ADVANCED COMPUTING IN NASS

Brian Carney, U.S. Department of Agriculture
1400 Independence Ave, SW, Washington, DC 20250

KEY WORDS: agricultural statistics, statistical computing

The survey and research programs of the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture, require collection and processing of data from a variety of sources: paper questionnaires, CATI and CAPI instruments, even remotely sensed satellite data and digital map data. To manage the analysis and dissemination of these data NASS staff use a wide variety of computing systems, from PCs to workstations to mainframes and supercomputers. A description of current research projects that use these different systems is the topic of this paper. The future directions indicated by our research are also presented.

Introduction

The National Agricultural Statistics Service (NASS) is responsible for collecting and disseminating agricultural statistics for the United States. This requires an elaborate survey program that involves NASS offices in 44 field offices. Managing the collection and dissemination of these data requires the use of computer systems that have become increasingly sophisticated. This paper describes some historic approaches to survey data processing in NASS, briefly describes current approaches, and focuses on some of the research projects in computing and survey data processing.

An Historic Perspective

The notion of 'advanced computing' is ridiculously fleeting these days. Systems I thought were quite advanced last year are now dinosaurs. So allowing for this rapid aging, consider some of the approaches that NASS has considered advanced.

NASS began, as did many survey organizations, by collecting information on paper, then tabulating the results. First calculations were performed by hand; later the work was done on tabulating machines and calculators. The agency continues to use paper forms for much of the survey data collection though the data is copied off the paper to the computer more quickly these days.

One imaginative method of tabulating increasing numbers of survey forms was the use of peg strips. A peg strip is a three-foot long metal bar, studded with pegs at regular intervals. The survey forms were punched along the top with holes that matched the spacing of the pegs. Forms were arrayed across these strips so that the response boxes were visible in rows across the stack of forms. The tabulator could then key in the response data reading across the neat rows of boxes. Forms attached to the peg strips were handled and stored as a unit.

NASS first used digital computers in the late 1960s, although punch cards had been used for tabulating since the late 1940s. The punch card served until the early 1980s when the last keypunch in the Washington, D.C. offices was removed.

The early use of an IBM 704 by the Agency evolved into the use in the 1970s of large commercial mainframe systems with access through a dedicated telecommunications network. The first computer center in the Washington USDA complex was an outgrowth of the first mainframe center set up by NASS.

NASS structure is based on offices in most of the states. These 44 state statistical offices (SSOs) operate in cooperation with the state departments of agriculture. Because of their joint funding between USDA and the states that they serve, the offices enjoy varying degrees of autonomy. This means that some offices have the opportunity to use a variety of systems to manage their data processing. Several of the SSOs have their own minicomputers or use state-operated mainframe systems in conjunction with the NASS mainframe systems. Important innovations such as the use of large database systems and the use of personal computers were first tried by SSOs before being introduced agency-wide. The first IBM PC was purchased by Headquarters about 1984. Previous work had been done on 8-bit CP/M systems such as those built by Radio Shack and Zorba as early as 1982. The first Macintosh I know about in NASS was purchased by the Iowa office in 1988.

NASS, like most organizations, recognizes how useful advanced technology can be to survey operations. And like most groups, technology advances and the accompanying changes are met with a mixture of excitement and foreboding. There is, however, a clear direction towards end-user computing; that is, that the person who needs an information product can initiate, control, and adapt the computer procedures that prepare the product.

Current Operational Processing

The operational side of NASS uses a variety of approaches for data entry. The SSOs use direct data entry (DDE), transcribing data from paper forms into key-to-disk systems or DDE programs running on multi-user microcomputers or PCs. Direct entry of data at the point of collection features computer-assisted telephone interviewing (CATI) and computer-assisted personal interviewing (CAPI). Both use the CASES system developed in cooperation with the University of California at Berkeley. Fourteen SSOs are operating CATI systems now, with most telephone surveys conducted by NASS using CATI. NASS expects to implement CATI in all SSOs over the next few years using it in virtually all the telephone surveys. The CATI systems operate on multiuser UNIX microcomputers and PC networks. The work with computer-assisted personal interviewing (CAPI) will be covered later in this paper.

Paper questionnaires are still in wide use. Methods for preparing them have progressed to the use of computer document processing systems for creating camera-ready copy. Future work will allow electronic distribution of the copy for printing at each SSO. The electronic systems will lead to customizing questionnaires for individual respondents, including historic information and tailoring the questions to the specific situation.

