdam; Evarist Giné, U. Connecticut; Jonathan Taylor, Stanford U.; Don Richards, Penn State; Jianqing Fan, Princeton; Piet Groeneboom, Delft; Friedrich Götze, Bielefeld; Brenda MacGibbon, UQAM Montreal; John Rice, UC Berkeley (FDA connections); Greg Rempala, U. Louisville, KY (genomics tie in); Hua Xu (student of Christian Houdré, Math, Georgia Tech); Steve Portnoy, U. Illinois; Olivier Lévêque (Communications, EPFL, Lausanne); Bin Yu (possible) UCB; Liza Levina, U. Michigan; Debashis Paul, Stanford U.; Sandrine Péché, Grenoble; Noureddine El Karoui, Stanford.

EOFs in Climatology: Ben Santer; Myles Allen; Gabi Herbeul

Nonlinear / Topological Approaches To Dimensional Reduction: Josh Tenenbaum, MIT; Lawrence Saul, U. Penn; Mikhail Belkin, C.S., Chicago; Vin de Silva, Stanford; Raphy Coifman,Yale; Partha Niyogi (U.Chicago); Steve Smale (UCB).

Machine Learning: John Lafferty (CMU); Martin Wainwright (UCB EE/Stat.); Peter Bartlett (possible UCB (CS/Stat.); John Shawe-Taylor (Southhampton); Michael Jordan (CS/Stat, UC Berkeley); Nello Cristianini (Stat, UC Davis).

APPENDIX E – Workshop Participants

TWO tables of participants for most of the SAMSI workshops follow. The first table lists only the individuals who received support. The second table lists all workshop participants. The minority status of each participant is available, but we do not include the information here because of privacy issues; the summaries in Section I.H. were compiled from this data.

The key to **Status** entry is as follows:

NRG – New Researcher or Graduate Student
FP – Faculty/Professional
S – Student (Education & Outreach)
A – Faculty (Education & Outreach)

Data Mining and Machine Learning Program
Closing Workshop
NISS-SAMSI Building
Supported Workshop Participants
May 17-18, 2004

Name	Gender	Affiliation	Department	Status
Chipman, Hugh	M	U of Waterloo	Statistics & Actuarial Science	FP
DuMouchel, William	M	AT&T Labs	Research & Development	FP
Hawkins, Douglas	M	U of Minnesota		FP
Ju, Wen-Hua	M	Avaya Labs	Research & Development	FP
Khattree, Ravindra	M	Oakland U	Mathematics & Statistics	FP
Lahiri, Soumendra	M	Iowa State U	Statistics	FP
Madigan, David	M	Rutgers U	Statistics	FP
Yang, Lijian	M	Michigan State U	Statistics & Probability	FP
Zhang, Ping	M	Avaya Labs	Research & Development	FP
Zhu, Ji	M	U of Michigan	Statistics	FP

Data Mining and Machine Learning Program
Closing Workshop
NISS-SAMSI Building
Workshop Participants
May 17-18, 2004

Name	Gender	Affiliation	Department	Status
Ahn, Jeongyoun	F	U of North Carolina	Statistics & Operation Research	NRG
Banks, David	M	Duke U	Statistics	FP
Berger, Jim	M	SAMSI		FP
Brooks, Atina	F	North Carolina State U	Statistics	NRG
Chipman, Hugh	M	U of Waterloo	Statistics & Actuarial Science	FP
Chu, Jen-hwa	M	Duke U	Statistics	NRG
Clarke, Bertrand	M	Duke U	Statistics	FP
Clyde, Merlise	F	Duke U	Statistics	FP
DuMouchel, William	M	AT&T Labs	Research & Development	FP
Fokoue, Ernest	M	SAMSI		NRG
Genton, Marc	M	North Carolina State U	Statistics	FP
Goel, Prem	M	Ohio State U	Statistics	FP
Haran, Murali	M	NISS		NRG
Hawkins, Douglas	M	U of Minnesota		FP
House, Leanna	F	Duke U	Statistics	NRG
Hughes-Oliver, Jacqueline	F	North Carolina State U	Statistics	FP
Hurtado, Gerardo	M	SAS Institute	Research & Development	FP

Ju, Wen-Hua	M	Avaya Labs	Research & Development	FP
Karr, Alan	M	NISS		FP
Khattree, Ravindra	M	Oakland U	Mathematics & Statistics	FP
Lahiri, Soumendra	M	Iowa State U	Statistics	FP
Liang, Feng	F	Duke U	Statistics	FP
Lin, Xiaodong	M	SAMSI & NISS		NRG
Liu, Fei	F	Duke U	Statistics	NRG
Liu, Peng	M	North Carolina State U	Statistics	NRG
Madigan, David	M	Rutgers U	Statistics	FP
Marron, Steve	M	SAMSI & U of North Carolina	Statistics	FP
Ouyang, Haojun	M	North Carolina State U		NRG
Palomo, Jesus	M	SAMSI		NRG
Park, Cheolwoo	M	SAMSI		NRG
Rempala, Greg	M	U of Louisville	Mathematics	FP
Sanil, Ashish	M	NISS		FP
Sun, Dongchu	M	U of Missouri-Columbia	Statistics	FP
Truong, Young	M	U of North Carolina	Biostatistics	FP
van Rhee, Michiel	M	ICAGEN	Discovery Chemistry	NRG
Yang, Lijian	M	Michigan State U	Statistics & Probability	FP
Young, Stan	M	NISS		FP
Zhang, Helen	F	North Carolina State U	Statistics	FP

Zhang, Ping	M	Avaya Labs	Research & Development	FP
Zhu, Ji	M	U of Michigan	Statistics	FP

MAA PREP Workshop: Mathematics Meets Biology
(co-sponsored by SAMSI)
University of Louisiana
Workshop Participants
May 26-29, 2004

Name	Gender	Affiliation	Department	Status
Ackleh, Azmy	M	U of Louisiana	Mathematics	FP
Alcock, Darren	M	McNeese State U	Mathematics, Comp Sci and Statistics	FP
Allen, Linda	F	Texas Tech U		FP
Atkinson, Graham	M	Atkinson Consulting	Health Economics	FP
Ayati, Bruce	M	Southern Methodist U	Mathematics	FP
Ball, Karen	F	Indiana U	Mathematics	FP
Banks, H.T.	M	North Carolina State U	CRSC	FP
Carter, Cathy	F	Christian Brothers U	Mathematics	FP
Comar, Timothy	M	Benedictine U	Mathematics	FP
Cunningham, Ellen	F	Saint Mary-of-the-Woods College	Mathematics	FP
Cushing, Jim	M	U of Arizona		FP
Friedman, Jane	F	U of San Diego	Mathematics	FP
Herod, Jim	M	Georgia Institute of Technology	Mathematics	FP
Hukle, Marian	F	U of Kansas	Biological Sciences	FP

Jang, Sophia	F	U of Louisiana		FP
Kersten, Thomas	M	Normandale Community College	Mathematics	FP
Landers, Mary	F	Weatherford College	Mathematics	FP
Lewand, Robert	M	Goucher College	Mathematics and Computer Sci	FP
Long, Andrew	M	Northern Kentucky U	Mathematics and Computer Sci	FP
Mackin, Gail	F	Northern Kentucky U	Mathematics and Computer Sci	FP
Oppenheimer, Seth	M	Mississippi State U	Mathematics and Statistics	FP
Pearson, Michael	M	MAA		FP
Porter, Jack	M	U of Kansas	Mathematics	FP
Reitter, Nicholas	M	Cooper Union for Adv of Science & Art		FP
Slay, David	M	McNeese State U	Mathematics, Comp Sci and Statistics	FP
Tolosa, Juan	M	Stockon College	Mathematics	FP
Wortman, Dennis	M	U of Massachusetts, Boston	Mathematics	FP
Yurekli, Osman	M	Ithaca College	Mathematics and Computer Sci	FP
Zachariah, Thomas	M	Loyola Maramount U, Los Angeles	Mathematics	FP

Education and Outreach Program
CRSC-SAMSI Interdisciplinary Workshop for Undergraduates
North Carolina State University & NISS-SAMSI Building
Supported Workshop Participants
May 31-June 4, 2004

Name	**Gender**	**Affiliation**	**Department**	**Status**
Davis, Darcy	F	Youngstown State U		S

Faulkner, James	M	U of North Carolina-Greensboro		S
Garrison, Derek	M	Taylor U		S
Gonzales, Chad	M	Colorado State U		S
Gorritz, Magdaliz	F	U of Texas-San Antonio		S
Grove, Rebecca	F	Youngstown State U		S
Hanson, Michael	M	U of North Carolina-Wilmington		S
Lambdin, Jennifer	F	Salisbury U		S
Mayumi, Sakamoto	F	U of North Carolina-Asheville		S
Salcedo, Maria	F	Youngstown State U		S
Schwartz, Scott	M	Trinity U		S
Scott, Michelle	F	Meredith College		S
Stapleton, James	F	Clemson U		S
Wright, Jabari	M	Benedict College		S

Education and Outreach Program
CRSC-SAMSI Interdisciplinary Workshop for Undergraduates
North Carolina State University & NISS-SAMSI Building
Workshop Participants
May 31-June 4, 2004

Name	**Gender**	**Affiliation**	**Department**	**Status**
Davis, Darcy	F	Youngstown State U	Applied Math & Computer Science	S
Faulkner, James	M	U of North Carolina-Greensboro	Computer Science	S
Gant, Raymond	M	Benedict College		S

Garrison, Derek	M	Taylor U	Math	S
Glasgow, Shane	M	Benedict College		S
Gonzales, Chad	M	Colorado State U	Applied Math	S
Gorritz, Magdaliz	F	U of Texas-San Antonio	Statistics & Math	S
Grove, Rebecca	F	Youngstown State U	Math & Engineering	S
Hanson, Michael	M	U of North Carolina-Wilmington	Statistics	S
Lambdin, Jennifer	F	Salisbury U	Statistics	S
Mayumi, Sakamoto	F	U of North Carolina-Asheville	Statistics	S
Salcedo, Maria	F	Youngstown State U		S
Schwartz, Scott	M	Trinity U	Computer Science & Math	S
Scott, Michelle	F	Meredith College	Computer Science	S
Stapleton, James	F	Clemson U	Applied Math	S
Thomas, Keith	M	U of North Carolina-Greensboro		S
Wright, Jabari	M	Benedict College	Physics	S
Zarif, Naim	M	Benedict College	Physics & Applied Math	S

Network Modeling for the Internet Program
Closing Workshop
NISS-SAMSI Building
Supported Workshop Participants
June 25-26, 2004

Name	**Gender**	**Affiliation**	**Department**	**Status**
Gonzalez, Barbara	F	U of Louisiana	Mathematics	NRG

Hannig, Ian	M	Colorado State U	Statistics	NRG
Hohn, Nicolas	M	Cen for Ultra-Broadband Information Networks	Electrical & Electronic Engineering	NRG
Lawrence, Earl	M	U of Michigan	Statistics	NRG
Maulik, Krishanu	M	EURANDOM		FP
Rolls, David	M	U of North Carolina-Wilmington	Mathematics & Statistics	NRG
de Oliveira, Rosario	F	Vienna Technical U	Mathematics	FP
Todorov, Diman	M	Vienna U of Technology	Research Industrial Software Engineering	NRG
Willinger, Walter	M	AT&T Labs	Research	FP

Network Modeling for the Internet Program
Closing Workshop
NISS-SAMSI Building
Workshop Participants
June 25-26, 2004

Name	**Gender**	**Affiliation**	**Department**	**Status**
Berger, Jim	M	SAMSI		FP
Buche, Robert	M	North Carolina State U	Mathematics	FP
Dinwoodie, Ian	M	Duke U	Institute of Statistics & Decision Sciences	FP
Ghosh, Arka	M	U of North Carolina & SAMSI	Statistics & Operations Research	NRG
Gonzalez, Barbara	F	U of Louisiana	Mathematics	NRG
Hannig, Ian	M	Colorado State U	Statistics	NRG
Hernandez-Campos, Felix	M	U of North Carolina	Computer Science	NRG
Hohn, Nicolas	M	Cen for Ultra-Broadband Information Networks	Electrical & Electronic Engineering	NRG

Karr, Alan	M	NISS		FP
Kaur, Jasleen	F	U of North Carolina	Computer Science	NRG
Lawrence, Earl	M	U of Michigan	Statistics	NRG
Marron, Steve	M	U of North Carolina & SAMSI		FP
Maulik, Krishanu	M	EURANDOM		FP
Papadopouli, Maria	F	U of North Carolina	Computer Science	FP
Park, Cheolwoo	M	SAMSI		NRG
Pipiras, Vladas	M	U of North Carolina	Statistics & Operations Research	FP
Rewaskar, Sushant	M	U of North Carolina	Computer Science	NRG
Rolls, David	M	U of North Carolina-Wilmington	Mathematics & Statistics	NRG
de Oliveira, Rosario	F	Vienna Technical U	Mathematics	FP
Smith, Don	M	U of North Carolina	Computer Science	FP
Todorov, Diman	M	Vienna U of Technology	Research Industrial Software Engineering	NRG
Willinger, Walter	M	AT&T Labs	Research	FP
Xu, Peng	M	North Carolina State U	Electrical & Computer Engineering	NRG
Zhang, Lingsong	M	U of North Carolina	Statistics & Operations Research	NRG
Zhu, Zhengyuan	M	U of North Carolina	Statistics & Operations Research	NRG

NPCDS & SAMSI Workshop on the Design and Analysis of Computer Experiments for Complex Systems

Banff, Canada

Supported Workshop Participants

July 13-17, 2004

Name	Gender	Affiliation	Department	Status
Higdon, Dave	M	Los Alamos National Laboratory		FP
Lemieux, Christine	F	U of Calgary		FP
Lin, Chunfang		Simon Fraser U		NRG
Linkletter, Crystal	F	Simon Fraser U		NRG
Mease, David	M	U of California Berkeley		FP
Nakhleh, Charlie	M	Los Alamos National Laboratory		FP
Notz, Bill	M	Ohio State U		FP
Reese, Shane	M	Brigham Young U		FP
Rilett, Larry	M	U of Nebraska		FP
Santner, Tom	M	Ohio State U		FP
Spiegleman, Cliff	M	Texas A&M U		FP
Sudjianto, Agus	M	Bank of America		FP
Xu, Jiaying		U of Western Ontario		NRG
Ye, Kenny	M	SUNY Stoneybrooke		FP

NPCDS & SAMSI Workshop on the Design and Analysis of Computer Experiments for Complex Systems

Banff, Canada

Workshop Participants

July 13-17, 2004

Name	Gender	Affiliation	Department	Status
Banks, H.T.	M	North Carolina State U	CRSC	FP
Bayarri, M.J.	F	U of Valencia	Statistics & Operations Research	FP
Berger, Jim	M	SAMSI		FP
Bingham, Derek	M	Simon Fraser U	Statistics	FP
Booker, Andrew	M	Boeing		FP
Brewster, John	M	U of Manitoba	Statistics	FP
Cafeo, John	M	General Motors		FP
Chipman, Hugh	M	Acadia U	Mathematics & Statistics	FP
Garcia-Donato, Gonzalo	M	SAMSI		NRG
Hengartner, Nick	M	Los Alamos National Laboratory		FP
Higdon, Dave	M	Los Alamos National Laboratory		FP
Lemieux, Christiane	F	U of Calgary	Mathematics	FP
Lin, Chunfang		Simon Fraser U	Statistics	NRG
Linkletter, Crystal	F	Simon Fraser U	Statistics	NRG
Loeppky, Jason	M	Simon Fraser U	Statistics	NRG
Lu, Wilson	M	Simon Fraser U	Statistics	NRG
Macleod, Robert	M	U of Manitoba		NRG

Mease, David	M	U of Pennsylvania		FP
Meckesheimer, Martin	M	Boeing		FP
Michailidis, George	M	U of Michigan	Statistics	FP
Nakleh, Charlie	M	Los Alamos National Laboratory		FP
Notz, Bill	M	Ohio State U	Statistics	FP
Palomo, Jesus	M	SAMSI		NRG
Paulo, Rui	M	NISS & SAMSI		NRG
Pepin, Jason	M	Los Alamos National Laboratory		FP
Ranjam, Pritam	M	Simon Fraser U	Statistics	NRG
Reese, Shane	M	Brigham Young U	Statistics	FP
Rillet, Larry	M	U of Nebraska		FP
Sacks, Jerry	M	NISS		FP
Santner, Tom	M	Ohio State U	Statistics	FP
Sitter, Randy	M	Simon Fraser U	Statistics	FP
Spiegelman, Cliff	M	Texas A&M U	Statistics	FP
Stafford, Jamie	M	U of Toronto	Statistics	FP
Steinberg, David	M	Tel Aviv U	Statistics	FP
Sudjianto, Agus		Bank of America		FP
Welch, William	M	U of British Columbia	Statistics	FP
Xu, Jiaying		U of Guelph	Mathematics & Statistics	NRG
Ye, Kenny	M	SUNY Stoneybrook	Applied Mathematics & Statistics	FP

Education and Outreach Program
SAMSI-CRSC Industrial Mathematical and Statistical Modeling Workshop
For Graduate Students
North Carolina State University
Supported Workshop Participants
July 26-August 3, 2004

Name	Gender	Affiliation	Department	Status
Almond, Natalie	F	Western Carolina U		S
An, Jung-Ha	F	U of Florida		S
Boubakari, Ibrahimou	M	U of South Florida		S
Breen, Miyuki	F	Case Western Reserve U		S
Burgess, Richard	M	U of Tennessee		S
Chen, Zheng	M	Florida State U		S
Cheng, Jing	F	Purdue U		S
Ettinger, Bree	F	U of Georgia		S
Fernando, Harshini	F	Texas Tech U		S
Hariharanath, Kavuri	M	U of Tennessee		S
Kandala, Sampath	M	U of Tennessee		S
Khoujmane, Ali	M	Texas Tech U		S
Li, Xiaochuan	M	U of Southern Mississippi		S
Liu, Xingtao	M	State U of New York		S
Liu, Yuan	F	U of North Carolina-Wilmington		S
Nandi, Subhrangshu	M	U of Massachusetts		S
Ramirez, Ismael	M	UAM-I		S

Ren, Kui	M	Columbia U		S
Rusnica, Steven	M	North Carolina State U		S
Santos, Brenda	F	Centro de Investigacion		S
Vaughan, Tamara	F	U of Alabama		S
Wu, Meng	M	U of North Carolina-Wilmington		S

Education and Outreach Program
SAMSI-CRSC Industrial Mathematical and Statistical Modeling Workshop
For Graduate Students
North Carolina State University
Workshop Participants
July 26-August 3, 2004

Name	**Gender**	**Affiliation**	**Department**	**Status**
Almond, Natalie	F	Western Carolina U	Applied Mathematics	S
An, Jung-Ha	F	U of Florida	Applied Mathematics	S
Boubakari, Ibrahimou	M	U of South Florida	Mathematics	S
Breen, Miyuki	F	Case Western Reserve U	Genetic Epidemiology	S
Burgess, Richard	M	U of Tennessee	Aerospace Engineering	S
Chen, Zheng	M	Florida State U	Applied Mathematics	S
Cheng, Jing	F	Purdue U	Statistics	S
Ding, Wandi	F	U of Tennessee	Applied Mathematics	S
Ettinger, Bree	F	U of Georgia	Mathematics	S
Fernando, Harshini	F	Texas Tech U	Statistics	S
Hariharanath, Kavuri	M	U of Tennessee	Mechanical Engineering	S

Jackson, Billy	M	U of Georgia	Applied Mathematics	S
Kandala, Sampath	M	U of Tennessee	Mechanical Engineering	S
Khoujmane, Ali	M	Texas Tech U	Statistics	S
Li, Xiaochuan	M	U of Southern Mississippi	Scientific Computing	S
Liu, Xingtao	M	State U of New York	Applied Mathematics	S
Liu, Yuan	F	U of North Carolina-Wilmington	Statistics	S
Nandi, Subhrangshu	M	U of Massachusetts	Applied Mathematics & Industrial Eng	S
Ramirez, Ismael	M	UAM-I	Mathematics	S
Ren, Kui	M	Columbia U	Applied Mathematics	S
Rothstein, Ivan	M	Virginia Tech		S
Rusnica, Steven	M	North Carolina State U		S
Santos, Brenda	F	Centro de Investigacion	Mathematics	S
Sweetingham, Kelly	F	Auburn U	Applied Mathematics	S
Vaughan, Tamara	F	U of Alabama	Applied Mathematics	S
Walker, Matthew	M	Texas Tech U		S
Ward, Carrie	F	North Carolina State U	Applied Mathematics	S
Wilder, Mike	M	U of North Carolina-Greensboro	Statistics	S
Wilson, Heather	F	North Carolina State U	Applied Mathematics	S
Wu, Meng	M	U of North Carolina-Wilmington	Statistics	S

Latent Variables in the Social Sciences Program
Opening Workshop
Radisson Hotel Research Triangle Park
Supported Workshop Participants
September 11-14, 2004

Name	Gender	Affiliation	Department	Status
du Toit, Stephen	M	Scientific Software International	LISREL	FP
Feng, Shibao	M	Georgetown U Medical Center	Biomathematics & Biostatistics / Oncology	FP
Heckman, James	M	U of Chicago	Economics	FP
Ji, Ming	M	San Diego State U	Graduate School of Public Health	FP
Joreskog, Karl	M	Uppsala U	Information Science	FP
Kamata, Akihito	M	Florida State U	Educational Psychology	FP
Kreuter, Frauke	F	U of California, Los Angeles		FP
Lin, Xiaodong	M	SAMSI/NISS & Univ of Cincinnati	Mathematical Sciences	FP
Loh, Chung-Ping	M	U of North Florida	Economics & Geography	FP
Masyn, Katherine	F	U of California, Los Angeles		FP
Richardson, Thomas	M	U of Washington	Statistics	FP
Sanders, Liz	F	U of Washington	Educational Psychology	NRG
Sayer, Aline	F	U of Massachusetts, Amherst	Psychology	FP
Shakiban, Cheri	F	U of St. Thomas	Mathematics	FP
Skinner, Chris	M	U of Southampton	Statistical Sciences	FP
Wang, Xiaohui	F	U of Virginia	Statistics	FP

Latent Variables in the Social Sciences Program
Opening Workshop
Radisson Hotel Research Triangle Park
Workshop Participants
September 11-14, 2004

Name	Gender	Affiliation	Department	Status
Abraham, Todd	M	Iowa State U	Institute for Social & Behavioral Research	NRG
Aldridge, Arnie	M	U of North Carolina	Economics	NRG
Baker, Robin	M	MetaMetrics, Inc.	Research & Development	FP
Balasubramanian, Siva	M	Southern Illinois U	Marketing	FP
Banks, David	M	Duke U	Statistics	FP
Banks, H.T.	M	NCSU	Mathematics/CRSC	FP
Bann, Carla	F	RTI International	Statistics Research Division	FP
Barker, William	M	Indiana U of Pennsylvania	Educational & School Psychology	FP
Bauer, Dan	M	U of North Carolina	Psychology	FP
Baydoun, May	F	U of North Carolina	Epidemiology	NRG
Bender, Randy	M	RTI International	Statistics Research Division	FP
Berger, James	M	SAMSI		FP
Biemer, Paul	M	U of North Carolina & RTI	Odum Institute	FP
Bollen, Ken	M	U of North Carolina	Sociology	FP
Boswell, Gracie	F	U of North Carolina	Institute on Aging	NRG
Boye, Mark	M	Pfizer, Inc.	WWOR Ann Arbor Development Site	FP
Browne, Michael	M	Ohio State U	Psychology & Statistics	FP

Burdick, Don	M	MetaMetrics, Inc.		FP
Chantala, Kim	F	U of North Carolina	Carolina Population Center	FP
Chen, Mario	M	Family Health International	Biostatistics	FP
Cheng, Mariah Mantsun	F	U of North Carolina	Carolina Population Center	FP
Cheuk, Michelle	F	U of North Carolina	Sociology	NRG
Christ, Sharon	F	U of North Carolina	Sociology	NRG
Clarke, Bertrand	M	UBC/SAMSI/Duke	Statistics	FP
Cook, Richard	M	U of Waterloo	Statistics & Actuarial Science	FP
Cooper, Alexandra	F	Duke U	Social Science Research Institute	FP
Cross, Catherine	F	U of North Carolina	Carolina Population Center	FP
de Leon-Wong, Emelita	F	Family Health International	Biostatistics	FP
du Toit, Mathilda	F	Scientific Software International		FP
du Toit, Stephen	M	Scientific Software International	LISREL	FP
Duclos, Rod	M	U of North Carolina	Marketing	NRG
Dunson, David	M	NIEHS	Biostatistics	FP
Edwards, Lloyd	M	U of North Carolina	Biostatistics	FP
Feng, Shibao	M	Georgetown U Medical Center	Biomathematics & Biostatistics / Oncology	FP
Fisher, William	M	MetaMetrics, Inc.	Research & Development	FP
Flora, Dave	M	U of North Carolina	Psychology	FP
Ghosal, Subhashis	M	North Carolina State U	Statistics	FP
Giddings, Bethany	F	U of Waterloo	Statistics & Actuarial Science	NRG

Gu, Jiezhun	F	North Carolina State U	Statistics	NRG
Guha, Abhijit	M	Duke	Fuqua School of Business	NRG
Guo, Guang	M	U of North Carolina	Sociology	FP
Handcock, Mark	M	U of Washington	Statistics	FP
Hardin, James	M	U of South Carolina	Epidemiology and Biostatistics	FP
Heckman, James	M	U of Chicago	Economics	FP
Hendricks, Charlene	F	NIH/NICHD	Child & Family Research	FP
Hipp, John	M	U of North Carolina	Sociology	NRG
Huang, Reiping	F	U of Minnesota	Sociology	NRG
Jang, Woncheol	M	Duke U	Statistics	FP
Ji, Ming	M	San Diego State U	Graduate School of Public Health	FP
Joreskog, Karl	M	Uppsala U	Information Science	FP
Kaarsemaker, Eric	M	Radboud U Nijmegen	Nijmegen School of Management	NRG
Kainz, Kirsten	F	U of North Carolina	School of Education	NRG
Kamata, Akihito	M	Florida State U	Educational Psychology	FP
Karr, Alan	M	NISS		FP
Kelly, Christopher	M	U of North Carolina	Institute on Aging	NRG
Kenney, Melissa	F	Duke U	Nicholas Sch of Env & Earth Sci	NRG
Kim, Hae-Young		UNC	Dental Ecology	NRG
Kim, Su Hyun		U of North Carolina	School of Nursing	NRG
Kinlaw, Zack	M	U of North Carolina	Economics	NRG

Kinney, Saki	F	Duke U	Statistics	NRG
Kolenikov, Stanislav	M	UNC	Statistics	NRG
Kosinski, Andrzej	M	Duke	Biostatistics & Bioinformatics	FP
Kovacevic, Milorad	M	Statistics Canada	Methodology	FP
Kreuter, Frauke	F	U of California, Los Angeles		FP
Lada, Emily	F	SAMSI		NRG
Lance, Peter	M	U of North Carolina	Carolina Population Center	FP
Leach, Diane	F	NIH/NICHD	Child & Family Research	FP
Lewis, Terri	F	U of North Carolina	Biostatistics	FP
Li, Lei	M	RTI	Statistics Research Division	FP
Lin, Xiaodong	M	SAMSI/NISS & Univ of Cincinnati	Mathematical Sciences	FP
Liu, Ying		U of North Carolina	Carolina Population Center	NRG
Loh, Chung-Ping	M	U of North Florida	Economics & Geography	FP
Lovett, Mitchell	M	Duke	Marketing	NRG
Lu, Irene	F	York U	School of Administrative Studies	FP
Ma, Renjun	M	U of New Brunswick	Mathematics & Statistics	FP
MacKinnon-Tucker, Dorene	F	U of North Carolina	School of Education	NRG
Malone, Patrick	M	Duke	Center for Child and Family Policy	FP
Masyn, Katherine	F	U of California, Los Angeles		FP
McClain, Jill	F	U of North Carolina	Carolina Population Center	NRG
Medhin, Negash	M	NCSU	Mathematics	FP

Meekins, Brian	M	Bureau of Labor Statistics	Office of Survey Methods Research	FP
Mels, Gerhard	M	Scientific Software International Inc		FP
Monda, Keri	F	U of North Carolina	Carolina Population Center	NRG
Mroz, Thomas	M	U of North Carolina	Economics	FP
Munk, Tom	M	U of North Carolina	School of Education	NRG
Murphy, Nicole	F	Duke	Fuqua School of Business	NRG
Narayanan, Sriram	M	U of North Carolina	Kenan-Flagler Business School	NRG
Nguyen, Hoan	F	SAMSI		NRG
Nylund, Karen	F	U of California, Los Angeles		NRG
Palermo, Tia	F	U of North Carolina	Economics	NRG
Palomo, Jesus	M	SAMSI		NRG
Pennell, Michael	M	NIEHS	Biostatistics	NRG
Pepin, Kate	F	U of North Carolina		NRG
Pieper, Carl	M	Duke U Medical Center	Biometry & Bioinformatics	FP
Preacher, Kristopher	M	U of North Carolina	Psychology	FP
Reiter, Jerry	M	Duke U	Statistics	FP
Richardson, Thomas	M	U of Washington	Statistics	FP
Rigdon, Edward	M	Georgia State U	Marketing	FP
Russell, Daniel	M	Iowa State U	Institute for Social & Behavioral Research	FP
Saha, Shampa		RTI International	Statistics Research Division	FP
Samuels, Jr., Johnny	M	NCSU & SAMSI		NRG

Sanders, Liz	F	U of Washington	Educational Psychology	NRG
Sanford, Eleanor	F	MetaMetrics, Inc.	Research & Development	FP
Sayer, Aline	F	U of Massachusetts, Amherst	Psychology	FP
Scharoun, Melissa	F	U of North Carolina	Carolina Population Center	NRG
Shakiban, Cheri	F	U of St. Thomas	Mathematics	FP
Shvydko, Tetyana	F	U of North Carolina	Economics	NRG
Siemsen, Enno	M	U of North Carolina	Kenan-Flagler Business School	NRG
Skinner, Chris	M	U of Southampton	Statistical Sciences	FP
Sloane, Rick	M	Duke U	Center for the Study of Aging	FP
Stenner, Jack	M	MetaMetrics, Inc.		FP
Tao, Betty	F	U of North Carolina	Economics	NRG
Thomas, Roland	M	Carleton U	Sprott School of Business	FP
Thompson, Mary	F	U of Waterloo		FP
Tucker, Clyde	M	Bureau of Labor Statistics	Office of Survey Methods Research	FP
Tudor, Gail	F	UNC	Biostatistics	FP
Van Horn, Elizabeth	F	U of North Carolina	School of Nursing	NRG
Wang, Liqun	M	U of Manitoba	Statistics	FP
Wang, Xiaohui	F	U of Virginia	Statistics	FP
Wang, Zilin	F	Wilfrid Laurier U	Mathematics	FP
Williamson, Gary	M	MetaMetrics, Inc.	Research & Development	FP
Wu, Changbao	M	U of Waterloo	Statistics & Actuarial Science	FP

Zavisca, Jane	F	SAMSI		NRG
Zayats, Yaraslau	M	U of North Carolina	Economics	NRG

Genomes to Global Health: Computational Biology of Infectious Disease Program Opening Workshop

Radisson Hotel Research Triangle Park

Supported Workshop Participants

September 19-22, 2004

Name	Gender	Affiliation	Department	Status
Bajaria, Seema	F	U of Michigan	Microbiology & Immunology	NRG
Bortz, David	M	U of Michigan	Mathematics	FP
Chaloner, Kathryn	F	U of Iowa	Biostatistics	NRG
Chang, Stewart	M	U of Michigan	Microbiology and Immunology	NRG
Chang, Xiaoguang	M	U of Rochester	Biostatistics	NRG
Chen, Jianwei	M	U of Rochester	Biostatistics & Computational Biology	NRG
De Boer, Rob	M	Utrecht U	Ost - Pathogen	FP
De Gruttola, Victor	M	Harvard U	School of Public Health	FP
Feng, Changyong	M	U of Rochester	Biostatistics & Computational Biology	NRG
Fuhrman, Kseniya	F	U of Wisconsin - Milwaukee	Mathematical Sciences	NRG
Goldstein, Byron	M	Los Alamos National Laboratory	Theoretical Biology & Biophysics	FP
Halloran, Betz	F	Emory U	Biostatistics	FP
Huang, Yangxin	M	U of Rochester	Biostatistics & Computational Biology	NRG
Joshi, Hem Raj	M	Xavier U	Mathematics and CS	NRG

Kesmir, Can	F	Utrecht U		FP
Khattree, Ravindra	M	Oakland U	Mathematics & Statistics	FP
Khoujmane, Ali	M	Texas Tech U	Mathematics & Statistics	NRG
Kirschner, Denise	F	U of Michigan	Microbiology & Immunology	FP
Liang, Hua	M	St. Jude Children's Research Hospital	Biostatistics	FP
Liu, Dacheng	M	U of Rochester	Biostatistics & Computational Biology	NRG
Ma, Jingming	M	U of Rochester	Biostatistics & Computational Biology	NRG
Mallick, Bani	M	Texas A&M U	Statistics	FP
Marino, Simeone	M	U of Michigan	Microbiology & Immunology	NRG
Mubayi, Anuj		Arizona State U - Tempe	Mathematics & Statistics	NRG
Ratnam, Ponmalar	F	U of Wisconsin - Milwaukee	Mathematics	NRG
Ray, Christian	M	U of Michigan	Microbiology & Immunology	NRG
Velasco-Hernandez, Jorge	M	Instituto Mexicano del Petróleo	Matemáticas Aplicadas y Computación	FP
Wang, Haonan	M	Colorado State U	Statistics	NRG
Wang, Lily	F	Vanderbilt Universtity	Biostatistics	NRG
Wu, Hulin	M	U of Rochester	Biostatistics & Computational Biology	FP
Zhang, Cun-Quan	M	West Virginia U	Mathematics	FP

Genomes to Global Health: Computational Biology of Infectious Disease Program Opening Workshop

