National Institute of Statistical Sciences

PO Box 14006 Research Triangle Park, NC 27709–4006 Tel 919.685.9300
FAX 919.685.9310

Annual Report of the Director

Alan F. Karr
December 27, 2000

Contents

1 Affiliates Program

The new Affiliates program is an extraordinary show of support for NISS by the statistical sciences community. Thanks especially to the energy of Mary Ellen Bock, Jon Kettenring, Jerome Sacks and George Williams, it is off to a vigorous start. Within one year, 17 corporations, 9 government agencies and national laboratories and 17 university departments have joined. The first two classes pay $10,000 annually; university departments pay $3000.

In addition to immediate programmatic and financial impact on NISS, the early success of the Affiliates program was the essential factor in NISS' receiving additional unrestricted funds from the Research Triangle Foundation.

Activities in 2000. These include:

- *Kickoff Meeting*, held at NISS on March 3, 2000, at which initial program activities — bioinformatics, computer model evaluation, data confidentiality and data quality — were selected.
- *Workshop on Gene Expression Data*, held at NISS on July 13, 2000. In light of the success of this event, workshops on focused topics in bioinformatics will continue on a semi–annual basis.
- *Affiliates Meeting at JSM*, held in Indianapolis on August 13, 2000, at which progress of the program and plans for future activities were discussed.
- *Workshop on Data Quality*, held at Telcordia Technologies (Morristown, NJ) on November 30 – December 1, 2000.

Details (including presentations) for many of these events are available from the Affiliates Web site: `www.niss.org/affiliates/affiliatesmain.html`.

An *Affiliates Advisory Committee* has been formed, to provide input and assistance to the Director. Initial members are David Banks (Bureau of Transportation Statistics), Richard Beckman (Los Alamos National Laboratory (LANL)), Nancy Flournoy (American University), James Landwehr (Avaya Labs), David Siegmund (Stanford University) and George Williams (Merck & Company).

An announcement about the program, listing charter affiliates and inviting others to join, appeared in the December, 2000, issue of *AMSTAT News*.

Plans for 2001. A full program of Affiliates' activities is planned for 2001, including:

- *Workshop on Pharmacogenomics* to be held in February 12–13 at Amgen (Thousand Oaks, CA).
- *Workshop on Network Data*, to be held at NISS in March or April. Stephen Marron of the University of North Carolina at Chapel Hill (UNC) is playing the lead role in organizing this activity. An organizing committee is in place, and the program is being planned.
- *Workshop on Data Confidentiality in the Internet Age*, to be held at NISS in May, with RTI playing a central role.
- *Summer Internship Program*, which will bring graduate student interns to NISS and Affiliates who wish to participate. Participating interns will meet late in the summer at NISS to share their experiences.
- *Workshop on Web and E-Commerce Data*, to be held at NISS in the Fall. In part, this will be a follow-on to the March workshop on network data.

Meetings of the Affiliates will be held in March at NISS and August at JSM 2001 in Atlanta.

Other activities, especially those focusing on moving beyond workshops to proposals and research projects, are under development.

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

Current projects (continuing and new) are summarized in Table 1.

There are two large ongoing projects.

Transportation. This $5.9 million project, funded by the Directorate for Physical and Mathematical Sciences at the NSF, is in its final year. During 2000, research was completed on (1) Validation of regional–level models for mode and route selection; (2) Models that link the transportation to inter-regional commodity flows; and (3) The interaction between permeability and deterioration of concrete.

Work led by Nagui Rouphail (North Carolina State University (NCSU)) and Jerome Sacks (NISS) is continuing on the use of computer simulation models — specifically, CORSIM model developed by the Federal Highway Administration — to design and (prospectively) evaluate plans for traffic signalization. NISS–designed signal settings have been implemented on an experimental basis in part of Chicago, and succeeded at improving traffic flow.