Most editing and summarizing of survey data is done in batch mode on the mainframe. The mainframe systems were first special-purpose programs written to process a single survey or type of survey. This evolved into what NASS calls *generalized systems*. A generalized system uses a parameter language to direct the operation of an edit or summary program. This innovation was designed to make it unnecessary to rewrite programs when survey designs changed. As it turned out, the parameter languages did not look much like programming languages, and for large complex surveys it was nearly as difficult to rewrite the parameters as it would have been to rewrite the software. It also became difficult to use the generalized systems to process increasingly complex survey designs and compute some of the more elaborate estimators. In response, the agency has begun using SAS for operational processing. SAS saw its first use in NASS Research in the mid 1970s. Now it is used to process most of the major surveys. Most of the NASS PCs run SAS too, so there is inexpensive statistical processing available on the desktop. The work of editing the survey data should be distributed to the SSOs soon, both so there is more local control over the processing, and also to make better use of the inhouse computing capability.

In 1984 NASS began exploring the use of an agency-wide database system residing on the mainframe as a repository for statistical data. The project of implementing such a system is well underway, with a first survey having been processed and analyzed using the mainframe database system this past March. The database system has made possible interactive maintenance of the list sampling frame and overlap/nonoverlap resolution for the multiple-frame surveys.

Much of the effort of the operational data processing has been to shorten the time between collection of the data and the publication of the survey results. Interactive processing is being explored in conjunction with the usual batch edit and summary to speed up analysis and publication. Interactive computing should also bring more sophisticated analytical tools, especially graphical tools, to the survey statistician.

The NASS direction for computing is being built on one central idea: that there will be a computer at the desk of each staff member, and through that screen there is access to all the services and information required to do the job. All the NASS SSOs have at least one PC, and equipping each NASS staff member with a PC is an Agency goal. Right now, though, we are far from having a PC at every desk. All the PCs and other computers will be connected on a network to provide high-speed access to the many different data and computers required by the NASS programs.

Research Systems

NASS Research is in the awkward position of having to work on both sides of the technology fence. The operational systems never seem to be quite as advanced as researchers might like. NASS Research attempts to handle this problem by setting up data communications and staying knowledgeable about the operational systems used by the agency. We make an effort to fit new technology and methods requiring new formulations, sampling techniques, quality control systems, and graphics displays into the operational scheme. This way, innovations can be implemented as quickly as possible. Of course it can be very difficult, if not practically impossible, to implement new methods using the operational systems. However, when the product that the research offers is compelling enough, the operational systems can catch up.

The systems in use by NASS Research are quite varied. They include IBM and Compaq PCs, Sun workstations, a DEC minicomputer and a Zilog microcomputer. Considerable use is made of an IBM mainframe and a Cray supercomputer. Incidentally, the Cray gets its big workout processing satellite data. This reflects the varied requirements of the research program. Computing resources are scaled, that is, the best adapted system is used for each requirement. The diversity of the computing systems results from the diversity of the research areas.

CATI and CAPI both began in NASS within the research group. The CATI project began with an agreement with the University of California at Berkeley who developed the software and included capabilities required by NASS (Tortora, 1985). At the time NASS was ready to begin inhouse CATI research, multiuser UNIX microcomputers were just becoming affordable. In 1984 NASS purchased a 16-user Zilog microcomputer, the agency's first UNIX system. The experience with the Berkeley software and the UNIX platform has led to work in portable software design. The knowledge of UNIX made it possible to embrace the technology of Sun workstations which opened up the area of affordable interactive graphical analysis for research.

Aerospace remote sensing research has been conducted by NASS since the late 1970s (Hanuschak et al., 1982). The demands of processing satellite imagery, both in terms of data volume and computational complexity, led to the use of the Cray supercomputer (Ozga, 1984). NASS has also done research with a variety computer architectures including massively parallel processors, array processors and RISC. Today, the Cray, as well as systems from Sun, DEC, and IBM, are used for the remote sensing processing.

Remote sensing research and the use of satellite data for area sampling frame construction led to applications of sophisticated image display and image processing systems. Researchers at NASA's Ames Research Center were instrumental in helping NASS configure an image processing system using a Raster Technologies display and a Forward Technologies workstation in 1983. This system has been replaced with Sun computers, but the software has been enhanced to facilitate research into the construction of area sampling frames using digital satellite and map materials.

Research systems have been important to NASS by demonstrating how advanced computers can be used effectively in a survey organization. Research has managed inhouse computers including minicomputers and networked workstations quite successfully without full-time operations staff. Sophisticated data communications systems are being operated and managed. NASS Research is introducing engineering workstations into statistical analysis, with an emphasis on interactive, graphical data analysis. An important element of research is the ability to embrace new technologies quickly. This has been possible in NASS because of the emphasis of portable software systems based on standard operating systems and standard languages. The remote sensing

program has demonstrated the ability to process and analyze very large datasets, datasets much larger than the survey procedures typically create.

Current Research

Some of the projects that involve computing and computer systems are described in this section.