Radisson Hotel Research Triangle Park

Workshop Participants

September 19-22, 2004

Name	Gender	Affiliation	Department	Status
Adams, Brian	M	North Carolina State U	Mathematics	NRG
Ahn, Chaehyung		U of North Carolina	Biostatistics	NRG
Alker, Alisa	F	U of North Carolina	Epidemiology	NRG
Atchley, William	M	North Carolina State U	Genetics	FP
Bajaria, Seema	F	U of Michigan	Microbiology & Immunology	NRG
Banks, H.T.	M	SAMSI & North Carolina State U	Mathematics	FP
Berger, James	M	SAMSI		FP
Blaser, Martin	M	New York U	School of Medicine	FP
Bortz, David	M	U of Michigan	Mathematics	FP
Carr, Thomas	M	Southern Methodist U	Mathematics	NRG
Chaloner, Kathryn	F	U of Iowa	Biostatistics	NRG
Chang, Stewart	M	U of Michigan	Microbiology and Immunology	NRG
Chang, Xiaoguang	M	U of Rochester	Biostatistics	NRG
Chen, Jianwei	M	U of Rochester	Biostatistics & Computational Biology	NRG
Chen, Xi		North Carolina State U	Statistics	NRG
Clark, Leona	F	Bennett College	Mathematics	FP
Clarke, Bertrand	M	UBC/SAMSI/Duke U	Statistics	FP

Cooke, Ben	M	Duke U	Mathematics	NRG
Cowell, Lindsay	F	Duke U	Biostatistics & Bioinformatics	FP
Crooks, James	M	U of North Carolina	Statistics	NRG
Crosslin, David	M	Duke U	Center for Clinical & Genetic Economics	FP
Curto, Carina	F	Duke U	Mathematics	NRG
Datta, Sujay	M	Northern Michigan U	Mathematics, Statistics & Computer Science	FP
De Boer, Rob	M	Utrecht U	Ost - Pathogen	FP
De Groot, Anne	F	Brown U	Bio Med Community Health	FP
De Gruttola, Victor	M	Harvard U	School of Public Health	FP
DePuy, Venita	F	Duke U	Center for Clinical & Genetic Economics	FP
Doughty, Sabrina	F	Virginia Tech U	Biology	NRG
Eckstrand, Irene Anne	F	National Institute of General Medical Sci	Cen for Bioinformatics & Comp Bio	FP
Ellwein, Laura	F	North Carolina State U	Mathematics	NRG
Elston, Timothy	M	U of North Carolina	Mathematics	FP
Feng, Changyong	M	U of Rochester	Biostatistics & Computational Biology	NRG
Feng, Sheng		North Carolina State U	Statistics	NRG
Fisher, William	M	MetaMetrics, Inc	Research & Development	FP
Fitch, Walter	M	U of California - Irvine	Eco-Evo	FP
Fueyo, Joanna	F	IBM	Information Based Medicine	NRG
Fuhrman, Kseniya	F	U of Wisconsin - Milwaukee	Mathematical Sciences	NRG
Goldstein, Byron	M	Los Alamos National Laboratory	Theoretical Biology & Biophysics	FP

Gupta, Mayetri	F	U of North Carolina	Biostatistics	NRG
Halloran, Betz	F	Emory U	Biostatistics	FP
Haney, Rich	M	Independent Software Developer		FP
He, Min	M	Duke U	Biostatistics & Bioinformatics	NRG
Huang, Yangxin	M	U of Rochester	Biostatistics & Computational Biology	NRG
Ji, Chuanshu	M	U of North Carolina	Statistics	FP
Joshi, Hem Raj	M	Xavier U	Mathematics and CS	NRG
Karr, Alan	M	NISS		FP
Kepler, Thomas	M	Duke U	Bioinformatics & Computational Biology	FP
Kesmir, Can	F	Utrecht U		FP
Khattree, Ravindra	M	Oakland U	Mathematics & Statistics	FP
Khoujmane, Ali	M	Texas Tech U	Mathematics & Statistics	NRG
Kirschner, Denise	F	U of Michigan	Microbiology & Immunology	FP
Kulasingam, Shalini	F	Duke U	Obstetrics and Gynecology	FP
Liang, Hua	M	St. Jude Children's Research Hospital	Biostatistics	FP
Liu, Dacheng	M	U of Rochester	Biostatistics & Computational Biology	NRG
Liu, Yufeng		U of North Carolina	Statistics	FP
Lloyd, Alun	M	North Carolina State U	Mathematics	FP
Lu, Jun	M	Duke U	Biostatistics & Bioinformatics	NRG
Ma, Jingming	M	U of Rochester	Biostatistics & Computational Biology	NRG
Mallick, Bani	M	Texas A&M U	Statistics	FP

Marino, Simeone	M	U of Michigan	Microbiology & Immunology	NRG
Markward, Nathan	M	V.A. Metrics		FP
Martin, William	M	EPIVAX		FP
Mitha, Faheem	M	Duke U	Biostatistics & Computational Biology	NRG
Mubayi, Anuj		Arizona State U - Tempe	Mathematics & Statistics	NRG
Neville, Padraic	M	SAS		FP
Nguyen, Hoan	F	SAMSI & North Carolina State U	Mathematics	NRG
Nobel, Andrew	M	U of North Carolina	Statistics	FP
Oduola, Ayoade	M	World Health Organization	Basic & Strategic Research	FP
Park, Soyoun	F	U of North Carolina	Statistics	NRG
Pittman, Jennifer	F	Duke U	Statistics	FP
Pourbohloul, Babak	M	U of British Columbia	Centre for Disease Control	FP
Ratnam, Ponmalar	F	U of Wisconsin - Milwaukee	Mathematics	NRG
Ray, Christian	M	U of Michigan	Microbiology & Immunology	NRG
Ray, Surajit	M	SAMSI		NRG
Rodriguez, Abel	M	Duke U	Statistics	NRG
Root, Morgan	M	North Carolina State U	Mathematics	NRG
Schmidler, Scott	M	Duke U	Statistics	FP
So, Anthony	M	Duke U	Terry Sanford Institute of Public Policy	FP
Thorne, Jeff	M	North Carolina State U	Bioinformatics Research Center	FP
Todd, Abby	F	U of North Carolina	Mathematics	NRG

Tomfohr, John	M	Duke U	Biostatistics & Bioinformatics	NRG
Truong, Young	M	SAMSI & U of North Carolina	Biostatistics	FP
Velasco-Hernandez, Jorge	M	Instituto Mexicano del Petróleo	Matemáticas Aplicadas y Computación	FP
Wagener, Diane	F	RTI International	Survey Research Div	FP
Wang, Haonan	M	Colorado State U	Statistics	NRG
Wang, Lily	F	Vanderbilt Universtity	Biostatistics	NRG
Wang, Xiao	M	U of North Carolina	Statistics & Operations Research	NRG
Wu, Hulin	M	U of Rochester	Biostatistics & Computational Biology	FP
Wu, Yichao		U of North Carolina	Statistics	NRG
Young, Stanley	M	NISS		FP
Zavisca, Jane	F	SAMSI		NRG
Zhang, Cun-Quan	M	West Virginia U	Mathematics	FP
Zuo, Peiying		U of North Carolina	Mathematics	NRG

Multiscale Model Development and Control Design Program
Closing Workshop
NISS-SAMSI Building
Supported Workshop Participants
September 27-28, 2004

Name	**Gender**	**Affiliation**	**Department**	**Status**
Gleser, Leon	M	U of Pittsburgh	Statistics	FP
Krener, Arthur	M	U of California, Davis	Mathematics	FP
Weiland, Lisa	F	Virginia Tech	Mechanical Engineering	NRG

Zabaras, Nicholas	M	Cornell U	Mechanical & Aerospace Engineering	FP

Multiscale Model Development and Control Design Program
Closing Workshop
NISS-SAMSI Building
Workshop Participants
September 27-28, 2004

Name	Gender	Affiliation	Department	Status
Banks, H.T.	M	North Carolina State U	CRSC	FP
Bokil, Vrushali	F	North Carolina State U	CRSC	NRG
Braun, Thomas	M	North Carolina State U	Mathematics	NRG
Ellwein, Laura	F	North Carolina State U	Mathematics	NRG
Ernstberger, Jon	M	North Carolina State U	Mathematics	NRG
Friedman, Jack	M	U of Chicago	Committee on Human Development	FP
Gelfand, Alan	M	Duke U	ISDS	FP
Gibson, Nathan	M	North Carolina State U	CRSC	NRG
Gleser, Leon	M	U of Pittsburgh	Statistics	FP
Krener, Arthur	M	U of California, Davis	Mathematics	FP
Lada, Emily	F	SAMSI		NRG
Matthews, Jessica	F	North Carolina State U	Mathematics	NRG
Mattingly, Jonathan	M	Duke U	Mathematics	FP
Mossi, Karla	F	Virginia Commonwealth U	Mechanical Engineering	FP
Newell, Andrew	M	North Carolina State U	Mathematics	NRG

Nguyen, Hoan	F	SAMSI & North Carolina State U	CRSC	NRG
Oates, William	M	North Carolina State U	Mathematics	NRG
Ounaies, Zoubeida	F	Virginia Commonwealth U	Mechanical Engineering	NRG
Smith, Ralph	M	North Carolina State U	Mathematics	FP
Vance, Eric	M	Duke U	ISDS	NRG
Ward, Carrie	F	North Carolina State U	Mathematics	NRG
Weiland, Lisa	F	Virginia Tech	Mechanical Engineering	NRG
Zabaras, Nicholas	M	Cornell U	Mechanical & Aerospace Engineering	FP

Workshop on Data Mining Methodology
The Fields Institute, Toronto, Canada
Supported Workshop Participants
October 28-30, 2004

Name	**Gender**	**Affiliation**	**Department**	**Status**
Cutler, Adele	F	Utah State U	Mathematics & Statistics	FP
Rosset, Saharon	M	IBM	TJ Watson Research Center	NRG
Shi, Bin	M	Georgia Tech U	Industrial & Systems Engineering	NRG
Zhu, Ji	M	U of Michigan	Statistics	NRG

Workshop on Data Mining Methodology
The Fields Institute, Toronto, Canada
Workshop Participants
October 28-30, 2004

Name	**Gender**	**Affiliation**	**Department**	**Status**
Aldave, Roberto	M	Generation 5		FP

Al-khasawneh, Mohanad		U of Windsor	Mathematics & Statistics	NRG
Allenger, Dennis	M	Canada Revenue Agency		FP
An, Lihua		U of Windsor	Mathematics & Statistics	NRG
Andreopoulos, William	M	York U	Computer Science	NRG
Andritsos, Periklis		U of Toronto	Computer Science	NRG
Aul, Chhad		Hillsdale Investment Management Inc.		FP
Banks, David	M	Duke U	Institute of Statistics and Decision Sciences	FP
Bengio, Yoshua		U de Montreal	IRO	FP
Benn, Alfred	M	TD Bank Financial Group		FP
Beyene, Joseph	M	Hospital for Sick Children	Population Health Sciences	FP
Boudreau, Francois	M	Dofasco Inc.	Process Automation Technology	FP
Bucur-Drugan, Cristina	F	TDBFG	GRM	FP
Bukhman, Yury	M	Protana Inc.		FP
Cao, Jiguo		McGill U	Mathematics and Statistics	NRG
Cao, Shelley Yun	F	U of Toronto	Statistics	NRG
Catsiliras, Kathy	F	Exchange Solutions Inc.		FP
Chamandy, Nicholas	M	McGill U	Mathematics and Statistics	NRG
Chau, Douglas	M	OPSEU Pension Trust	Investments	FP
Chen, Huarong	M	McMaster U	Computing Science	NRG
Chenouri, Shoja'eddin	M	U of Waterloo	Statistics and Actuarial Science	NRG
Chipman, Hugh A.	M	Acadia U	Mathematics and Statistics	FP

Choulakian, Vartan	M	U of Moncton	Mathematics and Statistics	FP
Chu, Peter	M	Protana Inc.		FP
Ciampi, Antonio	M	McGill U	Epidemiology and Biostatistics	FP
Clyde, Merlise A.	F	Duke U	Institute of Statistics and Decision Sciences	FP
Cutler, Adele	F	Utah State U	Mathematics and Statistics	FP
Dean, Jonathan	M	Hillsdale Investment Management Inc.		FP
Depoutvitch, Alex	M	Generation 5		FP
Duiella, Gianluigi		Exchange Solutions Inc.	Management Sciences	FP
Ewing, Rob	M	Protana	Sci Computing	FP
Fan, Chun-Po Steve		U of Toronto	Statistics	NRG
Fan, Guangzhe		U of Waterloo	Statistics and Acturial Science	FP
Fan, Jing		RBC Financial Group	Client & Marketing Strategy, Customer Info Management	FP
Friedman, Jerome	M	Stanford U	Statistics	FP
Galluzzi, Aaron	M	Corel Corporation	eMarketing Strategy and Intelligence	FP
Gashaw, Asrat Fikre		York U	Mathematics	NRG
Ghori, Razi Uddin		U of Windsor	Mathematics and Statistics	NRG
Gluzman, Simon	M	Generation 5		FP
Gu, Hong	F	Dalhousie U	Mathematics and Statistics	FP
Hossain, Md. Shakhawat		U of Windsor	Mathematics & Statistics	NRG
Howrigan, Shaun		Lakehead U	Independent Researcher	FP
Huang, Jiayuan	F	U of Waterloo	School of Computer Science	NRG

Huang, Wenxue		Generation5 Incorporated	Data Modeling & Statistical Analysis	FP
Huynh, Wyn		HIllsdale Investment Management Inc.		FP
Iglesias-Gonzalez, Sigfrido		U of Toronto	Statistics	NRG
Islam, Mohammad	M	U of Western Ontario	Statistical and Actuarial Sciences	NRG
Jarosz, Tracey	F	Loyalty Group		FP
Karakoulas, Grigoris		U of Toronto	Computer Science	NRG
Khan, Hafiz M. R.		U of Western Ontario	Statistical & Actuarial Sciences	NRG
Kourti, Theodora	F	McMaster U	Chemical Engineering	FP
Krneta, Milorad	M	Generation5 Incorporated	Data Modeling & Statistical Analysis	FP
Kröger, Helmut	M	Universite' Laval	Physics, Engineering Physics and Optics	FP
Larocque, Denis	M	HEC Montreal	Management	FP
Lee, Sophia	F	U of Toronto	Public Health Sciences	NRG
Li, Jeff	M	Generation 5		FP
Li, Mei		Carleton U	Mathematics and Statistics	NRG
Li, Xuli		The Loyalty Group	Database and Analytical team	FP
Lim, Sooyeol		Hospital for Sick Children	Genetics and Genomic Biology	FP
Lin, Jen-Wen		U of Western Ontario	Statistical and Actuarial Sciences	FP
Liu, Xianping		Generation5 Incorporated	Production and Research	FP
Liu, Xiwu		U of Toronto	Statistics	NRG
Loeppky, Jason	M	U of British Columbia	Statistics	NRG
Lu, Wilson	M	Simon Fraser U	Statistics and Actuarial Science	NRG

Ma, Guoxuan	M	McMaster U	Mathematics and Statistics	NRG
Makos, Rick	M	Teradata		FP
Markatou, Marianthi	F	Columbia U	Biostatistics, Mailman School of Public Health	FP
McCulloch, Robert	M	U of Chicago	Graduate School of Business	FP
McLeod, Ian A.	M	U of Western Ontario	PSU Chair	FP
Meré, Joaquin Ordieres	M	Universidad de la Rioja	Edificio Departamental	FP
Mersov, George	M	Generation 5		FP
Miletic, Ivan	M	Dofasco, Inc.	Process Automation Technology	FP
Mills, Kathryn	F	Carleton U	Mathematics and Statistics	NRG
Mills, Shirley	F	Carleton U	Mathematics and Statistics	FP
Mosesova, Sofia	F	U of Waterloo	Statistics and Actuarial Science	NRG
Nettel-Aguirre, Alberto	M	U of Calgary	Mathematics and Statistics	NRG
Norminton, Ted	M	Carleton U	Mathematics and Statistics	FP
Nwankwo, Anastesia	F	Enugu State U (Esutech)	Statistics & Computer Science	FP
Pan, Kevin	M	U of Western Ontario	Mathematics	NRG
Parkhomenko, Elena	F	U of Toronto	PHS	NRG
Payandeh, Amir T.		U of New Brunswick	Mathematics and Statistics	NRG
Peng, Jiming		McMaster U	Computing and Software	NRG
Preobrajenski, Serguei	M	Bell Canada		FP
Robinson, David	M	Canadian Revenue Agency	Compliance Programs Branch	FP
Robinson, Mark	M	U of Toronto	Banting and Best Dept of Medical Research	NRG

Rosset, Saharon		IBM	TJ Watson Research Center	FP
Saarenvirta, Gary	M	Loyalmetrics Inc.		FP
Safyallah, Hossein		McMaster U	Computing and Software	NRG
Sanjel, Deepak		U of Western Ontario	Statistical and Actuarial Sciences	FP
Shen, Hui		U of British Columbia	Statistics	NRG
Shi, Bin	M	Georgia Institute of Technology	Industrial and Systems Engineering	NRG
Shi, Xiaofei		Dalhousie U	Mathematics and Statistics	NRG
Staicu, Ana-Maria	F	U of Toronto	Statistics	NRG
Steele, Russell	M	McGill U	Mathematics and Statistics	FP
Su, Wanhua		U of Waterloo	Statistics and Actuarial Science	NRG
Sun, Tao		York U	Mathematics and Statistics	NRG
Toussaint, Godfried	M	McGill U	Computer Science	FP
Tritchler, David	M	U of Toronto/OCI/PMH	Medical Biophysics	FP
Vainder, Michael	M	Generation 5	Director G5 Lab	FP
Wang, Grant	M	Norkom Technologies		FP
Wang, Steven Xiaogang	M	York U	Mathematics and Statistics	NRG
Wang, Xu		U of Waterloo	Statistics and Actuarial Science	NRG
Wei, Yu	M	McMaster U	Computing and Software	NRG
Welch, Will	M	U of British Columbia	Statistics	FP
Wu, Jianhong		York U	Mathematics and Statistics	FP
Wu, Long Yang		Mount Sinai Hospital	SLRI	FP

Yan, Guohua	M	U of British Columbia	Statistics	NRG
Yin, Ling		U of Toronto	Statistics	NRG
Young, Stan	M	National Institute of Statistical Sciences	SAMSI	FP
Zamar, Reuben	M	U of British Columbia	Statistics	FP
Zhang, Jianguo		U of Toronto	Statistics	NRG
Zhang, Jiaqi		U of Toronto	Statistics	NRG
Zhang, Kun		U of Toronto	Statistics	NRG
Zheng, Zheng		U of Toronto	Statistics	NRG
Zhou, Wu		U of Waterloo	Statistics and Actuarial Science	NRG
Zhu, Ji	M	U of Michigan	Statistics	FP
Zhu, Mu		U of Waterloo	Statistics and Actuarial Science	FP
Zighed, Djamel	M	Université Lumière Lyon 2		FP
Zolotovitski, Alex	M	Generation 5		FP
Zurkowski, Victor Daniel	M	Generation 5		FP

Data Assimilation for Geophysical Systems Program
Opening Workshop
Radisson Hotel Research Triangle Park
Supported Workshop Participants
January 23-26, 2005

Name	**Gender**	**Affiliation**	**Department**	**Status**
Anderson, Jeffrey	M	National Center for Atmospheric Research	Data Assimilation Initiative	FP
Bishop, Craig	M	Naval Research Laboratory		FP

Fitzmaurice, Jean	F	Massachusetts Institute of Technology		NRG
Gelaro, Ron	M	NASA	Global Modeling and Assimilation Office	FP
Genton, Marc	M	Texas A&M U		FP
Herbei, Radu	M	Florida State U	Statistics	NRG
Kalnay, Eugenia	F	U of Maryland		FP
Kaufman, Cari	F	Carnegie Mellon U	Statistics	NRG
Khoujmane, Ali	M	Texas Tech U	Mathematics and Statistics	NRG
Kim, Sangil	M	U of Arizona	Applied Mathematics Program	NRG
Lam, Chen Quin "Eric"	M	Ohio State U	Statistics	NRG
Lekien, Francois	M	Princeton U	Mechanical & Aerospace Engineering	NRG
Leonard, Naomi Ehrich	F	Princeton U	Mechanical and Aerospace Engineering	FP
Linfoot, Andy	M	U of Arizona	Applied Mathematics	NRG
Matsuo, Tomoko	F	National Center for Atmospheric Research	Inst for Mathematics Applied to Geosciences	NRG
McKeague, Ian	M	Columbia U	Biostatistics	FP
Miller, Robert	M	Oregon State U	Oceanic and Atmospheric Sciences	FP
Nychka, Doug	M	National Center for Atmospheric Research		FP
Rodriguez-Yam, Gabriel	M	Colorado State U	Statistics	NRG
Speer, Kevin	M	Florida State U	Oceanography	NRG
Stroud, Jonathan	M	U of Pennsylvania	Statistics	FP
Toth, Zoltan	M	NCEP/NWS/NOAA	Environmental Modeling Center	FP

Wang, Haonan	M	Colorado State U	Statistics	NRG
Wang, Ke	F	Colorado State U	Statistics	NRG
Wang, Xiaohui Amanda	F	U of Virginia	Statistics	NRG
Wikle, Christopher	M	U of Missouri-Columbia	Statistics	FP
Wunsch, Carl	M	Massachusetts Institute of Technology	EAPS	FP

Data Assimilation for Geophysical Systems Program
Opening Workshop
Radisson Hotel Research Triangle Park
Workshop Participants
January 23-26, 2005

Name	**Gender**	**Affiliation**	**Department**	**Status**
Anderson, Jeffrey	M	National Center for Atmospheric Research	Data Assimilation Initiative	FP
Apte, Amit	M	SAMSI & U of North Carolina	Mathematics	NRG
Bach, Jr., Walter	M	U.S. Army Research Office		FP
Banks, H.T.	M	SAMSI & CRSC-NC State U		FP
Bayarri, M.J.	F	U of Valencia	Statistics and Operations Research	FP
Berger, Jim	M	SAMSI		FP
Berman, Spring	F	Princeton U	Mechanical and Aerospace Engineering	NRG
Bishop, Craig	M	Naval Research Laboratory		FP
Budhiraja, Amarjit	M	U of North Carolina	Statistics and Operations Research	FP
Choi, Jungsoon	F	North Carolina State U	Statistics	NRG
Choudhury, Roy	M	U of Central Florida	Mathematics	FP

Crooks, James	M	U of North Carolina	Statistics	NRG
Davis, Jerry	M	North Carolina State U	Marine, Earth & Atmospheric Sciences	FP
Dinwoodie, Ian	M	Duke U	Institute of Statistics and Decision Sciences	FP
Drewry, Darren	M	Duke U	Civil and Environmental Engineering	NRG
El Moghraby, Amal	F	UNC & Brown U	Math & Applied Math	NRG
Finkelstein, Peter	M	U.S. EPA		FP
Fitzmaurice, Jean	F	Massachusetts Institute of Technology		NRG
Foley, Kristen	F	North Carolina State U	Statistics	NRG
Foster, Steven	M	U of North Carolina	Mathematics	NRG
Gelaro, Ron	M	NASA	Global Modeling and Assimilation Office	FP
Genton, Marc	M	Texas A&M U		FP
Ghosh, Sujit	M	North Carolina State U	Statistics	FP
Herbei, Radu	M	Florida State U	Statistics	NRG
Holland, David	M	U.S. EPA	National Exposure Research Laboratory	FP
Howard, Louis	M	Duke U	Mathematics	FP
Ide, Kayo	F	U of California Los Angeles	Atmospheric and Oceanic Sciences	FP
Jefferys, William	M	U of Texas	Astronomy	FP
Jones, Chris	M	U of North Carolina	Mathematics	FP
Kalnay, Eugenia	F	U of Maryland		FP
Kaufman, Cari	F	Carnegie Mellon U	Statistics	NRG
Khare, Shree	M	SAMSI		NRG

Khoujmane, Ali	M	Texas Tech U	Mathematics and Statistics	NRG
Kim, Sangil	M	U of Arizona	Applied Mathematics Program	NRG
Koller, Josef	M	Los Alamos National Laboratory	Space Science and Applications	FP
Kyung, Minjung	F	North Carolina State U	Statistics	NRG
Lam, Chen Quin "Eric"	M	Ohio State U	Statistics	NRG
Lehman, Thomas	M	U of North Carolina	Mathematics	FP
Lekien, Francois	M	Princeton U	Mechanical & Aerospace Engineering	NRG
Leonard, Naomi Ehrich	F	Princeton U	Mechanical and Aerospace Engineering	FP
Li, Zhijin	M	Jet Propulsion Laboratory & Raytheon		FP
Lin, Zhi	M	U of North Carolina	Mathematics	NRG
Linfoot, Andy	M	U of Arizona	Applied Mathematics	NRG
Liu, Liyan	F	U of North Carolina	Mathematics	NRG
Lopes, Brian	M	U of North Carolina	Statistics & Operations Research	NRG
Lozier, Susan	F	Duke U	Earth and Ocean Sciences	FP
Maroulas, Vasileios	M	U of North Carolina	Statistics and Operations Research	NRG
Marron, J.S.	M	U of North Carolina	Statistics & Operations Research	FP
Martinsen-Burrell, Neil	M	U of North Carolina	Mathematics	NRG
Matsuo, Tomoko	F	National Center for Atmospheric Research	Inst for Mathematics Applied to Geosciences	NRG
McGuire, Jeff	M	Woods Hole Oceanographic Inst	Geology and Geophysics	FP
McKeague, Ian	M	Columbia U	Biostatistics	FP
Mich, Nicole	F	Duke U & SAMSI		NRG

Miller, Robert	M	Oregon State U	Oceanic and Atmospheric Sciences	FP
Naehr, Stephen	M	Rice U	Physics and Astronomy	NRG
Nguyen, Hoan	F	SAMSI		NRG
Nychka, Doug	M	National Center for Atmospheric Research		FP
Otte, Tanya	F	NOAA	Atmospheric Sciences Modeling Division	FP
Pai, Prashant	M	U of North Carolina	Carolina Environmental Program	NRG
Palomo, Jesus	M	SAMSI		NRG
Peng, Shiqiu	M	North Carolina State U		NRG
Pensky, Marianna	F	SAMSI		FP
Ray, Surajit	M	SAMSI		NRG
Restrepo, Juan	M	U of Arizona	Mathematics	FP
Rodriguez-Yam, Gabriel	M	Colorado State U	Statistics	NRG
Salman, Hayder	M	U of North Carolina	Mathematics	NRG
Scott, Sherry	F	U of North Carolina	Mathematics	NRG
Smith, Leonard	M	London School of Economics		FP
Smith, Richard	M	U of North Carolina	Statistics and Operations Research	FP
Snyder, Chris	M	National Center for Atmospheric Research		NRG
Song, Hae-Ryoung	F	North Carolina State U	Statistics	NRG
Speer, Kevin	M	Florida State U	Oceanography	NRG
Stephens, Monica	F	Spelman College	Mathematics	FP
Stroud, Jonathan	M	U of Pennsylvania	Statistics	FP

Swall, Jenise	F	NOAA & EPA	Atmospheric Sciences Modeling Division	FP
Toth, Zoltan	M	NCEP/NWS/NOAA	Environmental Modeling Center	FP
Wang, Haonan	M	Colorado State U	Statistics	NRG
Wang, Ke		Colorado State U	Statistics	NRG
Wang, Xiaohui Amanda	F	U of Virginia	Statistics	NRG
Wikle, Christopher	M	U of Missouri-Columbia	Statistics	FP
Wunsch, Carl	M	Massachusetts Institute of Technology	EAPS	FP
Xu, Ke	F	U of North Carolina	Mathematics	NRG
Yu, Jie	F	U of Manchester	Mechanical, Aerospace & Civil Engineering	FP
Zavisca, Jane	F	SAMSI		NRG
Zhang, Xuejin	M	North Carolina State U	Marine, Earth & Atmospheric Sciences	NRG
Zhou, Jie		U of North Carolina	Mathematics	NRG
Zhu, Zhengyuan	M	U of North Carolina	Statistics and Operations Research	FP

Genomes to Global Health: Computational Biology of Infectious Disease
Mid-Term Workshop on Mathematical Modeling of Infectious Disease
NISS-SAMSI Building
Supported Workshop Participants
January 31-February 1, 2005

Name	**Gender**	**Affiliation**	**Department**	**Status**
Castillo-Chavez, Carlos	M	Arizona State U	Mathematics & Statistics	FP
Chowell, Gerardo	M	Los Alamos National Laboratory	Mathematical Modeling and Analysis	FP
Feng, Zhilan	F	Purdue U	Mathematic	FP

Galvani, Alison	F	Yale U	Epidemiology	FP
Greenwood, Priscilla	F	Arizona State U	Mathematics and Statistics	FP
Kribs-Zaleta, Christopher	M	U of Texas at Arlington	Mathematics	FP

Genomes to Global Health: Computational Biology of Infectious Disease
Mid-Term Workshop on Mathematical Modeling of Infectious Disease
NISS-SAMSI Building
Workshop Participants
January 31-February 1, 2005

Name	**Gender**	**Affiliation**	**Department**	**Status**
Adams, Brian	M	North Carolina State U	CRSC	NRG
Alker, Alisa	F	U of North Carolina		NRG
Banks, H.T.	M	SAMSI & CRSC-NC State		FP
Barzohar, Efrat	F	Arizona State U	Mathematics & Biology	NRG
Castillo-Chavez, Carlos	M	Arizona State U	Mathematics & Statistics	FP
Chowell, Gerardo	M	Los Alamos National Laboratory	Mathematical Modeling and Analysis	FP
Cintron-Arias, Ariel	M	Arizona State U	Mathematics & Statistics	NRG
Cooke, Ben	M	Duke U	Mathematics	NRG
Davis, Jimena	F	North Carolina State U	Mathematics & CRSC	NRG
Feng, Zhilan	F	Purdue U	Mathematic	FP
Galvani, Alison	F	Yale U	Epidemiology	FP
Greenwood, Priscilla	F	Arizona State U	Mathematics and Statistics	FP
Grove, Sarah	F	North Carolina State U	Mathematics	NRG

Hu, Shuhua		North Carolina State U		NRG
Kang, Yun	F	Arizona State U	Mathematics and Statisitics	NRG
Kribs-Zaleta, Christopher	M	U of Texas at Arlington	Mathematics	FP
Kwon, Hee-Dae		Inha U & North Carolina State U	CRSC	NRG
Lloyd, Alun	M	North Carolina State U	Mathematics and Biomath	FP
Nguyen, Hoan	F	SAMSI & North Carolina State U		NRG
Nuno, Miriam	F	Cornell U	Biological Statistics and Computational Biology	NRG
Perelson, Alan	M	Los Alamos National Laboratory		FP
Root, Morgan	M	North Carolina State U	Mathematics	NRG
Samuels, Johnny	M	SAMSI & North Carolina State U	Mathematics & CRSC	NRG
Shim, Alicia	F	Arizona State U	Mathematics and Statistics	NRG
Tennenbaum, Stephen	M	Cornell U		NRG
Toivanen, Jari	M	North Carolina State U	CRSC	FP
Tran, Hien	M	North Carolina State U	Mathematics	FP

Education and Outreach Program
Two-Day Undergraduate Workshop on Computational Biology & Social Sciences
NISS-SAMSI Building
Supported Workshop Participants
February 18-19, 2005

Name	**Gender**	**Affiliation**	**Department**	**Status**
Algai, Shiran	M	Clemson U	Computer Engineering	S
Andreas, Ashley	F	U of Kansas	Math Education	S

Bates, Vivian	F	Benedict College		F
Burns, Coryna	F	U of San Diego	Psychology & Mathematics	S
Chandler, Adam Daniel	M	Duke U	Mathematics & Chemistry	S
Doyle, Jimmy	M	Clemson U	Mathematics & Biology	S
Gant, Raymond	M	Benedict College		S
Hong, Pauline	F	Duke U	Mathematics	S
Ip, Sarah	F	Clemson U	Mathematical Sciences	S
Katz, Lindsey Ann	F	U of North Carolina-Wilmington	Statistics	S
Kaut, Erika	F	U of Kansas	Human Biology	S
Kembell, Cara	F	U of Kansas	Mathematics	S
Layne, Lori	F	Clemson U	Mathematics	S
Lee, Sae-kyoung	M	Cornell U	Mathematics & Economics	S
Linder, Elizabeth	F	U of Dayton		S
Lintner, Elizabeth	F	Indiana U	Mathematics & Economics	S
McFerrin, Lisa	F	Virginia Tech		S
Mian, Matthew	M	Duke U		S
Oduor, Brian	M	Connecticut College	Mathematics & Computer Science	S
Park, Min-Hyong	F	Cornell U	Mathematics & Economics	S
Poe, Sarah	F	U of Dayton	Mathematics	S
Prier, David	M	U of Dayton	Mathematics	S
Ridenhour, Jamie	F	North Carolina State U	Statistics	S

Rodriguez, Nancy	F	U of San Diego	Applied Mathematics	S
Rogers, Chanda	F	Benedict College	Chemistry & Mathematics	S
Sklodowski, Joanna	F	U of Dayton	Mathematics	S
Somen, Bostjan	M	Indiana U	Mathematics & Business	S
Taft, Shaun	M	North Carolina State U	Statistics	S
Welch, Lucia Marie	F	Clemson U	Mathematics & English	S
Wheeler, Micah	M	Christian Brothers U	Mathematics	S
Whittle, Elizabeth	F	Northern Kentucky U	Mathematics & Computer Science	S

Education and Outreach Program
Two-Day Undergraduate Workshop on Computational Biology & Social Sciences
NISS-SAMSI Building
Workshop Participants
February 18-19, 2005

Name	**Gender**	**Affiliation**	**Department**	**Status**
Algai, Shiran	M	Clemson U	Computer Engineering	S
Andreas, Ashley	F	U of Kansas	Math Education	S
Banks, H.T.	M	North Carolina State U & SAMSI	CRSC	A
Bates, Vivian	F	Benedict College		A
Bollen, Ken	M	U of North Carolina		A
Burns, Coryna	F	U of San Diego	Psychology & Mathematics	S
Chandler, Adam Daniel	M	Duke U	Mathematics & Chemistry	S
Cowell, Lindsay	F	Duke U		A

