AGRICULTURAL SURVEYS AND TECHNOLOGY

R. D. TORTORA* and G. A. HANUSCHAK**, U.S. Department of Agriculture
R. D. Tortora, Room 4801 South Bldg., Washington, D.C. 20250

ABSTRACT

From the crop meter to earth resource observation satellites, from postcard surveys of farmers to computer-assisted telephone interviewing (CATI), from peg strips and listing sheets to the CRAY X-MP supercomputer; technology continues to have an impact on agricultural surveys. After a brief historical review of some of the important technologies, this paper describes how the National Agricultural Statistics Service (NASS) is using technology for its current surveys and how technology might affect future agricultural surveys.

INTRODUCTION

The mission of the Agency has always been to provide timely and accurate agricultural statistics for the United States. However, the technology used to accomplish the Agency's mission has changed rather dramatically over its first 125 years (1863-1988).

This paper will present the technology (past, present and a brief look to the future) used by NASS in accomplishing its mission of collecting, analyzing, and disseminating accurate and timely agricultural statistics for the United States.

GENERAL DATA PROCESSING

The U.S. Department of Agriculture's (USDA) National Agricultural Statistics Service has been collecting and summarizing data on crop conditions reported by farmers since 1863. However, for the first 99 years this was done without the aid of a computer. For many years, large listing sheets were used by clerks and statisticians to list the data (from postcards or questionnaires filled out by farmers, agribusinesses, and agricultural experts) for edit, and review. Then the listing sheets were also used to do hand or machine summaries (using adding machines, comptometers, and calculators: electro-mechanical and electronic). From the simplest techniques of manual partial sums and counts to the more advanced use of electro-mechanical and electronic calculators, these listing sheets were edited and summarized for nearly a century without the aid of any computer. Another technique used, for many years, was attaching questionnaires to a pegstrip board. Holes were punched in the questionnaires so that questionnaires could be attached to and displayed side by side on a metal pegstrip board. The data was then edited, reviewed, and summarized without having to do a separate hand listing onto a large listing sheet. Clerks and statisticians became very proficient at these rather manual but very important techniques for edit, review and summary of agricultural statistics data.

The Agency initiated the use of computers in 1962 for keypunching data onto paper punch cards and then utilizing computer programs to verify, edit, and summarize the data. In July 1965, the Agency was given funding and the mandate to establish the Washington Data Processing Center to serve NASS and other USDA agencies. In 1973, Headquarters and several of the 44 state statistical offices were connected, on a contract basis, to a telecommunications network for processing Agency survey data at a central location. Shortly thereafter, the network was extended to all 44 state statistical offices. The Agency is still using a high speed telecommunications network plus centralized mainframes on a 5 year contract basis to do the bulk of the data processing for its survey data.

The current contract is held by Martin Marietta Data Systems (MMDS) which has a large IBM mainframe computer facility in Orlando, Florida. Utilizing the MMDS mainframe and telecommunications network, NASS's Headquarters units and 44 state statistical offices accomplish their daily data processing mission. The Agency is in the middle of a major mainframe data base effort using ADABAS and NATURAL on the MMDS system.

The Agency also acquired a substantial and growing inventory of micro computers, local area networks (LANS), and minis in the 1970's and 1980's, and several supermicros in 1987-1988. The research staff of the Agency now does nearly all of its data processing on local and owned microcomputers and the two supermicros (with the exception of rented time on a Boeing Corp CRAY supercomputer for analyzing large areas of earth resource satellite data).

The inventory of microcomputers and LANS's peripherals such as laser printers is increasing in virtually all the 45 office sites in NASS. There are four major potential benefits: 1) general use of office automation software such as word processing, spreadsheets, and data bases, 2) access to the Agency data base and on-line processing when it's the most cost efficient route, 3) downloading some number crunching modules to fast (386 chip based plus adequate disk) micros for efficient and more timely processing during short survey periods, for research and for more detailed or computationally intensive statistical or graphical analysis, and 4) the use of local micros and LANS for computer-assisted telephone interviewing (CATI).

*R. D. Tortora is Director of the Research and Applications Division of the National Agricultural Statistics Service.

