



National Institute of Statistical Sciences
 PO Box 14006, Research Triangle Park, NC 27709-4006
 Tel: 919.685.9300 FAX: 919.685.9310
 www.niss.org

Annual Report of the Director

Alan F. Karr
 October 23, 2001

Contents

1	Affiliates Program	2
1.1	Management	2
1.2	Retention and Recruitment	2
1.3	Affiliate-Related Research	2
1.4	Workshops	2
1.5	Communication and Information Dissemination	3
1.6	Professional Development	3
2	Projects and Project Development	3
2.1	Ongoing Projects	4
2.2	New Projects	4
2.3	Completed Projects	5
2.4	Proposals	5
2.5	Planned Proposals	6
2.6	Initiatives with Other Research Organizations	6
3	Personnel	7
3.1	Postdoctoral Fellows	7
3.2	Staff	7
4	Communication and Outreach	7
5	Summary and Personal Comments	8

1 Affiliates Program

Now in its second year, the program is developing along multiple lines. The level of activity is not what it should be, principally because a Director for the program has yet to be appointed.

1.1 Management

Efforts to recruit a full-time Director for the program have not yet succeeded. Alternative means of meeting the needs — for example, via a person on sabbatical, someone who has retired recently, or a combination of individuals — are being explored. *At the moment, this is the most serious challenge facing NISS.*

1.2 Retention and Recruitment

Two new affiliates, Boeing and SPSS, have joined the program; two others have committed to do so, pending budgetary approval. Unfortunately, five charter affiliates did not renew their status. What effect the events of September 11 and their economic consequences will have on retention is not known, of course, but seems unlikely to be positive.

1.3 Affiliate-Related Research

During 2001, NISS was engaged in research funded by, or in collaboration with (in most cases, both) nine corporate and government affiliates: the Bureau of Labor Statistics (BLS), Bureau of Transportation Statistics (BTS), Census Bureau (Census), US Environmental Protection Agency (EPA), General Motors (GM), GlaxoSmithKline (GSK), Los Alamos National Laboratory (LANL), National Center for Education Statistics (NCES) and Visual Insights (VI). University affiliates involved in NISS projects and proposals during the past year include Carnegie Mellon University (CMU), Duke, North Carolina State University (NCSU), Penn State, Purdue and the University of North Carolina at Chapel Hill (UNC).

Thus, the expectation that the affiliates program will shape the research agenda at NISS is being realized, especially in regard to data quality, a problem brought to NISS by the affiliates, and in which we are engaged in significant and expanding research (§2.2 and 2.4).

1.4 Workshops

Three workshops have been held since October of 2000.

Data Quality. This workshop was held at Telcordia Technologies (Morristown, NJ) on November 30 – December 1, 2000, with more than 50 attendees, principally from corporate and government agency affiliates. An authoritative workshop report was produced, which informed the preparation of proposals for NISS research on data quality (§2.2 and 2.4).

Pharmacogenomics. The workshop was held on February 12–13, 2001 at Amgen (Thousand Oaks, CA). Affiliate attendance was approximately 40. Talks were presented by statistical and other scientists from Amgen and GSK.

Internet Data. This workshop, organized by J. S. Marron of UNC, took place at NISS on March 9–10, 2001. Attendance was more than 40. Speakers spanned corporations, universities and research centers.

Details (including presentations and reports) from the workshops are available from the affiliates Web site: www.niss.org/affiliates/affiliatesmain.html.

1.5 Communication and Information Dissemination

Planning Meetings. The March 2, 2001 planning meeting at NISS was attended by representatives of approximately 25 affiliates. As a result of group discussions, five scientific thrusts for the program during 2001–02 were identified — bioinformatics, customer behavior (especially but not exclusively on Web sites), data confidentiality, data quality and large data sets. Each will be the subject of a workshop, be pursued as a research activity of NISS (§2), or both.

The August 5, 2001 JSM affiliates meeting had an attendance of nearly 30. It reviewed progress and plans for the program, especially in connection with SAMSI (§2.4).

Job Listings. A Web-based job listing service for all affiliates was initiated in October, 2001. It is available at www.niss.org/affiliates/ajls.html.