Doyle, Jimmy	M	Clemson U	Mathematics & Biology	S
Gant, Raymond	M	Benedict College		S
Hong, Pauline	F	Duke U	Mathematics	S
Ip, Sarah	F	Clemson U	Mathematical Sciences	S
Katz, Lindsey Ann	F	U of North Carolina-Wilmington	Statistics	S
Kaut, Erika	F	U of Kansas	Human Biology	S
Kembell, Cara	F	U of Kansas	Mathematics	S
Lada, Emily	F	SAS Institute		A
Layne, Lori	F	Clemson U	Mathematics	S
Lee, Sae-kyoung	M	Cornell U	Mathematics & Economics	S
Linder, Elizabeth	F	U of Dayton		S
Lintner, Elizabeth	F	Indiana U	Mathematics & Economics	S
McFerrin, Lisa	F	Virginia Tech		S
Mian, Matthew	M	Duke U		S
Nguyen, Hoan	F	SAMSI		A
Oduor, Brian	M	Connecticut College	Mathematics & Computer Science	S
Park, Min-Hyong	F	Cornell U	Mathematics & Economics	S
Poe, Sarah	F	U of Dayton	Mathematics	S
Prier, David	M	U of Dayton	Mathematics	S
Ridenhour, Jamie	F	North Carolina State U	Statistics	S
Rodriguez, Nancy	F	U of San Diego	Applied Mathematics	S

Rogers, Chanda	F	Benedict College	Chemistry & Mathematics	S
Sklodowski, Joanna	F	U of Dayton	Mathematics	S
Smith, Ralph	M	CRSC-NC State U		A
Somen, Bostjan	M	Indiana U	Mathematics & Business	S
Taft, Shaun	M	North Carolina State U	Statistics	S
Welch, Lucia Marie	F	Clemson U	Mathematics & English	S
Wheeler, Micah	M	Christian Brothers U	Mathematics	S
Whittle, Elizabeth	F	Northern Kentucky U	Mathematics & Computer Science	S
Zavisca, Jane	F	SAMSI		A

APPENDIX F – Workshop Programs and Abstracts

I. DATA MINING AND MACHINE LEARNING

A. Closing Workshop Program & Abstracts
May 17-18, 2004

Monday – May 17, 2004

RADISSON HOTEL RESEARCH TRIANGLE PARK, ROOM F-G

8:30-9:00 am Registration Check-In

9:00-9:15 am Welcome and Introductions
Alan Karr, Director of NISS & DMML Program Leader

9:15-10:45 am Working Group I: Bioinformatics

9:15-9:45 Summary
Stanley Young, NISS

9:45-10:15 Technical Highlight
Jacqueline Hughes-Oliver, North Carolina State University

10:15-10:30 Problem List
Young Truong, University of North Carolina-Chapel Hill

10:30-10:45 Discussion

10:45-11:00 am Coffee Break

11:00-12:30 am Working Group II: Large P, small n

11:00-11:30 Summary
Bertrand Clarke, Duke University and SAMSI

11:30-12:00 Technical Highlight
Ernest Fokoue, SAMSI

12:00-12:15 Problem List
Xiaodong Lin, SAMSI and NISS

12:15-12:30 Discussion

12:30-1:30 pm Lunch

1:30-3:00 pm Working Group III: Support Vector Machines

1:30-2:00 Summary

Marc Genton, North Carolina State University

2:00-2:30 Technical Highlight
Helen Zhang, North Carolina State University

2:30-2:45 Problem List
Ernest Fokoue, SAMSI

2:45-3:00 Discussion

3:00-3:30 pm Coffee Break

3:30-5:00 pm Working Group IV: Theory and Methods

3:30-4:00 Summary
David Banks, Duke University

4:00-4:30 Technical Highlight
Merlise Clyde, Duke University

4:30-4:45 Problem List
Feng Liang, Duke University

4:45-5:00 Discussion

5:00-6:30 pm "Open Mike" Discussion

Tuesday – May 18, 2004

RADISSON GOVERNORS INN, ROOM F-G

8:30-9:00 am Registration Check-In

9:00-10:00 am Panel Discussion I
Working Group Leaders

10:00-10:15 am Coffee Break

10:15-11:30 am Graduate Student Presentations

- **Atina Brooks**, North Carolina State University
- **Jen-hwa Chu**, Duke University
- **Leanna House**, Duke University
- **Fei Liu**, Duke University
- **Peng Liu**, North Carolina State University

11:30-12:30 pm Panel Discussion II
Selected Outside Panelists

12:30-1:30 pm Lunch

1:30-3:00 pm Birds of a Feather Sessions (Optional)

3:00 pm Adjourn

Atina Dunlap Brooks
North Carolina State University
Department of Statistics
adbrook2@stat.ncsu.edu

Automated SVM Tuning on HTS Datasets

It is well understood that tuning Support Vector Machine (SVM) parameters, including kernel selection, to the dataset being analyzed will increase performance. However, High Throughput Screening (HTS) datasets often have data features which are difficult to model, so determining the best SVM parameters for a particular HTS dataset can be difficult. Our idea is to use ten fold cross validation to systematically compare different SVM kernels, kernel parameters and cost factors. Code was developed to wrapper the SVMLight (Joachims, 1999) SVM software and perform automated comparisons. These results allow easy use of SVMs on complicated datasets requiring very little knowledge of the dataset or even SVM theory. This code is evaluated on several different HTS datasets. Additionally the tuned SVM's performance is compared against a default SVM.

Jen-hwa Chu
Duke University
Institute of Statistics and Decision Sciences
jenhwa@stat.duke.edu

Choice of Statistics in LPSN Settings

We describe a method to reduce the dimension in 'Large P, small n' settings by taking summary statistics over groups of variables and regularizing on them instead of the full data. We have tested our proposed method on a variety of simulated data. Our results suggest that regularization on summary statistics can give smaller predictive errors in several realistic settings.

Bertrand Clarke
Duke University
Institute of Statistics and Decision Sciences
bertrand@stat.duke.edu

Short, Fat Data: Old Techniques and New Perspectives

We review some of the established techniques for dealing with 'Large P, small n' data and then discuss some recent ideas and contributions by members of the working group from the past year. These include basis considerations, connections with other working groups, and a discussion of recent efforts to combine various ideas in a general strategy.

Leanna House
Duke University

Institute of Statistics and Decision Sciences
house@stat.duke.edu

Cherry Picking: Multidimensional Perspective

No abstract submitted

Fei Liu
Duke University
Institute of Statistics and Decision Sciences
fei@stat.duke.edu

Mining Temporal Pattern

No abstract submitted

Peng Liu
North Carolina State University
Department of Statistics
pliu3@ncsu.edu

Mining Spatio-Temporal Data

Ionomeric polymer membranes exhibit electromechanical and chemoelectric transduction when plated with conductive metal electrodes. In this poster we will overview our recent efforts to model the physics of ionomeric transducers using electrostatic models of sensing and actuation and phenomenological models of electromechanical coupling. Experimental results will be presented for a variety of ionomers, cation types, and solvents.

Jacqueline Hughes-Oliver
North Carolina State University
Department of Statistics
hughes01@stat.ncsu.edu

"Twins and High-Dimensional Data: Can Leave-one-out Cross Validation Survive?"

High-dimensional low-sample-size problems cause many difficulties for analysis and interpretation. Leave-one-out cross validation is routinely recommended as a technique for assessing a model's predictive power without overfitting. Unfortunately, when the dataset contains virtually identical cases (which we call twins), leave-one-out cross validation may give an overly optimistic assessment of a model's predictive power. Using a recently published study for investigating whether in vitro gene expression profiles of drug efficacy can predict therapeutic classes of compounds, we demonstrate the existence of twins and their impact on several cross-validation studies.

Young Truong
University of North Carolina-Chapel Hill
Department of Biostatistics
truong@bios.unc.edu

Future Problems for Bioinformatics Team

No abstract submitted

Helen Zhang
North Carolina State University
Department of Statistics
hzhang2@stat.ncsu.edu

Technical Highlight for Support Vector Machine

The use of kernels is one key factor in the success of many classification algorithms by allowing nonlinear decision surfaces. We investigate the radial basis function (RBF) kernels and show that, one mathematical formulation permits an interesting modification of any RBF kernel yielding a compactly supported kernel which has a sparse Gram matrix. Different degrees of sparsity lead to different severity in information loss during the modification. For any compactly supported kernel, we propose quantitative measures for evaluating its similarity to the original kernel and its sparsity. Furthermore, a systematic way is suggested to tune the thresholding parameter which controls the similarity-sparsity tradeoff. We illustrate an approach on support vector machines, least squares support vector machines, and kernel principal component analysis. Simulations show that, properly-tuned compactly supported kernels give comparable performances while enjoying more efficient computation.

II. NETWORK MODELING FOR THE INTERNET PROGRAM

A. Closing Workshop Program & Abstracts
June 25-26, 2004

Friday – June 25, 2004
NISS-SAMSI Building, Room 104

8:30-9:10 am Registration and Continental Breakfast

9:10-9:15 am Welcome
Jim Berger, SAMSI

9:15-9:30 am SAMSI Internet Overview
J.S. Marron, University of North Carolina & SAMSI

9:30-10:25 am **SAMSI RESEARCH I**
Chair: **J.S. Marron**

9:30-9:45 "Long-Range-Dependence in a Changing Internet Traffic Mix"
Don Smith, University of North Carolina

9:50-10:05 "Dependent SiZer: Goodness of Fit Tests for Time Series Models"
Cheolwoo Park, SAMSI

10:10-10:25 "Wavelet-based Synthesis of the Rosenblatt process"
Vladas Pipiras, University of North Carolina

10:30-11:00 am Coffee Break

11:00-12:15 pm **SAMSI RESEARCH II**
Chair: **Cheolwoo Park**

11:00-11:15 "Algebraic Equations for Blocking Probabilities in Asymmetric Networks"
Ian Dinwoodie, Duke University & SAMSI

11:20-11:35 "Adaptive Scheduling Using Online Measurements for Efficient Delivery of Quality of Service"
Peng Xu, North Carolina State University

11:40-11:55 "Semi Experiment Analysis of the Shifting Knee Wavelet Spectrum"
Felix Hernandez-Campos, University of North Carolina & SAMSI

12:00-12:15 "Long Range Dependence Analysis of Internet Traffic"
J. S. Marron, University of North Carolina & SAMSI

12:15-1:45 pm Lunch

1:45-3:00 pm **NEW DIRECTIONS I**
Chair: **Don Smith**

1:45-2:00 "Splitting and Merging of a Traffic Model: Validation"
Nicolas Hohn, University of Melbourne

2:05-2:20 "Traffic Utilization Profiles of Internet Users"
M. Rosario de Oliveira, Instituto Superior Tecnico and CEMAT

2:25-2:40 "Delay Tomography"
Earl Lawrence, University of Michigan

2:45-3:00 "File Sizes Reloaded"
Diman Todorov, Vienna Technical University

3:00-3:30 pm Coffee Break

3:30-4:45 pm **NEW DIRECTIONS II**
Chair: **Robert Buche**

3:30-3:45 "Biased Sampling from Heavy Tailed Data"
Barbara Gonzalez, University of Louisiana

3:50-4:05 "Advanced Distribution Theory for SiZer"
Jan Hannig, Colorado State University

4:10-4:25 "Inferring Queuing Behavior From Passive TCP Traces"

Jasleen Kaur, University of North Carolina

4:30-4:45 "Analysis of Wireless Information Locality and Association Patterns in a Campus"
Maria Papadopouli, University of North Carolina

4:50-5:05 pm **OPEN DISCUSSION**: "Wished I had said..." Session

Saturday – June 26, 2004
NISS-SAMSI Building, Room 104

8:30-9:10 am Registration and Continental Breakfast

9:10-10:25 am **SAMSI RESEARCH III**
Chair: **Vladas Pipiras**

9:10-9:25 "Thresholded Log-Log Correlation Analyses of TCP Characteristics"
Felix Hernandez-Campos, University of North Carolina & SAMSI

9:30-9:45 "Impact of the Sampling Rate on the Estimation of the Parameters of Fractional Brownian Motion"
Zhengyuan Zhu, University of North Carolina

9:50-10:05 "Queueing Analysis of Network Traffic"
David Rolls, University of North Carolina at Wilmington

10:10-10:40 am Coffee Break

10:40-11:55 am **SAMSI RESEARCH IV**
Chair: **Felix Hernandez-Campos**

10:40-10:55 "Visualization and Inference Based on Wavelet Coefficients, SiZer and SiNos"
Cheolwoo Park, SAMSI

11:00-11:15 "Identification of Periodic and Cyclic Fractional Stable Motions"
Vladas Pipiras, University of North Carolina

11:20-11:35 "Robust Estimation of Self-similarity Parameter Using the Wavelet Transform"
Zhengyuan Zhu, University of North Carolina

11:40-11:55 "Heavy Traffic Working Group"
Robert Buche, North Carolina State University

12:00-1:30 pm Lunch

1:30-4:00 pm **ISERC**

Chair: **David Rolls**

1:30-2:20	Overview and Future Directions **J. S. Marron**, University of North Carolina & SAMSI **Walter Willinger**, AT&T Labs
2:20-3:30	Breakout Discussion Groups & Break
3:30-4:00	Discussion Results Presentations

Robert Buche
North Carolina State University
Department of Mathematics
rtbuche@unity.ncsu.edu

Heavy Traffic Working Group

Heavy traffic methods for queuing systems (in particular Brownian approximations) applied to the internet are in the beginning stages. In [1], which was the subject of Ruth William's talk at the SAMSI heavy traffic workshop, work in this area is introduced with several conjectures (problems) for further investigation pointed out. The heavy traffic working group has started to look at one of these conjectures: state-space collapse for a diffusion approximation for the number of flows in a route. We will describe the model in [1] and the state space collapse conjectured there, the results in [2] proving state space collapse for multiclass queuing networks (this was pointed out in [1] as the likely framework for proving the conjecture), and comment briefly on the current focus of the group.

[1] William, R.J. and Kelly, F.P. "Fluid Model for a Network Operating under a Fair Bandwidth-Sharing Policy", Preprint, 2004.
[2] Bramson, M. " State space collapse with application to heavy traffic limits for multiclass queuing networks, Queuing Systems, Vol. 30. (1998), pp. 89-148.

M. Rosario de Oliveira
Instituto Superior Tecnico and CEMAT
Department of Mathematics
rosario.oliveira@math.ist.utl.pt

Traffic Utilization Profiles of Internet Users

M. Rosário de Oliveira(1),
Rui Valadas(2), António Pacheco(1), Paulo Salvador(2)

(1)Instituto Superior
Técnico - UTL
Department of Mathematics and CEMAT

Av. Rovisco Pais, 1049-001
Lisboa, Portugal

(2)University of Aveiro / Institute of Telecommunications

Aveiro
Campus de Santiago, 3810-193 Aveiro, Portugal

The identification of profiles of user utilization can be of significant importance in traffic engineering and tariffing. The knowledge of groups with similar traffic utilization can help to define better ISP marketing policies but can also be useful to propose tariffing policies that may promote internet access during the least busy hours, resulting in better quality services provide to internet users.

In this work we analyze one day traffic utilization, on two different ISPs, measured every half-hour. Making use of cluster analysis, three profiles (clusters) with homogeneous behaviour are identified in each ISP. The profiles are validated using discriminant analysis. The descriptive study and the principal component analysis performed on the data give a coherent interpretation of the identified profiles, leading to the conclusion that similar patterns and interpretations of the profiles were found in the two ISPs. Finally, the applications, flow durations and transfer rates of each profile are analyzed.

Ian Dinwoodie
Duke University
Institute of Statistics & Decision Sciences
ihd@stat.duke.edu

Algebraic Equations for Blocking Probabilities in Asymmetric Networks

Polynomial systems for blocking probabilities on networks with dynamic routing are derived. With capacity parameters on the edges, we seek parameter regions that give two meaningful probability solutions, a situation known to exist in the symmetric network. The technique of traceforms solves the problem in many cases.

Barbara Gonzalez
University of Louisiana
Department of Mathematics
babara@louisiana.edu

Biased Sampling from Heavy Tailed Data

Random sampling from heavy tailed data seems to erase some of the heavy tailed properties that are usually of interest. We develop a different sampling method to deal with this problem and then study the properties of the sample we obtain with this method. Some graphical goodness of fit procedures are discussed.

Jan Hannig
Colorado State University
Department of Statistics
hannig@stat.colostate.edu

Advanced Distribution Theory for SiZer

SiZer is a powerful method for exploratory data analysis. In this paper the distributions underlying the statistical inference are investigated, and large improvements are made in the

approximation. This results in improved size, and in an improved global inference version of SiZer. The main points are illustrated with simulated examples.

Felix Hernandez-Campos
University of North Carolina
Department of Computer Science
fhernand@cs.unc.edu

Semi-experiment Analysis of the Shifting Knee Wavelet Spectrum

No abstract submitted

Thresholded Log-Log Correlation Analyses of TCP Characteristics

We explore the use of the joint distribution of TCP flow durations and sizes as a means to characterize the dependency between user behavior and network performance. In particular, we uncover some pitfalls in the use of the log-log correlation on thresholded data, which can result in misleading conclusions.

Nicolas Hohn
University of Melbourne
Department of Electrical & Electronic Engineering
n.hohn@ee.mu.oz.au

Splitting and Merging of a Traffic Model: Validation

This talk concerns the modelling of Internet packet traffic. In previous work, we showed that the Bartlett-Lewis point process is a very good model of packet arrivals with physically motivated parameters. It is based on the central empirical observation that flows can be considered independent for the purpose of modelling packet arrival times. In this talk, we extend these results in two ways by using a unique dataset obtained from an experimental setup where all the packets crossing a router are captured. First we check that the fundamental assumptions underlying our physical packet traffic model are verified on a variety of links, with a large range of speeds and utilizations. Second, we show how the model can capture the splitting and merging of traffic streams through a router. This has far reaching consequences since it proves that the model is very versatile: it is not limited to individual backbone links but can instead potentially be applied through an entire network.

Jasleen Kaur
University of North Carolina
Department of Computer Science
jasleen@cs.unc.edu

Inferring Queuing Behavior From Passive TCP Traces

Packet queuing at routers and switches impacts the end-to-end performance received by an application running over the Internet. Not much is known about the nature of packet queuing in today's Internet --- on what kind of nodes do packets get queued, how many space is allocated for queues in today's routers, at what time-scales do queues grow and shrink and by how much? It is the goal of our research to infer the queuing behavior experienced by packets of TCP connections

by examining passively-collected traces of real Internet connections. In this talk, we will briefly outline ongoing work.

Earl Lawrence
University of Michigan
Department of Statistics
earl@umich.edu

Delay Tomography

No abstract submitted

Steve Marron
University of North Carolina & SAMSI
Department of Statistics & Operations Research
marron@email.unc.edu

Long Range Dependence Analysis of Internet Traffic

No abstract submitted

Maria Papadopouli
University of North Carolina
Department of Computer Science
maria@cs.unc.edu

Analysis of Wireless Information Locality and Association Patterns in a Campus

Our goal is to explore characteristics of the environment that provide opportunities for caching, prefetching, coverage planning, and resource reservation. We conducted a one-month measurement study of locality phenomena among wireless web users and their association patterns on the entire UNC-CH campus using the IEEE 802.11 wireless infrastructure.

We evaluated the performance of different caching paradigms, such as single user cache, cache attached to an access point (AP), and peer-to-peer caching. In several settings such caching mechanisms could be beneficial. Unlike other measurement studies in wired networks in which 25\% to 40\% of documents draw 70\% of web access, our traces indicate that 13\% of unique \textsc{url}s draws this number of web accesses. In addition, the overall ideal hit rates of the user cache, cache attached to an access point, and peer-to-peer caching paradigms (where peers are coresident within an AP) are 51\%, 55\%, and 23\%, respectively.

We distinguished wireless clients based on their inter-building mobility, their visits to APs, their continuous walks in the wireless infrastructure, and their wireless information access during these periods. We modeled the associations as a Markov chain using as state information the most recent AP visits. We can predict with high probability (86\%) the next AP with which a wireless client will associate. Such measurements can benefit protocols and algorithms that aim to improve the performance of the wireless infrastructures by load balancing, admission control, and resource reservation across APs.

You can find more info about my research at http://www.cs.unc.edu/~maria.

Cheolwoo Park
SAMSI
cwpark@samsi.info

Dependent SiZer: Goodness of Fit Tests For Time Series Models

In this paper, we extend SiZer (SIgnificant ZERo crossing of the derivatives) to dependent data for the purpose of goodness of fit tests for time series models. Dependent SiZer compares the observed data with a specific null model being tested by adjusting the statistical inference using an assumed autocovariance function. This new approach uses a SiZer type visualization to flag statistically significant differences between the data and a given null model. The power of this approach is demonstrated through some examples of time series of Internet traffic data. It is seen that such time series can have even more burstiness than is predicted by the popular, long range dependent, Fractional Gaussian Noise model.

Visualization and Inference Based on Wavelet Coefficients, SiZer and SiNos

SiZer (SIgnificant ZERo crossing of the derivatives) and SiNos (SIgnificant NOnStationarities) are scale-space based visualization tools for statistical inference. They are used to discover meaningful structure in data through exploratory analysis involving statistical smoothing techniques. Wavelet methods have been successfully used to analyze various types of time series. In this paper, we propose a new time series analysis approach, which combines the wavelet analysis with the visualization tools SiZer and SiNos. We use certain functions of wavelet coefficients at different scales as inputs, and then apply SiZer or SiNos to highlight potential non-stationarities. We show that this new methodology can reveal hidden local non-stationary behavior of time series, that are otherwise difficult to detect.

Vladas Pipiras
University of North Carolina
Department of Statistics & Operations Research
pipiras@email.unc.edu

Wavelet-based Synthesis of the Rosenblatt Process

The Rosenblatt process is an important example of self-similar stationary increments stochastic processes whose finite-dimensional distributions are non-Gaussian with all their moments finite. We introduce a wavelet-based simulation of the Rosenblatt process. The simulation is computationally fast and provides approximations, which converge to the Rosenblatt process exponentially fast and almost surely uniformly on compact intervals.

Identification of Periodic and Cyclic Fractional Stable Motions

Self-similar stable mixed moving average processes can be related to nonsingular flows through their minimal representations. Self-similar mixed moving averages are called periodic, resp. cyclic, fractional stable motions if their minimal representations are generated by periodic, resp. cyclic, flows. These processes, however, are often defined by a non-minimal representation. We provide a way to detect whether they are periodic or cyclic even if their representation isnon-minimal. By using these identification results, we obtain a more refined decomposition of self-similar mixed moving averages.

David Rolls
University of North Carolina - Wilmington
Department of Mathematics & Statistics
rollsd@uncw.edu

Queuing Analysis of Network Traffic: a Theoretical Framework and Visualization Tools

Trace-driven queuing simulation is discussed as an appropriate technique for exploring internet traffic data. Metrics and visualizations are discussed. In a particular example using data from the UNC Computer Science network testbed, traces that look similar under various statistical measures are shown to exhibit rather different behavior under queuing simulation.

Don Smith
University of North Carolina
Department of Computer Science
smithfd@cs.unc.edu

Long-Range-Dependence in a Changing Internet Traffic Mix

This talk describes a deep analysis of long-range dependence in a continually evolving Internet traffic mix by employing a number of recently developed statistical methods. Our study considers time-of-day, day-of-week, and cross-year variations in the traffic on an Internet link. Surprisingly large and consistent differences in the behavior of packet-count time series were observed between data from 2002 and 2003. A careful examination, based on stratifying the data according to protocol, revealed that the large difference was driven by a single UDP application that was not present in 2002. Another result was that the observed large differences between the two years showed up only in packet-count time series, and not in byte counts (while conventional wisdom suggests that these should be similar). We also found that a number of traces exhibited more “bursty” characteristics than could be modeled as Fractional Gaussian Noise. We explored some of the implications resulting from these findings for buffering and queue management in routers.

Diman Todorov
Vienna University of Technology
Department of Research Industrial Software Engineering
diman.todorov@chello.at

File Sizes Reloaded

A rigorous analysis of a recent survey shows that file sizes on computer systems come from the same distribution like ten years ago.

Several papers (A. B. Downey, ``The structural cause of file size distributions", Hernandez-Campos, Marron, Samorodnitsky, Smith, ``Variable Heavy Tails in Internet Traffic.", and others) have shown parallels between the distribution of file sizes on different Unix machines. The results of those papers are based on Gordon Irlam's ``Unix File Size" survey which has been made in 1993. I reproduced this survey on a smaller scale and compared the new data with the data so broadly analyzed. Since both datasets are very large (several billion samples) much attention has been paid to time efficient analyzis by combining two very different calculation evironments (C and R). I performed the comparison with Monte Carlo simulated chi squared tests. The yielded

results strongly support the hypothesis that the analyzed distribution hasn't changed significantly since 1993.

Based on this result I have developed several theories on why nothing has changed in the structure of the data in spite of the advances in technology in the past ten years. One of the ideas is that the structure is not inherent to computer systems but to the users operating those systems. Further research has to be done to support or falsify those ideas.

Peng Xu
North Carolina State University
Department of Electrical Computer Engineering
pxu3@ncsu.edu

Adaptive Scheduling Using Online Measurements for Efficient Delivery of Quality of Service

In this presentation, we investigate in detail the performance properties of a dynamic resource allocation scheme that utilizes online measurements to optimally adjust scheduling weights and to achieve the required QoS, including packet loss ratio and queuing delay, under a given pricing structure. The proposed scheme can be implemented under various differentiated network frameworks, such as IntServ, DiffServ, and Traffic Engineering with MPLS. Our scheme adjusts the weights of the underlying packet scheduling algorithm, such as weighted round robin or weighted fair queuing, by utilizing the information contained in traffic measurements obtained over successive time windows.

Our objective is to formally investigate optimal settings that guarantee an improved QoS performance in both stable and unstable cases. Furthermore, we aim to develop fundamental insights based on a detailed, case-by-case mathematical model that takes into account the relevant QoS parameter constraints. A complete optimal solution algorithm for the stable (the capacity of the system exceeds the total input rate) and the unstable case is presented in this paper. Extensive simulation results validate the proposed scheme and show that it exhibits an excellent performance regarding provisioning of available resources for delivering the required QoS, compared to static schemes.

Zhengyuan Zhu
University of North Carolina
Department of Statistics & Operations Research
z-zhu@unc.edu

Impact of the Sampling Rate on the Estimation of the Parameters of Fractional Brownian Motion

No abstract submitted

Robust Estimation of Self-Similarity Parameter Using the Wavelet Transform

No abstract submitted

III. MULTISCALE MODEL DEVELOPMENT AND CONTROL DESIGN

A. Closing Workshop Program & Abstracts

September 27-28, 2004

Monday-September 27, 2004
NISS-SAMSI Building, Room 104

8:00 - 9:00 am Registration and Continental Breakfast.

9:00 - 12:00 pm **SESSION 1***: Homogenization and Controls Group*

- **Nathan Gibson,** North Carolina State University
- **Vrushali Bokil**, North Carolina State University
- **Art Krener**, University of California- Davis

12:00 - 1:00 pm LUNCH

1:00 - 4:15 pm **SESSION 2***: Paradigms and Ionic Polymer Groups*

- **Emily Lada,** North Carolina State University
- **Lisa Weiland,** Virginia Tech University
- **Andrew Newell,** North Carolina State University
- **Joe Lucas**, Duke University

Tuesday-September 28, 2004
NISS-SAMSI Building, Room 104

8:00 - 9:00 am Registration and Continental Breakfast.

9:00 - 12:00 pm **SESSION 3**: *Presentations on Multiscale Issues and Applications*

- *9:00-9:45* Nicholas Zabaras, Cornell University
- *10:00-10:45* Jonathan Mattingly, Duke University
- *11:00-11:45* Open Discussion

12:00 pm Adjournment

Vrushali Bokil
North Carolina State University
Department of Mathematics/CRSC
vabokil@ncsu.edu

Multiscale and Homogenization in Dielectrics

No abstract submitted

Nathan Gibson
North Carolina State University
Department of Mathematics/CRSC

ngibson@ncsu.edu

Multiscale and Polarization in Dielectrics

No abstract submitted

Arthur Krener
University of California, Davis
Department of Mathematics
ajkrener@ucdavis.edu

Control of Multiscale Models

No abstract submitted

Emily Lada
SAMSI
elada@samsi.info

Monte Carlo Simulation of a Solvated Ionic Polymer with Cluster Morphology

A multiscale method for describing the macroscopic material properties of the ionc polymer Nafion is presented. Traditional rotational isomeric state theory is applied in combination with a Monte Carlo methodology to develop a simulation model of the formation of Nafion polymer chains on a microscopic level from which a large number of end-to-end chain lengths are generated. The probability density function of end-to-end distances is then estimated and used as an input parameter to enhance existing energetics-based macroscale models of ionic polymer behavior. Several methods for estimating the probability density function are compared, including estimation using Johnson distributions, Bezier distributions, and cubic splines.

Jonathan Mattingly
Duke University
Department of Mathematics
jonm@math.duke.edu

Scale in Stochastic Partial Differential Equations

I will try to highlight some different questions of scale which arise in Stochastic Partial differential equations. I will talk about the differences between the typical questions asked in the randomly forced and random medium settings. I will talk about the different types of scaling limits often considered and their underlying assumptions. Then I'll contrast that will the questions in large scale forced systems. I will definitely ask more questions than I will answer and try to keep the discussion elementary and connected to some of the issues raised in the working groups during the past year.

Lisa Weiland
Virginia Tech
Department of Mechanical Engineering
lmauck@vt.edu

Relating Rotational Isomeric State Theory to Ionomer Stiffness

No abstract submitted

Nicholas Zabaras
Cornell University
Materials Process Design and Control Laboratory
Sibley School of Mechanical and Aerospace Engineering
zabaras@cornell.edu

A Computational Statistics and Stochastic Modeling Approach to Multiscale Modeling and Design Of Materials

We will discuss a number of diverse problems that highlight the important role that stochastic modeling, Bayesian computation and statistical learning techniques can play in the analysis, design and control of materials systems. Topics to be introduced include:

- Development of a variational multiscale approach with subgrid scale modeling for stochastic continuum transport systems.
- Bayesian statistical inference and advanced simulation techniques for multiscale and multiphysics transport phenomena in porous media.
- An information theoretic approach to uncertainty propagation across length scales in the deformation of polycrystal materials.
- Hierarchical statistical (machine) learning techniques for representation, modeling, design and control of materials across length scales.

IV. LATENT VARIABLES IN THE SOCIAL SCIENCES

A. Opening Workshop Program & Abstracts
September 11-14, 2004

Saturday – September 11, 2004: Tutorials
Radisson Hotel Research Triangle Park
Room DE (2nd Floor)

1:00-3:00 pm	An Overview of Structural Equation Models with Latent Variables **Kenneth Bollen,** University of North Carolina at Chapel Hill
3:00-3:30 pm	Break
3:30-5:30 pm	An Overview of Multilevel Models **Aline Sayer**, University of Massachusetts, Amherst

Sunday – September 12, 2004: Workshop
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

9:30-9:45 am	Welcome and Introductions **James Berger,** SAMSI **Kenneth Bollen,** University of North Carolina at Chapel Hill

9:45-10:45 am Measurement Error in Generalized Linear Models
James Hardin, University of South Carolina

10:45-11:00 am Break

11:00-12:00 pm Variables in Heterogeneous Response Models
James Heckman, University of Chicago

12:00-1:00 pm Lunch

1:00-2:00 pm New Researcher Presentations

- **Frauke Kreuter**, University of Maryland
  Applications of Growth Mixture Modeling to Non-Normal Outcomes
- **Katherine Masyn**, University of California, Los Angeles
  Multivariate Extensions in Discrete-Time Survival Analysis using Latent Variables
- **Jerome Reiter**, Duke University
  Multiple Imputation for Missing Data in Surveys with Complex Sampling Designs

- **Jane Zavisca**, SAMSI
  Latent Class Analysis of Cultural Fragmentation in the U.S.

2:00-2:45 pm Break

2:45-5:00 pm Session on Analysis pf Complex Survey Data & Latent Variables: New Developments & Challenges

- **Paul Biemer**, Odum Institute, University of North Carolina at Chapel Hill & RTI

- **Stephen du Toit,** Scientific Software (LISREL)

- **Chris Skinner**, University of Southampton

- **Mary Thompson**, University of Waterloo

5:00-5:20 pm Introductions to Evening Posters

Monday – September 13, 2004: Workshop
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

9:00 am- 12:00 pm Session on Structural Equation Models: New Development & Challenges

- **Ken Bollen**, University of North Carolina at Chapel Hill

- **Michael Browne**, Ohio State University
- **Karl Joreskog**, University of Uppsala
- **Thomas Richardson**, University of Washington

12:00-1:00 pm Lunch

1:00-3:30 pm Session on Longitudinal Models: New Developments & Challenges

- **Richard Cook**, University of Waterloo
- **David Dunson**, National Institute of Environmental Health Sciences
- **Lloyd Edwards**, University of North Carolina at Chapel Hill
- **Guang Guo**, University of North Carolina at Chapel Hill

3:30-4:00 pm Break

4:00-5:00 pm Birds-of-a-Feather Sessions

Tuesday – September 14, 2004: Workshop and Working Group Meetings
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

9:30 -12:00 pm Session on Multilevel Models: New Developments & Challenges

- **Dan Bauer,** University of North Carolina at Chapel Hill
- **Mark Handcock,** University of Washington
- **Aki Kamata,** Florida State University
- **Aline Sayer,** University of Massachusetts, Amherst

12:00-1:00 pm Lunch

1:00-1:30 pm Group Discussion to Identify Initial Working Groups

1:30-3:00 pm Working Group Meetings

3:00-3:30 pm Working Group Reports

3:30 pm Workshop Adjourns

Dan Bauer
University of North Carolina & North Carolina State University
Department of Psychology
dan_bauer@ncsu.edu

Title and abstract not submitted

Paul Biemer
Odum Institute, University of North Carolina & RTI International
ppb@rti.org

Title and abstract not submitted

Ken Bollen
University of North Carolina
Department of Sociology
bollen@unc.edu

An Overview of Structural Equation Models with Latent Variables

No abstract submitted

Michael Browne
Ohio State University
Department of Psychology
browne.4@osu.edu

Estimation or Approximation in Structural Equation Modeling

A distinction is made between estimation with tests of model fit under the assumption of a correct model and approximation with measures of model adequacy when no model is regarded as correct. Previous research is reviewed. An example is presented in which measures of model fit derived under the assumption of a correct model give paradoxical results when applied in situations where an approximation is required. Suggestions for future research are given.

