UPDATE ON REMOTE SENSING PROGRAM OF THE NATIONAL AGRICULTURAL STATISTICS SERVICE

BY
George A. Hanuschak *
Chief, Survey Research Branch
National Agricultural Statistics Service
U.S. Department of Agriculture

OCT. 1989

ABSTRACT

This paper discusses two recent additions to the remote sensing and technology research programs of the National Agricultural Statistics Service of the USDA. The first research effort is the construction of area sampling frames using all digital data (Landsat Thematic Mapper Data and Digital Line Graph Data). The second project discussed is a sensor comparison study of Landsat Thematic mapper and French SPOT data for winter wheat acreage estimation in the state of Kansas.

I. COMPUTER ASSISTED AREA SAMPLING FRAME CONSTRUCTION

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) has the primary responsibility of providing statistics for domestic crop and livestock production. These statistics are derived from data collected through a variety of sampling techniques and surveys. The computer-aided stratification (CAS) system has been developed, jointly between NASS and NASA Ames Research Center (ARC), to automate the stratification and sampling-unit delineation portion of area frame based sample survey procedures that are currently performed manually. The CAS stratifies sampling units by land-use and land cover type, using image processing hardware and software. The system provides coverage areas and boundaries of stratified sampling units which are used as inputs for the subsequent sampling procedures from which agricultural statistics are developed.

Area sampling frame - The current manual procedure employed by NASS to develop and edit sampling frames was designed to ensure a statistically valid sample and operational efficiency. Some of the key steps in the procedure, are as follows:

1. Stratification: the technique of dividing the land in an area, for example, a county or a state, into land-use and land-cover groups (known as strata) based on photo-interpretation processes. The strata are defined mostly by the intensity of agricultural activities or cultivation. This step included preconstruction analysis; procurement of stratification materials, for example, Landsat Multispectral Scanner (MSS) and Thematic mapper (TM) imagery 1:250,000 - scaled false - color composite prints and/or high altitude photography and maps. The materials are then used for land-use/land-cover stratification into primary sampling units (PSUs). PSUs are groups of six to ten ultimate sampling units or segments bounded by permanent features of the landscape, such as roads and rivers.

* PRESENTED AT EUROPEAN UNION CONFERENCE ON THE APPLICATION OF REMOTE SENSING TO AGRICULTURAL STATISTICS, VARESE, ITALY.

2. Digitization: PSUs are electronically digitized for area measurements that are used to allocate the samples.

3. Multi-step sampling: PSUs are delineated from the strata based on land use, and a random sample of PSUs is chosen for further breakdown into segments. After the PSU has been delineated into segments, a segment is randomly chosen and is visited in the June Agricultural Survey to collect information concerning, for example crop types and livestock inventories.

4. Analysis and quality assurance: the land-use strata definitions, the number of PSUs, the size of the segments, etc. are periodically analyzed to eliminate errors of commission and omission and, thus, to improve the accuracy of the statistics produced.

Objectives

The existing NASS manual procedures for creating and editing area sampling frames is somewhat labor-intensive. The CAS procedure attempts to perform area frame functions with display hardware and software on a microprocessor-based workstation. The goals of the current work on the automated procedure are to complete the software and implement an operational system by 1991.

This study concentrates on the automation of the stratification and sampling unit delineation procedures, and has three basic objectives which are :

1. Select or implement a software package that will supply the tools necessary for compiling and editing area frame land producing stratified sampling unit boundaries using Landsat TM and U.S. Geological Survey (USGS) Digital Line Graph (DLG) data. The software will operate on a image display and graphics workstation.

2. Design a procedure for boundary compilation and editing based on on-screen image/photo interpretation of Landsat TM digital data and DLG data. The procedure must be user friendly, that is, the hardware and software systems must be easy to operate so that the procedure can be followed efficiently by a staff.