Publicity. An announcement about the program, listing current affiliates and inviting others to join, will appear in the January, 2002 issue of *AMSTAT News*. A copy of last year's announcement, listing charter affiliates, is displayed in the NISS building.

1.6 Professional Development

Especially at the March, 2001 planning meeting, the need and potential for the program to support employee recruiting by affiliates emerged strikingly. The Job Listings (§1.5) are one effort in this direction. Two more important initiatives are:

Postdoctoral Program for Federal Agency Affiliates (FAAs). A proposal for a joint postdoctoral program of NISS and its FAAs rests with the agencies for their approval. Under the auspices of the program, NISS will appoint postdoctoral fellows for 2–3 year terms, who will be placed in challenging assignments at the FAAs, principally in Washington. In addition, NISS will conduct activities fostering interaction among these postdocs, and between them and NISS postdocs based in Research Triangle Park (RTP).

Benefits to the FAAs include immediate help on pressing research issues from talented young researchers, positive exposure to potential employees, including non-US citizens, opportunity for in-depth evaluation of potential employees; and continuing attention to research problems, which could persist after postdocs complete their assignments. Benefits to the postdoctoral participants are the intellectual and career benefits of time spent on challenging applied statistics problems with strong cross-disciplinarity and deep theoretical roots, valuable contacts, in terms of securing other positions and future research support, and interaction with one another and other NISS postdocs. The value to NISS includes greater visibility, increased service to our FAAs (and the statistical sciences community generally), and opportunity for research relationships complementing or spawned by the program.

Summer Internship Program. The NISS Affiliates Internship Program (NAIP) is designed to match graduate students from NISS university affiliates to exciting internships in the summer of 2002 at corporate, government and university affiliates (especially those without strong existing internship programs of their own), as well as at NISS itself. NISS' role will be to assemble and make available job descriptions and to assemble and forward to employers resumes and statements of interest from student applicants. A Closing Conference of all interns will be held at NISS in August of 2002.

2 Projects and Project Development

The research program is in a healthy state of flux: projects are being completed, initiated and proposed.

2.1 Ongoing Projects

Digital Government (DG). This three-year project funded by the Computer and Information Sciences and Engineering (CISE) directorate at the NSF merges statistics, information technology and domain knowledge to build Web-based systems (and the underlying abstractions, theory and methodology) that alleviate the tension imposed on Federal statistical agencies by the mandate to disseminate information and the responsibility to protect the privacy of individuals and establishments described by the data.

Participants come from NISS, CMU and the LANL. Partner Federal agencies are the BLS, Census, National Agricultural Statistics Service (NASS), NCES and National Center for Health Statistics (NCHS).

During the past year, algorithms for geographic aggregation were created and incorporated in a software prototype delivered to the NASS, which plans to implement it as part of their data dissemination system. These were accompanied by a detailed analysis of statistical implications of aggregation.

Currently, table servers are a principal focus of project research at NISS. A first prototype, capable of handling tables with as many as 20 variables, has been constructed. Scalable methods for computation of bounds for table entries (in terms of released marginals) have been developed and implemented for the special case that the released marginals constitute a decomposable model.

Work continues at CMU and LANL to develop a Bayesian framework for confidentiality protection, accounting for the value as well as the risk of releasing information.

The project Web site — www.niss.org/dg — contains additional information.

Computer Model Evaluation. This research consists of two complementary projects. Participants come from NISS, GM, Duke University (Duke) and NCSU, led by Jerome Sacks (Duke) and James Berger (Duke). Additional information is available at www.niss.org/frg.

Statistical Framework for Evaluation of Complex Computer Models. This three-year project funded by the Focused Research Group program of the Division of Mathematical Sciences (DMS) at NSF is creating a unifying statistical framework for model evaluation.

Mathematically/Statistically Based Validation Systems. This is an 18-month project funded by NISS affiliate GM. Its goal is to develop a strategy for the evaluation of GM computer models, in cooperation with GM scientists, and to implement the strategy on testbed problems. Attention to date has focused on finite element models to predict the size of spot welds and for vehicle behavior in crash tests.