**G. A. Hanuschak is Chief of the Survey Research Branch of the National Agricultural Statistics Service.

Software needs and availability continue to change rapidly. The Agency began with in-house COBOL programs for generalized data verification, edit and summary systems. Mainframe statpackages grew popular in the 1970's. In the last several years, many of these have been converted to the mainframe version of the Statistical Analysis System Inc. (SAS). Another recent development (1986-1988) is the use of PC based statistical packages such as PC-SAS, SYSTAT, PC-CARP, STAT-GRAPH, AUTO-BOX time series, and supermicro interactive math/statistics and graphics packages such as S and S-PLUS especially in the research arena.

These recent developments, plus those covered in the other sections of this paper, present an exciting and challenging era for the government statistician in agriculture. Generations of technology seem to be of shorter and shorter duration and require rather rapid and demanding training of the users.

USE OF REMOTELY SENSED DATA AND ASSOCIATED TECHNOLOGY

In the early 1970's, the challenge of using earth resource satellite data, along with conventional ground-gathered data, for crop inventories pushed the Agency into the new technology frontier of supercomputers. Earth resource observing satellites, such as the first Landsat satellite launched in 1972, provide substantial information about the earth's surface but also generate enormous volumes of data to be processed. The satellite data is briefly described by Vogel (1988). Just to store the original Landsat satellite data (observes 60 x 80 meter objects on ground using 4 types of energy readings) takes approximately 40 mega-bytes of disk storage for an area of 12,500 square miles on the ground. It currently takes 290 mega-bytes of disk storage to store the new improved resolution Landsat scene (which observes 30 X 30 meter objects on the ground using seven different energy readings) called the Thematic Mapper.

A supercomputer, such as a CRAY-XMP, is needed to perform multivariate clustering and classification on these very large data sets, even after using statistical data reduction techniques. Ozga best describes this in a paper entitled "Experience With the Use of Supercomputers to Process Landsat Data" (1984).

Although the Agency has been involved with the processing of earth resource satellite data for only the past 16 years, major technological developments in this subject area have already occurred. Two such developments are presented in the remaining paragraphs of this section.

First, the process of electronically recording (digitization) the crop and land use fields for remote sensing research has paid big dividends for the Agency. The process enabled ground truth data to be located on the satellite coordinate system. The ground truth data was then used statistically to develop crop "signatures" for multivariate classification. The ground truth was then also used as the y variable in a regression estimator where the x,X were, respectively, the sample mean and population mean

for the classified satellite data. In 1972, an electronic data recording tablet (called a digitizer) with .01 or .001 inch cartesian coordinate system accuracy was used interactively, over a network, with a rented DEC-10 minicomputer to record crop field boundaries. This process was initially a bottleneck and somewhat labor intensive. Since then, there have been several new generations of hardware and software improvements. One major generation of improved hardware was the conversion, in 1981, to the use of a video camera, and an image processing system to capture and record crop and land use field boundaries. This process eliminated much of the need for manually digitizing each vertex of each polygon. A recent development, in 1987, has been to use the digitizing tablets along with owned microcomputers not a rented minicomputer or mainframe computer. The most recent development is to overlay the crop field boundaries onto a graphic representation (color or black and white) of the satellite data, using a SUN supermicro computer workstation, and to align them using a graphics mouse.

The second major area of the Agency's program to benefit from digitization was area sampling frame construction, use and storage. Starting in 1979, area sampling frame land use strata and primary sampling unit (PSU) boundaries were electronically digitized. Prior to 1979, the strata and PSU boundaries were measured by the use of a planimeter (manual and hand driven area measuring device). After the areas were planimetered (measured), they were recorded and then summarized initially by hand and then in the late 1960's and 1970's by using computers. The process of digitization has positively impacted the Agency's area sampling frame development in these major ways:

1) more accurate and cost efficient area measurements and built-in quality control graphics for checking the measurements.

2) safe storage of area frames on computer files (with backups) as well as the "paper version."

3) future control data for area sampling frames (such as earth resource satellite data classified into crop and land use types) can become part of a computer based geographic information system.