Digital Government. This three-year $1.25 million project funded by the Directorate of Computer and Information Sciences and Engineering at the NSF is seeking to use statistics, information technology and domain knowledge, in collaboration, to alleviate the tension imposed on Federal statistical agencies between the mandate to disseminate information and the responsibility to protect the privacy of individuals and establishments described by the data.

Participants come from NISS, Carnegie Mellon University, Ohio State University, LANL and MCNC. Partner Federal agencies are the Bureau of Labor Statistics (BLS), the Census Bureau, the National Agricultural Statistics Service (NASS), the National Center for Education Statistics (NCES) and the National Center for Health Statistics (NCHS). Alan Karr (NISS) is project director.

Thrusts of the research during the first year of the project have been (1) Algorithms for geographic aggregation, incorporated in a NASS prototype (available at `niss.cnidr.org`); (2) Analysis of statistical implications of aggregation; (3) A prototype and preliminary design specifications for table servers; (4) Scalable methods to compute bounds for table entries and (5) 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.

Smaller ongoing projects are several.

Travel Demand. The transportation project is complemented by a small project funded by the North Carolina Department of Transportation, led by John Stone (NCSU) and Alan Karr (NISS), that is examining the use of property tax information to estimate travel demand.

Large Data Sets. Work on drug discovery, in collaboration with GlaxoWellcome, is continuing. Current activity, led by Jerome Sacks (NISS) and Stanley Young (GlaxoWellcome), is focusing on three–way interactions among chemical structure, biological activity and target proteins.

CEIS. NISS involvement in environmental research continues, in the form of advice to the EPA concerning their Center for Environmental Information and Statistics.

Postdoc Support. The existing NSF grant for "generic" postdoc support, awarded in the 1997 Group Infrastructure Grants competition, is in its last year. Over the past year, postdocs working on digital government, computer model evaluation, drug design and bioinformatics have been supported, the latter in

Project	Agency	**Duration**	**End Date**	**Funding**
Transportation	NSF	6 years	9/30/01	$5,900,000
Digital Government	NSF/BLS/Census/GSA/NASS/NCES	3 years	9/30/02	$1,400,000
Travel Demand	NCDOT	1 year	6/30/01	$27,000
Large Data Sets	NSF	4 years	8/31/01	$800,000
CEIS	EPA	1 year	9/30/01	$50,000
Postdoc	NSF (GIG)	4 years	8/31/01	$400,000
Computer Model Evaluation	NSF (FRG)	3 years	7/31/03	$900,000
Computer Model Evaluation	General Motors	1.5 years	4/30/02	$435,000

Table 1: *Current Projects.* Total funding amounts, which include both direct and indirect costs, are approximate. Depending on scientific circumstances, some projects may be extended beyond the dates shown.

conjunction with Duke University (Duke). This award and its predecessor have been instrumental in maintaining a rich scientific environment at NISS.

2.2 New Projects

Two new projects are addressing the evaluation of complex computer models of physical phenomena.

Statistical Framework for Evaluation of Complex Computer Models. This three–year, $900,000 project funded by the Focused Research Group program of the Division of Mathematical Sciences at NSF, began on September 1, 2000, and is creating a unifying statistical framework for model evaluation. The foundations of the framework are quantification and the creation of theory and methods that allow dual use of data in both estimation of model inputs and evaluation of outputs.

The project team consists of researchers from NISS, Duke, and NCSU, with Jerome Sacks (NISS) as principal investigator. Additional information is available on the project Web site: `www.niss.org/frg`.

Mathematically/Statistically Based Validation Systems. This is an 18–month, $435,000 project funded by NISS affiliate General Motors (GM). Its goals are to define and develop a strategy for the evaluation of GM computer models, in cooperation with GM scientists, and to implement the strategy on testbed problems. Central research issues are association of confidence limits to predictions of computer models; uncertainty estimates for predictions "beyond the data;" either in the sense of predicting over a new range of inputs or predicting with a variant of already studied models; and determination of sensitivities (or possibly faults) in model components.