Richard Cook
University of Waterloo
Statistics & Actuarial Science
rjcook@uwaterloo.ca

Cure-Rate Models for Current Status Data

A method for fitting cure-rate models to current status data are described via the EM algorithm. Indentifiability issues are considered. This problem is motivated by work on an orthopedic study of patients undergoing hip and knee replacement.

Stephen du Toit
Scientific Software International
LISREL
sdutoit@ssicentral.com

Latent Variable Models and Complex Survey Data

There has been a growing interest in recent years in fitting models to data collected from longitudinal surveys that use complex sample designs. This interest reflects expansion in requirements by policy makers and researchers for in-depth studies of social processes over time. An important feature of software for the analysis of structural equation models is their facility to deal with a wide class of models for the analysis of latent variables (LVs). In the social sciences, and increasingly in biomedical and public health research, LV models have become an indispensible statistical tool. There are basically three major reasons for the utility of LV models. First, this kind of model can summarize information contained in many response variables by a few latent variables. Second, when properly specified, a LV model can minimize the biasing effects of errors of measurement in estimating treatment effects. Third, LV models investigate effects between primary conceptual variables, rather than between any particular set of ordinary response variables. This means that a LV model is often viewed as more appropriate theoretically than is a simpler analysis with response variables only. In order to address concerns regarding the appropriate analyses of survey data, new features have been added to a number of SEM software packages. The present paper deals with the use of design weights to fit SEM models to continuous data with or without missing values with optional specification of stratum and/or cluster variables. It also deals with the issue of robust standard error estimation and the adjustment of the chi-square goodness of fit statistic.

Research on the longitudinal analysis of complex survey data with LISREL is supported by SBIR grant R43 AA014999-01 from NIAAA to Scientific Software International.

David Dunson
NIEHS
Department of Biostatistics
dunson1@niehs.nih.gov

Semiparametric Bayesian Latent Variable Methods

The incorporation of latent variables within hierarchical models has proven to be a useful approach for addressing a wide variety of statistical problems. For example, in the context of longitudinal data analysis, random effects are commonly incorporated to account for heterogeneity among subjects. A common criticism of latent variable analyses is possible sensitivity to model structure and distributional assumptions. To address the first concern, we propose methods for latent variable selection in the context of normally distributed random effects models. To relax distributional assumptions, we consider semiparametric Bayesian methods. Methods based on Dirichlet process priors are reviewed and new methods are proposed using generalized stick-breaking priors, which allow the unknown distributions to change in shape with predictors. These priors are very flexible and also facilitate posterior computation via Polya urn or blocked Gibbs sampling algorithms. Some applications of the approach are considered, including flexible modeling of covariate effects on a high-dimensional response variable. The methods are illustrated using simulated data examples and real data from a genotoxicology study.

Lloyd Edwards
University of North Carolina
Department of PBiostatistics
Lloyd_Edwards@unc.edu

Title and abstract not submitted

Guang Guo
University of North Carolina
Department of Sociology
gguo@email.unc.edu

Title and abstract not submitted

Mark Handcock
University of Washington
Department of Statistics
handcock@stat.washington.edu

Title and abstract not submitted

James Hardin
University of South Carolina
Department of Epidemiology & Biostatistics
jhardin@gwm.sc.edu

Measurement Error in Generalized Linear Models

Measurement error can be addressed within the context of generalized linear models using 3 different methods: regression calibration, instrumental variables, and simulation extrapolation. We will highlight these three methods illustrating the derivations, discussing construction of estimated variance matrices for the coefficients, and presenting computational output from an add-on product for the Stata software system. Time permitting, we will also discuss the application of the methods outside the context of GLMs.

James Heckman
University of Chicago
Department of Economics
jheckman@uchicago.edu

Instrumental Variables in Heterogeneous Response Models

No abstract submitted

Karl Joreskog
Uppsala University
Department of Information Science
karl.joreskog@dis.uu.se

Structural Equation Models: New Developments

No abstract submitted

Akihito Kamata
Florida State University
Department of Educational Psychology

kamata@coe.fsu.edu

Title and abstract not submitted

Frauke Kreuter
University of California, Los Angeles
Department of Statistics
fkreuter@stat.ucla.edu

Applications of Growth Mixture Modeling to Non-Normal Outcomes

No abstract submitted

Katherine Masyn
University of California, Los Angeles
kmasyn@ucla.edu

Multivariate Extensions in Discrete-Time Survival Analysis Using Latent Variables

No abstract submitted

Jerry Reiter
Duke University
Institute of Statistics & Decision Sciences
jerry@stat.duke.edu

Multiple Imputation for Missing Data in Surveys with Complex Sampling Designs

Model-based multiple imputation approaches, though used to handle nonresponse in surveys, seldom account for complex sampling design features, such as stratification and clustering. As a result, analyses of these multiply-imputed data sets may yield biased estimates from the design-based perspective. In this talk, I explore the extent of this bias using simple simulation studies. The simulations indicate that the bias can be severe when the design features are related to the survey variables of interest. The simulations also suggest that inferences based on imputation models that condition on irrelevant design features are conservative. These results lead to a useful prescription for imputers: the safest course of action is to include design variables in the specification of imputation models. The simulations suggest this can be done effectively and simply by including dummy variables for stratum or cluster effects in the imputation models. A more efficient approach is to use an hierarchical model where the effect of clustering is incorporated using random effects and the effect of stratification using fixed effects.

Thomas Richardson
University of Washington
Department of Statistics
tsr@stat.washington.edu

Statistical Challenges presented by Structural Equation Models

I will describe some of the challenges that Structural Equation Models pose, when viewed as statistical models, particularly in the areas of interpretation, likelihood inference, model selection

and testing. I will outline research in the field of graphical models, which attempts to address some of these problems.

Aline Sayer
University of Massachusetts, Amherst
Department of Psychology
sayer@psych.umass.edu

An Overview of Multilevel Models

No abstract submitted

Chris Skinner
University of Southampton
Department of Statistical Sciences
cjs@socsci.soton.ac.uk

Title and abstract not submitted

Mary Thompson
University of Waterloo
methompson@uwaterloo.ca

Title and abstract not submitted

Jane Zavisca
SAMSI
janez@samsi.info

Latent Class Analysis of Cultural Fragmentation in the U.S.

POSTER ABSTRACTS

Siva K. Balasubramanian and Songpol Kulviwat, Hofstra University
Southern Illinois University
Department of Marketing
siva@cba.siu.edu
mktszk@hofstra.edu

The Determinants of Online Purchase

Methodology
The paper investigates online shopping behavior with an emphasis on online search activities as predictors of online purchasing. Perceived benefit, perceived risk, capability, and opportunity variables are hypothesized to be antecedents of online information search that ultimately determines online purchase. Data were obtained from the 10th WWW User Survey conducted by Georgia Institute of Technology's Graphic, Visualization & Usability Center (GVU) with 381 usable respondents. The data were analyzed using a two-step SEM approach: measurement model (internal consistency, convergent and discriminant validity) and structural model (path coefficients). Finally, a number of overall fit measures were assessed (χ^2, GFI, AGFI, CFI, and RMSEA)

Our theoretical SEM model, suitably modified per Wald and Lagrange mulitiplier (LM) tests indicated a good fit. Results show that online search strongly predicts online purchase. Direct and indirect relationships of perceived benefit and perceived risk with online purchase were also evidenced. The proposed positive relationship between perceived benefit and online search was supported. Further, a significant relationship was observed between capability and online search. As expected, opportunity and online search were also related. However, a negative relationship between perceived risk and online search was not found in the present study. Understanding the factors determining online information search and online purchase is crucial for designing effective marketing communication and marketing strategy. Our results provide a better understanding of the interrelationships among the determinants of online information search that is valuable to both marketing practitioners and scholars.

Mark Boye and Sunny Kim, Ohio University
Pfizer, Inc.
WWOR Ann Arbor Development Site
mark.boye@pfizer.com

Background: The sponsor developed and included into the acute post-operative pain clinical development program two instruments: the modified Brief Pain Inventory Short Form (mBPI-sf) and the Opioid-related Symptom Distress Scale (OR-SDS). The mBPI-sf includes 6 Pain Intensity/Relief items and 8 Pain Interference with Function items. The mBPI-sf was adapted from the BPI-sf, which has been used in pain states other than acute pain. Validation of the BPI-sf has previously been conducted within the tradition psychometric/effect indicator framework.

Included in the OR-SDS are 10 adverse effects known to be associated with opioid medication; 3 indicators are used to assess each effect and include symptom frequency, severity, and level of bother. The OR-SDS was developed by the sponsor and has yet to undergo formal instrument evaluation.

Objective: To examine the rationale for specifying mBPI-sf and OR-SDS causal indicator versus effect indicator models.

Methods: Traditional effect indicator methods are used to construct a model for each of these instruments. Thereafter, each model is recast using a causal/formative indicator conceptualization.

Results/Conclusion: The model requirements and interpretations change when switching from an effect to a causal indicator model. Based on the clinical intent underlying the development of these models, the causal indicator specification provides the most meaningful analytic framework for both the mBPI-sf and OR-SDS. Future research for this program is outlined.

Stephen du Toit, Mathilda du Toit and Gerhard Mels
Scientific Software International
LISREL
sdutoit@ssicentral.com
mels@ssicentral.com

Fitting Models to Data Collected from Longitudinal Studies Based on Complex Survey Designs

There has been a growing interest in recent years in fitting models to data collected from longitudinal surveys that use complex sample designs. This interest reflects expansion in requirements by policy makers and researchers for in-depth studies of social processes over time.

Although structural equation modeling allows for a tremendous flexibility in modeling error structures, it is in general not straightforward to analyze nested data structures with it. This, on the other hand, is a strong point of multilevel modeling which is also more flexible than structural equation modeling when repeated measurement occasions vary between individuals. In order to address concerns regarding the appropriate analyses of survey data, new features have been added to a number of software packages such as SPSS 12, SAS 9.1, HLM 6, MPLUS 3, LISREL 8.70 and MLWIN. Some of these features are as follows:

- The use of design weights to fit SEM models to continuous data with or without missing values with optional specification of stratum and/or cluster variables. Correct parameter estimates and robust standard error estimates, using a Taylor linearization approach, are produced under single stage sampling.
- The analysis of hierarchical linear models with the option to include design weights on levels 1, 2 or 3 of the hierarchy. Robust standard error estimates are produced using a Taylor linearization procedure.
- The analysis of generalized linear models under complex sampling designs. Users can select from various sampling distributions, for example, the multinomial, Poisson and negative binomial as well as various link functions such as logit, probit and complementary log-log.
- Simulation of complex datasets with software packages such as SAS PROC SURVEYSELECT. Simulation studies enable a researcher to evaluate important estimation issues such as bias, coverage and goodness of fit statistics.

The authors will illustrate their contributions to the advancement of these methods as implemented in LISREL 8.70 for Windows.

Research on the longitudinal analysis of complex survey data with LISREL is supported by SBIR grant R43 AA014999-01 from NIAAA to Scientific Software International.

Subhashis Ghosal
North Carolina State University
Department of Statistics
ghosal@stat.ncsu.edu

Bridging Maximum Likelihood And Bayes, And Beyond

It is well known that posterior distributions are consistent if (i) Kullback-Leibler neighborhoods of the true distribution get positive probabilities; (ii) an exponentially consistent test exists for testing the true value of the parameter against the complement of neighborhoods. In a recent paper, Walker and Hjort (2001, JRSS B, 811--821) found that the pseudo posterior distribution defined by taking a weighted average of the square root of the likelihood according to a prior distribution as in the Bayes theorem, is consistent only under the prior positivity condition (i).

We consider a whole spectrum of pseudoposterior distributions constructed by averaging the likelihood to the power α with respect to the prior for each $\alpha>0$ and investigate the consequences of altering the posterior distribution on Bayes estimates, credibility intervals, Bernstein-von Mises phenomenon and convergence rates for various parametric and

nonparametric families. In particular, it will be seen that raising the likelihood to the power α, $0<\alpha<1$, does help the posterior converge at the usual rate in terms of the Hellinger distance without any size restriction on the model. Extensions to independent, non-identically distributed observations are also considered.

Eric Kaarsemaker
Radboud University Nijmegen
Nijmegen School of Management
e.kaarsemaker@nsm.kun.nl

Employee Ownership from a Strategic Human Resource Management Perspective

This PhD project is about the antecedents and consequences of employee ownership in firms (for example through broad-based stock (option) plans, like ESOPs), and ultimately the effects of employee ownership on firm performance. Employee ownership is approached as a human resource management (HRM) practice and as such it is embedded within the work system that consists of all the HRM practices. This is where a first type of "fit" pops up: internal fit of employee ownership with the other HRM practices/the work system. This is a fundamental problem, as the HRM practices are directly linked to the firm's employees and as such determine their behavior.

From a theoretical viewpoint, the existence of an "ownership work system" seems plausible. Within the strategic HRM literature on work systems, a solution to the theoretical problem of the choice of HRM practices and their allocation to internally consistent work systems has not been found yet. One possibility is the idea that certain configurations of work systems show an underlying logic, in which so-called core HRM practices serve as a stable anchor for employee identification with the firm. This logic is connected with the image or anthropology of the employee in the eyes of the employer, which ultimately determines the work system.
Based on the SHRM literature and the work systems that have been discussed, two logics can be distinguished: a control or hierarchical logic, and a commitment or market logic. These are related to a certain anthropology. In case of the control logic, this anthropology is rather negative: employees should be controlled. In case of the commitment logic, this anthropology is more positive: employees should be committed to the firm. These anthropologies could be labeled as outsider vs. insider and they are related to McGregor's (1960) X and Y factors, respectively. However, it is as yet unclear which are the core HRM practices within work systems that are built on these logics.

While theorizing on the antecedents and consequences of employee ownership as a HRM practice, it seemed to be plausible to assume that employee ownership is such a core HRM practice, but of work systems with a different type of logic. If employees are at the same time owners, this seems a bit paradoxical. It signals that employees are being approached as equals, hence not any longer as outsiders or insiders. This is also related to Rosen & Young's (1991) O (ownership) factor. It can be shown that work systems that have been built on this O factor as stable anchor are potentially high performance work systems: "ownership work systems."

An important assumption is that the work system should be consistent in its signals to the employees, in order to be the most effective. This is also related to the legitimacy of the work system in the eyes of the employees. Approaching employees as "equals who need to be controlled" will not contribute to for example organizational citizenship behaviors by those employees. Contradictory signals indicate lack of (internal) fit. As this is about the underlying latent dimensions of the work systems and the influence on employee behaviors (HRM

outcomes), some of which are also latent (commitment, motivation), with the former on the organizational level and the latter on the individual level, multilevel structural equation modeling would appear to be appropriate to analyze the data that I collected and am collecting.

Next to the internal fit of employee ownership within the work system, three other types of fit are distinguished: organizational, strategic, and environmental fit. Employee ownership is also embedded with the firm, its structure, and history. Is employee ownership as a HRM practice more appropriate for certain types of firms, or certain groups of employees than for others, and does it fit the long-term relationships of the firm with its employees? Does a work system that has been built around employee ownership better fit certain firm strategies, e.g. those strategies aimed at innovation? And are there any cross-cultural differences? Would configurations of work systems with employee ownership, as we see them for example in the US, be equally effective in for example the Netherlands? For these questions the same kind of argumentation will be used as with regard to the problem of internal fit.

Melissa A. Kenney, Kenneth H. Reckhow, George B. Arhonditsis, Sandra J. McBride
Duke University
Nicholas School of the Environment & Earth Sciences
m.kenney@duke.edu

Selecting Appropriate Eutrophication Criteria using Structural Equation Modeling

Eutrophication is the process, fueled by excess nitrogen and phosphorus, which causes problems such as anoxia, noxious algal blooms, and fish kills. These eutrophication symptoms result in noncompliance of a waterbody's water quality standards. Water quality standards are set using a qualitative designated use and a quantitative criterion. The *criterion* is a numeric measure that serves as a scientific surrogate for the designated use. The *designated use* is a narrative statement that describes the water quality goal. The water quality standards are designed to protect the designated use; however, they are indirectly measured and assessed using the water quality criterion. The use of an indirect measure to assess attainment of the water quality goal theoretically is not a problem; the problem may arise if the relationship between the designated use and the criterion was never rigorously tested to determine if the judgment used to set the criterion was appropriate to protect the designated use of a waterbody. To properly place emphasis on criterion selection statistically linked to the designated use, our research focuses on *prediction* of designated use attainment. Designated use attainment is determined by exploring the relationships between causal variables (nitrogen and phosphorus) and response variables (algae and turbidity) using expert elicitation, water quality data, and structural equation modeling (SEM). *Expert elicitation* is the systematic process of extracting subjective judgments provided by a recognized specialist. This method is used when there is no other means to collect the data necessary to answer a question. In our research, expert elicitation is used to define the narrative designated use, to establish what variables should be measured to explore designated use attainment, and to predict waterbody compliance. Expert elicitation is used to fill-in the translation gaps from our rich water quality data sets to designated use compliance. There are abundant cross-sectional water quality data sets that can be used to address the question of appropriate criteria to predict designated use attainment. We have data that cover extensive geographic regions, a variety of water quality parameters, different socioeconomic situations, and a range of ecological conditions. However, issues of non-normality and time series are important questions that must be addressed because of the nature of the data. Our research could provide an excellent case study to investigate important methodological questions including the development of robust methods for SEM and SEM with time series data. Water quality data are particularly well suited to explore these questions because they contain both extreme values (which provide

essential information regarding water quality degradation) and measurements that have temporal autocorrelation problems. Our research methods could additionally explore techniques such as the elicitation of a Bayesian prior for use in a SEM. The ultimate goal of this research is to establish a method that could be applied to set appropriate water quality criteria which are predictive of the designated use.

Irene R. R. Lu and D. Roland Thomas, Sprott School of Business, Carleton University, Canada
York University
School of Administrative Studies
rlu@sprott.carleton.ca

A Monte-Carlo Comparison of Conventional Two-Step Score Regression and Structural Equation Modeling: Examining Biases in the Latent Variable Regressions

The study of the relationships among latent variables is very common in social science research, where many researchers use a two-step approach, i.e., latent variable scores are first obtained by some method and then used directly in subsequent statistical analyses. In this study, scoring methods based on classical test theory (CTT) and item response theory (IRT) were examined. It was found that, irrespective of the scoring technique used, the two-step approach has the serious drawback of suffering from finite item bias caused by the measurement error in the scores.

The primary goal of the study was to empirically compare and quantify the bias in various approaches to latent regression analyses when the indicators are discrete. Besides the two-step methods, two approaches based on the application of discrete structural equation methodology were studied, referred to as simultaneous IRT-SEM and constrained IRT-SEM. A preliminary investigation of the popular PLS methodology was also carried out.

The Monte Carlo simulation results show that the biases in R^2 and regression coefficients obtained using IRT- and CTT-scores in regression are very similar in general, with the two-step IRT regression approach being slightly more robust than the CTT regression approach. Unfortunately, however, both of the two-step methods yield biases that are too large to be ignored, particularly for scales based on small numbers of items. Both simultaneous and constrained IRT-SEM approaches yield appreciably smaller bias than either of the two-step approaches. Preliminary results suggest that the parameter bias in PLS is very similar to that observed in the two-step approaches. The different bias properties of the two-step and IRT-SEM methods will be illustrated using data from Youth In Transition Survey (YITS) and National Work-Life Conflict Survey.

Recommendations are as follows: First, use IRT for scale design. Second, use the conventional two-step approach to latent variable analysis with caution. Third, do not choose the easy way out; consider IRT-SEM instead. Fourth, for a small sample size, use constrained IRT-SEM whenever the IRT item parameter values are available.

Tom Munk
University of North Carolina
School of Education
tmunk@email.unc.edu

Strong Variables Affecting Mathematics Achievement: Toward a Structural Equation Model

As increasing importance is attached to mathematics scores, the research community must work to understand the causes of these scores. The National Assessment of Educational Progress (NAEP) includes an extensive background survey of students, teachers, and principals that facilitates investigation of these causes. An online data tool allows easy analysis of the relationships between these variables and NAEP mathematics scale scores, including the moderating effects of powerful background variables such as parental education. All of the variables most strongly predictive of NAEP mathematics scores, were shown to be at least doubly determined. The relationships with math scores were generally attenuated when variance in parental education was limited, most dramatically at lower levels of parental education. The strength on the relationship is, in every case, enhanced by a tendency for children of parents with higher levels of education to have more advantageous values on the other strong variables. In particular, segregation and tracking serve to strengthen these variables. The proposed structural model, incorporates all of the strongest variables into five constructs, and highlights structural features of our educational system that lead to widening academic gaps. Specification of this model would provide a strong background for studies of the effectiveness of educational techniques such as those proposed by the National Council of Teachers of Mathematics.

Michael Pennell and David B. Dunson
NIEHS
Department of Biostatistics
pennell@niehs.nih.gov
dunson1@niehs.nih.gov

Many biomedical studies collect data on times of occurrence for a health event that can occur repeatedly, such as infection, hospitalization, recurrence of disease, or tumor onset. To analyze such data, it is necessary to account for within-subject dependency in the multiple event times. Motivated by data from studies of palpable tumors, this article proposes a dynamic frailty model and Bayesian semiparametric approach to inference. The widely used shared frailty proportional hazards model is generalized to allow subject-specific frailties to change dynamically with age while also accommodating non-proportional hazards. Parametric assumptions on the frailty distribution are avoided by using Dirichlet process priors for a shared frailty and for multiplicative innovations on this frailty. By centering the semiparametric model on a conditionally-conjugate dynamic gamma model, we facilitate posterior computation and lack of fit assessments of the parametric model. Our proposed method is demonstrated using data from a cancer chemoprevention study.

Edward Rigdon
Georgia State University
Department of Marketing
erigdon@gsu.edu

Structural Equation Modeling of Data Sampled from a Finite Population without Replacement

Drawing a fixed-size sample from a finite population without replacement induces dependence across observations. Implications for structural equation modeling include overly optimistic fit assessment, based on either the chi-square statistic or alternative fit indices, and inflated standard errors. Organizational and business marketing research are especially vulnerable. A correction is proposed.

Cheri Shakiban
University of St. Thomas

Department of Mathematics
c9shakiban@stthomas.edu

Classification of objects Using Latent Semantic Analysis

This poster provides details on the attempts of using Latent Semantic Analysis to categorize two or three-dimensional objects according to their Euclidean signature curves as an application in computer vision. After calculating the signature curves for several objects, we use use Latent Semantic Analysis to build correlation matrices, which reveal the similarities between the objects. We will provide the details on the method and the procedure used, along with the results.

Liqun Wang
University of Manitoba
Department of Statistics
liqun_wang@umanitoba.ca

Estimation Of Nonlinear Regression Models With Berkson-Type Measurement Errors

Berkson-type measurement errors arise frequently in epidemiological, environmental and health sciences. A typical example is the individual exposure to certain air pollutants which are measured at monitoring stations in a city. I study a general nonlinear regression model where the predictor variable is subject to Berkson measurement error. It is assumed that the measurement error has a general parametric distribution which is not necessarily normal, while the distributions of the unobserved predictor variable and the random error in the regression equation are nonparametric. I propose a minimum distance estimator based on the first two conditional moments of the response variable given the observed predictor variable. To overcome the possible computational difficulty of minimizing an objective function which involves multiple integrals, a simulation-based estimator is constructed. Consistency and asymptotic normality for both estimators are derived under fairly general regularity conditions.

V. GENOMES TO GLOBAL HEALTH: COMPUTATIONAL BIOLOGY OF INFECTIOUS DISEASE

A. Opening Workshop Program & Abstracts
September 19-22, 2004

Sunday – September 19, 2004: Tutorials
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

8:00-9:00 am Registration and Continental Breakfast

9:00-12:30 pm **Mathematics and Statistics Tutorials**

9:00-9:45 Statistics and Data Analysis
Tom Kepler, Duke University

9:45-10:30 Bioinformatics and Genomics
Tom Kepler, Duke University

10:30-11:00 Coffee Break

11:00-11:45 Epidemiology
Alun Lloyd, North Carolina State University

11:45-12:30 Mathematical Modeling & Simulation
Tom Kepler, Duke University

12:30-1:30 pm Lunch

1:30-5:00 pm **Biology Tutorials**

1:30-2:30 Molecular Evolution
Jeff Thorne, North Carolina State University

2:30-3:00 Coffee Break

3:00-4:00 Microbiology
Denise Kirschner, University of Michigan

4:00-5:00 Immunology
Lindsay Cowell, Duke University

Monday – September 20, 2004: Workshop
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

8:30-9:00 am Registration and Continental Breakfast

9:00-9:15 am Welcome and Introductions
James Berger, SAMSI
Denise Kirschner, University of Michigan
Lindsay Cowell, Duke University

9:15-9:45 am Opening Address
Tom Kepler, Duke University

SESSION 1: *Biostatistics, Epidemiology, and Public Health*
Chair: **Andrew Nobel**, University of North Carolina

9:45-10:45 am Markov Models for Characterizing the Development of HIV Resistance Mutations
Victor De Gruttola, Harvard University

10:45-11:00 am Coffee Break

11:00-12:00 pm Stochastic Models and Analytic Issues for Interventions Against Pandemic Influenza
James Heckman, University of Chicago

12:00-1:00 pm Lunch

1:00-2:00 pm From Immunome to Vaccine: Using Immunoinformatics to Design New Vaccines
Anne De Groot, Brown University

2:00-3:00 pm New Paradigm for Management of Tropical Diseases
Ayoade Oduola, World Health Organization

3:00-3:30 pm Coffee Break

SESSION 1 continued
Session Chair: **Tim Elston**, University of North Carolina

3:30-4:30 pm DISCUSSION: *Mathematical Challenges in Biostatistics and Public Health*

Discussion Leaders:
Alun Lloyd, North Carolina State University
Anthony So, Duke University

4:30-5:30 pm Introductions to Evening Posters
Denise Kirschner, University of Michigan

Tuesday – September 21, 2004: Workshop
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

8:30-9:00 am Registration and Continental Breakfast

SESSION 2: *Microbiology, Microbal Ecology & Evolution and Microbal Genomics*
Session Chair: **Andrew Nobel**, University of North Carolina

9:00-10:15 am Use of Repetitive DNA in Helicobacter Pylori for Persistence in Human Hosts
Martin Blaser, New York University

10:15-10:45 am Coffee Break

10:45-12:00 pm Predicting the Future Evolution of Human Influenza Viruses
Walter Fitch, University of California-Irvine

12:00-1:00 pm Lunch

SESSION 2 continued
Session Chair: **Scott Schmidler**, Duke University

1:00-2:00 pm Bioinformatic Methods for Virulence Gene Prediction in Microbal Pathogens
Joanna Fueyo, IBM

2:00-3:00 pm A Mathematical Model for Nutrient Depletion and Detachment in a Hetergeneous Biofilm
Jorge Velasco-Hernandez, Instituto Mexicano del Petroleo

3:00-3:30 pm Coffee Break

3:30-4:30 pm Modeling Long-Term HIV Dynamics and Antiviral Treatment with Consideration of Pharmacokinetics, Adherence and Drug Susceptibility
Hulin Wu, University of Rochester

4:30-5:30 pm PANEL DISCUSSION: *Mathematical Challenges in Microbiology*

Discussion Leader:
Denise Kirschner, University of Michigan

Wednesday – September 22, 2004: Workshop
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

8:30-9:00 am Continental Breakfast

SESSION 3: *Immunology*
Session Chair: **Alun Lloyd**, North Carolina State University

9:00 -10:00 am Monoclonal Anitbodies in the Treatment of Disease: Modeling How They Couple Target Cell to Killer Cells
Byron Goldstein, Los Alamos National Laboratory

10:00-11:00 am MHC Polymorphism and Peptide Diversity
Rob De Boer, Utrecht University

11:00-11:30 am Coffee Break

11:30-12:30 pm A Bioinformatics Approach to the Antigen Presentation and Processing Pathways
Can Kesmir, Utrecht University

12:30-1:30 pm Lunch

SESSION 3 continued
Session Chair: **Scott Schmidler**, Duke University

1:30-2:30 pm PANEL DISCUSSION: *Mathematical Challenges in Immunology and Systems Biomedicine*

Discussion Leaders:
Lindsay Cowell, Duke University
Denise Kirschner, University of Michigan

2:30-3:00 pm	Coffee Break
3:00-3:15 pm	Charge to the participants and final words **Tom Kepler**, Duke University
3:15-5:00 pm	Working Group Formation and Planning

Martin Blaser
New York University
School of Medicine
Martin.blaser@med.nyu.edu

Use of Repetitive DNA in Helicobacter Pylori for Persistence in Human Hosts

No abstract submitted

Rob De Boer
Utrecht University
Department of Ost – Pathogen
r.j.deboer@bio.uu.nl

MHC Polymorphism and Peptide Diversity

The genes encoding the major histocompatibility (MHC) molecules are among the most polymorphic genes known in vertebrates. Since MHC molecules play an important role in the induction of immune responses, this polymorphism is probably due to selection for increased protection of hosts against pathogens. In contrast to the large population diversity of MHC molecules, each individual expresses only
a limited number of different MHC molecules. This is widely believed to represent a trade-off between maximizing the detection of foreign antigens, and minimizing the loss of T cell clones during self tolerance induction in the thymus. Here we review three theoretical models that we have developed to study the diversity of MHC molecules, both at the individual and at the population level. We have found that thymic selection does not limit the individual MHC diversity. Expression of extra MHC types decreases the number of clones surviving negative selection, but increases the number of positively selected clones. The net effect is that the number of clones in the functional T cell repertoire would increase if the MHC diversity within an individual were to exceed its normal value.

It has been proposed that the large population diversity of the MHC is due to selection favouring MHC heterozygosity. Since MHC heterozygous individuals can present more peptides to the immune system, they are better protected against infections than MHC homozygous individuals. Using a population genetics model, we found however that this heterozygote advantage is insufficient to explain the large degree of MHC polymorphism found in nature. Only if all MHC alleles in the population were to confer unrealistically similar fitness contributions to their hosts, could heterozygote advantage account for an MHC polymorphism of more than ten alleles. Thus, additional selection pressures seem to be involved. Using a computer simulation model we found that frequency-dependent selection by host--pathogen coevolution provides such an additional selection pressure that can account for realistic polymorphisms of the MHC. The polymorphism of the MHC thus seems a result of host--pathogen coevolution, giving rise to a large population diversity despite the limited degree of MHC diversity within individuals.

Finally we studied whether the peptides of nine amino acids (9-mers) that are typically used in MHC class I antigen presentation are sufficiently unique for self:non-self discrimination. The human proteome contains 28783 proteins comprising a total of 10^7 distinct 9-mers, most of which (76%) occur only once in the human proteome. Enumerating all distinct 9-mers for a variety of viral and bacterial proteomes we found that the average overlap, i.e., the percentage of foreign 9-mers that are also a human self peptide, is about 0.2%.

Victor DeGruttola*, Andrea Foulkes**, David Loecke*
*Department of Biostatistics, Harvard School of Public Health **Department of Biostatistics, University of Pennsylvania
victor@sdac.harvard.edu

Markov Models for Characterizing the Development of HIV Resistance Mutations

Development of resistance of Human Immunodeficiency Virus Type 1 to antiretroviral therapies is a serious medical and public health concern. A wide variety of mutations have been identified that either singly or in combination reduce the susceptibility of the virus to available therapies. This presentation will describe methods for characterizing the genetic pathways that lead to high level drug resistance under selective drug pressure, as well as for estimating the rate which viral populations progress along these pathways. These methods can be used to determine whether the presence of certain mutations among drug-sensitive viruses predispose a patient under a particular treatment to develop patterns of mutation that confer high-level drug resistance. Our approach assumes that viral genotypes can be characterized as belonging to discrete states, defined by patterns of viral mutations; we considers two approaches to modeling the rates of transition between these states. The first approach treats the state at a given time point as known while the second treats this as a latent variable. We also consider incorporation of covariates, such as genetic diversity among clones of virus from a single patient, which may impact on the type of mutations that develop, as well as on the speed and the order of their occurrence. We apply our methods to genetic sequences of virus cloned from plasma of 170 patients who participated in three phase II clinical studies of efavirenz combination therapy (DMP 266-003, DMP 266-004 and DMP 266-005). Multiple viral clones are available from each plasma sample at each time of measurement, allowing for consideration of the effect of minority species on the evolution of the viral populations infecting patients; the availability of such information motivates the second analytic approach. The sequences can be found in the Stanford HIV Resistance Website.

Walter Fitch and Geoffrey Graham
University of California – Irvine
Department of Eco-Evo
wfitch@uci.edu

Predicting the Future Evolution of B-Type Human Influenza Virus

In 1999, [Bush et al., Science 286, 1921-1925 (1999)] we culminated a study in which we determined eighteen residues in the A-type influenza hemagglutinin that were under positive selection to change their amino acid. We also reconstructed a putative evolutionary tree which had the usual single long lineage (trunk) that shows all but one branch dying out in the course of four or five years. Assuming that the positively selected positions were changing under selection pressure to evade, or at least mitigate the effects of, the human immunological counterattack, we asked whether we might use t changes in these amino acid positions to predict from which group of influenza viruses that currently cause influenza, future epidemics will arise. We predicted

these groups for eleven consecutive years. Our predictions were correct in nine of the cases. This work now examines the B-type influenza to see if we get similar results. This could be important for helping to choose strain for inclusion in the vaccines against influenza.