A second example of a technology breakthrough in remote sensing research was the process (registration) of locating the area frame sample units (approximately 1 square mile in size) on the satellite data tapes. In 1972 the process of registration was also a major bottleneck. At first, the parameters from the polar orbiting satellite such as roll, yaw and pitch were used to get within a few miles of the correct area. In addition, a bivariate polynomial regression was developed between a map coordinate system and the satellite coordination system using ground control points such as road intersections. In 1972, this involved printing a graphic representation of the satellite data (called a greyscale) for 30-50 square miles and then, using manual photo interpretation techniques, finding the correct crop field patterns on the greyscale. The process was labor intensive and required

many large graphic paper plots, light tables, and visual interpretations. Since 1972, there have been several generations of hardware and software improvements. First of all, the earth resource satellite parameters of roll, yaw, and pitch are more accurate than 1972. Also, the previously discussed major improvements in the process of digitization and improved statistical techniques led to major improvements in the accuracy and cost efficiency of the process of registration. Today researchers routinely get within 60 meters or less of the correct location before using a computer screen to lock in the exact location of the ground truth data using a graphics mouse.

Technological improvements in computers and peripherals have dramatically reduced the costs associated with using the original Landsat satellite data to estimate crop acreages. In 1975, the first full state research and development project cost $750,000. In 1978, the first full state applications project cost $300,000. In 1987, when eight major producing states were involved, the cost per state was $129,000 and the vast majority of the cost savings were due to rapidly advancing computer technology. If the 1987 dollars were converted to 1975 or 1978 dollars, the inflation adjusted savings is even more dramatic.

The agency is currently in the process of evaluating the new and improved satellite sensors such as the U.S. Landsat Thematic Mapper (30 meter resolution) and French SPOT satellite (20 and 10 meter resolution) for area frame construction, crop acreage inventories and crop condition assessments. These new generation sensors have more information but also 7 to 10 times the amount of raw data to be processed. Thus statistical data reduction techniques and advanced computer technology will be key to developing a cost effective information system for the Agency.

COMPUTER-ASSISTED SURVEY METHODS

Switching back to conventional ground surveys, the dominant source of Agency information, this section describes some of NASS's major current and future uses of the computer for survey processes. Activities are currently underway for increased usage of the computer in survey preparation, data collection, and pre-analysis coding of survey data. For survey preparation, NASS is planning to introduce computer-assisted self-administered training and computer-assisted area frame construction and maintenance. The data collection activities are computer-assisted telephone interviewing (CATI), computer-assisted personal interviewing (CAPI), and computer-assisted objective yield laboratories. CATI is now operational in 14 of NASS's field offices. A small CAPI research project was started in the spring of 1988. A prototype computerized objective yield laboratory, that automates the counts and measurements necessary to forecast yields, was also started in the spring of 1988. Finally, for pre-analysis coding of multiple frame surveys, NASS is replacing a manual microfiche look-up and coding system with an on-line look-up and automatic status coding system. The most important factors associated with these systems are not the technology that NASS is using, but rather that technology is being

applied to surveys to 1) increase data quality by reducing nonsampling errors and to 2) reduce the amount of resources to conduct surveys. We will briefly discuss each of these computer-aided methods in the order they occur in the survey process.

COMPUTER-ASSISTED AREA FRAME CONSTRUCTION

NASS uses area sampling frames in each State for major crop acreage estimates as part of dual frame sampling and to provide complete coverage for agricultural statistics (see Fecso, et.al (1986) for an overview of the historic development and current use of area and list sampling frames and their integration through the use of dual frame sampling). The construction and maintenance of area frames, even with the benefits of digitization previously covered in this paper, remains somewhat labor intensive. It takes over 20 staff years to construct a new frame. Because of resource limitations, they are used for 15-20 years before being redone. Usually sampling errors increase because of land use change, often in dispersed geographic areas. In addition, land use change causes enumeration problems since enumerators have a difficult time using outdated frame materials to locate sampled segments. So an increase in sample sizes is not an option to control error. To reduce both the increase in total survey error and the resources needed for area frame construction and maintenance, NASS is investigating the integration of Thematic Mapper digital satellite data, digital line graph data - 1:100,000 U.S. topographic maps with digitized transportation and water networks, using image display software and SUN workstations with raster graphics display devices, to build, update and maintain area frames (Carney, et.al, 1987). This approach will give NASS a digital area frame (strata and primary sampling units). Periodically, each PSU can be digitally classified for current land use and reassigned to its proper land use stratum. This technology will be evaluated through comparative studies for area frames in Michigan and Missouri in 1988 and 1989. This project will combine technologies in several areas: remote sensing, area frame construction, a digitized topographic data base, scientific workstations and advanced graphics displays. The project, three years in duration, is a cooperative effort with the National Aeronautics and Space Administration.