3. Implement the CAS procedures in an operation environment.

Procedures

To emulate the current manual stratification and sampling-unit delineation procedures as performed by NASS, the CAS system uses on-screen digitization techniques. The work under the CAS system will be performed on a graphic workstation display using digital imagery, instead of on false-color composite prints. The new procedure begins by entering the PEDITOR system. CAS is accessed through the "Area Frame Development" menu selection of PEDITOR. The general work flow involves displaying a TM image, overlaying it with DLG data, delineating PSU boundaries, editing, and then saving the results.

Study Area

Three Missouri counties were chosen as the study area for the test application of the CAS procedures. They were chosen partially because of the availability of USGS DLG data, which served as a reference data plane for the area frame construction procedures.

Conclusions

At the conclusion of this study, the CAS procedures were evaluated relative to the manual method of PSU delineation, and a number of advantages of using the CAS method of automated stratification were identified:

1. The manual labor required to construct area frames was reduced substantially. For example, the construction time required for one operator to complete Macon County was 2.5 days (including digitizing time) using the CAS system, and 12.5 days plus 2.5 days for digitizing using the conventional manual method. The PSU delineation, performed manually by transferring boundaries between map and image media, can be done electronically. Much of the manual work is eliminated.

2. The ability to update an area frame is improved with the CAS system. Because the PSU boundaries can be easily edited based on the changes of the land-use pattern, the area frames can be updated with greater frequency.

3. Precision of the surveys may be enhanced. The CAS capability of viewing Landsat and map attribute data together may enable the operator to more accurately assign the PSUs to strata.

4. Cost comparisons of the two methods are summarized in figure 1.

References

COMPUTER-AIDED BOUNDARY DELINEATION OF AGRICULTURAL LANDS, Thomas D. Cheng, Gary L. Angelici, Robert E. Slye, and Matt Ma, NASA Technical Memorandum 102243 , November 1989. (Major excerpts from).

REMOTE SENSOR COMPARISON FOR CROP AREA ESTIMATION, James M. Harris, Sherman B. Winings, and Martin S. Saffell, Submitted For Presentation At IGARSS 1989 Symposium in Vancouver, Canada 1/13/89, National Agricultural Statistics Service, U. S. Department of Agriculture. (Excerpted).

AREA FRAME DESIGN FOR AGRICULTURAL SURVEYS. By Jim Cotter and Jack Nealon. Area Frame Section, Research and Applications Division, National Agricultural Statistics Service, U.S. Department of Agriculture, August 1987.

Figure I

COST COMPARISON OF CASS VERSUS OPERATIONAL STRATIFICATION

The following compares the cost for the work produced in one year by the stratification unit using the current methods to the estimated cost for the same quantity of work in one year using computer-assisted stratification and sampling (CASS) methods. Three new frames are worked on each year currently with at least two new frames implemented in the next June Survey. Costs for materials for three frames are used although about five frames are begun in two years.

It is not suggested that the staff reductions shown here to be possible under CASS should actually be done. The staffing and organization under CASS is a separate issue to be addressed later. This report only makes cost comparisons which logically must be done before further decisions on staffing and goals are made.

CURRENT		CASS	
1. Personnel			
strat --			
1 supervisor	$24,000	1 supervisor	$24,000
9.5 FTE carto-aids	$133,000	4 FTE carto-aids	$64,000
digitizing --			
1 supervisor	$22,000		
2 FTE carto-techs	$32,000	1 FTE carto-tech	$16,000
2. Materials (3 states)			
NAPP aerial photos	$45,000	Digital Landsat	$119,000
1:100,000 frame maps	$ 1,000	Digital Line Graph	$ 2,000
Quad maps	$ 1,000	Quad maps	$ 1,000
Satellite Imagery	$35,000		
County Overlays	$ 1,000		
3. Equipment			
Light Tables & misc (replace 1/yr)	$ 1,000	SUN Server (maint/replace)	$ 5,000
Maintenance -- 4 PC's	$ 2,000	HP Workstations (3*50,000 - 5 yrs)	$30,000
TOTAL	$297,000	TOTAL	$261,000

II. REMOTE SENSOR COMPARISON FOR CROP AREA ESTIMATION

The second part of this paper will discuss an initial pilot level research project in Kansas designed to compare Landsat Thematic Mapper data and French SPOT data in a regression estimator situation.