Joanna Fueyo
IBM
Information Based Medicine
jfueyo@us.ibm.com

Bioinformatic Methods for Virulence Gene Prediction in Microbial Pathogens

No abstract submitted

Byron Goldstein
Los Alamos National Laboratory
Theoretical Biology & Biophysics
bxg@lanl.gov

Monoclonal Antibodies in the Treatment of Disease: Modeling How They Couple Target Cells to Killer

One emerging strategy in drug development is to take advantage of immune effector mechanisms to destroy cancer cells or over-reactive immune cells. The drug, usually a monoclonal antibody, is designed to couple the target cells to cells of the immune system that express Fc receptors on their surface and mediate cell killing. As an example we look at in vitro experiments that study how a drug (Alefacept) used in the treatment of psoriasis and psoriatic arthritis mediates the elimination of a subset of T cells that drive the autoimmune disease. We develop a model to predict the concentration range of the drug over which it couples T cells to killer cells and we use the model to analyze the experiments. The model reveals what properties of the drug, target cell and killer cell determine the lowest concentration of drug that will be effective.

M. Elizabeth Halloran
Emory University
Department of Biostatistics
mehallo@sph.emory.edu

Stochastic Models and Analytic Issues for Interventions against Pandemic Influenza

We present an overview of topics related to stochastic modeling of and evaluation of interventions for pandemic influenza.

Can Kesmir
Utrecht University
Department of Biology
c.kesmir@bio.uu.nl

A Bioinformatics Approach to the Antigen Presentation and Processing Pathways

In the recent years we have been developing a number of tools to predict the specificity of several steps involved in the antigen presentation and processing pathways. I will start my talk by giving an overview of the the tools developed by us and the others. While developing these tools, we

learned a lot about the biology of these pathways. For example, we now can show that the specificity of the enzymes/molecules involved in the pathways have co-evolved to optimize the antigen presentation.

In the second part of my talk, I will use HIV as a model pathogen to demonstrate how we study the interactions between hosts and pathogens using these bioinformatics tools.

Aoyade Oduola
World Health Organization
Basic & Strategic Research
oduolaa@who.int

New Paradigm for Management of Tropical Diseases

No abstract submitted

Jorge Velasco-Hernandez
Instituto Mexicano del Petróleo
Department: Matemáticas Aplicadas y Computación
velascoj@imp.mx

A Mathematical Model for Nutrient Depletion and Detachment in a Heterogeneous Biofilm

No abstract submitted

Hulin Wu
University of Rochester
Biostatistics and Computational Biology
hwu@bst.rochester.edu

Modeling Long-Term HIV Dynamics and Antiviral Treatment with Consideration of Pharmacokinetics, Adherence and Drug Susceptibility

No abstract submitted

POSTER ABSTRACTS

Thomas Carr
Southern Methodist University
Department of Mathematics
tcarr@smu.edu

Migration-with-Delay Coupled Patch Models for Epidemics

Patch models are used to describe disease dynamics in population centers and infectious transport by vector migration couples the populations. We consider two populations coupled by migration, while also considering the effect of delay to examine disease outbreaks. The delayed-migration coupling between the populations induces time-oscillating outbreaks where there was only a steady-state. We show how the migration-rate and delay-time affect the dynamics of the epidemics.

Kathryn Chaloner and Cong Han, TAP Pharmaceutical Products, Inc., IL
University of Iowa
Department of Biostatistics
kathryn-chaloner@uiowa.edu

Bayesian Analysis and Design for a Non-linear Mixed Effects Model of HIV Dynamics

Non-linear mixed-effects models arise from mathematical and biological models of the dynamics of HIV replication in the presence of anti-viral therapy. The derivation of one such model, for protease inhibitor therapy, will be illustrated. A data set from a 1996 paper by Perelson and colleagues (Science 271:1582-1586) will be used to demonstrate how a Bayesian analysis can be used to improve upon other methods for estimation. The prior distribution is based on the literature prior to 1996. A Markov chain Monte Carlo algorithm is implemented to estimate the posterior distribution of the rate of decay of the free HIV virus and the rate of decay of the virus producing cells. The same data set will be used to illustrate the problem of experimental design for such experiments and a case study will be presented comparing candidate designs.

Some of this work is in collaboration with Alan Perelson, Los Alamos National Laboratories. Additional material can be found in:

1. Han C, Chaloner K, Perelson AS (2002). Bayesian analysis of a population HIV dynamic model. In Case Studies in Bayesian Statistics VI, C Gatsonis et al eds, Spinger-Verlag, 223-237.
2. Han C, Chaloner K (2004). Bayesian experimental design for nonlinear mixed-effects models with application to HIV dynamics. Biometrics, 60, 25-33.

Min He
Duke University
Department of Biostatistics & Bioinformatics
min.he@duke.edu

WeSpA: Web-accessible Spectratype Analysis: Data Management, Statistical Analysis and Visualization

Summary: WeSpA, a Web-accessible system for the management, visualization and statistical analysis of T-cell receptor and immunoglobulin spectratype data, was developed. Users upload data from their spectratype analyzer to WeSpA, which saves the raw data and user-defined and user-supplied supplementary covariates to a secure database. The analysis engine performs several data analyses, providing estimated relative frequencies, and summary statistics. The visualization engine presents analyzed histogram results in a Java applet and an image. Specialized statistical tools, developed in our group for hypothesis testing and modeling for multiple spectratypes, is also available through the WeSpA interface.

Availability: The service is accessible at no cost for academic users via web-interface at https://cbcb.duke.edu/weSpA/. Additional technical support and specialized statistical analysis and consultation are available by arrangement with the authors.

Babak Pourbohloul a,b, Lauren A. Meyers c,d
University of British Columbia
Centre for Disease Control
babak.pourbohloul@bccdc.ca

Respiratory-borne Outbreaks in Populations: Contact Networks and the Spread of Disease

A large class of infectious diseases spread through direct person-to-person contact. Respiratory-borne diseases like influenza, tuberculosis and SARS, spread through the exchange of respiratory droplets between people in close physical proximity to each other. The patterns of these contacts tend to be highly heterogeneous. Explicit models of the patterns of contact among individuals in a community, contact network models, underlie a powerful approach to predicting and controlling the spread of such infectious disease. Effective control of respiratory infectious diseases requires quantitative comparisons of quarantine, infection control precautions, case identification and isolation, and immunization interventions. We use contact network epidemiology to predict the impact of various control policies for both a mildly contagious disease such as SARS and a more highly contagious disease such as smallpox. The success of an intervention depends on the transmissibility of the disease and the contact pattern among people within a community. We illustrate that contact network epidemiology can provide detailed and valuable insight into the fate and control of an outbreak. Integrating these tools into public health decision-making should facilitate more rational strategies for managing newly emerging diseases, bioterrorism and pandemic influenza in situations where empirical data are not yet available to guide decision making.

a University of British Columbia Centre for Disease Control, 655 West 12th Avenue, Vancouver, British Columbia V5Z 4R4, Canada

b Department of Health Care & Epidemiology, Faculty of Medicine, University of British Columbia, 5804 Fairview Avenue, Vancouver, BC, V6T 1Z3, Canada

c Section of Integrative Biology and Institute for Cellular and Molecular Biology, University of Texas at Austin, 1 University Station C0930, Austin, Texas 78712, USA

d Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico, USA

Ali Khoujmane
Texas Tech University
Department of Mathematics & Statistics
akhoujma@math.ttu.edu

Improving Regression Function Estimators

This research is concerned with estimation problems regarding nonparametric regression functions that are not necessarily directly observed. Practical examples are abundant; for instance one may want to infer about the weight distribution of a cable of which only the shape is known. In general one may think of an input-output system, where one wants to recover an unknown parameter of the input. At least two measurement designs can be employed: random design, where the points at which the output function is observed are chosen according to a random mechanism; or deterministic design where these points are chosen essentially error free by the observer (for instance equally distant points in the unit interval). The random design model leads statistically to independent and identically distributed observations. This is no longer true for the deterministic design where the data are independents but not identically distributed. Therefore the latter situation is mathematically somewhat harder to deal with than the former, and most of the results in the literature, whenever available at all in this rather complicated model, are usually formulated and proved for the independent and identically distributed case. The parameter of particular interest to us is a linear functional of the input, like for instance a Fourier coefficient in

an on expansion of this function. The traditional estimator has an exact and asymptotic variance that both can be improved when a suitable modification is applied. This may lead to improvement of the entire regression function.

Abel Rodriguez
Duke University
Institute of Statistics & Decision Sciences
abel@stat.duke.edu

Bayesian Structural Analysis of Proteins (joint with Scott Schmidler, ISDS)

Understanding the 3D structure of proteins is a key issue in molecular biology since their functionality depends mainly on its folding. Also, the 3D structure is more conserved in evolution than sequence, providing for a better classification mechanism. In this work we pretend to develop a Bayesian method to perform the structural alignment of proteins, both for the order-preserving and non order-preserving case. Among other advantages, our method provides with a straightforward mechanism to measure the significance of the matches and to explore multiple alternative alignments. In this poster we present the current state of our research as well as the challenges we face.

Bayesian Selection of Haplotypes Predictive of DNA Damage and Repair (joint with David Dunson, NIEHS)

Chemical insults may be more likely to result in long term adverse outcomes, such as cancer, for individuals having polymorphisms in genes involved in base excision and repair. Our goal is to select haplotypes that predict baseline DNA damage, susceptibility to induced damage, and rate of DNA repair. The number of possible haplotypes of candidate genes can be large and baseline damage, susceptibility, and rate of repair cannot be measured directly. Instead, we measure these latent traits indirectly by obtaining surrogates of the frequency of DNA strand breaks before and after exposure to a genotoxic agent for individual cells. The distribution of the surrogates tends to vary substantially for different individuals and follow-up times, and standard hierarchical models are not appropriate. Instead, we propose a finite mixture of normals approach in which the mixture weights are modeled using a hierarchical latent factor structure. Bayesian methods are developed for selection of the factors to be included and predictors of these factors. These methods rely on mixture priors and stochastic search variable selection algorithms. Using data from a recent study, we show that individuals tend to vary substantially in rate of DNA repair, and select genes predicting the repair rate.

Abby Todd
University of North Carolina
Department of Mathematics
atodd@email.unc.edu

Modeling the Interruption of Immune Processes

Certain drugs have the unwanted side effect of suppressing the immune system while some drugs are used specifically for their immunosuppressant properties, particularly in the treatment of auto immune disorders such as chronic asthma and rheumatoid arthritis. The exact location of where immunosuppressants interrupt communication or signaling within the immune system greatly determines the effectiveness of an immune response. High levels of an immune response are

undesirable due to possible harm to the host while low levels of an immune response may be dangerous as the host system is unable to protect itself from harmful foreign invasion. This project explores the development of a mathematical model reflecting the effects of an immunosuppressant on the host system at a transcriptome level.

Cun-Quan Zhang
West Virginia University
Department of Mathematics
cqzhang@mathwvu.edu

Defining and Detection of Communities -- a New Algorithm for Clustering and Network Partition

Let $G = (V, E)$ be a graph and H be a subgraph of G. The dynamic density of H is the greatest integer k that

$$\min_{P}\left\{\frac{|E(H/P)|}{|V(H/P)|-1}\right\} > k$$

where the minimum is taken over all possible partitions P of the vertex set of H, and H/P is the graph obtained from H by contracting each part of P as a single vertex. And, as a default, a single vertex is of dynamic density k for any integer k.
A subgraph H of G is a k-level community if H is a maximal subgraph of G with dynamic density at least k.

We have found a polynomial algorithm for finding all k-level communities of a graph G.

Joint work with Y.-B. Ou, Dept. Statistics, WVU and B. Yuan, Dept. Biomedical Informatics, OSU. US patent application pending.

Peiying Zuo
University of North Carolina
Department of Mathematics
pyzuo@email.unc.edu

Modeling the Extracellular Nucleotide and Nucleoside Metabolism

Nucleotide metabolism plays a critical role in controlling mucus transport in the airway surface of the human lung. For example ion transport, ASL volume, ciliary beat frequency and mucin secretion are all regulated by specific cell surface receptors, such as P2Y2_R and A2b, that respond to changes in ATP and adenosine concentrations. We have constructed a mathematical model of nucleotide metabolism that consists of a set of rate equations based on the (1) the Kai's and Vmax's of the enzymes identified on human airway surfaces that metabolize purine and pyrmidine nucleosides; (2) rates of ATP/UTP release by airway epithelia; and (3) rates of purine/pyrmidine uptake by human airway epithelia. The capacity of the model to predict the measured concentration of purine/pyrmidine nucleotides/nucleosides on human airway surface was tested and the predicted patterns compared well with the patterns measured with 3H-ATP and etheno-derivatization techniques. A mathematical analysis of the model was performed to generate testable hypotheses that can be confirmed experimentally. Future directions for this work are to expand the model to include ion fluxes and water transport across the cell membrane

regulated by the purinergic system. The long term goal of this project is to develop a quantitative understanding of the biochemical mechanisms that regulate mucus transport.

B. *Mid-Term Workshop on Mathematical Modeling of Infectious Diseases Program*
January 31-February 1, 2005

Monday – January 31, 2005
NISS-SAMSI Building
Room 104

9:00-9:30 am	Registration and Continental Breakfast.
9:30-10:15 am	Welcome and Discussion
10:15-10:45 am	Coffee Break
10:45-12:00 pm	Discussion
12:00-1:00 pm	Lunch
1:00-2:30 pm	Discussion
2:30-3:00 pm	Coffee Break
3:00-3:30 pm	Discussion
3:30-4:30 pm	Break and make way to MCNC for Distinguished Lecture
4:30-5:30 pm	***SAMSI Distinguished Lecture*** Alan Perelson, Los Alamos National Laboratory *"Modeling Viral Infections"*
5:30-6:00 pm	Reception in MCNC lobby

Tuesday – February 1, 2005
NISS-SAMSI Building
Room 104

8:30-9:00 am	Continental Breakfast
9:00-10:00 am	Discussion
10:00-10:30 am	Coffee Break
10:30-12:00 pm	Discussion
12:00-1:00 pm	Lunch
1:00-2:30 pm	Discussion

2:30-3:00 pm Coffee Break

3:00-4:00 pm Discussion

VI. DATA ASSIMILATION FOR GEOPHYSICAL SYSTEMS

A. Opening Workshop Program & Abstracts
January 23-26, 2005

Sunday – January 23, 2005
Radisson Hotel RTP, Room H (3rd floor)

11:30-1:00 pm Registration and Lunch

1:00-5:00 pm **Tutorial Session I**

1:00-2:30 Brief Introduction to Bayesian Statistics and Applications to Data Assimilation (part 1)
Christopher Wikle, University of Missouri

2:30-3:00 Coffee Break

3:00-4:00 Topics in Data Assimilation (tentative title) (part 1)
Robert N. Miller, Oregon State University

4:00-5:00 Brief Introduction to Bayesian Statistics and Applications to Data Assimilation (part 2)
Christopher Wikle, University of Missouri

Monday – Januray 24, 2005
Radisson Hotel RTP, Room H (3rd floor)

8:00-8:30 am Registration and Continental Breakfast

8:30-12:30 pm **Tutorial Session II**

8:30-10:00 Ensemble Filters for Geographical Data Assimilation (1)
Jeff Anderson, NCAR

10:00-10:30 **Coffee Break**

10:30-11:30 Topics in Data Assimilation (tentative title) (part 2)
Robert N. Miller

11:30-12:30 Ensemble Filters for Geographical Data Assimilation (2)
Jeff Anderson, NCAR

12:30-1:45 pm Lunch

1:45-2:00 pm Welcome

- **Jim Berger**, SAMSI
- **Chris Jones**, University of North Carolina

2:00-5:00 pm **Research Session**: Oceanic Data Assimilation
Co-Chairs: **Zhijin Li**, Jet Propulsion Laboratory
Susan Lozier, Duke University

2:00-2:30 **Carl Wunsch,** Massachusetts Institute of Technology

2:30-3:10 Discussion

3:10-3:30 Coffee Break

3:30-4:00 **Naomi Leonard,** Princeton University

4:00-4:40 Discussion

4:40-5:00 pm **Poster Session Preview**
Each poster presenter is invited to present a preview of their poster. Each preview should be no longer than two minutes and using only one overhead transparency slide.

6:30-8:30 pm **Poster Session and Reception at SAMSI**
Continuous shuttle service for the Poster Session will be provided by Carolina Livery. A shuttle will depart from the Radisson at 6:25pm and the last shuttle will leave SAMSI 8:35pm. Poster Presenters: A shuttle will leave at 5:45pm to bring you to SAMSI.

Tuesday – January 25, 2005
Radisson Hotel RTP, Room H (3rd floor)

8:15-8:30 am Registration and Continental Breakfast

8:30-10:50 pm **Research Session:** *Statistical Approaches*
Co-Chairs: **Amarjit Budhiraja**, University of North Carolina
Sujit Ghosh, North Carolina State University

8:30-9:00 **Doug Nychka,** NCAR

9:00-10:00 Discussion

10:00-10:30 Coffee Break

10:30-11:00 **Ian McKeague**, Columbia University

11:00-12:00 Discussion

12:00-1:00 pm Lunch

1:00-4:30 pm **Research Session:** Atmospheric Data Assimilation
Co-Chairs: **Leonard Smith**, London School of Economics
Chris Snyder, NCAR

1:00-1:30 **Craig Bishop**, Naval Research Laboratory Monterey

1:30-2:30 Discussion

2:30-3:00 **Ron Gelaro**, NASA Goddard Space Flight Center

3:00-4:00 Discussion

4:00-4:30 pm Coffee Break

4:30-5:30 pm **SAMSI Distinguished Lecture**
Chair: **Jim Berger**, SAMSI

"Data Assimilation and Ensemble Forecasting: Two Problems with the Solution?"
Eugenia Kalnay, University of Maryland

Wednesday – January 25, 2005
Radisson Hotel RTP, Room H (3rd floor)

8:00-8:30 am Registration and Continental Breakfast

8:30-9:40 am **Research Session:** Problems and Challenges
Co-Chairs: **Kayo Ide**, University of California, Los Angeles
Chris Jones, University of North Carolina

8:30-9:00 **Juan Restrepo,** University of Arizona

9:00-9:40 Discussion

9:40-10:00 am Coffee Break

10:00-11:30 am **Closing Lecture**
Co-Chairs: **Kayo Ide**, University of California, Los Angeles
Chris Jones, University of North Carolina

10:30-11:00 **Zoltan Toth,** National Centers for Environmental (NCEP)

11:00-11:30 Discussion

11:30-2:00 pm Working Groups and Lunch

11:30-11:45 Choice of Working Groups and Their Charge

11:45-1:00 Lunch and Working Group Meetings

1:00-2:00 Working Group Reports and Finalization

Jeffrey Anderson
National Center for Atmospheric Research
Data Assimilation Initiative
jla@ucar.edu

Ensemble Filters for Geophysical Data Assimilation: A Tutorial

This tutorial aims to present a simple, comprehensive introduction to a variety of ensemble filtering algorithms starting from Bayes rule. Initially, methods for dealing with a single observation of a single variable are derived in great detail. Next, ensemble methods for a single observed variable and a single unobserved variable are outlined. These simple methods extend transparently to large geophysical models with many variables, but require the addition of error correction algorithms. Finally, a global atmospheric data assimilation produced by an ensemble filter is described.

Craig H. Bishop
Naval Research Laboratory
bishop@nrlmry.navy.mil

Observational Network Design and Data Assimilation

Observational networks of geophysical systems comprise fixed measurements whose location and frequency is fixed and adaptive or targeted observations whose location and/or frequency can be varied. Techniques for optimizing the fixed and adaptive components of geophysical observing networks will be briefly discussed. It turns out that the questions one needs to accurately answer to optimize the observational network are very similar to those that must be answered to optimize data assimilation. In particular, one needs to be able to predict how the distribution of state estimation errors propagates through time and how current and future observations change the distribution. In the on-going adaptive sampling program run by the National Centers for Environmental Prediction (NCEP) called the Winter Storms Reconnaissance Program an Ensemble Transform Kalman Filter (ETKF) method is used to select flight paths for weather reconnaissance aircraft to improve forecasts of high impact winter weather. Difficulties associated with the current approach employed include (a) the routine network is not accounted for in the selection of regions where large forecast errors are likely, and (b) the ensemble based error covariance model used by the ETKF is inconsistent with that used in NCEP's data assimilation studies. Results of idealized Observation System Simulation Experiments (OSSEs) will be used to illustrate how these difficulties can significantly reduce the potential value of observations from weather reconnaissance aircraft.

Ron Gelaro
NASA Goddard Space Flight Center
Global Modeling and Assimilation Office
ron.gelaro@nasa.gov

Sensitivity Analysis in Atmospheric Data Assimilation

The number and types of atmospheric observations, especially from satellites, will increase dramatically over the next decade. It is unlikely that even our next generation data assimilation systems will be capable of accommodating more than a fraction of the available observations. There is thus a growing need to understand the sensitivity to observations in these systems and to assess the impact of observations on relevant measures of skill. Such information is critical, e.g., to the development of intelligent strategies for data selection and utilization.

While it is true in an average sense that observations improve the quality of analyses and forecasts, there are significant differences in the value added by individual observations or subsets of observations depending on their quality, distribution, and state of the atmosphere. Traditional methods for estimating observation impact, such as observing system experiments (OSEs) or observation system simulation experiments (OSSEs), provide some insight into these issues but are in general limited by their expense and broad-brush diagnostic value.

Mathematical adjoints derived from the complex equations comprising the data assimilation and forecast model systems have proven to be effective tools for estimating sensitivities in many contexts. With the adjoint of the analysis component of a data assimilation system, sensitivities of aspects of either forecasts or the analyses themselves can be efficiently estimated. These can be determined with respect to observational data, background fields or assimilation parameters, all computed simultaneously. This permits arbitrary aggregation of the sensitivities, e.g., by data type, data channel, data location, etc. It also allows for estimation of the impacts of any subset of data on standard forecast measures and has proven useful for monitoring observation quality. The results so far show great promise, but also raise interesting questions, such as: Is it better to assimilate fewer observations with large individual sensitivity (impact), or assimilate large numbers of observations with less individual sensitivity? What effect does nonlinearity have on both the sensitivity estimate and the analysis solution itself?

SAMSI Distinguished Lecture

Eugenia Kalnay
University of Maryland
Department of Meteorology
ekalnay@atmos.umd.edu

Data Assimilation and Ensemble Forecasting: Two Problems with the Same Solution?

Until 1991, operational numerical weather prediction models utilized a single control forecast representing the best estimate of the state of the atmosphere at the initial time. In 1992, operational NWP models began to utilize ensembles of forecasts from slightly perturbed initial conditions. Such ensemble forecasts provide human forecasters with a range of possible solutions, whose average is generally more accurate than the single deterministic forecast, and whose spread gives information about the forecast errors. It also provides a quantitative basis for probabilistic forecasting.

The two essential problems in the design of an ensemble forecasting system are how to create effective initial perturbations, and how to handle model deficiencies, which make the ensemble forecast spread smaller than the forecast error. In this talk we present a brief historic review of ensemble forecasting and current methods to create perturbations. We point out that the promising approach of ensemble Square Root Kalman Filtering for data assimilation can solve, at the same time, the problems of obtaining optimal initial ensemble perturbations, and possibly

estimating the impact of model errors. We also discuss the problem of coupled systems with instabilities that have very different time scales.

Naomi Leonard
Princeton University
Department of Mechanical & Aerospace Engineering
naomi@princeton.edu

Adaptive Sampling in the Ocean

The Autonomous Ocean Sampling Network (AOSN) project and the Adaptive Sampling and Prediction (ASAP) project both aim to coordinate the sampling effort of high-frequency radars, ships, airplanes, satellites, buoys and underwater vehicles for the purpose of improving observation, nowcast and prediction of ocean processes. The first coordinated AOSN experiment was run in August 2003 in Monterey Bay, CA. Another coordinated experiment will be run as part of ASAP in August 2006 also in Monterey Bay, CA.

The objective is to learn how to deploy, direct and utilize autonomous vehicles and other mobile sensing platforms most efficiently to sample the ocean, assimilate the data into numerical models in real or near-real time and predict the current and future state with minimal error.

We view the optimal strategy for the mobile sensors as the global minimizer of a synoptic metric. The presence of currents in the domain modifies the optimal trajectories of the least powered vehicles. The adaptive sampling plan implies that the optimal or near optimal strategies must be revised as new data and forecasts become available. I will discuss the coordination of multiple assets to optimize sampling at the different physical scales of interest. I will also discuss the problem of approximating optimal trajectories in the presence of currents.

Ian McKeague
Columbia University
Department of Biostatistics
im2131@columbia.edu

Statistical Inversion of Oceanographic Tracer Data

This talk will discuss open issues and new research directions in the development of statistical methods to improve our understanding of deep ocean circulation. A Bayesian inversion approach will be discussed in detail, and illustrated by analyzing tracer data collected from oceanographic field work in the South Atlantic. Further research is needed to refine the combination of physical and statistical models that underlie the inversion, and to find better ways of assessing the uncertainties involved.

Robert N. Miller
Oregon State University
College of Oceanic and Atmospheric Sciences
miller@coas.oregonstate.edu

Topics in Data Assimilation

No abstract submitted

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

A Non-Gaussain Filter for Assimilating Data into a Numerical Model

Although the ensemble Kalman filter appears to work well with highly nonlinear systems it is also acknowledged that the basic assumption of multivariate normality that support the Kalman filter is dubious. We provide an extension of this linear filter to nonGaussian distributions. Although mixtures of normal distributions are used as a nonparametric starting point, the form of the estimator is very similar to a local linear regression estimator. This nonGaussian approach produces strikingly better assimilations for the Lorenz 1963, three variable system for longer time intervals. In addition, this model can be extended to high dimensional systems. The basic idea is that individual observations often have a local impact on a small number of components of the state vector. A nonGaussian update is applied to these components and then other components are updated using the more conventional ensemble Kalman filter. This hybrid approach is shown to give improved assimilations in the Lorenz 40 variable model and motivates some basic theoretical questions about when nonGaussian structure needs to be accounted for and when more elaborate methods will be effective.

Juan Restrepo
University of Arizona
Mathematics Department and Physics Department
restrepo@physics.arizona.edu

Accelerating Sampling via Nonlocal Filters

By casting stochastic optimal estimation of time series in path integral form one can apply analytical and computational techniques of equilibrium statistical mechanics. In particular, one can use standard or accelerated Monte Carlo methods for smoothing, filtering and/or prediction. The applicability and efficiency of the generalized (nonlocal) hybrid Monte Carlo and multigrid methods applied to optimal estimation will be discussed and compared computationally to more traditional methodologies. We will show that the nonlocal hybrid Monte Carlo is statistically accurate, we demonstrate a significant speedup compared with other strategies, thus making it a practical alternative to smoothing/filtering and data assimilation on problems with state vectors of fairly large dimensions, as well as a large total number of time steps. When the problem is linear it is clear how to build the nonlocal filter. The challenge at present is to find an optimal nonlinear filtering scheme for nonlinear problem, or at the very least, guiding principles that lead to the construction of nonlocal filters that create a significant speedup in the decorrelation times of the Monte Carlo process.

Zoltan Toth
National Centers for Environmental Prediction (NCEP)
Environmental Modeling Center
Zoltan.Toth@noaa.gov

Open Questions in Atmospheric Data Assimilation

This presentation will review some open scientific questions about atmospheric data assimilation that, if solved successfully, have a promise of improving the skill of operational weather forecasts. Connections with THe Observing system Research, and Predictability EXperiment (THORPEX), a long-term research program aimed at accelerating 1-14 day forecast improvements, will also be covered. Topics discussed will include the choice of variational vs. ensemble data assimilation, issues related to the assimilation of remotely sensed data, and challanges posed by the use of imperfect numerical models.

Christopher K. Wikle
University of Missouri
Department of Statistics
wikle@stat.missouri.edu

Brief Introduction to Bayesian Statistics and Applications to Data Assimilation

No abstract submitted

Carl Wunsch
Massachusetts Institute of Technology
Department of Earth, Atmospheric and Planetary Sciences
cwunsch@ocean.mit.edu

Global Oceanographic State Estimation---In Practice

Oceanographers have achieved the practical, if still computationally marginal, ability to do global scale decadal three-dimensional, time-dependent state estimation. Although the underlying principles are the same, the emphasis and techniques differ greatly from those usually labeled "data assimilation", a terminology I will reserve for the specific meteorological forecasting methodology. The oceanographic focus is primarily on "smoothing", not forecasting; the data types differ markedly; and the need for understanding of model/data errors is much more serious. Much of the mathematical focus on inverse problems and state estimation vanishes the instant that the problem is put onto a computer, becoming a problem indistinguishable from finite-dimensional least-squares, albeit one nearly overwhelming in its dimensionality. The size of the problem drives one towards focusing on computational load, automatic differentiation (AD) tools, the perhaps central problem of developing means for inferring and describing state estimate and control vector uncertainties; controllability and observability as a function of time of integration.

POSTER ABSTRACTS

Darren Drewry
Duke University
Department of Civil and Environmental Engineering
dtd2@duke.edu

Scalar Profile Assimilation Into a Multi-Layer Model of Canopy-Atmosphere Exchange: Toward Optimal Estimation of Net Ecosystem Exchange

A major focus of current research efforts in land-atmosphere exchange is the modeling and quantification of carbon dioxide, water vapor and energy fluxes between forested ecosystems and the atmosphere. Mechanistic models of the coupled physical and biological processes that

determine the magnitude of scalar fluxes have been developed and tested at many sites under a variety of environmental conditions. High frequency eddy covariance measurements of scalar fluxes are often used to test these canopy exchange models. Concurrent observations of carbon dioxide concentration profiles within the canopy airspace are frequently left unutilized in such modeling studies. We explore the assimilation of the information contained in these concentration profile measurements to constrain forward model estimates of net ecosystem exchange (NEE). A high-resolution, one-dimensional multi-layer model of canopy-atmosphere dynamics, including turbulent transport, vertical radiation interception, photosynthesis, stomatal conductance and respiration, is implemented and tested against eddy covariance flux measurements taken at the Duke Forest. A simple assimilation scheme is used to compare the results of forward model integrations both with and without the assimilated profile information. Implications are discussed for the optimal merger of data and models for the estimation of NEE.

Amal El Moghraby
Brown University and University of North Carolina-Chapel Hill
Department of Mathematics
amal@unc.edu

Extended Kalman Filter and a 2- tracer Rossby wave

True state observations are of the essence in data assimilation. And to avoid collecting redundant information from the velocity field using expensive equipment, the choice of tracer launch coordinates is of great significance. Here we study two passive tracers that are launched sufficiently close in our Eulerian field, close but in dynamically different regions. Although these tracers started initially close, they will gather different Lagrangian information because of their dissimilar fate. So the question becomes, when the data gathered from the two tracers is assimilated, will it recover similar Eulerian fields, or different? The tool we will use in our quest to answer this question is the Extended Kalman Filter, and the velocity field we will consider is the Rossby wave.

Joint work with Chris K.R.T. Jones.

Kristen Foley
North Carolina State University
Department of Statistics
kmfoley@unity.ncsu.edu

Statistical Data Assimilation for Coastal Ocean Prediction

The South-Eastern United States has had a tremendous residential and commercial investment in coastal areas during the past 10 years. Rapid development of the coast produces a stressed ecosystem and an exponential increase in human and property exposure to storm hazards. Our goal is to improve the prediction of coastal ocean processes that are associated with tropical storms and hurricanes. Specifically we are interested in predicting storm surge, the onshore rush of seawater caused by the high winds of a land-falling cyclone. Statistical methods are used to improve the inputs needed to initiate a numerical ocean model for the coastal Carolinas and Georgia. Data assimilation methods are used to combine output of the numerical model with observed data from water level stations along the coast. Results from Ensemble Kalman Filtering are presented with a focus on computational issues.

Joint work with Montserrat Fuentes and Lian Xie.

Josef Koller
Los Alamos National Laboratory
Space Science and Applications
jkoller@lanl.gov

Space Weather, Energetic Particle Storms, and Radiation Belt Modeling

The dynamic processes in the radiation belt, especially relativistic electron acceleration and transport, are not fully understood yet. Currently, neither data-based statistical models nor physics based models alone can capture all of the observed dynamics. Especially important parameters for describing the radiation belt evolution are the diffusion coefficients which are considered as a proxy for all the complicated physics that might be going on in the radiation belt.

We present here a preliminary study to test if a combination of a physics based model and data assimilation could recover the true diffusion coefficients. Based on a simple one-dimensional radial diffusion code and a Kalman filter, we explore how one could estimate the most likely diffusion coefficient. However, in reality this will be much more complicated since the diffusion coefficients are highly non-linear with a dependency on all three-dimensional coordinates.

Our goal for the future is to combine the extensive observational data of energetic electrons in the inner magnetosphere with a physical model of transport processes in the radiation belt using the diffusion code Salammbo. Initial data will come from the LANL GEO and GPS constellation and the Polar satellite. The results can be used to actively assess and validate acceleration and loss processes – the balance of which ultimately controls the dynamics of energetic particles. In addition, the results will prove beneficial in the design of future spacecraft missions. The study will provide the space industry with a tool to predict space weather in the radiation belts and to determine the threat to satellites, humans in space, and the systems that rely on them.

Francois Lekien, Spring Berman and Naomi Leonard
Princeton University
Department of Mechanical & Aerospace Engineering
lekien@princeton.edu, sberman@princeton.edu, naomi@princeton.edu

Adaptive Sampling in the Ocean and Coordinated Mobile Sensors

The Autonomous Ocean Sampling Network (AOSN) project and the Adaptive Sampling and Prediction (ASAP) project both aim to coordinate the sampling effort of high-frequency radars, ships, airplanes, satellites, buoys and underwater vehicles for the purpose of improving observation, nowcasting and prediction of ocean processes. The first coordinated AOSN experiment was run in August 2003 in Monterey Bay, CA. Another experiment in Monterey Bay will be run as part of ASAP in August 2006.

The objective is to learn how to deploy, direct and utilize autonomous vehicles and other mobile sensing platforms most efficiently to sample the ocean, assimilate the data into numerical models in real or near-real time and predict the current and future state with minimal error.

The optimal strategy for the mobile sensors is viewed as the global minimizer of a synoptic metric. The presence of currents in the domain modifies the optimal trajectories of the least powered vehicles and the sampling strategy must be revised as new data and forecasts become available.