Computer-Assisted Personal Interviewing

Computer-assisted personal interviewing (CAPI) is an outgrowth of the NASS CATI operations. Just how practical is it to conduct our personal interviews using CATI instruments on laptop computers? We are researching the use of the Berkeley CASES software on IBM PC compatible laptops. This provides consistency with current agency CATI operations and since the PC is widely used in NASS there is considerable expertise in programming and using these computers and their software. These laptops have 24-line screens, so questions do not have to be abbreviated and can be consistently presented to the respondent by the enumerator. With 20 megabyte disks, there may be enough storage on a laptop for a day's work as well.

The research is uncovering considerations relating to the size of the survey instrument, the amount of data storage required, the speed of the computer during an interview, the technical problems with transmitting the data electronically or by diskette, and the level of training required of enumerators using the laptops. In the current demonstration study, two enumerators are collecting monthly livestock price data using laptops. The project should be expanded to include about twenty interviewers next year.

Computer-Assisted Area Frame Construction

Area sampling frames are developed as an adjunct to the list sampling frame, to provide complete coverage in NASS sample surveys. The current process for constructing these frames is tedious and time consuming, requiring something over twenty staff years to complete. With frames this expensive to construct, they tend to remain in place without being updated for at least a decade, and their efficiency surely degrades as agricultural patterns change.

Now NASS uses paper prints of satellite images of agricultural areas to assist with building area sampling frames. The Thematic Mapper (TM) data provides more current information than is usually available from the aerial photography that is also used in area frame construction. The paper TM products do not, however, provide enough resolution to pick out roads and all the needed sampling unit boundaries easily.

To attempt to solve some of these difficulties, NASS is experimenting with combining digital TM data with digital line graph (DLG) map data from the U.S. Geological Survey (Carney et al., 1987). The DLG data has roads, waterways, and other transportation features in digital form. Overlaying the satellite and map data on a high-resolution computer display system allows both sources to be considered simultaneously for frame construction: the TM data for clues to agricultural use, and the TM and DLG combined to determine sampling unit boundaries. The sampling units can be digitally outlined on screen using a mouse pointing device. This system produces an entirely digital area frame. As new digital information becomes available for a state, it will be possible to update required portions of the frame instead of reconstructing it entirely.

The digital area frame research builds on the extensive software system called PEDITOR that is designed to process remote sensing data (Ozga et al, 1986). Additional modules to display and manipulate the data TM satellite image data and to digitize the sampling unit boundaries on screen have been written. The new modules integrate the processing capability of PEDITOR with the computing and graphical analysis power of scientific workstations.

Once the frame construction process is available in digital form other work can follow. It will be possible to integrate area frame survey information with other geographic information into the frame construction process through ties to a geographic information system. New methods of area frame construction can be considered and tested, methods that are not limited by the requirements of paper-and-pencil methods. There is the potential for quicker construction of special-purpose area frames. It should be possible to build an expert system to assist in setting up strata definitions and selecting boundaries.

This three-year project is being conducted with the cooperation of the National Aeronautics and Space Administration.

Yield Laboratory Automation

Predicting crop yield is an important element of the NASS estimation program. Surveys of crop yield are conducted to gather the information required for these estimates. Some of the needed measurements on crops are made in the field, but many are made in NASS yield laboratories. Traditionally, measurements such as corn ear weight or moisture content were made by hand and recorded on paper forms. The forms were then keypunched before the data could be checked, analyzed and summarized. A project is underway to examine systems for automating the data collection.

Adding computers to the laboratories has been a challenge. The laboratories are extremely dusty so the usual office-style computer equipment and terminals are not usable. Because samples are processed in no particular order, each processing station has to operate more or less independently, but each station must be able to know what processing has already been done on a sample. NASS is trying a multi-user computer connected to industrial-style terminals that have bar code readers and membrane keyboards, and electronic scales and moisture meters that are designed for dusty conditions.

The manual scales are replaced with those having digital output ports. The standard moisture meters use electromagnetic methods of estimating grain moisture content. These meters require the technician to correct meter readings for temperature by reading from a calibration table. New moisture meters use near-infrared sensors with digital outputs so that measurements are made quickly, and automatically corrected for temperature.

The yield laboratory has several stations at which certain measurements are made on each grain sample. As a sample arrives at each station, the technician scans the identification tag with a bar code reader and the computer verifies that this is the correct station for that sample. Once the measurement is made it is sent electronically to the computer from devices that have digital output. When counts of some element of the sample have to be made by the technician, these data are keyed into the industrial terminals. Consistency checks flag improper or unlikely data so

that measurements can be made again immediately. Procedures are built in for monitoring the progress of the samples in the laboratory. Periodically, the completed sample measurements are sent electronically to the mainframe computer where the results are combined with those from other laboratories and summarized.

The system will be evaluated according to the changes in data quality, and staff and time required to conduct the laboratory measurements. A reduction in nonsampling errors is likely because error checking is performed at the point of data collection.