COMPUTER-ASSISTED SURVEY TRAINING

Until recently, NASS's usual method of training for a survey consisted of lectures presented to surveys statisticians from each field office, at national training schools, by Headquarters staff. These statisticians, in turn, returned to their office and lectured their enumerators on the need, content, and procedures of the survey. Besides turning to adult education techniques, workshops and learn-by-doing, NASS has begun to use the Berkeley CATI software in its survey training. Self-administered training instruments (SAI's) have been developed that a) measure the pre-training knowledge level of each survey statistician, b) develop survey concepts, and c) measure post-training and post-survey knowledge. The pre-training knowledge levels allow NASS to better tailor survey training to target audiences.

Thus, the veteran may participate in different workshop sessions than those for the novice participant. Self-administered instruments that develop survey concepts are intended to 1) assure a minimal basic knowledge of concepts prior to national training and 2) extend the training to staff not able to attend the national training. With post training and post-survey SAI's, the NASS training staff is able to evaluate the effectiveness of its training procedures and survey manuals. In the future, the biggest jump in training ability using SAI will occur with CAPI by extending the SAI training to each face-to-face enumerator. We now move from a discussion of the pre-survey activities to a discussion of the computer-assisted methods that NASS is introducing in the data collection area.

COMPUTER-ASSISTED TELEPHONE INTERVIEWING (CATI)

Naturally, CATI development occurred first. NASS and the University of California at Berkeley signed a cooperative research agreement in 1981. The main goals of these projects were to test and evaluate the applicability of the Berkeley CASES software for agricultural surveys and to compare CATI with conventional telephone interviewing (Tortora, 1985). In 1982 and 1983, CATI surveys were started in California and Nebraska, respectively (House and Morton, 1983, and Morton and House, 1983 for details of some of the research). In 1985, NASS decided to accept CATI as the operational mode for data collection by telephone and started phasing in hardware to cover 14 of its field offices. There are 185 stations for calling in these offices. (This hardware serves several other purposes besides CATI: direct data entry of mail and face-to-face data, remote job entry and office automation applications during a.m.-calling hours.)

NASS uses CATI on the Quarterly Agricultural Surveys, the Quarterly Labor Surveys, Semi-Annual Cattle and Sheep Surveys, Annual Farm Land Values Survey, Monthly Milk Production Surveys and several other individual state surveys. In six separate survey periods, June 1987 through June 1988 , over 79,000 interviews were conducted by CATI for the Quarterly Agricultural and Cattle and Sheep Surveys. In three Labor Surveys, another 13,000 plus interviews were conducted. CATI has many well known advantages that improve the quality of surveys. However, two other aspects of CATI should not be overlooked. First, it facilitates research. NASS has used CATI to study methods of using past interview data in a current interview, to test different CATI instrument designs (questionnaire designs), measure interviewer variances (Pafford, 1988), evaluate different administrative structures for data collection (Bass and Tortora, 1988), and improve the timeliness of response variance studies (report to be published in late 1988 by NASS). Second, CATI can often be used to provide a quick response for a critical issue. For example, NASS fielded an August Acreage Update Survey in less than 6 weeks this year. This special survey was conducted because of the 1988 drought. It helped USDA obtain more information about farmers opinions of the acres they expected to harvest. Seven

regional CATI locations covered 24 states affected by the drought in this survey.

Finally, NASS is in the process of revising the hardware it uses for CATI. Beginning in 1989, NASS will begin installing Local Area Networks (as opposed to multi-user systems) in each field office for CATI. In 1992, all field offices will be equipped and NASS will have about 750 CATI stations throughout the United States for collecting agricultural survey data. With the advent of the portable computer, computer-assisted data collection can be extended to the face-to-face interviewers.