While completing their initial sampling objectives, the vehicles collect data that is assimilated both in ocean models and coarser assimilation schemes. This provides a better understanding of the ocean dynamics and the correlation in the domain. As a result, the adaptive sampling scheme creates a complex control problem where the robustness also depends on the unknown flow field and ocean statistics.

Joint work with Derek Paley and Eddie Fiorelli.

Zhijin Li
Jet Propulsion Laboratory
zhijin.li@jpl.nasa.gov

A Three-Dimensional Variational Data Assimilation Scheme for the Regional Ocean Modeling System

A three dimensional variational data assimilation (3DVAR) scheme, which is associated with the Regional Ocean Modeling System (ROMS) and thus called ROMS-DAS, has been developed. ROMS-DAS emphasizes the capability for predicting coastal oceanic meso- to small-scale variations with temporal scales from hours to days. To cope with particular difficulties that arise from complex coastline and bottom topography, unbalanced dynamics and sparse observations, ROMS-DAS is incorporated with novel strategies.

These strategies include the implementation of three-dimensional anisotropic and inhomogeneous error correlations, application of weak dynamic constraints to the increments, and implementation of efficient and reliable minimization methods. During real-time experiments, ROMS-DAS has demonstrated satisfactory predictive skill, as well as high efficiency and reliability.

Joint work with Yi Chao (Jet Propulsion Laboratory), James C. McWilliams (UCLA), and Kayo Ide (UCLA).

Liyan Liu
University of North Carolina-Chapel Hill
Department of Mathematics
liuliyan@email.unc.edu

Lagrange Data Assimilation for Two Layer Point Vortex system

This paper provides a summary of Lagrange Data Assimilation results for two-layer point vortex system and parameter estimation of vortex positions and circulations for both one layer and two layer cases.

Robert N. Miller
Oregon State University
College of Oceanic and Atmospheric Sciences
miller@coas.oregonstate.edu

Generalized Inversion of Thermistor Chain Data and a Layer Model of Lake Kinneret

Lake Kinneret is Israel's only fresh water lake. During the stratified period from May to September, the lake exhibits low horizontal mode diurnal and semi-diurnal motions in response to the diurnal westerly winds. We applied a weak-constraint variational method to the thermistor chain data and a two-layer model in order to investigate the dynamical origins of the diurnal and semi-diurnal motions.

Joint work with Guillaume Vernieres and Ayal Anis.

Susan Minkoff
University of Maryland, Baltimore County
Department of Mathematics & Statistics
sminkoff@math.umbc.edu

Coupled Flow and Mechanics for Time-Lapse Seismic Modeling

To accurately model hydrocarbon recovery in compactible reservoirs, we must numerically simulate both flow and mechanical deformation. To monitor producing reservoirs over time, we take seismic snapshots of the field at periodic intervals. Combining these three types of physics (flow, deformation, and wave propagation) is a major undertaking but can reveal interesting nonlinear phenomena which are not readily apparent from single-process modeling. We will describe use of our staggered-in-time flow and mechanics simulator to time-lapse seismic studies of weak-formation reservoirs. We highlight adaptive time stepping, use of an approximate rock compressibility for improved convergence, and validation against fully-coupled simulation results.

Steve Naehr
Rice University
Department of Physics and Astronomy
naehr@rice.edu

Radiation Belt Data Assimilation with an Extended Kalman Filter

This poster explores the application of the extended Kalman filter to specify and forecast the distribution of relativistic electrons within the Earth's radiation belts. A data assimilation algorithm is derived for a simple radiation belt forecast model driven by radial diffusion. The model assimilates particle flux measurements from spacecraft near the magnetic equatorial plane, using an external magnetic field model to calculate adiabatic invariants and phase space density. The algorithm is tested in a series of virtual experiments, with data from an idealized geomagnetic storm simulation. Compared to assimilation by direct insertion of data, the extended Kalman filter more accurately reconstructs the global particle distribution from sparse observational data. The response of the filter to errors in the observations, magnetic field model, and forecast model is examined, in anticipation of application to more realistic models and data sets.

Chris Synder
National Center for Atmospheric Research
chriss@ucar.edu

Topics in ensemble filtering

I will outline work with collaborators on non-Gaussian filters based on mixtures, radar data assimilation with ensemble (Kalman) filters, and lateral boundary conditions for the forecast step of ensemble filters in limited-area models.

Radu Herbei
Florida State University
Department of Statistics

Statistical Inversion of Oceanographic Tracer Data

No abstract submitted

B. Mid-Term Workshop on Issues, Challenges and Interdisciplinary Perspectives Program and Abstracts
IPAM – February 22-26, 2005

Tuesday – February 22, 2005
IPAM Building, Room 1200

8:00-8:45 am	Registration and Continental Breakfast
8:45-9:00 am	Welcome and Opening Remarks
9:00-10:00 am	Waves, Information and Local Predictability **Joseph Tribbia**, National Center for Atmospheric Research
10:00-10:30 am	Coffee Break
10:30-11:30 am	An Ensemble Filtering Pot-pourri **Chris Snyder**, National Center for Atmospheric Research
11:30-1:30 pm	Lunch
1:30-2:30 pm	Dynamical predictability and initialization: a statistical prediction perspective **Richard Kleeman**, New York University
2:30-3:00 pm	Coffee Break
3:00-4:00 pm	Model Based Smoothing of Linear and Nonlinear Processes **Arthur Krener**, University of California-Davis
4:00-5:00 pm	Graphical Models for Sequential Data Modeling and Forecasting **Padhraic Smyth**, University of California-Irvine
5:00-6:30 pm	Poster Session

Wednesday – February 23, 2005
IPAM Building, Room 1200

8:00-9:00 am	Registration and Continental Breakfast
9:00-9:45 am	Entropy-Based Ensemble Prediction Schemes **Greg Eyink**, Johns Hopkins University
9:45-10:15 am	Coffee Break
10:15-11:00 am	Adaptive Observing **Carolyn A. Reynolds**, Naval Research Laboratory
11:00-11:45 am	Is Galerkin Projection The Right Scheme? **Edriss Titi**, University of California-Irvine and the Weizmann Institute of Science
11:45-1:30 pm	Lunch
1:30-2:15 pm	Indistinguishable States and Data Assimilation **Leonard A. Smith**, London School of Economics
2:15-2:45 pm	Coffee Break
2:45-4:00 pm	Open Discussion
4:00-4:45 pm	Diffusely-Perturbed Transport of Passive Scalars in Pseudoperiodic Flows **Richard Sowers**, University of Illinois, Urbana-Champaign

Thursday – February 24, 2005
IPAM Building, Room 1200

8:00-9:00 am	Registration and Continental Breakfast
9:00-9:45 am	Efficient assimilation of atmospheric data: a Local Ensemble Transform Kalman Filter **Brian Hunt**, University of Maryland
9:45-10:15 am	Coffee Break
10:15-11:00 am	Conditional Path Sampling of SDEs and Data Assimilation **Andrew Stuart**, University of Warwick
11:00-11:45 am	Adaptive Sampling, Data assimilation and Adaptive Modeling **Pierre F.J. Lermusiaux**, Harvard University
11:45-1:30 pm	Lunch
1:30-2:15 pm	Two topics in coupling probabilistic and dynamical systems approaches in analysis of complex system behavior **Igor Mezic**, University of California-Santa Barbara

2:15-2:45 pm	Coffee Break
2:45-3:45 pm	A View of Earth Observing Systems, Earth System Modeling, and Challenges for Data Assimilation **James McWilliams**, University of California-Los Angeles
3:45-5:00 pm	Open Discussion

Friday, February 25, 2005
IPAM Building, Room 1200

8:00-9:00 am	Registration and Continental Breakfast
9:00-10:00 am	Assessing a Local Ensemble Kalman Filter **Istvan Szunyogh**, University of Maryland
10:00-10:30 am	Coffee Break
10:30-11:30 am	Cross-scale interaction in seismic inversion **William W. Symes**, Rice University
11:30-1:30 pm	Lunch
1:30-2:30 pm	Parameter estimation: To what extent can data assimilation techniques correctly uncover stochasticity? **Jim Hansen**, Massachusetts Institute of Technology
2:30-3:00 pm	Coffee Break
3:00-4:00 pm	Some Aspects of Validation of Assimilation Algorithms **Oliver Talagrand**, Ecole Normale Superieure

Greg Eyink
Johns Hopkins University

Entropy-Based Ensemble Prediction Schemes

Statistical distributions of geophysical systems are often non-Gaussian (multi-modal, skewed, fat-tailed, etc.) However, many conventional ensemble prediction schemes based on Bayesian methods, such as the Ensemble Kalman Filter, assume normal statistics when conditioning upon observations and this can lead to poor performance. On the other hand, systematically converging ensemble schemes may require very large numbers of samples to properly represent non-Gaussian statistics, more than can be practically obtained for large-scale geophysical models. I shall introduce ensemble prediction methods that model system statistics as maximum-entropy/minimum information distributions, that can be quite non-Gaussian. The basic idea is to represent system statistics by a distribution which, subject to moment constraints, minimizes the relative entropy or Kullback-Liebler information with respect to a model of the prior statistics. When the latter is taken to be a mixture of Gaussians, this scheme gives a natural generalization of the Ensemble Kalman Filter for non-normal statistics. I'll discuss the method, its

implementation and its advantages and disadvantages. Finally, I'll present results for a simple stochastic PDE model of the ocean thermohaline circulation, which has bimodal statistics associated to two distinct stable states, and compare the maximum-entropy method in its cost and performance with other standard ensemble filtering methods.

Jim Hansen
Massachusetts Institute of Technology

Parameter estimation: To what extent can data assimilation techniques correctly uncover stochasticity?

Information about some of the ways in which models are both parametrically and structurally inadequate can be inferred by dynamically altering model parameters using a data assimilation framework. Instead of (or in addition to) altering model states to minimize a model-data misfit, one alters model parameters during the minimization process. Structural model inadequacy implies that one should not search for a global ``best'' set of parameter values, but rather allow the parameter values to change as a function of state; different parameter values will be needed to compensate for the state-dependent variations of realistic model inadequacy. Care must be taken when interpreting results. It is shown that when the system of interest is stochastic, it is {\it{impossible}} to uncover the correct form of stochasticity even when the model is taken to be stochastic. The interpretation errors are especially egregious when the model is deterministic. Of course, this need not imply that the resulting parameter estimatates are not useful. The parameter estimation approach can be used both to attempt to quantitatively identify model inadequacy (with an eye towards model improvement), and to develop strategies for accounting for model inadequacy (with an eye towards forecast improvement).

Brian Hunt
University of Maryland

Efficient assimilation of atmospheric data: a Local Ensemble Transform Kalman Filter

Ensemble Kalman filters work within the iterative Bayesian framework of the classical Kalman filter: assimilate new data by statistically interpolating between it and a prior (background) estimate of the system state, then use the result to forecast the background state for the next assimilation cycle. They allow for nonlinear model dynamics by using an ensemble of forecasts (rather than a linearized model) to assess the background uncertainty, and the assimilation essentially determines what linear combination of ensemble states best fits the data. In practice, some modification to the Kalman filter formalism (including covariance localization and inflation) is necessary in order to compensate for the problems of small sample (ensemble) size, nonlinearity, and model error. I will describe an approach that attempts to solve the "best fit" problem in a manner that is both computationally efficient and effective for atmospheric models. Formally it combines elements of the Ensemble Transform Kalman Filter (Bishop et al. 2001) and the Local Ensemble Kalman Filter (Ott et al. 2004). The filter allows for nonlinear, nonlocal dependence of observations on model state variables, and is able to deal easily with asynchronous data: it can assimilate batched data accumulated over a period of time with very little loss of accuracy relative to assimilating data more frequently, as it arrives.

Richard Kleeman
New York University
Courant Institute

Dynamical predictability and initialization: a statistical prediction perspective

The prediction process can be viewed as the relaxation of the initial condition probability distribution toward the climatological or equilibrium distribution (the degree of disequilibrium measures the "usefulness" of the prediction). We use a realistic atmospheric model and very large ensembles to study this process in detail. A particular focus will be on the utilization of information theoretic functionals on these two distributions. We shall also discuss various measures of information flow and discuss in particular the information flow between the initial conditions and predictions. At the beginning of the presentation a personal perspective on challenges in model initialization and prediction from a mathematical viewpoint will be given.

Arthur Krener
University of California-Davis
Department of Mathematics

Model Based Smoothing of Linear and Nonlinear Processes

We introduce the problems of filtering and smoothing of data based on linear and nonlinear models. There are several approaches both deterministic and stochastic including the Kalman filter, the least squares filter and the min-max filter. We discuss the extension of these to the smoothing problem and the complications due to nonlinearities and/or boundary data.

Pierre F.J. Lermusiaux
Harvard University

Adaptive Sampling, Data assimilation and Adaptive Modeling

Data assimilation allows the efficient acquisition of multidisciplinary data via adaptive sampling and efficient selection of model dynamics via adaptive modeling. Quantitative adaptive sampling uses data-driven dynamical forecasts and their uncertainties to predict in space and time the observation system that is optimum for regional coverage, dynamical study and/or uncertainty reduction. Multidisciplinary adaptive sampling is especially challenging because of the multiple interdisciplinary correlations. Quantitative adaptive modeling uses model-data misfits and their uncertainties to evolve and select the model dynamics that are most suited to the rapidly evolving ocean dynamics. Both structural as well as parametric adaptation are possible. Conceptual issues and ongoing methodological developments, as well as computational and numerical considerations, will be outlined. Research with real ocean data is carried out with the Harvard Ocean Prediction System (HOPS) and Error Subspace Statistical Estimation (ESSE) data assimilation system. Real-time uncertainty forecasting, data assimilation, adaptive sampling, dynamical investigations, multi-model estimation and adaptive biogeochemical modeling will be illustrated for the Autonomous Ocean Sampling Network-II (AOSN-II) experiment in Monterey Bay. Environmental-acoustical uncertainty estimation and transfers, and coupled data assimilation for physical-acoustical-seabed predictions and inversions will be illustrated in the Middle Atlantic Bight shelfbreak front region (PRIMER).

Igor Mezic
University of California-Santa Barbara
Mechanical Engineering

Two topics in coupling probabilistic and dynamical systems approaches in analysis of complex system behavior

Large dynamical systems with complex behavior, such as those arising in climate and weather modeling have over the years been treated with a variety of approaches from both deterministic and probabilistic perspective, perhaps the most expansive of these being Lorenz' 1963 work that lead to a massive amounth of work both within the field and outside of it. In this talk we advocate an approach that stems from von Neumann's and Koopman's work on operator methods in classical mechanics, where the key object is not the phase-space representation of the system and the associated geometry (e.g. of the attractors), but an operator representing the system on the space of observables, called the Koopman operator. Two topics are treated within this context: The first is the issue of decomposition of dynamics into "natural modes". The so-called Proper Orthogonal Decomposition (POD) (or Karhunen-Loeve, Singular Value Decomposition) is a popular method of achieving this. We study the relationship of the spectral properties of the dynamics on the attractor of the original, high-dimensional dynamical system within the context of the spectral properties of aforementioned Koopman operator. We apply this theory to obtain a decomposition of the process. This allows us to extract the {\it almost periodic} part of the evolving process. The remainder of the process has continuous spectrum and can be decomposed using POD. We call this the Mixed Orthogonal Decomposition (MOD). The second is the issue of uncertainty of process outputs. We define a notion of uncertainty specifically designed for the fact that the process is described as a dynamical system. Specifically, a dynamical system that asymptotes to a stable equilibrium point from any initial condition has asymptotically a perfectly certain output. The notion of uncertainty that we propose measures deviations from that case. We discuss differences of this measure and other measures of uncertainty such as variance and entropy, and its connection with the properties of the Koopman operator.

Carolyn A. Reynolds
Naval Research Laboratory

Adaptive Observing

Effective observing network design is an important component of the data assimilation problem. The observing network is composed of both fixed and adaptive components, with the adaptive component usually being deployed with the goal of decreasing the forecast error for a particular event. The fact that we are dealing with a non-normal, time-dependent system adds considerable complexity to the problem. Evolved forecast errors don't necessarily look like the initial errors that spawned them, and the perturbations that grew rapidly yesterday are not necessarily the ones that will grow rapidly tomorrow. Because errors evolve structurally, and propagate faster than particular features (with the group speed, rather than the phase speed), model dynamics have been incorporated into adaptive observing techniques. Thus, adaptive methods have advanced beyond observing features of interest (such as when "Hurricane Hunters" observe in and around the hurricane itself), to include observations of upstream or precursor features. There are different methods for incorporating model dynamics into observing system design; including adjoint sensitivity-based techniques and ensemble transform Kalman filter techniques. A desirable capability of observing system design is the quantitative prediction of the impact of the hypothetical observing network on the forecast error. As such, adaptive observing techniques have been extended beyond consideration of the model dynamics only, to include the characteristics of the data assimilation system itself. In this presentation, aspects of observing network design, with the primary focus on adaptive observing, will be presented in the context of both simple and operational systems. The impact of limiting assumptions (such as linear perturbation growth), will be discussed.

Leonard A. Smith
London School of Economics

Indistinguishable States and Data Assimilation

The roles of observational uncertainty and model inadequacy are contrasted in the case of physical simulation models when our best forecast models are chaotic (deterministic models with exponential-on-average sensitivity to initial condition). In the perfect model scenario, the framework of indistinguishable states provides an (?effectively Bayesian?) algorithm for constructing accountable (ideal) probabilistic forecasts. Within the perfect model scenario, the model-class in-hand will admit a model trajectory which shadows the observations: specifically, a trajectory which is consistent with the observations given the observational noise model. Outside the perfect model scenario, it can be proven that the set of indistinguishable states is empty, suggesting that no algorithm exists which can provide accountable probability forecasts. Practical implications differ for weather-like forecast applications and climate modelling. Adaptive observations are considered in this context, and it is noted that state-estimation might be profitably distinguished from forecast initialisation. Open questions of data assimilation in climate modelling are also touched upon.

Padhraic Smyth
University of California-Irvine
Information and Computer Science

Graphical Models for Sequential Data Modeling and Forecasting

The first part of this talk will provide a tutorial overview of a general class of probabilistic models known as graphical models. Graphical models provide a general language for both (a) representing complex sets of dependency relationships among random variables, and (b) providing a systematic framework for computing quantities of interest with such models such as conditional expectations and forecasts. We will see how state-based models, that sequentially combine observations at time t with model-based predictions from time t-1 (including Kalman filters and hidden Markov models), fit naturally within this framework and computational algorithms for such models arise "automatically". In the second part of the talk we will look at more recent research developments in graphical models for sequential data, including development of models with richer representations and development of faster computational methods. Examples and illustrations from the geosciences will be used for motivation and illustration as appropriate.

Chris Snyder
National Center for Atmospheric Research

An ensemble filtering pot-pourri

It has long been speculated that a Kalman filter for numerical weather prediction (NWP) models would provide significant improvements in both the analyses (i.e., the initial conditions for the model) and forecasts. While a naive implementation of the Kalman filter is impossible with present computers, work over the last decade suggests that a Monte-Carlo approach, in which the required covariances are estimated from an ensemble of forecasts, may be both feasible and effective. This approach is usually known as the ensemble Kalman filter (EnKF). Surprisingly, the EnKF appears to work with ensembles of a few 10's of members even for systems, such as in

NWP, with dimension of 10^7 or more. I will discuss possible reasons for this surprising performance, along with other topics in ensemble filtering, including sampling error, non-Gaussian effects, and error in the forecast model.

Richard Sowers
University of Illinois at Urbana-Champaign

Diffusely-Perturbed Transport of Passive Scalars in Pseudoperiodic Flows

We consider the behavior of a passive scalar being transported by a 2-dimensional diffusively-perturbed pseudoperiodic flow. Pseudoperiodic flows, introduced by Arnold in 1991, have several physically-relevant types of behavior. They are ergodic in parts of the phase space, and they have periodic orbits in other parts (and homoclinic orbits in between). We are interested in small diffusive perturbations of such dynamics. In particular, we are interested in transitions between the periodic orbits and the ergodic orbits. We adapt the theory of stochastic averaging (in particular, some work of Freidlin) to study these questions when we look at long periods of time. We find that, in an appropriate sense, a Markov process on a graph asymptotically gives the important statistics.

Andrew Stuart
University of Warwick

Conditional Path Sampling of SDEs and Data Assimilation

The problem of data assimilation can be formulated as the sampling of solutions of time-dependent equations given noisy observations. We introduce a stochastic PDE based approach to sampling paths of differential equations, conditional on observations. The stochastic SPDEs are derived by generalising the Langevin MCMC method to infinite dimensions. Various applications are described, including sampling paths subject to two end-point conditions (bridges), nonlinear filter/smoothers and a toy model for data assimilation.

William W. Symes
Rice University

Cross-scale interaction in seismic inversion

The inverse problem of reflection seismology differs in many ways from data assimilation problems arising in meteorology and oceanography, but exhibits some intriguing similarities as well. Seismic inversion shares a model-based data-fitting goal with many data assimilation problems. Also, multiscale nonlinearities present both obstacles, theoretical and practical, and critical opportunities for teasing information about the Earth's structure from seismic observations. This talk will overview seismic reflection imaging as an inverse problem, and explore the consequences of intra-scale and cross-scale interactions between seismic waves and the Earth's mechanical heterogeneity.

Istvan Szunyogh
University of Maryland

Assessing a Local Ensemble Kalman Filter

The accuracy of the recently proposed Local Ensemble Kalman Filter (LEKF) data assimilation scheme is investigated on a state-of-the-art operational numerical weather prediction model using simulated and real observations. The model selected for this purpose is the T-62 horizontal- and 28-level vertical-resolution version of the Global Forecast System (GFS) of the National Centers for Environmental Prediction (NCEP). The performance of the data assimilation system is assessed for different configurations of the LEKF scheme. It is shown that a modest size (40-member) ensemble is sufficient to track the evolution of the atmospheric state with high accuracy. The analyses are extremely accurate in the mid-latitude storm track regions. The largest analysis errors, which are typically much smaller than the observational errors, occur where parameterized physical processes play important roles. Since these are also the regions where model errors are expected to be the largest, limitations of a real-data implementation of the ensemble based Kalman filter may be easily mistaken for model errors. In light of these results, the importance of testing the ensemble based Kalman filter data assimilation systems on simulated observations is stressed.

In collaboration with E. J. Kostelich, G. Gyarmati, D. J. Patil, B. R. Hunt, E. Kalnay, E. Ott, and J. Yorke

Oliver Talagrand
Ecole Normale Superieure

Some Aspects of Validation of Assimilation Algorithms

The theory of assimilation is now well established on the basis of Bayesian estimation. As such, assimilation requires the a priori specification, either explicit or implicit, of the probability distributions of the errors affecting the various data to be assimilated (i. e., observations proper, and assimilating model). Most present assimilation algorithms can be described as particular applications of Best Linear Unbiased Estimation, or BLUE, and require as such the specification of only the first- and second-order statistical moments of the errors (i. e., expectations and covariances). Even nonlinear algorithms, such as ensemble sequential assimilation, most often use only expectations and covariances of data errors. Objective assessment of the quality of an assimilation algorithm can only be performed by comparaison against unbiased and statistically independent data. However, the unbiasedness and the independence have to be assumed, and cannot be objectively verified. Another type of validation can bear on the specification of the error statistics. Within assimilation itself, the only objective source of information on the data error lies in the overdeterminacy of the data, i. e. in the innovation vector. The only possible objective check of whether the error statistics have been properly specified is therefore comparison of the a priori specified, and a posteriori observed statistics of the innovation. That can be done on the innovation itself, or equivalently on the difference between the assimilated fields and the data that have been used in the assimilation. However, the number of degrees of freedom required for defining the error statistics is larger than the number of parameters required for actually performing the assimilation, with the consequence that the problem of estimating the error statistics from the innovation statististics is totally undetermined. Independent hypotheses, which cannot be validated within the assimilation procedure itself, are always necessary. A number of diagnostics, which bear directly or indirectly on the innovation, are presented and discussed. They include in particular the classical so-called c2-test. The link of those diagnostics with the informative content of the various of data is discussed, as well as a number of recent applications. In the linear case, an objective test of the optimality of an assimilation system can be based on the fact that optimality is equivalent to decorrelation between the innovation and the estimation error. That can be checked against independent data. Relatively few works have been performed so far along that line. These are presented and discussed.

Edriss Titi
University of California-Irvine and the Weizmann Institute of Science
Department of Mathematics

Is Galerkin Projection The Right Scheme?

In this talk, we will show that, to leading order, the correct approximative scheme is actually the post-processing Galerkin method, and not the standard Galerkin method as is commonly believed. The post-processing Galerkin scheme was introduced as an alternative to, and more efficient than, the nonlinear Galerkin method which is based on the theory of approximate inertial manifolds. Inspired by this observation, we raise the question about the validity of the Galerkin type projection methods which are based on energy norms. In particular, the Galerkin method based on the Proper Orthogonal Decomposition (Karhunen Leove expansion). Similar questions can be asked in the context of data assimilation schemes.

Joseph Tribbia
National Center for Atmospheric Research

Waves, Information and Local Predictability

This work will re-examine the question of local predictability which was brought to the fore by Lorenz and Emanuel in the context of the utility of adaptive observations. This study approaches the questions of local predictability and expected forecast skill using the tools of two formalisms which heretofore have not been widely incorporated into such problems: wave propagation and information theory. Because the questions of adaptive observations are most naturally posed in physical space and not in a modal phase space, wave propagation ideas are the most convenient context within which to ask such questions as: Where should one observe the atmosphere today to increase our confidence in a 3 day forecast over Los Angeles? An interpretation of the adaptive observing problem using the concepts of group velocities of barotropic and baroclinic waves will be given. Additionally, it will be asserted that the relative information metric is the most useful measure for posing and gauging the efficacy of adaptive observing strategies. Examples of the usefulness of these concepts will be given using simple atmospheric models.

VII. EDUCATION AND OUTREACH PROGRAM

A. SAMSI-CRSC Interdisciplinary Workshop for Undergraduates Program
May 31-June 4, 2004

Monday – May 31, 2004
Harrelson Hall, NCSU & SAMSI

9:15-9:30 am	Welcome **H.T. Banks**, Director of CRSC and Associate Director of SAMSI
9:30-10:30 am	Introduction and Background **H.T. Banks**
10:30-10:45 am	Coffee Break

10:45-11:45 am	Introduction to the *forward problem*: solving the harmonic oscillator equation **Stacey Ernstberger**
12:00 pm	Vans leave for SAMSI
12:45-1:30 pm	Lunch at SAMSI
1:30-4:00 pm	Presentation from SAMSI programs - Network Modeling for the Internet - Multiscale Model Development and Control Design - Data Mining and Machine Learning

Tuesday – June 1, 2004
Harrelson Hall, NC State University

9:00-10:15 am	MATLAB Tutorial with a brief intro to UNIX **Jon Ernstberger**
10:15-10:30 am	Coffee Break
10:30-12:00 pm	Linear Inverse Problems: a MATLAB Tutorial **Jimena Davis**
12:00-1:00 pm	Lunch
1:00-2:15 pm	Basic Statistical Concepts and Some Probability Essentials **Atina Brooks**
2:15-2:30 pm	Coffee Break
2:30-4:00 pm	Statistical View of Linear Least Squares: a MATLAB Tutorial **Emily Lada**
4:00-4:15 pm	Coffee Break
4:15-5:30 pm	Deterministic Models, Stochastic Models and Monte Carlo Integration **Peng Liu**

Wednesday – June 2, 2004
Harrelson Hall, NC State University

9:00-11:30 am	*Vibrating Beam* Data Collection at CRSC Laboratory **John David & Sarah Grove**
11:30-12:30 pm	Lunch
12:30-1:15 pm	Reflection on the data collection experience

	Karen Chiswell
1:15-1:30 pm	Coffee Break
1:30-3:00 pm	Solving the *Vibrating Beam* inverse problem **Lara Dick**
3:00-3:15 pm	Coffee Break
3:15-4:45 pm	Statistical Analysis for the *Vibrating Beam* Inverse Problem **Emily Lada**
4:45-5:00 pm	Coffee Break
5:00-5:30 pm	Help Session

Thursday – June 3, 2004
Harrelson Hall, NC State University

9:00-10:30 am	Teams work on their inverse problem
10:30-10:45 pm	Coffee Break
10:45-12:00 pm	Teamwork continues
12:00-1:00 pm	Lunch
1:00-2:30 pm	What could we do better? Alternative Models/Statistical Methods **Jon Ernstberger, Stacey Ernstberger, Sarah Grove, Emily Lada & Karen Chiswell**
2:30-3:00 pm	Coffee Break
3:00-4:30 pm	Teamwork continues and report preparation begins

Friday – June 4, 2004
Harrelson Hall, NC State University

9:30-10:30 am	Report Presentations and Discussion
10:30-10:45 pm	Coffee Break
10:45-11:45 am	Presentation & Discussion continues
11:45-12:00 pm	Closing Remarks & Workshop Evaluation **H.T. Banks**
12:00-1:00 pm	Lunch
1:00 pm	Participants depart NC State

B. *SAMSI-CRSC Industrial Mathematical and Statistical Modeling Workshop for Graduates Program & Abstracts* -- July 26-August 3, 2004

Sunday – July 25, 2004

Participants arrive at NC State University

Monday – July 26, 2004
Harrelson Hall, NC State University

8:45-12:30 pm	Workshop Overview, Welcome, Presentation of Problems **H.T. Banks**, Director of CRSC & Associate Director of SAMSI **Jean-Pierre Fouque**, Department of Mathematics
12:30-1:30 pm	Lunch
1:30-5:00 pm	Working Session

Tuesday – July 27, 2004
Harrelson Hall, NC State University

8:30-5:00 pm	Working Session

Wednesday – July 28, 2004
Harrelson Hall, NC State University

8:30-5:00 pm	Working Session

Thursday – July 29, 2004
Harrelson Hall, NC State University

8:30-5:00 pm	Working Session

Friday – July 30, 2004
Harrelson Hall, NC State University

8:30-12:00 pm	Working Session
12:00-1:00 pm	Lunch
1:00-5:00 pm	Tour of Centennial Campus & Math Lab

Saturday – July 31, 2004
Harrelson Hall, NC State University

8:30-11:00 am Working Session

11:00 am Free Afternoon

Sunday – August 1, 2004
Harrelson Hall, NC State University

Free Day

Monday – August 2, 2004
Harrelson Hall, NC State University

8:30-5:00 pm Working Session

Tuesday – August 3, 2004
Harrelson Hall, NC State University

8:30-12:30 pm Working Session

12:30-1:30 pm Lunch

1:30-5:00 pm Formal Presentation of Results

Wednesday – August 4, 2004

Departure of participants from NC State University

C. Two-Day Undergraduate Workshop on Computational Biology and Social Sciences Program -- February 18-19, 2005

Friday – February 18, 2005

NISS-SAMSI BUILDING, ROOM 104

9:15 am Carolina Livery Shuttle departs the hotel

9:30-10:00 am Arrival at SAMSI and Continental Breakfast

10:00-11:00 am Welcome and Introduction to SAMSI
Jim Berger, Director of SAMSI
H.T. Banks, Director of CRSC and Associate Director of SAMSI

11:00-11:15 am Coffee Break

11:15-12:15 pm Immunology and Infectious Diseases
Lindsay Cowell, Duke University

12:15-1:00 pm Lunch

1:00-2:20 pm	Modeling in the Social Sciences **Ken Bollen**, University of North Carolina **Jane Zavisca**, SAMSI
2:20-2:40 pm	Coffee Break
2:40-4:20 pm	Modeling in the Social Sciences -- *continued*
4:30 pm	Carolina Livery Shuttle back to the hotel

Saturday – February 19, 2005
NISS-SAMSI Building, Room 104

8:30 am	Carolina Livery Shuttle departs the hotel
8:45-9:15 am	Arrival at SAMSI and Breakfast
9:15-10:30 am	Multiscale Materials and Control **Ralph C. Smith**, Associate Director of CRSC **Emily Lada**, SAS Institute
10:30-10:45 am	Coffee Break
10:45-12:00 pm	Multiscale Materials and Control -- *continued*
12:00 pm	Adjournment and Departure to Hotel/Airport

VIII. WORKSHOPS CO-SPONSORED BY SAMSI

A. *PREP Workshop on Mathematics Meets Biology: Epidemics, Data Fitting and Chaos Program* -- May 26-28, 2004

Wednesday – May 26, 2004
University of Louisiana

7:30 am	Bus departs hotel to workshop
7:45-8:30 am	Continental Breakfast
8:30-8:45 am	Welcoming Remarks **Dean Clark**
8:45-9:00 am	Announcements **Azmy Ackleh**
9:00-9:50 am	Session on Epidemics **Professor Allen**

9:50-10:00 am	Coffee Break
10:00-10:50 am	Session continues
10:50-11:00 am	Coffee Break
11:00-11:50 am	Session continues
11:50-1:30 pm	Lunch
1:30-4:30 pm	Project on Epidemics **Professor Allen**
4:45 pm	Bus departs workshop for hotel

Thursday – May 27, 2004
University of Louisiana

8:00 am	Bus departs hotel to workshop
8:15-9:00 am	Continental Breakfast
9:00-9:50 am	Session on Data Fitting **Professor Banks**
9:50-10:00 am	Coffee Break
10:00-10:50 am	Session continues
10:50-11:00 am	Coffee Break
11:00-11:50 am	Session continues
11:50-1:30 pm	Lunch
1:30-4:30 pm	Project on Data Fitting **Professor Banks**
4:45 pm	Bus departs workshop for hotel

Friday – May 28, 2004
University of Louisiana

8:00 am	Bus departs hotel to workshop
8:15-9:00 am	Continental Breakfast
9:00-9:50 am	Session on Chaos **Professor Cushing**

9:50-10:00 am	Coffee Break
10:00-10:50 am	Session continues
10:50-11:00 am	Coffee Break
11:00-11:50 am	Session continues
11:50-1:30 pm	Lunch
1:30-4:30 pm	Project on Chaos **Professor Cushing**
4:45 pm	Bus departs workshop for hotel

B. *NPCDS-SAMSI Workshop on the Design and Analysis of Computer Experiments for Complex Systems Program and Abstracts* -- July 13-17, 2004

Wednesday – July 14, 2004
Banff Centre

8:30-8:45 am	Introductory Remarks
8:45-10:00 am	SESSION 1: Computer Models for the Automotive Industry
10:00-10:30 am	Coffee Break
10:30-12:00 pm	SESSION 1 (continued)
12:00-1:00 pm	Lunch
1:00-2:30 pm	SESSION 2: Combining Information in Computer Experiments
2:30-3:00 pm	Coffee Break
3:00-4:30 pm	SESSION 3: Variable Selection
4:30-5:15 pm	DISCUSSANT: **Hugh Chipman**, Acadia University

Thursday – July 15, 2004
Banff Centre

8:30-10:45 am	SESSION 4: New Approaches to Computer Integration
10:45-11:05 am	Coffee Break
11:05-12:35 pm	SESSION 5: Applications in Industry
12:35-1:35 pm	Lunch

1:35-5:00 pm	Extended Break
5:00-6:30 pm	SESSION 6: Computer Simulation in Complex Systems
6:30-7:15 pm	DISCUSSANT: **Randy Sitter**, Simon Fraser University

Friday – July 16, 2004
Banff Centre

8:30-10:45 am	SESSION 7: Design of Computer Experiments
10:45-11:05 am	Coffee Break
11:05-12:35 pm	SESSION 8: Applications and Experiences at Los Alamos
12:35-1:35 pm	Lunch
1:35-5:00 pm	Extended Break
5:00-6:30 pm	SESSION 9: Application with Functional Data
6:30-7:15 pm	DISCUSSANT: **Nick Hentgartner**, Los Alamos National Laboratory

Saturday – July 17, 2004
Banff Centre

9:00-10:00 am	Planning Meeting
10:00-10:30 am	Coffee Break
10:30-12:30 pm	Planning Metting (continued)
12:30 pm	Adjournment

Jim Berger
SAMSI

Validation with Functional Data

Often the output of a computer model is quite irregular functional data. Dealing with such data within the SAVE validation strategy can be done by using basis representation of the functions, and applying the methodology to each basis coefficient. An important pre-processing step is to 'register' each function, so that the basis elements for the field data and the computer model data are matched and can be individually compared. Implementation of this strategy is considered in an application involving functional data arising from road load dynamics of vehicles.