Project participants come from NISS, GM, Duke and NCSU, led by Jerome Sacks (NISS) and James Berger (Duke).

2.3 Completed Projects

Three major efforts were completed during the year.

Software Engineering. The NISS–Lucent Technologies project *Code Decay in Legacy Software Systems* was completed March of 2000. Stephen Eick (Lucent) and Alan Karr (NISS) were project leaders.

This project quantified, measured and predicted the previously elusive phenomenon of code decay, that is, the increasing difficulty to change large software systems over time without negative consequences. Products of the research included one patent — for *Live Documents* that define new modes of interaction between readers of scientific documents and the underlying data — and nearly two dozen papers.

Statistically Based Activity Generation. NISS created the activity generation model for the TRANSIMS model developed by LANL. TRANSIMS, featured on the back page of the January, 2001, issue of *Wired* magazine, represents regional-level transportation from people who live there to the pollution consequences of the traffic they generate. The NISS–built *activity generator* synthesizes, for each person in the region, the daily activities that generate the need to travel. Statistical challenges were the scale of the activities to be generated (leading to millions of trips per day), capturing dependences (for example, shared rides) among patterns for different individuals and accommodating feedback to activities from TRANSIMS modules that generate routes for individual travelers.

Paul L. Speckman (Missouri) led the project.

Network Intrusion. This component of the Large Data Sets project (see also §2.1), a collaboration with AT&T Labs Research, examined statistical methods for detection of network intrusion into computer systems, focusing on a large experiment comparing multiple methods. A secondary activity was development of models to predict defection of ISP (Internet service provider) customers, which entailed particular challenges to deal with massive data sets in multiple relational databases.

This research was led by Daryl Pregibon (AT&T), William DuMouchel (AT&T) and Alan Karr (NISS).

2.4 Proposals

A proposal *A Visualization Infrastructure for Science and Engineering Indicators* was submitted to the NSF program on Analysis of Science and Technology Resources on September 15, 2000. The research addresses Web-based dissemination and visualization of science and engineering indicators. Alan Karr is principal investigator.

Two major proposals are in progress.

IDAP. NISS is developing a proposal for EPA, in response to the call for proposals on *Particulate Matter (PM) Supersites Integrative Data Analysis Project*. The focus is on management, integration, visualization and analysis of data from EPA's eight PM supersites (monitoring airborne particulate matter), as well as other data sources. Personnel from Clarkson University, NCSU and elsewhere are involved. Alan Karr (NISS) will be project director.

SAMSI. Duke, NCSU, UNC and NISS, as a consortium, are submitting a proposal to the current NSF solicitation for Mathematical Sciences Research Institute to establish a *Statistical and Applied Mathematical Sciences Institute* (SAMSI).

SAMSI will focus on problems at the interface of statistics, applied mathematics and other disciplines, addressing the very hardest and most important data– and model–driven scientific problems of our times. Necessarily, these problems are of a scope and scale such that they cannot be adequately addressed by individuals, groups or single disciplines. SAMSI will conduct programs that identify the scientific areas most in need of attention and most amenable to progress, bring together established and young researchers from academia, industry, national laboratories and government to define and formulate the central problems, and catalyze the research needed to address those problems.

The principal effects on NISS will be reduction of costs: SAMSI will be housed at NISS, and will pay for the space and services, including staff, that it uses. SAMSI will also provide potent input to the project generation process at NISS.

In addition, a proposal to NSF is being prepared for a postdoctoral program featuring strong involvement of the NISS affiliates.

A follow-on proposal to the Digital Government program at the NSF is planned for submission in July of

Topic	Agency	Status	Start Date	Duration	Amount
S&E Indicators	NSF (ASTR)	Submitted 9/15/00	4/1/01	1.5 years	$130,000
IDAP	EPA	Due 1/17/01	7/1/01	4 years	$900,000
SAMSI	NSF	Due 1/16/01	7/1/02	5 years	$1,500,000
Postdoc	NSF	Due Spring 01			
DG Follow-On	NSF	Due 7/01			

Table 2: *Pending and Planned Proposals.* The SAMSI proposal will be submitted by Duke; the amount shown is for the NISS sub-award only. Omitted items are to be determined.