COMPUTER-ASSISTED PERSONAL INTERVIEWING (CAPI)

In 1988, NASS started a small demonstration project with CAPI. Using laptop portable computers and the Berkeley CASES software, an instrument was coded to collect data on livestock prices. Two enumerators are currently collecting monthly livestock prices (the price of cattle, hogs, and sheep sold for slaughter) on these portable computers. The purpose of this project is to identify potential problems and benefits that face-to-face interviewers may have using portable computers in data collection. This information should help NASS prepare for a 15-20 interviewer test in one or two state statistical offices during 1989. NASS staff also was consulted by USDA's Human Nutrition Information Service on a full scale national nutrition survey in 1987 where CAPI was used.

COMPUTER-ASSISTED OBJECTIVE YIELD LABORATORIES

A major part of the NASS program is crop estimation. This includes estimation of intentions to plant, planted acres, harvested acres, yield, and production. In the past, technology has affected this program. The 1920's crop meter (Bosecker, 1988) was developed to improve planted acreages estimates. The remote sensing techniques (described in this paper) also improved planted acreage estimates. To improve yield estimates, NASS has developed a prototype automated objective yield laboratory.

Objective yield laboratories make counts and measurements of crops during the latter part of the growing season. Enumerators harvest these crops, even before they mature, send them to NASS objective yield laboratories where counts and measurements are made for use as independent variables to forecast final crop yield. For example, for the May 1 wheat forecast, the laboratory, after threshing, counts kernels and heads, weighs and converts each objective yield wheat sample to a standard moisture content.

The manual process has the potential to introduce many nonsampling errors. Miscounts, scales out of adjustment, moisture meters that are not calibrated, and incorrect or inaccurate moisture conversion look-up are some potential sources of error. To minimize these types of errors, NASS has developed a computer-assisted

objective yield laboratory. A bar code reader, scale, and moisture meter communicate with and are under the control of a computer. The laboratory technician, after checking-in the sample using the bar code reader, moves the sample from counter, to scale, to moisture meter and measurements are automatically posted to the data record for each sample unit. These computer-assisted laboratories are intended to replace a manual and time consuming process that often introduces potentially serious errors into the survey.

ON-LINE SURVEY OPERATIONS

Real-time survey operations improve survey quality by minimizing the potential for nonsampling errors during the very limited time frame when the survey operations are conducted. Examples are editing, removing duplication in the sample, coding correct frame overlap status, coding questionnaires, etc. NASS is in the process of developing real-time computer systems (Tortora, et.al, 1986) to improve these operations. A portion of this on-line system has been implemented in 10 field offices. It improves data quality, timeliness and consistency. It also enhances list frame update activities, ensuring that new or corrected names, addresses and control data are promptly and properly posted to the frame. Finally, it efficiently implements a procedure to provide respondents feedback in the form of survey results. Of course, there are other survey operations and procedures that can be semi-automated. However, some care must be taken to appropriately introduce these techniques into the survey program. As pointed out by Shanks and Tortora (1985), an important part of effectively instituting these computer-assisted methods will be system integration. All of these systems will have to "talk" to each other efficiently and in a user friendly manner.

OTHER TECHNOLOGY

With a vast amount of other new technology quickly being developed, NASS (and other survey organizations) must be careful to allocate scarce resources in an efficient manner. Some thought should be given to what technologies should be investigated and how many resources are allocated to the investigation. The philosophy that NASS will use evolves around the principle of working on "important problems". That is, problems, which if solved, will increase the accuracy, timeliness or relevance of our surveys for a constant or reduced amount of resources. For example, a technology such as a statistical expert system appears to have the potential to solve "important problems" in NASS when applied to the editing and imputation of survey data and linking of records for our list frame.
Powerful microcomputers and scientific workstations are proving to be a very effective research tool. Fast processing and interactive statistics, mathematics and graphics at the fingertips of the researcher are speeding up research analysis. This allows "more computationally intensive algorithms and graphics" to be evaluated.