Andrew J. Booker

Computer Experiments at Boeing

Computer experiments are an integral part of scientific investigation on the aerospace industry. In this talk, I will give an overview of factor screening in high dimensions and function fitting in this setting.

John F. Brewster (Part I)
Robert G. McLeod (Part II)
University of Manitoba

Applications of Computer Experiments Involving Functional Data

The Institute of Industrial Mathematical Sciences at the University of Manitoba has recently become involved in a number of collaborative research projects that use computer experiments to investigate models for complex systems. Two of these projects will be described in the talk. The first deals with the flow of personnel through a human resource system. The second is an uncertainty and sensitivity analysis of an epidemiological model for the spread of an infectious disease. In both projects the response variable of interest is a time-dependent curve. The longitudinal (or "functional data analytic") nature of the output (and input) raises interesting questions about the design of the experiments and the analysis of the resulting data. Suggestions will be sought from participants at the workshop.

John A. Caffo
General Motors Research and Development

Industrial Perspective on Vehicle Road Load Prediction Math Model Validation

We discuss the importance of math model validation from an industrial perspective. While the vehicle development process (VDP) may be viewed from many perspectives, we consider it to be a series of decisions. In the absence of uncertainty, this series of decisions can, in principle, be posed as a very complex multi-dimensional problem. Decisions, however, are actions taken in the present to achieve an outcome in the future. Because it is impossible to predict the outcomes of these decisions with certainty, the characterization and management of uncertainty in engineering design is essential to the decision-making that is the core activity of the vehicle development process. One of the main uncertainties we deal with is the confidence we have in the math model results we use.

Gonzalo Garcia-Donato, University of Castilla-La Mancha
Jesus Palomo, SAMSI

Validation with Multiple Inputs

Often, in a given computer model validation problem, several outputs are simultaneously of interest. When multiple outputs are analyzed independently, it is usually the case that different (and sometime contradictory) strategies for tuning/calibrating the same model inputs are obtained. Hence, a model which deals with the multivariate nature of the output is needed. We propose and implement such a model in the road load problem, in which bi-variate functional data has been observed.

Dave Higdon
Los Alamos National Laboratory

A Framework for Uncertainty Quantification Combining Detailed Computer Simulations and Experimental Data

Inference regarding complex physical systems (e.g. subsurface aquifers, charged particle accelerators, nuclear weapons) is typically plagued by a lack of information available from relevant, experimental data. What data is available is usually limited and informs indirectly about the phenomena of interest. However, when the physical system is amenable to computer simulation, these simulations can be combined with experimental observations to give useful information regarding calibration parameters, prediction uncertainty, and model inadequacy. This talk will discuss an approach for carrying out such analyses.

An encompassing framework for carrying out such analyses has the potential to shed light on a number of important issues in simulation-based predictive investigations:

- combining information from multiple experiments
- uncertainty quantification for simulation-based predictions
- high dimensional calibration/baselining of model parameters
- assessment of the value of various types of experimental data
- assessment of discrepancy between simulation output and experimental data

The details of any analysis will likely vary with a given application. However, common aspects of any analysis are bound to include the following components:

- selection of input settings over which to carry out simulation runs
- sensitivity analysis – i.e. understanding which simulation inputs are impacting the simulation output
- response surface modeling of the simulation output – finding parsimonious models in very high dimensional settings
- constraining the range of possible simulation output with experimental data
- accounting for systematic discrepancies between the simulation output and experimental data.

This talk will touch on the above issues in the context of ongoing work at LANL.

Stella Karuri, University of Waterloo
William J. Welch, University of British Columbia

A Practical Bayesian Implementation for the Analysis of Computer Experiments

"Kriging" models are often used for the analysis of data from a computer model. They are data adaptive and can handle complex, non-linear relationships with a modest number of runs relative to the number of input variables. Such models involve hyper-parameters, which have to be estimated from the data. The uncertainty from estimating these parameters can be taken into account by a Bayesian approach, but the usual Bayesian problems arise: (1) specifying the prior and (2) the computational burden of numerical integration and sampling from posterior distributions. We demonstrate that some simple approximations can speed up these computation and work well relative to methods like MCMC.

Christiane Lemieux
University of Calgary
Department of Mathematics and Statistics

Randomized Quasi-Monte Carlo Methods for Computer Experiments

Quasi-Monte Carlo (QMC) methods can be thought of as a deterministic version of the Monte Carlo (MC) method for multivariate integration. They are based on highly uniform point sets, i.e. point sets constructed so that their induced empirical distribution is as close as possible to the uniform distribution. QMC methods have been shown to outperform the MC method in various settings, but one drawback of these methods for practical purposes is that their deterministic nature does not allow for error estimates. With randomized quasi-Monte Carlo (RQMC) methods, it is possible to construct error estimates just like in the MC method, while still preserving the high uniformity – and thus high precision – of QMC methods. In this talk, we'll present the main ideas behind RQMC methods, with a focus on how they can be used for computer experiments. Our general goal will be to make connections between different works done in this area, and discuss ideas for future research.

Crystal Linkletter
Simon Fraser University
Department of Statistics and Actuarial Science

Inert Column Variable Selection

In many situations, simulation of complex phenomena requires a large number of inputs and is computationally expensive. Identifying which inputs most impacts the system can be a critical step in the scientific endeavor so that these factors can be further investigated. In computer experiments, it is common to use a Gaussian spatial process to model the output of the simulator. In this article, we introduce a new, simple method for identifying active factors in computer screening experiments. The approach is Bayesian and only requires the generation of a new inert variable in the analysis. The posterior distribution of the inert factor is used as a reference distribution with which we assess the importance of the experiment factors. The methodology is demonstrated on simulated examples as well as an application in material science. This is a joint collaboration with Derek Bingham, David Higdon, Kenny Ye and Nick Hentgartner.

David Meese, University of Pennsylvania and University of California Berkeley
Derek Bingham, Simon Fraser University

Latin Hyper-Rectangle Sampling for Computer Experiments

Latin hypercube sampling is a popular method for evaluating the expectation of functions in computer experiments. However, when the expectation of interest is taken with respect to a non-uniform distribution, the usual transformation to the probability space causes relatively smooth functions to become extremely variable in areas of low probability. Consequently, the equal probability cells inherent in hypercube methods tend to sample an insufficient proportion of the total points in these areas. In this paper we introduce Latin hyper-rectangle sampling to address this problem. Latin hyper-rectangle sampling is a generalization of Latin hypercube sampling which allows for non-equal cell probabilities. A number of examples are given illustrating the improvement of the proposed methodology over Latin hypercube sampling with respect to the variance of the resulting estimators. Extensions to orthogonal-array based Latin hypercube sampling, stratified Latin hypercube sampling and scrambled nets are also described.

George Michailidis
University of Michigan

Developing Efficient Simulation Methodology for Complex Queuing Networks

Over the years a fairly rich literature has been developed for determining the throughput capacity of complex queuing systems (e.g. drift analysis, fluid models, sample path analysis). However, it is a much harder task to obtain analytical results for other performance metrics such as backlogs and delays of complex queuing models. Simulation can provide insight to the behavior of a complex system by identifying the response surface of several performance measures (e.g. delays and backlogs) as functions of various control parameters (e.g. control policies, distributions of inputs and service times, buffer sizes, etc). However, simulation of large systems are expensive both in terms of CPU time and use of available resources (e.g. processors). Thus, it is of paramount importance to carefully select the inputs of simulation in order to (i) adequately capture the underlying response surface of interest and (ii) minimize the required number of simulation runs (experiment trials). In this study, we present a methodological framework for designing efficient simulations for complex networks. Our approach works in a sequential manner and combines the methods of CART (Classification and Regression Trees) and the design of experiments. The results of the proposed methodology are illustrated on a model of a generalized switch, where analytical results for the maximum throughput policy have been obtained through drift analysis. This is joint work with Derek Bingham and Ying-Chao Hung.

Charles W. Nakhleh
Los Alamos National Laboratory

Night Thoughts of a Numerical Experimenter

All realms of science and engineering are increasingly reliant on numerical simulation to design highly complex and expensive experiments; to predict the results of experiments in a usually semi-qualitative way; and to extrapolate beyond the data obtained to predict the performance of engineered systems in different regimes. At the same time, the complexity of the simulations involved has increase enormously: more comprehensive models are continuously being incorporated into codes; all codes draw on physical databases, which are themselves often the result of simulations to be much more faithful to the actual physical setup.

Given these developments, large and important questions have arisen within several numerical modeling communities as to the validity of the simulations; the correctness of the models and numerical methods used in codes; and the impact of the uncertainties involved in the carious stages of computation on overall predictions made using the codes. At the same time there is a dire need for quantitative methods of discriminating between models, methods, and codes themselves given a (usually sparse) set of validation data.

In this talk, I will discuss two representative applications of carrying degrees of complexity: the simulation of particle beams by particle-in-cell (PIC) codes; and the simulation of turbulent jetting phenomena by continuum radiation hydrodynamic codes. I argue that there are mutually beneficial opportunities and challenges for both the statistical and scientific communities in dealing with (i) output from numerical simulations, (ii) assessing quantitatively how sparse experimental data constrain the modeling, (iii) assessing the degree of confirmation of one computational model over another, (iv) and making quantitative predictions (with uncertainties) using inaccurate simulators.

Bill Notz
Ohio State University

A Sequential Strategy for Constrained Optimization of Computer Experiments

I will discuss sequential strategies for constrained optimization of output from computer experiments when there are two outputs (responses), each of which is expensive or time consuming to compute. I will consider the situation in which there are two types of inputs; control variable and noise variable. The objective is to find values of the control variable that optimize the mean (over the distribution of the noise variables) of one response subject to a constraint on the mean of the other response. I will describe a sequential strategy to select the values of the inputs at which to observe the responses. The methodology is Bayesian; the prior takes the responses as draws from a Gaussian stochastic process. At each stage, the strategy determines which response to observe and at what set of inputs so as to maximize a posterior expected "improvement" over the current estimate of the optimum. This is joint work with Jeffrey Lehman, Tom Santner and Brian Williams.

Rui Paulo
University of Bristol

Validation of a CRASH Model

When the output of the computer model is a fairly regular function of time, the general SAVE model validation strategy can be successfully applied by sampling the function output at a number of points and considering time as just another input to the computer model.

Unfortunately, the number of points at which the functions are sampled is severely limited by the numerical calculations involved in fitting the statistical model. To address this issue, we take advantage of the particular form of the covariance structure and design sets in order to be able to express covariance matrices using Kronecker products and hence considerably speed up calculation and allow for a larger number of grid points. This is illustrated with an application involving vehicle crashes against a barrier and the associated velocity curves.

Another interesting feature of this application is the presence of multiple barrier types, some of which with limited data. We develop a hierarchical model approach in order to be able to 'borrow strength' across different barrier types and compare it with the competing approach of treating the type of barrier as an added input to the computer model.

Jason E. Pepin
Francois M. Hemez
Los Alamos National Laboratory

Uncertainty Quantification for Applied Engineering at LANL

Current and future Department of Energy (DOE) mission will increasingly rely on large-scale simulations for scientific guidance supporting issues of national importance. Test ban treaties have placed a tremendous burden on the nation laboratories for assuring stockpile certification, whose fundamental objective is to maintain a high confidence in the safety, reliability, and performance of the existing U.S. nuclear weapons stockpile.

This presentation overviews some of the current capabilities and needs for Modeling and Simulation (M&S), codes and computing resources, Verification and Validation (V&V), and uncertainty analysis for complex engineered systems. Application in structural mechanics, vibration and shock response are emphasized. Tools for uncertainty analysis include the design of computer experiments (such as fractional factorial, central composite, and orthogonal array designs), analysis of variance and effect screening, and meta-modeling using polynomial response surfaces. Tools for uncertainty propagation rely on sampling performed using the NESSUS (Numerical Evaluation of Stochastic Structures Under Stress) software that simulates uncertainties in loads, geometry, material behavior, and other user-defined uncertainty inputs to compute reliability and probabilistic sensitivity measures.

The applications discussed illustrate the need for designs of computer experiments and uncertainty analysis methods capable of handling about twenty variable with a few hundred simulations or less. More advanced designs of experiments may also prove valuable for training meta-models such as neural networks and polynomials. Finally, statistical tests are needed to assess the confidence that analysts may have about using a particular design of experiments to develop a meta-model or study the reliability of an engineered system.

Shane Reese
Brigham Young University

Integration of Responses from Physical and Computer Experiments

We consider the problem of integration of computational and physical experiments. Two approaches, a hierarchical Bayesian response surface approach and a combination of a Gaussian process and a response surface approach are compared and contrasted. The two methods are illustrated on a (poorly designed) computer and physical experimental procedure. The data are collected on a food coatings application and both methods applied with comparable results.

Laurence R. Rilett, University of Nebraska
Cliff Spiegelman, Texas A&M University

Calibration and Validation Issues in Micro-Simulation Models for Traffic Modeling

Over the past twenty years a significant amount of work has been conducted on improving the quality and accuracy of transportation supply and demand models. Much of this work has concentrated on micro-simulation models because they have a greater ability to model the inherently complex stochastic and dynamic nature of transportation systems when compared to traditional macroscopic models. More recently this approach has been adopted in long-term planning models as demonstrated by the development of the Transportation Analysis and Simulation System (TRANSIMS) as part of the Travel Model Improvement Program (TMIP). While these micro-simulation models are more versatile with respect to representing transportation systems, they do require significantly more, and different, data for calibration and validation as compared to traditional macroscopic models. In addition, because of operating constraints, their sophistication and underlying theory can vary significantly. For example, the TRANSIMS highway micro-simulation model may be classified as low fidelity because it has few driver decision rules and uses a discrete, cell based representation of the traffic network. A low fidelity approach was adopted because of the need in TRANSIMS to models a much larger geographic area at an individual traveler level in a relatively short period of time.

The recent advances in ITS technologies and implementation also have had a profound effect on traffic models. One aspect of their effect is that these models have to be adapted for modeling the technology as well as drivers' reaction to it. From the perspective of calibration of the micro-simulation models ITS can provide large amounts of data that can be disaggregated by day, by time of day, by location, and, in certain situations, by individual vehicle.

The purpose of this presentation is to focus on a number of calibration/validation issues including:

1. What approaches are available for calibrating traffic micro-simulation models with ITS data
2. How can demand, a necessary input, be calibrated using ITS data
3. Applicability/portability of methods
4. Robustness/repeatability
5. How many parameters are appropriate for validation (or for that matter how many parameters should a traffic micro-simulation model have given the quality of the calibration data)
6. How does supply affect the demand

Jerry Sacks (Part I), Duke University and National Institute of Statistical Sciences
M.J. Bayarri (Part II), University of Valencia

Formulation and Approach

The use of mathematically based computer models for the study of scientific and engineering processes is wide-spread and becoming even more critical in industrial setting. The most basic question in evaluating such a model is: Does the computer model adequately represent reality?

Statistical methodology for addressing this question is described within the context of a test-bed problem. The proposed six-step strategy deals with major issues associated with a validation process: quantifying the typically multiple sources of error and uncertainty in computer models, designing experiments (both field and computer), combining the multiple sources of information from field experiments and computer runs, calibrating parameters of the computer model, and assessing model predictions. A combination of spatial and Bayesian statistical tools provides the technical apparatus.

Tom Santer
Ohio State University

Designing Computer Experiments to Determine Robust Control Variables

This talk describes a sequential method for designing a computer experiment to determine a robust set of control variables when the inputs consist of both control and environmental variables. Control variables, thought of as manufacturing or engineering design variables in this talk, are determined by a "product designer" while environmental variables, called noise variables in the quality control literature, are uncontrolled in the field but take values that are characterized by a given nominal probability distribution. The objective is to design the computer experiment so as to find a set of control variables which are "robust" in the sense that the mean response over the environmental variable distribution is as insensitive as possible to uncertainty in the nominal environmental variable distribution. The proposed algorithm for designing the computer experiment and the associated method for determining a robust design is Bayesian. The prior for

the response is a stationary Gaussian stochastic process. Given the previous information, the sequential algorithm computes, for each untested site, the "improvement" over the current guess of the optimal robust setting. The design selects the next site to maximize the expected improvement criterion. Joint work with Jeff Lehman and Bill Notz.

David M. Steinberg, Tel Aviv University
Tamir Reisin, Soreq Nuclear Research Center
Eyal Hashavia, Israel Atomic Energy Commission

Identifying Critical Parameters in Simulations: A Case Study of a Nuclear Waste Repository

An important issue in nuclear waste disposal is to assess the potential risk to human life over the very long time scales that are associated with the decay of radioactive isotopes. Typical risk analyses consider the migration of radio-nuclides into, for example, the food and water supply during tens of thousands of years. Field observations, however, are limited to much shorter time scales (tens of years). There is good understanding of the physics that govern decay and migration so that physical models can be derived and computer simulators are thus important tools in carrying out risk analyses. The physical models require as inputs a large number of parameters that govern the interaction between the isotopes and the repository site (for example, level of precipitation, pumping rate and distribution coefficients). The exact values of the parameters depend on the specifics of the repository site and the isotopes themselves. Prior knowledge of many of the parameter values is often weak. A concern in planning a repository is to identify the parameters that are most influential in controlling the risk.

The focus of the talk will be a case study to identify critical parameters for a nuclear waste repository using the RESRAD simulator, developed at Argonne National Laboratory. The study involved 23 different input parameters. Among numerous outcomes associated with risk, we focus on the maximal equivalent annual dose in the drinking water during a 10,000-year time frame. Many conceivable input settings lead to no migration at all. When migration does occur, the maximal doses can vary by orders of magnitude. Statistical analyses to identify influential parameters must take into account this combination of highly skewed, yet truncated, outcome data. We will present a two-phase analysis strategy that examines, first, whether or not migration is present and then the extent of the migration, assuming that it occurs.

Agus Sudjianto
Bank of America

Computer Experiments in Complex Product Development: Practice, Methodology, and Challenge

In the past few years, with the advance of computing technologies and numerical approaches, scientific and engineering disciplines have experienced tremendous usage growth of sophisticated computer models to assist scientific investigation. In product development such as automobile, the use of sophisticated engineering computer models are ubiquitous for many reasons such as to make significant upfront design decision making prior to the availability of physical experiment due to measurement system limitation or its practicality.

Because of the complexity of real physical systems represented in the models, usually, there is no simple analytic formula to sufficiently describe the phenomena and systematic experiments to interrogate the models are required. Exercising the models, however, can be a very expensive proposition because these models are commonly tedious to prepare, computationally expensive,

and time consuming to analyze. The presenter will share his professional experiences leading large-scale product development where his team extensively applied computer experiments (including design of computer experiment, analysis of computer experiment, sensitivity analysis, probabilistic analysis, design optimization including robust design, and design validation). The framework and methodology advancement will be illustrated using real world automotive engine design problems. Several challenges will be present to motivate further research.

Kenny Ye
State University of New York at Stony Brook

A Simple Two-stage Analysis of Designed Experiment with Complex Aliasing

Non-regular factorial designs with complex aliasing are popular choices for computer experiments. In this talk, I will propose a simple two-stage analysis for experiments using factorial design with complex aliasing. The first stage is factor screening through a simple Bayesian procedure and an even simpler frequentist alternative. In the second stage, we further select active factorial effects from those involving factors selected in the first stage. This can be done using common best subset selection procedures or stepwise procedures. I will demonstrate our approach through two classic case studies. Potential application of this analysis strategy will be discussed as well as design properties desired for this data analysis method.

C. NPCDS-SAMSI Workshop on Data Mining Methodology and Applications Program and Abstracts -- October 28-30, 2004

Thursday – October 28, 2004
Fields Institute, Toronto

8:30-9:00 am Registration and Continental Breakfast

9:00-9:20 am Opening Remarks

9:20-10:20 am Learning in neural networks with small-world architecture
Helmut Kröger

10:20-10:45 am Coffee Break

10:45-12:15 pm Session on Rare Target Problems

Linking and pattern matching in multiple large data two-way tables
Stan Young

An Adaptive Radial Basis Function Network Model for Statistical Detection
Mu Zhu

ROC-based Learning for Imbalanced Class Problems
Grigoris Karakoulas

12:15-1:45 pm Lunch

1:45-3:15 pm Session on Unsupervised methods I

Clustering Categorical Data Based on Distance Vectors
Steven Wang

Algebraic Geometry and Model Selection for Naive Bayes Networks
Russel Steele

Generation 5 Hybrid Clustering System and its Application
Xianping Liu

3:15-3:45 pm Coffee Break

3:45-4:45 pm Feature extraction

Prediction of Real Variables with Non-Polynomial Approximants
Roberto Aldave and **Simon Gluzman**

Dependence Degree and Feature Selection for Categorical Data
Wenxue Huang

4:45-5:15 pm Daily discussant: **William Welch**, UBC

Friday – October 29, 2004

8:30-9:00 am Continental Breakfast

9:00-10:00 am Importance Sampling: An Alternative View of Ensemble Learning
Jerome Friedman

10:00-10:30 am Coffee Break

10:30-12:00 pm SAMSI data mining theme year speakers

Scalability of Models in Data Mining
David Banks

Random Forests: Proximity, Variable Importance, and Visualization
Adele Cutler

Bayesian Perspectives on Combining Models
Merlise Clyde

12:00-1:30 pm Lunch

1:30-3:30 pm Session on Supervised methods I

The use of grid computing to speed up prediction
Alex Depoutovitch

Robust Methods and Data Mining
Reuben Zamar

Proximity Graph Methods for Data Mining
Godfried Toussaint

Automated Trade area analysis. Case study of G5 MWM software application
Alex Zolotovitski

3:30-4:00 pm Coffee Break

4:00-5:00 pm Is regularization: efficient and effective Piecewise linear SVM paths
Ji Zhu and **Saharon Rosset**

5:00-5:30 pm Daily discussant: **Hugh Chipman**, Acadia University

Saturday – October 30, 2004

8:30-9:00 am Continental Breakfast

9:00-10:00 am Statistical Learning from High Dimensional and Complex Data: Not a Lost Cause
Yoshua Bengio

10:00-10:30 am Coffee

10:30-12:00 pm Session on Mining industrial process data

Data-Mining for industrial processes
Joaquin Ordieres Meré

Data Mining in Industry for Process and Product Improvement
Theodora Kourti

12:00-1:30 pm Lunch

1:30-3:30 pm Panel Discussion

Tracey Jarosz, Loyalty group
Jerome Friedman, Stanford University
Theodora Kourti, McMaster University
Rick Makos, Teradata
Ivan Miletic, Dofasco Inc.
Milorad Krneta, Generation 5
Stan Young, National Institute of Statistical Sciences

3:30-4:00 pm Coffee Break

4:00-5:00 pm Constructing induction graphs

Djamel Zighed

Roberto Aldave and Simon Gluzman
Generation 5 Mathematical Technologies Inc

Prediction of Real Variables with Non-Polynomial Approximants

Multivariate Non-Polynomial Approximants have been applied for numerical prediction of real variables dependent on real variables. Calculations were performed for several databases. Accuracy of the method is compared with some known methods including different local regressions. We argue that the method is effective for discovery of non-linear dependencies.

David Banks
Duke University

Scalability of Models in Data Mining

The SAMSI data mining year examined a range of problems in the area, from use of unlabeled sample in classification to issues in overcompleteness. A major theme concerned the analysis of high-dimension, low sample-size datasets. This talk reviews various strategies for handling these problems, which are closely related to the Curse of Dimensionality and also relevant to data quality. Our work explores the effect of combining smart algorithms, aggressive feature selection, and combinatorial search. The ideas are illustrated through several examples, and we draw conclusions and give advice for those who must analyze such data.

Keywords: Curse of Dimensionality ; large p, small n ; regression; MDS ; data mining ; selection

Yoshua Bengio
Université de Montréal

Statistical Learning from High Dimensional and Complex Data: Not a Lost Cause

This talk will start from a sociological perspective on current trends in statistical machine learning, and with the claim that less and less research is attempting to address the really difficult question of generalization from high-dimensional and complex data sets, while the mathematical sophistication in analyzing and devising statistical learning algorithms has grown steadily in the last few years. To illustrate that issue I will present arguments to the effect that a large class of manifold learning and spectral embedding algorithms (e.g. kernel PCA, spectral clustering, LLE, Isomap, Laplacian Eigenmaps) that are essentially local will suffer from at least four generic problems associated with noise in the data, curvature of the manifold, dimensionality of the manifold, and inability to take advantage of the presence of many manifolds with little data per manifold. This analysis suggests investigating non-local manifold learning algorithms which attempt to discover shared structure in the tangent planes at different positions in data space. Unfortunately this opens to door to non-convex optimization problems, which have received strongly negative connotation in the statistical machine learning community in recent years. A simple (but non-convex) criterion for a non-local manifold learning algorithm is thus proposed and experiments estimating a tangent plane prediction function are presented, showing its advantages with respect to local manifold learning algorithms. For example, from examples of rotated digits the algorithm generalizes to rotations of other characters, whereas a local method such as kernel PCA or Isomap fails.

Merlise Clyde
Duke University

Bayesian Perspectives on Combining Models

Consideration of multiple models is routine in statistical practice. With computational advances over the past decade, there has been increased interest in methods for making inferences based on combining models. Examples include boosting, bagging, stacking, and Bayesian Model Averaging (BMA), which often lead to improved performance over methods based on selecting a single model. Bernardo and Smith have described three Bayesian frameworks for model selection known as the M-closed, M-Complete, and M-open perspectives. The standard formulation of Bayesian Model Averaging arises as an optimal solution for combining models in the M-closed perspective where one believes that the ``true" model is included in the list of models under consideration. In the M-complete and M-open perspectives the ``true" model is outside the space of models to be combined, so that model averaging using posterior model probabilities is no longer applicable. Using a decision theoretic approach, we present optimal Bayesian solutions for combining models in these frameworks. We illustrate the methodology with an example of combining models representing two distinct classes, prospective classification trees and retrospective multivariate discriminant models applied to gene expression data in advanced stage serous ovarian cancers. The goals of this analysis are two-fold: identifying molecular tumor characteristics associated with prognosis and determining if long-term survival can be predicted by features intrinsic to the molecular biology of the tumor.

Adele Cutler
Utah State University

Random Forests: Proximity, Variable Importance, and Visualization

The Random Forests algorithm, introduced by Breiman in 2001, has been shown to be competitive with some of the most accurate classification methods available. Unlike many other accurate methods, Random forests is interpretable via proximities and variable importance measures. In this talk, I will describe the methods we use to compute these measures and illustrate their use with a special-purpose java-based visualization package.

Joint work with Leo Breiman

Alex Depoutovitch
Generation 5

The use of grid computing to speed up prediction

Rapidly growing size of data becoming available to data mining applications creates a real challenge to process such volume of data within short period of time. CPU speeds are improving but yet not enough to address this challenge - it may take days and even weeks to run such things like predictions required by analytical and marketing departments. To address this demand we developed a technology that together with highly scalable method of prediction allows executing calculations in parallel by using a grid of computers. Our benchmarks on life systems show almost linear scalability as the number of computers in grid grows.

Jerome H Friedman

Stanford University

Importance Sampling: An Alternative View of Ensemble Learning

Learning a function of many arguments is viewed from the perspective of high-dimensional numerical integration. It is shown that many of the popular ensemble learning methods can be cast in this framework. In particular, bagging, boosting, and Bayesian model averaging are seen to correspond to Monte Carlo integration methods each based on different importance sampling strategies. This interpretation explains some of their properties and suggests modifications to them that can improve their accuracy and especially their computational performance.

Joint work with Bogdan Popescu

Wenxue Huang
Generation 5 Mathematical Technologies Inc.

Dependence Degree and Feature Selection for Categorical Data

Traditionally the measure of association for cross-clasification for categorical data takes a view of variance or divergence. We take an opposite angle of view: convergence. This point of view has certain advantages. It allows us to see more directly and clearly how a vairable is associated with another/others both locally (veritically) and globally (horizontally). From this point of view we introduce a new measure of association, referred to as *dependence degree*, and dicuss one of its applications in data mining technologies: feature selection.

Grigoris Karakoulas
University of Toronto
Computer Science

ROC-based Learning for Imbalanced Class Problems

In general, modelling techniques for learning probabilistic models from data can be categorized into discriminative (e.g. logistic regression) and generative (e.g. Normal Discriminant Analysis). In this paper we describe a new probabilistic modeling technique that combines the advantages of these two categories of techniques. More specifically, let us denote by X the vector of model input variables and by Y (0/1) the target variable. To learn the probability distribution P{X|Y} as in a generative model our technique is recursively searching for discriminatory projections in the input space that maximize the area under the Receiver-Operating-Characteristic (AUROC) curve. Although the latter is a well-accepted performance measure for evaluating probabilistic classification models, most of the generative and discriminative techniques in statistics and machine learning use other performance measures for building a model. In our case we use the same performance measure, AUROC, for both building and evaluating a model. Our experiments show that this new technique is particularly advantageous in imbalanced datasets due to the use of AUROC as the objective function.

Theodora Kourti
McMaster University
Chemical Engineering Department

Data Mining in Industry for Process and Product Improvement

With process computers routinely collecting data from on-line sensors on hundreds to thousands of variables every few seconds, large databases are accumulated in industry. The exploitation of these data is a critical component in the successful operation of any industrial process. However the task is difficult because of the nature of such data:

1) Large Data Sets
2) The data are highly correlated (many variables being collinear) and non-causal in nature.
3) The information contained in any one variable is often very small due to the low signal / noise ratios.
4) There are often missing measurements on many variables.

In order to utilize these databases, an empirical modelling method must be able to deal effectively with all these difficulties. Multivariate projection methods have gained rapid acceptance by industry for troubleshooting, on-line process monitoring, fault diagnosis and equipment maintenance. Successful applications are reported by many diverse industries such as pharmaceuticals, semiconductor manufactures, steel producers, pulp and paper producers, polymer plants and refineries

This paper will discuss the latest developments on latent variable methods, speech recognition methods and missing data treatment for successful data mining in industry.

Helmut Kroeger
Laval University

Learning in neural networks with small-world architecture.

I will introduce small-world (sw) and scale-free (sf) networks, giving a few examples from organisation of society, internet, and biology. I will discuss experimental observations in neuroscience: cat cortex, macaque visual cortex, human brain activity network from magnetic resonance imaging. Then I address computer simulations involving sw- and sf-architecture: (a) fast response and coherent oscillations in a Hodgkin-Huxley sw-network, (b) efficient associative memory and pattern retrieval in sw-architecture, (c) supervised learning in multi-layer feed-forward networks.

Learning algorithms like backpropagation are commonly used in regular networks where neurons are organized in successive fully-connected layers. Here, the effect of learning is studied as a function of connectivity architecture of neurons in multi-layer networks going from regular to small-world to random connectivity. One observes that sw-connectivity offers great advantages on the learning speed of the backpropagation algorithms, especially when the number of layers is high. I will also discuss an example of generalisation.

Xianping Liu
Generation 5 Mathematical Technologies Inc.

Generation 5 Hybrid Clustering System and its Application

Clustering is used in data mining for .nding the useful patterns in large high dimensional databases. Some challenges for large data sets are scalability, automation, any types of data, input order dependency, and understanding end-results. Generation 5 is developing a fully automatic hybrid clustering system Clus5, a part of G5MWM data mining suite, which addresses all the challenges above. Clus5 employs hybrid on-line and o.-line partitional clustering algorithms that

are more e.ective and input order-independent. It uses two-stage approach. In stage 1, it clusteres multiple samples with di.erent initial partitions to produce a near-optimal initial seed based on one of built-in validity indexes. In stage 2, it clusteres the whole dataset using result from stage 1. Clus5 can automatically determine the optimal number of clusters and handle any types of data. It also has a special algorithm for marketing applications which prefer equal size clusters. Clus5 has been successfully applied to solve many real-life problems of Gen5 clients.