2.2 New Projects

Collectively, these new projects represent a significant diversification of NISS' Federal funding: not one is supported by the National Science Foundation.

Building on the affiliates workshop and associated report, two projects focus on data quality (DQ).

Initial Research on Data Quality. This project, funded by the BTS, examined DQ issues — and paths to resolve them — for the BTS' Intermodal Transportation Database (ITDB). One recommendation is that BTS develop a software toolkit to automate generic aspects of DQ assessments and systematize application of relevant domain knowledge. Alan Karr and Ashish Sanil were the project personnel.

CEIS. NISS continues to provide advice to the EPA's Center for Environmental Information and Statistics. Research in 2001 focused on issues of DQ for EPA's Toxic Release Inventory (TRI) database. NISS participants were Alan Karr, Ashish Sanil and Jerome Sacks.

Two new projects supplement, either in spirit or explicitly, the ongoing DG project.

Confidentiality Edits. Under this NCES-funded project, NISS will review and evaluate methodologies for data swapping as a means of protecting confidentiality, and recommend to NCES strategies for data

swapping. The approach will be to use Bayesian methods and simulation to characterize and compare swapping strategies. This will allow assessment of the effects of parameters such as swapping rates and of other characteristics of swapping strategies and comparison of swapping to other disclosure limitation strategies.

Aggregation Supplement. Aggregation (for example, of neighboring geographical regions) is a principal means of ensuring confidentiality. The thrust of this supplement to the DG project is aggregation for multiple variables, with particular attention to tradeoffs between protection of confidentiality, which is maximized by high aggregation, and loss of information, which is minimized by low aggregation.

Two new projects represent continuing NISS activity in transportation.

Variability-Sensitive Measures of Transportation System Performance. This project, funded by BTS and led by Nagui Roupail (Civil Engineering, NCSU) and Jerome Sacks (Duke) has the goal of developing measures of transportation system performance that explicitly incorporate variability. The use of computer simulations to estimate variability links this effort with those on computer model evaluation (§2.1).

Traffic Forecasting for North Carolina. This project, funded by the North Carolina Department of Transportation (NCDOT) and led by John Stone (Civil Engineering, NCSU), builds on the travel demand project (§2.3) to address a variety of traffic forecasting issues. Examples include localization of national forecasts to North Carolina and use of computer models to develop project-level forecasts.

A final project built on relationships established during NISS' research on software engineering.

Segmentation of Web Site Visitors. This project was funded by Visual Insights, a Lucent Technologies spinoff specializing in visual software tools for analysis and reporting of E-commerce data. Its principal product was fast, scalable methodology for segmentation of Web site visitors, based on a coarse categorization of pages on the site; *K*-means clustering is used to do the segmentation. Alan Karr and Ashish Sanil were the project personnel.

2.3 Completed Projects

Two major efforts and one smaller effort were completed during the year.

Transportation. This \$5.9 million project, funded by the Directorate for Physical and Mathematical Sciences at the NSF, has concluded. During 2001, the final component of the research addressed validation of computerized traffic simulations and use of these models to design and evaluate plans for traffic signalization. The work was led by Nagui Roupail (Civil Engineering, NCSU) and Jerome Sacks (Duke).

Large Data Sets. This collaboration with GSK, led by Jerome Sacks (Duke) and Stanley Young (GSK), came to an end with statistical analysis of the three-way interactions among chemical structure, biological activity and target protein structure.

Travel Demand. Funded by the NCDOT and led by John Stone (Civil Engineering, NCSU), this project examined the use of property tax information in estimating travel demand. Its principal finding is that tax data alone are not rich enough to reproduce the costly but fine-grained condition assessments made during on-site "windshield surveys." Whether additional data (for example, vehicle ownership) would produce sufficient improvement remains an open question.

2.4 Proposals

Two major proposals and one smaller one remain pending:

SAMSI. The proposal of Duke, NCSU, UNC and NISS, as a consortium, to establish a Statistical and Applied Mathematical Sciences Institute (SAMSI) was submitted in January of 2001, and will receive a site visit on November 5-6, 2001. The principal effects of SAMSI on NISS will be scientific interaction and potent input to the project generation process. SAMSI will also stabilize costs, by being housed at NISS and paying for the space and services, including staff, that it uses.