2001, emphasizing theory and methodology questions generated by the current project, but not addressable within the scope of that project.

On longer time scale, preliminary discussions are underway regarding NISS responses to two anticipated NSF initiatives — mathematical sciences (Federal FY 2002) and social sciences (Federal FY 2003).

Proposals are summarized in Table 2

2.5 Other Research Initiatives

Discussions are taking place with the NCES to organize a set of evaluation workshops for NCES' new statistical standards (currently under development).

Negotiations have been completed with Visual Insights, a Lucent Technologies spinoff at which Stephen Eick, co-principal investigator for the Software Engineering project (§2.3)), for a small project involving statistical analysis of E-commerce transaction data.

3 Personnel

This has been a year of transition.

3.1 NISS Leadership

I replaced Jerome Sacks as Director on July 1, 2000. Since then, Sacks has remained active in the management and development of the Affiliates program. As of January, 2001, he will assume on a full-time basis his position as Professor of Statistics and Decision Sciences at Duke, but will retain a leadership role in NISS projects on computer model evaluation.

Currently, the Associate Directorship of NISS is vacant, as a number of issues are explored that impact the future nature of the position, the funds available and potential candidates.

3.2 Postdoctoral Fellows

As of December, 2000, there are five postdoctoral fellows. **Marc Kennedy** (Ph.D., Statistics, Nottingham, UK) arrived at NISS in October, 2000, and is involved in the projects on computer model evaluation. **Li Liu** (Ph.D., Statistics, University of Illinois at Urbana-Champaign) arrived at NISS in October, 2000, and is involved in research on computer model evaluation. **Brian Park** (Ph.D., Transportation Engineering, Texas A&M) has been at NISS since 1999, working on design and evaluation of traffic signalization plans for urban street networks, and associated issues of model evaluation. He will leave NISS in January of 2001, to

assume a faculty position at the University of Virginia. **Jennifer Pittman** (Ph.D., Statistics, Pennsylvania State University) arrived in May, 2000. She is involved in collaborative research with statisticians and scientists from GlaxoWellcome on the "three-way problem," as well as research on computer model evaluation. **Ashish Sanil** (Ph.D. Statistics, Carnegie Mellon), has been at NISS since 1998, working on the Digital Government project. He has assumed a key role in the project, developing geographical aggregation methods to disclose information derived from Federal databases, but maintain confidentiality of the data themselves, and in creation of table servers.

Jaeyong Lee (Ph.D., Statistics, Purdue University), who spent two years at NISS, became Assistant Professor of Statistics at Pennsylvania State University in August, 2000. During his last year at NISS, he worked on Digital Government, focusing on statistical consequences of geographical aggregation procedures. **Alex Stark** (Ph.D., Engineering, Cambridge, UK), who worked on bioinformatics as part of the large data sets project, let NISS in January of 2000 to assume a position in the private sector, at BOPS, Inc.

3.3 Summer Interns

For unfortunate reasons (the Los Alamos fire), but to the immense benefit of NISS, two graduate students who had planned to be interns at LANL spent the summer at NISS instead. **Karen Brady** (George Washington University) led a study of risk reduction methods for tabular data as part of the Digital Government project. **Christopher Holloman** (Duke) provided major statistical and simulation support for the Digital Governments project's study of the statistical consequences of geographical aggregation. As noted in §1, the experience was so positive that NISS will institute its own summer internship program in 2001.

3.4 Staff

Currently, the NISS staff consists of three extremely strong individuals. **Martha Williamson**, Administrative Assistant, in addition to providing strong support for the Director, is assuming duties as business manager of NISS. Not only does this resolve issues raised during the FY 1999 NSF audit of NISS, but also it removes some responsibility for financial management from the Director. **Katherine Kantner**, Staff Assistant, joined the staff in August of 2000, replacing Shirley Griffiths. Katherine has major responsibilities in connection with the Affiliates Program, the Board of Trustees and communications initiatives (§4). **Andrew Eberhart**, part-time computer system manager, provides system and network support.