Joaquín Ordieres Meré
University of la Rioja (Spain)

Data-Mining for industrial processes

The most common goal of the factory owner is to achieve better quality in the final product by means of improved process control. The significance and relevance of optimizing the existing control models is even greater in the open-loop control systems or in those governed by computational methods dependent on adjustable parameters.
This talk reviews some typical industrial environments and focuses on some parts of them in order to show the real interest of these improvements. We will identify some difficulties in obtaining these improvements and show how the optimal control model for the manufacturing process can be obtained from data provided by sensors. We will also discuss some technical problems that are related to the main goal, and will identify some topics concerning outliers, density and topology. Also, we will show how these techniques can be applied as an instrumental toolbox in addressing some environmental problems.

Keywords: industrial applications; neural networks; outliers; density; improvement process control; advanced quality control

Saharon Rosset and Ji Zhu
University of Michigan

II regularization: efficient and effective

We consider the general regularized optimization problem of minimizing loss+penalty, where the loss depends on the data and the model, and the penalty on the model only. We illustrate that the choice of l_1 (lasso) penalty, in combination with appropriate loss, leads to several desirable properties: (1) Approximate or exact l_1 regularization has given rise to highly successful modeling tools: the lasso, boosting, wavelets and 1-norm support vector machines. (2) l_1 regularization creates sparse models, a property that is especially desirable in high-dimensional predictor spaces. We formulate and prove sparsity results. (3) l_1 regularization facilitates efficient methods for solving the regularized optimization problem. The LARS algorithm takes advantage of this property. We present a general formulation of regularized optimization problems for which efficient methods can be designed. We show how we can create modeling tools which are robust (because of the loss function selected), efficient (because we solve the regularized problem efficiently) and adaptive (because we select the regularization parameter adaptively).

Joint work with Trevor Hastie and Rob Tibshirani

Russell Steele
McGill University

Algebraic Geometry and Model Selection for Naive Bayes Networks

Recent advancements in the machine learning community have been made in the application of algebraic geometry to Bayesian model selection for neural networks. Rusakov and Geiger (2003), using results from Watanabe (2001), present a new asymptotic approximation to the integrated likelihood for Naive Bayes networks (or, in more statistical language, finite mixture models). The key to the approximation is an attempt to correct the standard dimensionality penalty for the BIC to reflect the "true" effective dimensionality of the model. However, only limited work has been done to interpret these results with respect to current work on effective dimensionality in Bayesian research, particularly, in light of the currently widespread use of the DIC (Spiegelhalter, et al., 2002). In this talk, the speaker presents the basic idea behind the Rusakov/Geiger approximation, compares their approximation to the actual integrated likelihood, and links these results to what the DIC presumably is trying to estimate. The speaker will focus on how various standard information criteria perform for model selection for simple Naive Bayes networks.

Godfried T. Toussaint
McGill University
School of Computer Science

Proximity Graph Methods for Data Mining

In the typical approach to learning in data mining, random data (the training set of patterns) are collected and used to design a classification rule. One of the most well known such rules is the K-nearest-neighbor rule in which an unknown pattern is classified into the majority class among its K nearest neighbors in the training set.

Several questions related to this rule have received considerable attention over the past fifty years. Such questions include the following. How large should K be? How should a value of K be chosen? Should all K neighbors be equally weighted when used to decide the class of an unknown pattern? Should the features (measurements) be equally weighted? If not, how should these weights be chosen? What metric should be used?

How can the rule be made robust to outliers present in the training data? How can the storage of the training set be reduced without degrading the performance of the decision rule? How fast can the nearest neighbor of a query point be found? What is the smallest neural network that can implement the nearest neighbor rule? Geometric proximity graphs offer elegant solutions to most of these problems, in many cases, by discarding the classical methods altogether. In this talk we review and discuss recent work on the application of geometric proximity graphs in this area, propose some new solutions and mention some open problems.

Steven Wang
York University

Clustering Categorical Data Based on Distance Vectors

We introduce a novel statistical procedure to cluster categorical data based on distance vectors. The proposed method is conceptually simple and computationally straightforward as it does not require any specific statistical models or any convergence criteria. Moreover, unlike most existing algorithms that compute the class membership or membership probability for every data point at each iteration, our algorithm sequentially extracts clusters from the given data set. That is, at each iteration our algorithm only strives to identify one cluster, which will then be deleted from the

data set at the next iteration; this procedure repeats until there are no more significant clusters in the remaining data. Consequently, the number of clusters can be determined automatically by the algorithm. As for the identification and extraction of a cluster, we first locate the cluster center by using a Pearson Chi-square type statistic on the basis of categorical distance vectors. The partition of the data set produced by our algorithm is unique and robust to the input order of data points. The performance of the proposed algorithm is examined by both simulated and real world data sets. Comparisons with two well-known clustering algorithms, K-modes and AutoClass, show that the proposed algorithm substantially outperforms these competitors, with the classification rate or the information gain typically improved by several orders of magnitudes.

S. Stanley Young
National Institute of Statistical Sciences

Linking and pattern matching in multiple large data two-way tables

Drug discovery is coming into multiple large data sets, micro arrays, protein arrays, chemical structural descriptors. All of these data sets have many more columns than rows, n << p. We will explore various methods of pattern matching between multiple two-way tables. The benefit of the methods will be the discovery and possible validation of biological targets for drug discovery.

Joint work with Douglas M. Hawkins, U Minn and Li Liu, Aventis

Ruben Zamar
University of British Columbia

Robust Methods and Data Mining

We consider the role of various performance measures for supervised and unsupervised learning. Real world applications of data-mining typically require optimizing for non-standard performance measures that depend on the application at hand. For example, in spam mail filtering, error rate is a bad indicator of performance because there is a strong imbalance in cost between missing a good message and letting through a spam message. In our experience, the choice of a good subset of variables to perform a data mining task may be at least as important as the choice of the procedure. We will illustrate this situation in the case of the Protein Homology Prediction Task for this year KDD Cup Competition. We will also discuss the problem of detecting a relative small fraction of "interesting" cases (in the unsupervised learning setup) using robust Mahalanobis distances and distance based outliers.

Joint work with Yi Lin, Guohua Yan, Will Welch, Department of Statistics, UBC

Ji Zhu
University of Michigan

Piecewise linear SVM paths

The support vector machine is a widely used tool for classification. In this talk, we consider two types of the support vector machine: the 1-norm SVM and the standard 2-norm SVM. Both types can be written as regularized optimization problems. In all current implementations for fitting an SVM model, the user has to supply a value for the regularization parameter. To select an appropriate regularization parameter, in practice, people usually pre-specify a finite set of values

for the regularization parameter that covers a wide range, then either use a separate validation data set or use cross-validation to select a value for the regularization parameter that gives the best performance among the given set. In this talk, we argue that the choice of the regularization parameter can be critical. We also argue that the 1-norm SVM may have some advantage over the standard 2-norm SVM under certain situations, especially when there are redundant noise features. We show that the solution paths for both the 1-norm SVM and the 2-norm SVM are piecewise linear functions of the regularization parameter. We then derive two algorithms, respectively for the 1-norm SVM and the 2-norm SVM, that can fit the entire paths of SVM solutions for every value of the regularization parameter, hence facilitate adaptive selection of the regularization parameter for SVM models.

It turns out that the piecewise linear solution path property is not unique to the SVM models. We will propose some general conditions on the generic regularized optimization problem for the solution path to be piecewise linear, which suggest some new useful predictive statistical modeling tools.

This is joint work with Saharon Rosset (IBM T.J.Watson), Trevor Hastie (Stanford U.), and Rob Tibshirani (Stanford U.)

Mu Zhu
University of Waterloo

An Adaptive Radial Basis Function Network Model for Statistical Detection

We construct a special radial basis function (RBF) network model to detect items belonging to a rare class from a large database. Our primary example is a real drug discovery application. Our method can be viewed as modeling only the rare class but allowing for local adjustments depending on the density of the background class in local neighborhoods. We offer a statistical explanation of why such an approach is appropriate and efficient for the detection problem. Our statistical explanation together with our empirical success with this model has implications for a new paradigm for solving these detection problems in general.

This work is joint with Wanhua Su and Hugh Chipman.

Djamel A. Zighed
University of Lyon 2

Constructing induction graphs

The presentation will focus on tree issues concerning the construction of induction graphs from data.

The first one is how to obtain the optimal join-partitioning of the contingency tables. We propose a quasi-optimal algorithm for resizing contingency tables, which takes into account the type (nominal or ordinal) of the attributes crossed with the categorical variable to predict.

The second issue is how to use the results of the best-join-partitioning to build a new type of decision tree. The algorithm we propose is called Arbogodaï. It leads to a new structure of decision trees that could be seen as a generalisation of CART procedure. Indeed, instead of having for the predicted attribute two super classes, Arbogodaï defines the optimal number of super-classes. Also, the splits are not only binary, as in the CART algorithm, but their multiplicity

depends on the local situation. At each node, the algorithm looks for the best join-partitioning on the contingency tables crossing the predicted attribute over all predictive ones.

The third issue of this presentation is how to provide an axiomatic definition for a measure of the quality of partition. We propose leading to an algorithm for constructing induction graphs that can be seen as a generalisation of a decision tree algorithm. The framework will be common for tree or lattice structures.

Alex Zolotovitski
Generation 5

Automated Trade area analysis. Case study of G5 MWM software application

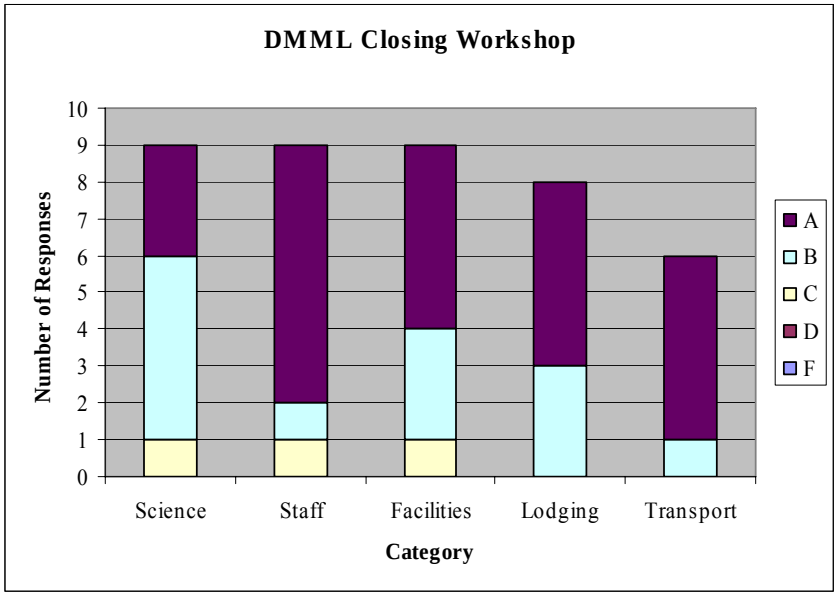
We created an automatized tool for estimating, at the Zip+4 level, the dollar amount that each US household (approximately 35 million zip+4's) spends in each US grocery store (approximately 50,000) for a specified group of about 60 products. These amounts were aggregated to estimate sales in each US grocery store for each specified group of products. We used non-parametric local and global regression; as data sources we used 1) panel data, which contained consumption of 290 brands of products for 37,000 Households in 717 groups of Stores; 2) attributes of 50,000 stores; and 3) G5 zip+4 level databases which contain almost 10,000 demographic, expenditure, and behavioral variables.

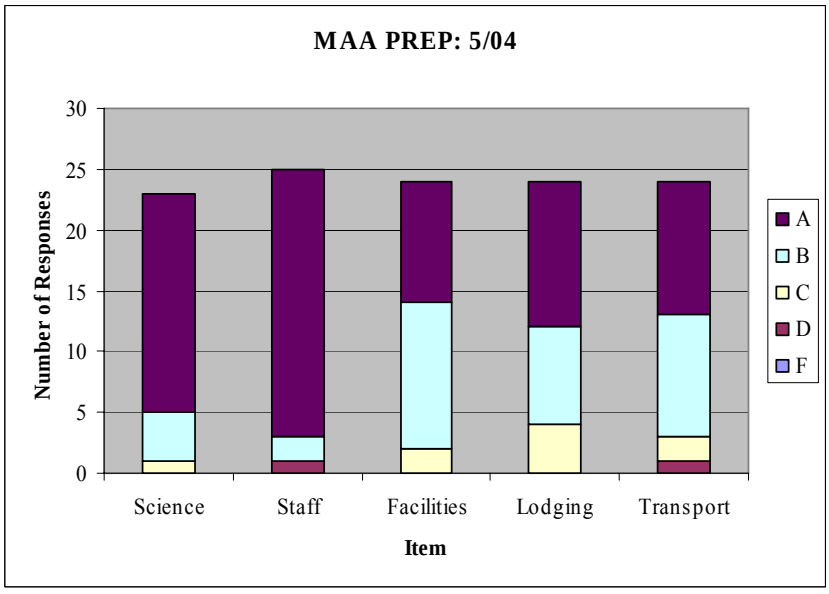
APPENDIX G – Workshop Evaluation Summaries

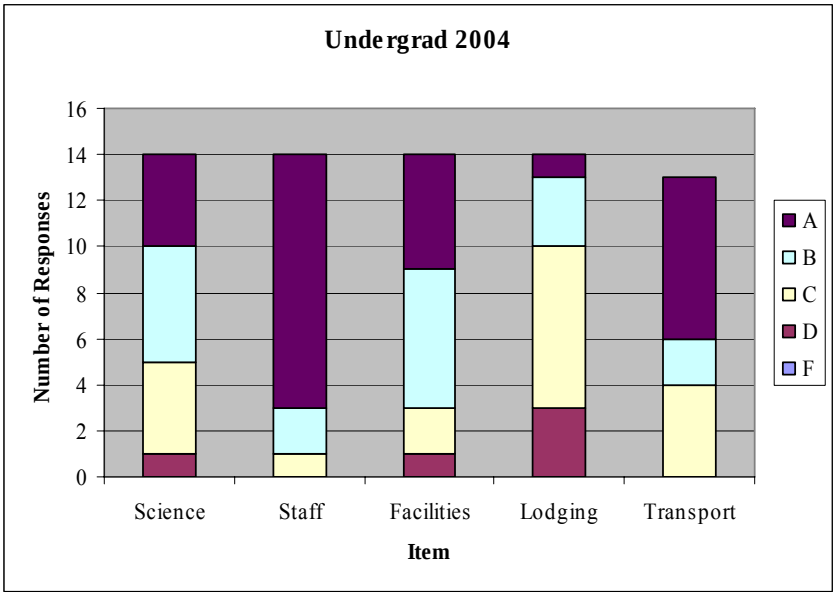
Workshop participants were given an evaluation questionnaire to complete in each of the SAMSI workshops. A sample questionnaire is given on the following pages.

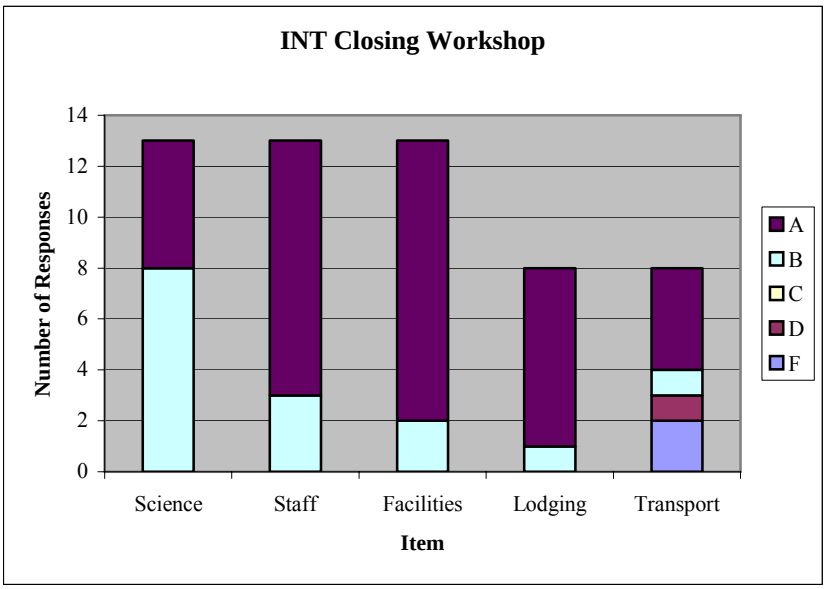
Below are the summaries of the participant evaluations for the four main scientific workshops held before February 28, 2005 (events held March-June 2005 will be reported in the 2005-06 Annual Report). The rating scale was 1-5 (lowest to highest). The five questions addressed in the table were:

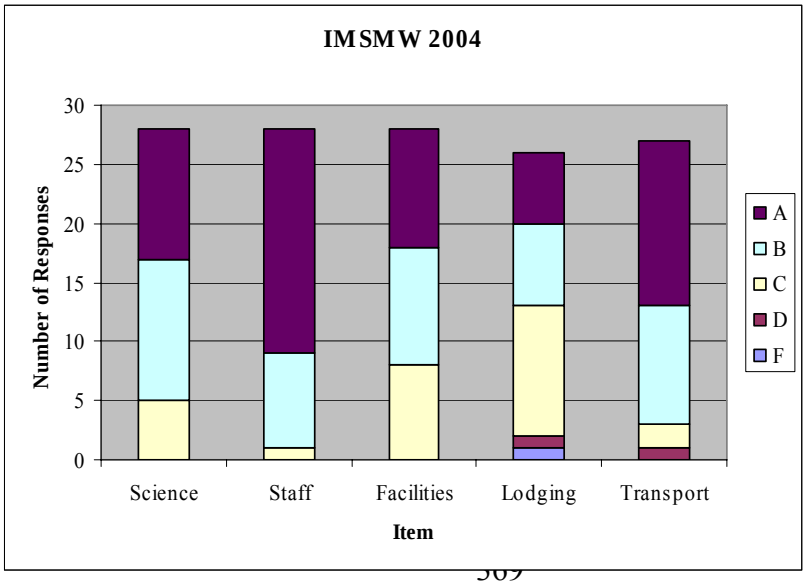
a. Scientific Quality
b. Staff Helpfulness
c. Meeting Room/AV Facilities
d. Lodging
f. Local Transportation

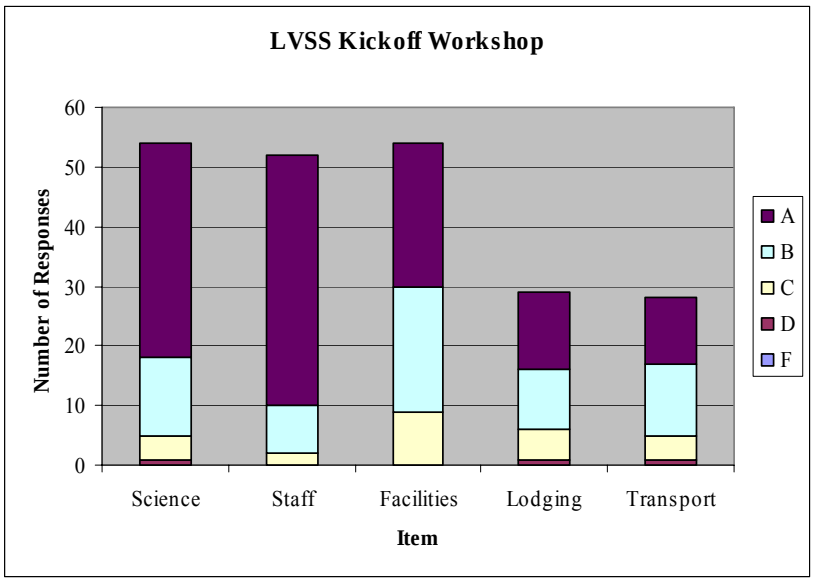


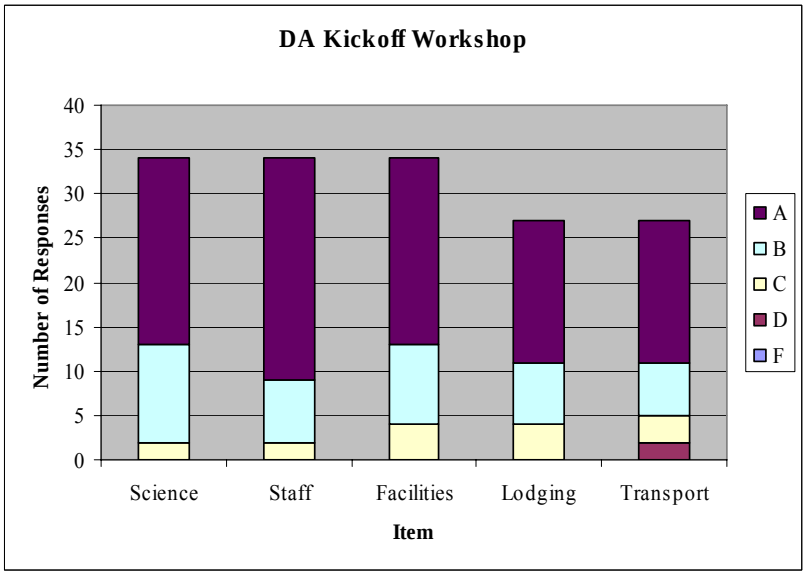


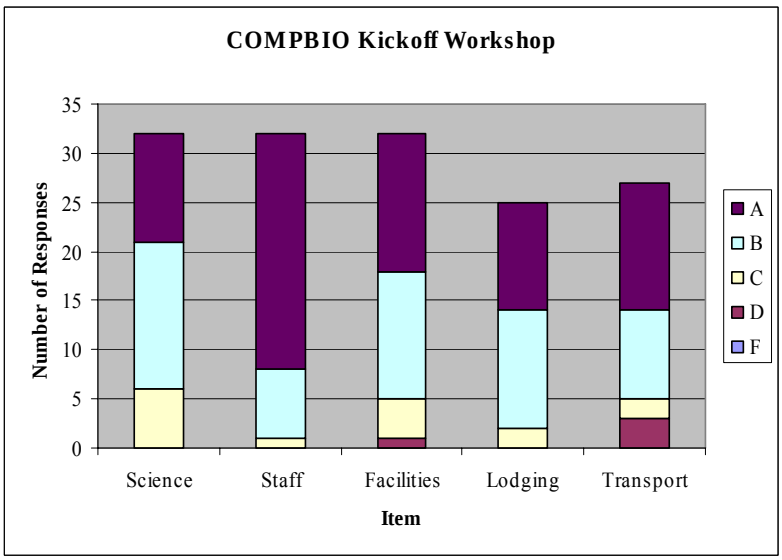

Undergrad 2004
Number of Responses
16
14
12
10
8
6
4
2
0
Science
Staff
Facilities
Lodging
Transport
Item
A
B
C
D
F

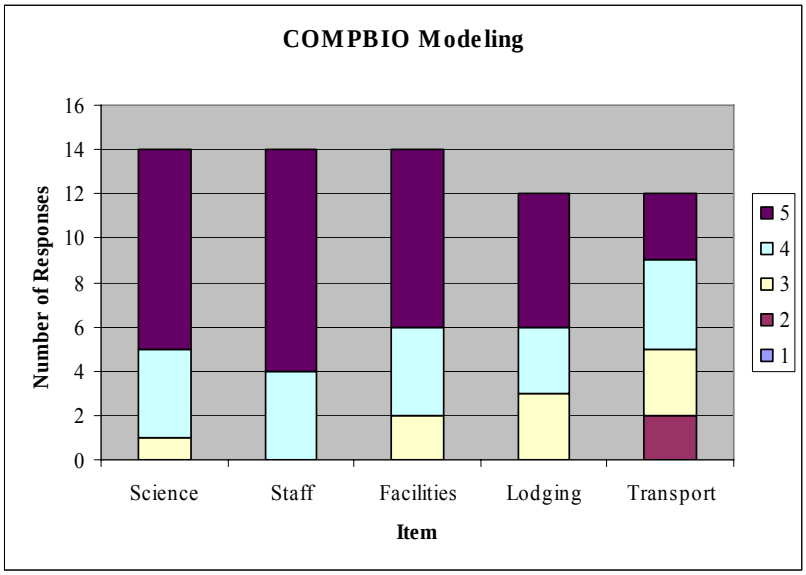

INT Closing Workshop
Number of Responses
14
12
10
8
6
4
2
0
Science
Staff
Facilities
Lodging
Transport
Item
A
B
C
D
F

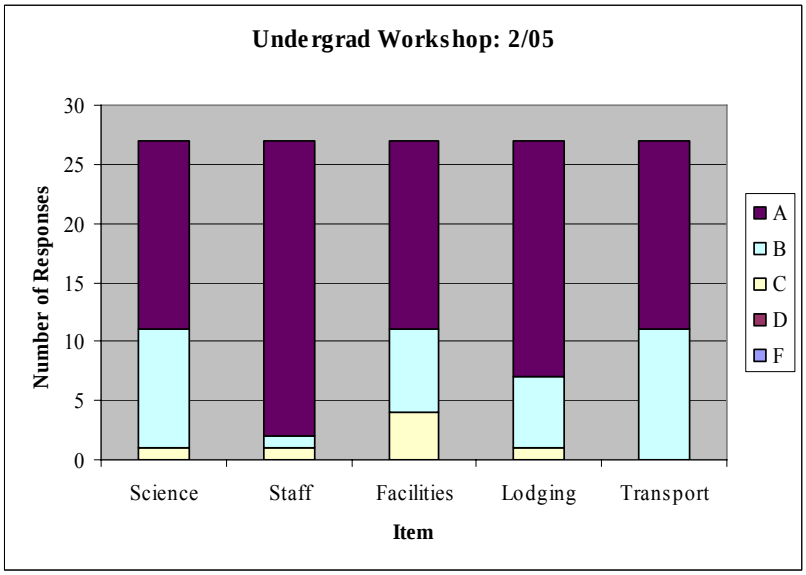

IMSMW 2004
Number of Responses
30
25
20
15
10
5
0
Science
Staff
Facilities
Lodging
Transport
Item
A
B
C
D
F

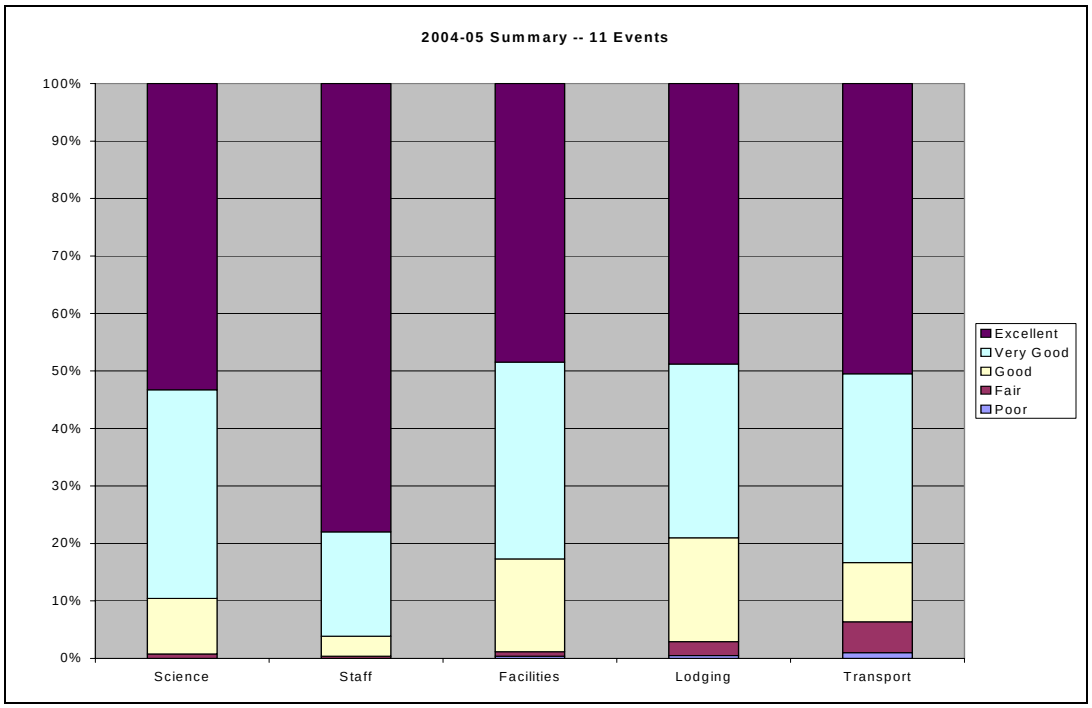

LVSS Kickoff Workshop
Number of Responses
60
50
40
30
20
10
0
Science
Staff
Facilities
Lodging
Transport
Item
A
B
C
D
F

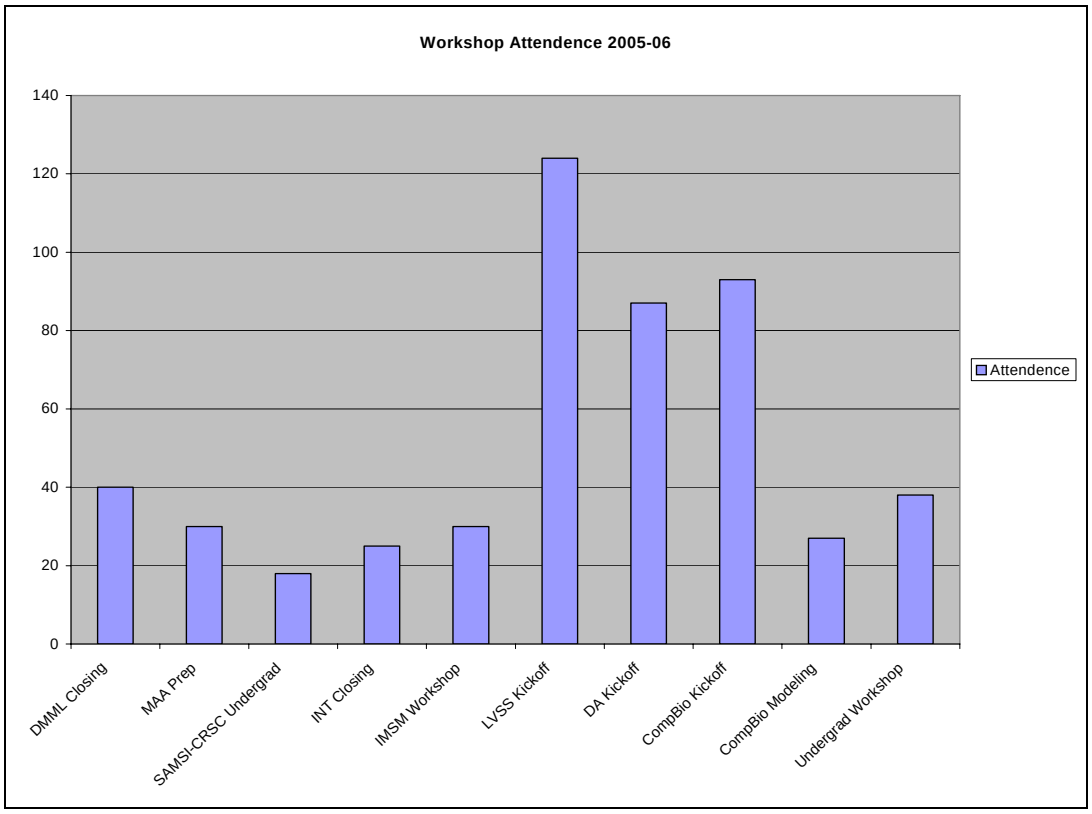

DA Kickoff Workshop
Number of Responses
40
35
30
25
20
15
10
5
0
Science
Staff
Facilities
Lodging
Transport
Item
A
B
C
D
F


COMPBIO Kickoff Workshop
Number of Responses
35
30
25
20
15
10
5
0
Science
Staff
Facilities
Lodging
Transport
Item
A
B
C
D
F


COMPBIO Modeling
Number of Responses
16
14
12
10
8
6
4
2
0
Science
Staff
Facilities
Lodging
Transport
Item
5
4
3
2
1


Undergrad Workshop: 2/05
Number of Responses
30
25
20
15
10
5
0
Science
Staff
Facilities
Lodging
Transport
Item
A
B
C
D
F


2004-05 Summary -- 11 Events
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Science
Staff
Facilities
Lodging
Transport
Excellent
Very Good
Good
Fair
Poor


Workshop Attendence 2005-06
140
120
100
80
60
40
20
0
DMML Closing
MAA Prep
SAMSI-CRSC Undergrad
INT Closing
IMSM Workshop
LVSS Kickoff
DA Kickoff
CompBio Kickoff
CompBio Modeling
Undergrad Workshop
Attendence

SAMSI Workshop Evaluation
Workshop Name
Workshop Dates

Your feedback on this workshop is requested by the National Science Foundation, who view it as important for assessing and improving the performance of institutes. Your feedback is also gratefully appreciated, because it will enable us to immediately improve SAMSI workshops. Please fill this out and hand it to a SAMSI Staff Member.

0. Personal Information:

a. Discipline (e.g. Statistics, Applied Math.) _____

b. Highest Degree: _____ Year: _____ Current Student: ____

1. General Ratings:	Poor	Fair	Good	Very Good	Excellent
a. Scientific Quality	1	2	3	4	5
b. Staff Helpfulness	1	2	3	4	5
c. Meeting Room/AV Facilities	1	2	3	4	5
d. Lodging	1	2	3	4	5
e. Local Transportation	1	2	3	4	5

2a. What were the positive aspects of the organization and running of this workshop?

_____

_____

2b. What parts of the organization and running need improvement?

_____

_____

3. Please comment on the Scientific Quality:

a. Innovation: _____

_____

b. Communication: _____

_____

c. Level: _____

_____

4. Additional Comments on the overall workshop / tutorial.

_____

5. An important goal of SAMSI is to create synergies between statistics, applied mathematics, and other disciplines. How well did this workshop further this goal?

_____

6. How did you learn of this workshop?

_____

7. Please suggest ideas / contacts for future SAMSI programs

_____

8. Personal Information:

Name: _____

Affiliation: _____

Email Address: _____

The following information is for reporting purposes only. It has no bearing whatsoever on future participation in SAMSI events. Please circle appropriate choice.

Gender: Male Female

Ethnicity: Hispanic Not Hispanic

Race:

White African American Asian Native American

Hispanic Native Hawaiian/Pacific Islander Other _____