High-Resolution Satellite Sensors

NASS has been using digital satellite data as an adjunct to acreage estimation surveys since the early 1970s. NASS is now beginning research with the newer high resolution Thematic Mapper (TM) and French SPOT satellite data to determine how useful these new sensors can be for estimating major crop acreages.

The new sensors offer great improvements in the amount of information available, both in terms of improved spatial resolution (as low as 10 meters for the SPOT panchromatic sensor) and spectral resolutions (seven channels of 8-bit data for TM). This improvement comes at a corresponding increase in cost, of course, since the quantity of data is seven to ten times greater than the earlier sensors. NASS uses inhouse computer systems to great advantage in controlling these computing costs. Unfortunately, in terms of computational complexity the increase in data volume and dimensionality is hurtful. Methods for data reduction and new clustering and classification algorithms are being examined.

Expert Statistical Systems

There is considerable interest in NASS in developing expert statistical systems. NASS processes large amounts of survey data under increasingly tight deadlines. Some analyses of the data may not always be done because time is so short and the data volume is so large. There is also a growing shortage of staff who are be both statistician and subject matter specialist. NASS is actively seeking applications of expert statistical systems to its program areas where large amounts of staff time is being spent on fairly well-defined statistical analyses, for example, in data editing and validation. The hope is that routine problems can be handled by the machine and the more difficult problems can be flagged for the statistician.

In addition to data editing and validation, NASS is examining the potential for expert systems in automating the overlap-nonoverlap determination in the multiple frame surveys.

Technical Workstations

An important element in the progress of NASS statistical research is the use of technical workstations and computer networks. The role and progress of this type of system is described in Crecine (1986). Currently, a network of Sun workstations provides high-resolution graphics and fast computing power for research statisticians. A DEC MicroVAX is to be installed in September. Advanced software for statistical analysis such as S-Plus is being used for highly interactive, advanced data analysis. Such advanced software on workstations provides powerful data manipulation tools and facilitates integrating expert systems with statistical analysis. This systems are being used extensively in projects such as the hog composite estimator research, the high-resolution satellite sensor research, analysis of drought patterns from satellite data, and the assessment of production and yield deficiencies brought on by the drought.

Computer networks and high-speed communications are laying the groundwork for providing researchers with immediate access to the Agency's data, all in digital form. NASS is collecting more and more of its data directly in digital form. Computer networks offer the promise of allowing different special-purpose computers to be connected with operational computers such as PCs. This may prove to be an excellent way to implement research computing advances without greatly disturbing the operational systems.

Future Directions

The role of advanced computing in this statistical agency is developing under a variety of pressures. We face reduced staff, a diminishing budget and increasing requirements for improved timeliness of survey processing. There are increasingly elaborate survey designs and estimators under consideration. To our advantage, computing becomes cheaper and more competent. Statistical software is more sophisticated and more accessible to statisticians and analysts.

NASS seeks greater computing resources available at each desk. We look for improved data communications through local and wide area networks, and are considering voice and data integration as a means to improve our CATI operations. Graphics systems are evolving, and the goal is to have them become a routine part of the analysis process, not just for displaying the final results, but for visualization and analysis.

Just as the survey sample sizes are increasing so are the complexity of the survey designs and the estimators required. We ask for more and more complicated processing of more and more data, and fortunately the capability of the computer systems improves.

Expert statistical systems can offer assistance to the statistician. These systems should play an increasing role in routine analyses. They should manage the basic statistical processing, but flag problem data for the analyst.

The most optimistic sign within NASS for our future is the fact that the computer is regarded more as an associate than adversary. This transformation has taken years, but is the solid foundation for embracing the growing computing power we eagerly anticipate.

References

Carney, J.B., et al. *Compiling and Editing Area Sampling Frames using Digital Data for Land Use Analysis and Boundary Definition.* Proposal submitted to NASA Research Announcement 87-0SSA-6, NASS, Washington, DC, 1987.

Crecine, John P. *The Next Generation of Personal Computers.* **Science** Vol. 231 (1986), pp. 935-943.

Hanuschak, G.A., Allen, R.D., and Wigton, W.H., *Integration of Landsat Data into the Crop Estimation Program of USDA's Statistical Reporting Service, 1972-1982.* Invited paper, 1982 Machine Processing of Remotely Sensed Data International Symposium, Purdue University, July 1982.

Ozga, Martin. *Experience with the Use of Supercomputers to Process Landsat Data.* 1984 machine Processing of Remotely Sensed Data International Symposium, Purdue University, June 1984.

Ozga, Martin, et al. *PEDITOR: A Portable Image Processing System,* 1986 International Geoscience and Remote Sensing Symposium, Zurich, Switzerland, September 1986.

Tortora, R.D. *CATI in an Agricultural Statistics Agency.* **Journal of Official Statistics** Vol. 1 (1985).