FUTURE TECHNOLOGY

Undoubtedly, the pace of technological developments involved in the survey process will continue to be rapid and a challenge to statistical organizations. It is important to keep in focus that technology is a tool to aid in survey processes. Not all new tools will turn out to be true improvements for the real NASS goal -- timely, relevant, and accurate agricultural statistics for the United States.

An organization staff must keep its "antenna" mechanism in functioning order. The challenge is, with limited or scarce resources, to pick those technological developments with the highest probability of actually improving the efficiency of the survey process or that reduce total survey error. Secondly, a strict evaluation of a new tool must be done to screen out technological developments that are not true improvements to the U.S. agricultural statistics program. In addition to technological developments involved with survey processes, there will be many other areas with technological changes that may affect the Agency's program in the future. This may include areas such as: biotechnology, animal genetics, aquaculture, agricultural chemicals and water quality, meteorology, satellites (land, water and weather), changing farm structure (number, type and size of farms), automated weather stations, need for spatial information, robotics and geographic information systems. These developments and challenges provide a stimulating environment for the agricultural statistician in the foreseeable future.

REFERENCES

Bass, R.T. and Tortora, R.D. "A Comparison of Centralized CATI Facilities for an Agricultural Labor Survey, " To appear in Proceedings of the International Conference on Telephone Survey Methodology, Groves, et.al Editors, 1988. J. Wiley and Sons, Inc. New York.

Bosecker, R.R. "Sampling Methods in Agriculture," Proceedings of the Survey Research Methods Section, ASA, 1988. Washington, D.C.

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 87-055A-6, 1987, NASS, Washington, D.C.

Cook, Paul W. "Landsat Registration Methodology Used by the U.S. Department of Agriculture's Statistical Reporting Service, 1972-1982," Staff Report, Statistical Reporting Service, May 1982.

Fecso, R., Tortora, R.D., Vogel, F.A. "Sampling Frames for Agriculture in the United States," Journal of Official Statistics, Vol. 2, No. 3 1986, pp. 279-279, Statistics Sweden.

Hanuschak, G.A., Allen, R.D., 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.

Hanuschak, G.A., et.al. "Obtaining Timely Crop Area Estimates Using Ground-Gathered and Landsat Data," U.S. Department of Agriculture, Technical Bulletin No. 1609, August 1979.

House, C.C. and Morton, B.T. "Measuring CATI Effects on Numerical Data," Proceedings of the Section on Survey Research Methods, ASA, Washington, D.C., 1983.

Morton, B.T. and House, C.C. "Training Interviewers for Computer Assisted Telephone Interviewing," Proceedings of the Section on Survey Research Methods, ASA, Washington, D.C., 1983.

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 and Sigman, Richard. "An Autodigitizing Procedure for Ground Truth Labeling of Landsat Pixels," Fifteenth International Symposium of Remote sensing of Environment, Environmental Research Institute of Michigan, May 1981.

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

Pafford, B.V. "The Influence of Using Previous Survey Data in the 1986 April ISP Grain Stocks Survey," National Agricultural Statistics Service Research Report Number SRB-88-01. Washington, D.C., 1988.

Shanks, J.M. and Tortora, R.D. "Beyond CATI: Generalized and Distributed Systems for Computer-Assisted Surveys," Proceedings of the Bureau of the Census First Annual Research Conference, 1985. Washington, D.C.

Tortora, R.D. "CATI in Agricultural Statistics Agency," Journal of Official Statistics, Vol. 1, No. 3, 1985.

Tortora, R.D., Vogel, F.A. and Shanks, J.M. "Computer-Aided Survey Methods," Proceedings of Computer Science and Statistics: 18th Symposium on the Interface. 1986, Washington, D.C.

U.S. Department of Agriculture, Statistical Reporting Service. The Story of U.S. Agricultural Estimates. Miscellaneous Publication No. 1088, 1969.

_____. Scope and Methods of the Statistical Reporting Service. Miscellaneous Publication No. 967, 1964.

_____. Miscellaneous Publication No. 1308, 1975.

_____. 1983

Vogel, F.A. "125 Years of Agricultural Estimates," Proceedings of the Survey Research Methods Section, ASA 1988. Washington, D.C.