INTRODUCTION: The National Agricultural Statistics Service (NASS) initiated this comparison of remote sensors to select a replacement for the Multispectral Scanner (MSS) LANDSAT data. NASS used MSS data in the crop area estimation program during the 1980 - 1987 time period. The operational remote sensing program processed between 70 and 90 LANDSAT MSS scenes per year. Ground truth data is combined with MSS data and processed through the USDA's PEDITOR system. Output from the PEDITOR system for each sampled unit, or segment, is the number of pixels classified to each cover. Reported acres for a crop in each sample segment, y_i the dependent variable, is combined with the classified number of pixels for a crop, x_i the auxiliary variate, to produce a regression estimator of total crop acres in a Landsat scene. Appendix A lists the formulas used to calculate total crop acres and the estimated variance for the total crop acres in one stratum.

Remotely sensed data types included in the comparison were LANDSAT Thematic Mapper (TM), French SPOT, and LANDSAT MSS data. MSS data were used as a comparison base. The choice of the "best" satellite type was based on the statistical performance of the regression estimator. This is different than most other remote sensing studies, in which percent correct classification is often used. It should be noted that the only product produced from the satellite data is an estimate of crop area and its estimated variance. The regression estimator needs the property of consistent classification to provide a reliable estimate. Thus, the NASS remote sensing program is looking for a data type that provides the most consistent classification. Other operational considerations are reviewed in this paper.

GROUND TRUTH, COVERAGE DATES, STUDY AREA: The study site was the Garden City area of western Kansas. The study included parts of Lane, Ness, Finney, Hodgeman, Gray, Ford, Meade, and Clark counties. The agricultural crop of interest was hard red winter wheat. Other prominent covers included pasture, fallow and bare soil. Overpass dates of the three sensors varied from April 20 to May 11, 1986. The SPOT scenes had overpass dates of May 1 or May 11. The TM scenes had overpass dates of April 20 and May 6. The MSS scenes had an overpass date of April 28. The different date combinations were grouped to produce three analysis areas or "analysis districts." See APPENDIX B for an overview of the study area, scene row/path, scene dates, locations and analysis district groupings.

Ground truth information is collected as a part of the June Enumerative Survey, an annual area-frame based sample survey. Ground data was from 234 sample segment sites in the eight county region. Each county is stratified into six land use strata. Two urban strata were excluded from the analysis due to small sample size and minimal crop presents. Another stratum, the rangeland stratum, had a target segment size of four square miles. The rangeland segments were included in the training process but were excluded from statistical analysis due to the small sample size and lack of crop of

interest. The remaining three strata definitions were greater than 75 percent cultivated, 50 - 75 percent cultivated, and 25 - 50 percent cultivated. These three strata had target segment sizes of one square mile. As noted earlier, during the operational program segments are surveyed during June. For this study, however, administrative records from the USDA were used for the crop acreage and field boundaries. The administrative records kept by the Agricultural Stabilization and Conservation Service list by field the program crop participation. Any errors in ground truth data are consistent in all three sensors. That is, as noted above, the y_i, reported crop acreage per segment, is the same for all three sensors.

OPERATIONAL CONSIDERATIONS FOR THE NEW SENSORS: A primary operational consideration is the increased processing requirements of SPOT and TM due to the increased data volume. The MSS scenes were processed on a rented mainframe and a Cray supercomputer. NASS is currently converting to a new processing environment. The new processing environment will network PC's and supermicros, VAX and SUN computers, and will include a link to a Cray supercomputer.