Digital Government II. A broadened follow-on proposal to the current DG project entitled *Data Confidentiality, Data Quality and Data Integration for Federal Databases: Foundations to Software Prototypes* was submitted to NSF/CISE in July of 2001. Stephen Fienberg of CMU and I are co-principal investigators.

The proposal is for a large-scale, cross-disciplinary research to create abstractions, theory, implementable methodology and software prototypes to meet three central, interacting, data-driven challenges facing Federal statistical agencies — data confidentiality (DC), DQ and data integration (DI). It will lead to effective, credible ways, ranging from theory and methodology to prototype software, to ensure DC in the face of strong, even competing, concern about DQ and the growing need and capability for DI.

The research will be carried out by statistical and computer scientists from NISS, CMU, the University of Maryland College Park (UMd), the Institute for Social Research at the University of Michigan, Purdue University, Southern Methodist University and LANL. Five leading Federal statistical agencies — BLS, BTS, Census, NASS and NCES — are partners in the project.

BTS Confidentiality. Building on NISS' increasing reputation for research on confidentiality, an initial proposal entitled *Disclosure Limitation for Tabular Data* was submitted to the BTS in August, 2001. Initially, NISS will review and report to the BTS on the current state of theory, computational implementation and practice regarding protection of the identity of subjects of tabular data. Then, based on the results of Task 1 and discussions with BTS, NISS will identify and pursue research issues at four levels — development of new theory and methodology, computational implementation, statistical consequences of disclosure limitation strategies, and empirical study of the behavior of methods and algorithms on BTS databases.

2.5 Planned Proposals

IDAP. A NISS-led proposal to EPA, in response to the call for proposals on *Particulate Matter (PM) Super-sites Integrative Data Analysis Project (IDAP)* was submitted in January, 2001. It involved personnel from Clarkson University, NCSU and NISS. All proposals submitted were declined by EPA. A revised call for proposals is anticipated in April, 2002, as is another NISS proposal in response.

Information Technology Research (ITR). Together with computer scientists from the UMd, Washington University and elsewhere, NISS is part of a proposal to be submitted in November, 2001 to NSF's ITR competition. The central focus of the research will be use of instrumented software applications to perform distributed user/usage/environment profiling, especially for embedded software systems.

Homeland Security. Discussions are underway to chart a path for NISS to support the statistical sciences community's response to needs for homeland security that is consistent with our mission and strengths and does not infringe on other groups.

2.6 Initiatives with Other Research Organizations

NCGBC. NISS is a charter member of the North Carolina Bioinformatics and Genomics Consortium (NCGBC) organized by the North Carolina Biotechnology Center (a TUCASI campus neighbor of NISS). The mission of the consortium is to focus attention on bioinformatics and genomics research in North Carolina and to stimulate collaborations among its members, which include universities, corporations ranging

from start-ups to well-established, and research organizations such as NISS. Benefits to NISS to date are visibility and an emerging relationship with the CIIT Centers for Health Research.

CIIT. Meetings have been held and are scheduled between NISS and the CIIT Centers for Health Research (in RTP) to explore NISS' providing statistical support for CIIT research in genomics (for example, microarray analysis or genetic effects of environmental toxins).

EURANDOM. Discussions continue with EURANDOM to establish and implement sensible forms of interaction. The two under most active consideration are a joint workshop on large data sets and a postdoctoral exchange program.

DIMACS. Through the efforts of Jon Kettenring, discussions have been initiated with the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) on potential joint efforts involving data and information quality.

3 Personnel

3.1 Postdoctoral Fellows

As of October, 2001, there are five postdoctoral fellows resident at NISS.

Adrian Dobra (Ph.D., Statistics, CMU) and **Shanti Gomatam** (Ph.D., Statistics, Florida State University) both arrived at NISS in the fall of 2001, and both are involved in the DG project, the NCES confidentiality edits project and the anticipated BTS project on confidentiality for tabular data.