4 Communication and Outreach

An essential step in improving relationships between NISS and the statistical sciences community is proactive, effective communication and outreach. Multiple initiatives have been launched this year:

NISS Newsletter. A quarterly newsletter has been instituted. Mailings will supplement Web distribution (`www.niss.org/newsletter.html`).

Monthly updates on activities at NISS, events of interest and issues faced by NISS are sent to the Board of Trustees and Members of the Corporation. A separate update is sent to the Affiliates.

NISS Web site. The Web site — `www.niss.org` — has been redesigned completely, to be easier to use and more effective at conveying the NISS story. Software to summarize usage will be installed soon.

JSM Events. At JSM 2000, Jon Kettenring and I met with the Board of Directors of the ASA, COPSS and the Executive Committee of the IMS, to provide updates about NISS. Nearly thirty NISS affiliates the

August 13 meeting (§1). More than 150 people attended the NISS Reception on Monday evening, August 14. This will become an annual event. The NISS session on "Statistics and Information Technology" had 75 attendees. The JSM 2001 session will be on lessons learned from the transportation project (§2.1).

AMSTAT News. The December, 2000, issue contains an interview of me, in my role as the new Director of NISS, as well as the announcement about the affiliates program (§1).

IMS/NISS Initiatives. Discussions are underway between the IMS and NISS regarding joint activities of interest to both organizations.

Tenth Anniversary Brochure and Poster. In recognition of the tenth anniversary of the founding of NISS (at a ceremony on December 3, 1990), and also for other communications and public relations purposes, a brochure celebrating important events and people during the first ten years of NISS, as well as the vision for the future, is being created, and is expected to be complete early in 2001. A companion poster will be produced as well.

5 Finances

The past year has seen aggressive efforts to diversify NISS' sources of income and to reduce costs. Income from the affiliates program and an additional grant of unrestricted funds from the Research Triangle Foundation provide a new financial stability for NISS.

6 The Future of NISS

In many ways, the future of NISS is much more positive than it seemed during the summer of 1999. At the moment, two interwined developments define my view of the future.

First, the Affiliates program is clearly the most important contributor to the improved outlook. It benefits NISS financially, will provide essential input to the project generation process, and is powerful evidence of the community's support for NISS.

Second, SAMSI, if successful, shares many of these characteristics. There will be financial benefit to NISS, but much more important are the input to project development and the freedom for NISS to concentrate on those activities at which we are best — enlarging the future of statistics by (as our mission statement articulates) identifying, catalyzing and fostering high-impact cross-disciplinary research involving the statistical sciences.

Nevertheless, daunting challenges remain.

We must continue to attract the best people to become involved with NISS, and on an even broader scale than in the past. One of two ultimate criteria on which we should be judged is how NISS benefits those who are involved (in whatever role) with us.

We need to develop a research strategy for NISS that combines long–term leadership positions in selected core research areas with the flexibility to respond opportunistically to emerging issues. The second principal criterion on which NISS should be judged is the quality, impact and relevance of its research (as well as the extent to which research in areas initiated by NISS is taken up by others in the community), and there must be relentless attention to each of these.

NISS must continue to improve our image and relationships with the statistical sciences community. Not only must NISS be of value to the community, but also it must be seen by the community to add value

to the profession. An essential element of this process is diversify NISS' activities in ways that serve more segments of the statistical sciences community.

Finally, this is the time to push hard to attain financial stability for NISS. To some extent, NISS will always be subject to the vagaries of Federal research funding, but with programs that generate the right resources for project development, we can aim high, with exciting, innovative proposals.

I am, and will remain, optimistic that NISS can and will meet the challenges.

Respectfully submitted,