Marc Kennedy (Ph.D., Statistics, Nottingham, UK) arrived at NISS in October, 2000, and is engaged in the two projects on computer model evaluation. **Li Liu** (Ph.D., Statistics, University of Illinois at Urbana-Champaign) arrived at NISS in October, 2000, and participates in research on analysis of gene microarray data, in collaboration with GSK. **Ashish Sanil** (Ph.D. Statistics, CMU), has been at NISS since 1998, working on the DG project, on which he now functions as a co-PI.

Jennifer Pittman, who arrived in May, 2000 and worked collaboratively with statisticians and scientists from GSK on the "three-way problem" (§2.3), has assumed a postdoctoral appointment in the Institute of Statistics and Decision Sciences at Duke, but remains associated with NISS.

Other initiatives are weekly meetings of the postdocs and me, at which we discuss items ranging from research to NISS initiatives such as SAMSI to current events, and explicit inclusion of proposal writing as part of the NISS postdoctoral experience.

3.2 Staff

Currently, the NISS staff consists of three uncommonly dedicated and efficient individuals. **Martha Williamson**, Administrative Assistant, in addition to providing strong support for me, has substantially assumed duties as business manager of NISS. **Katherine Kantner** has major responsibilities in connection with the Affiliates Program, the Board of Trustees and communication (§4). **James Thomas**, part-time computer system manager, provides outstanding system, software and network support.

4 Communication and Outreach

NISS Newsletter. The quarterly newsletter is distributed by both mail and the NISS Web site (specifically, www.niss.org/newsletter.html). Katherine Kantner now has editorial responsibility, and with James

Thomas oversees the NISS Web site.

Monthly updates from me are sent to the Board of Trustees and Members of the Corporation. Separate updates are sent to the Affiliates.

JSM Events. Continuing a practice instituted in 2000, at JSM 2001, Jon Kettenring and I met with the Board of Directors of the ASA, COPSS and the IMS Council, to provide updates about NISS. Nearly thirty NISS affiliates attended the August 5 meeting (§1). More than 150 people attended the now-annual NISS Reception on Monday evening, August 6. A NISS-sponsored invited session on August 8 summarized findings and lessons learned from the transportation project (§2.3).

Sacks Award. By action of the Board of Trustees in November, 2000, the *Jerome Sacks Award for Cross-Disciplinary Research* was created to bring visibility to NISS and to recognize Jerry's service as founding Director. The recipient receives a cash award of \$1,000 and a certificate. A plaque in the NISS building lists recipients.

A Board of Trustees committee (Alicia Carriquiry, Douglas Nychka, Frank Rockhold) selected Elizabeth Thompson of the University of Washington as inaugural recipient. The award was announced at the NISS reception at JSM 2001. Although Thompson was not able to be present in person to receive the award, she provided an eloquent statement of acceptance.

Through the efforts and generosity of many individuals (including a Board committee of Mary Ellen Bock, Ingram Olkin and Anne Petersen), nearly \$29,000 has been raised to date to endow the award.

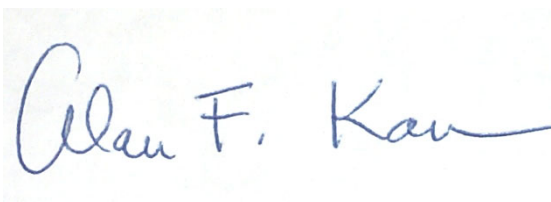
5 Summary and Personal Comments

NISS has made progress over the past year, and we remain on a path for continued progress. I am proud of what we are able to accomplish, both scientifically and for the professional development of NISS researchers (especially our postdocs), on a relatively small budget.

Many people (and I hope even more in the future) have contributed to this. First among these is Jon Kettenring, whose vision and faith in the affiliates program and extraordinary service as Chair of the Board of Trustees have revitalized and energized NISS. The Executive Committee of the Board has been a constant source of guidance and advice to me. Many others on the Board, through service on committees, involvement in NISS research or simply a positive attitude about NISS, have made an enormous difference.

Finally, I thank the NISS postdocs and staff, with whom I work on a daily basis, for their support. They help make the job fun and rewarding.

Respectfully submitted,

A handwritten signature in blue ink that reads "Alan F. Kaw". The signature is written in a cursive style with a long horizontal flourish at the end.