The amount of data from TM and SPOT is about seven times the amount of data from MSS for a given land area. One MSS pixel has 4 data points or 4 band readings. For SPOT to cover the same ground area as one MSS pixel it requires 27 data points, 3 channels X 9 SPOT pixels per MSS pixel. For TM, 28 data points are required, 7 channels X 4 TM pixels per MSS pixel, to cover the same ground area as MSS's 4 data points. One example of additional computer resource requirements is the Maximum Likelihood Classification program. CPU requirements for Maximum Likelihood Classification is a constant X number of channels X number of categories X number of pixels. In addition to the increased data volume, NASS's clustering method creates more categories with both TM and SPOT. Without care one could easily use 14 to 21 times the computer resources used for MSS.

A caveat about directable satellites is the number of looks per repeat cycle may be deceiving. If the study area is greater than a path (two paths for SPOT) the effective number is one look per nadir orbit. Thus programable satellites lose their advantage over non-pointables when there is a large continuous study area or a short time window.

Tape handling and file maintenance could become operationally difficult when compared to MSS. Ninety MSS tapes would translate into 270 TM tapes and as many as 810 SPOT tapes. An analysis district is created from adjacent scenes for a specific date. The number of analysis districts for TM should not exceed the number for MSS. The number of analysis districts for SPOT would approximately triple, thus tripling the number of files.

The basic aggregation unit used in the operational program with MSS was a county. When a scene splits a county, additional processing is required. SPOT scenes, which are smaller than MSS or TM scenes, require much more splitting of county masks and adjustment of frame units. Mask splitting and frame unit adjustment are machine and labor intensive. Much care must be used to maintain the integrity of the frame. The work is directly proportional to the number of splits.

Registration of SPOT may be accomplished by using the five points furnished to calculate first order registration parameters. MSS and TM require labor intensive map to image registration. Local registration of SPOT is conducted using the usual shifting method. Larger number of boundary pixels must be pulled for SPOT as the shifts tend to be the same magnitude with respect to the earth as MSS and TM.

ANALYSIS: The statistical analyses determined whether there were significant differences between the accuracies of the regression estimates produced by the three sensors. If the mean absolute value of the residuals from a regression with data from sensor A were less than the mean absolute value of residuals from sensor B, then sensor A would produce more accurate regression estimates. Linear regressions were performed by land use stratum within analysis district. So, there were nine models, three analysis districts X three strata, estimated for each of the three sensors. Since the ground truth are identical across sensors, the absolute value of residuals could be compared in the statistical analysis.

The first test was an analysis of variance. The null hypothesis was that there was no significant difference between the means of absolute values of the residuals from TM, SPOT and MSS regressions. To increase the power of the test for sensor effect, the variation in absolute residuals due to the analysis district, the stratum, the interaction between stratum and analysis district and segment within analysis district was blocked. Land use strata were created independently of analysis districts and are thus not nested within analysis district. The following treatments were tested with the F-statistic against the remaining error in the model: sensor, sensor-analysis district (sensor*AD) interaction, and sensor-stratum interaction. There were a total of 181 segments and thus 543 observations. The ANOVA table is shown below.

ANALYSIS OF VARIANCE TABLE
FOR THE EFFECT OF SENSOR ON ABSOLUTE VALUE OF RESIDUALS

Source	DF	Sum of Squares	Mean Square	F	Pr>F
Analysis district	2	6506.11	3253.06		
Stratum	2	2262.59	1131.30		
Stratum*AD	4	22490.69	5622.67		
Segment (Analysis District Stratum)	172	607172.44	3530.07		
Sensor	2	32161.74	16080.87	16.42	.001
Sensor*AD	4	6396.85	1599.21	1.63	.165
Sensor*Stratum	4	6952.93	1738.23	1.77	.133
Error	352	344740.80	979.38		
TOTAL (Corrected)	542	1028684.15	3599.70		

The sensor effect was significant at alpha = .01, and the interaction effects were not significant. This made it possible to determine which sensor had the smallest absolute values of residuals. The mean absolute residuals from TM, SPOT and MSS were 38.39, 55.76 and 53.41, respectively. The values for SPOT and MSS were close, and the choice of future sensor is between SPOT and TM. Consequently, a t-test was conducted in which SPOT and TM absolute residuals were paired by segment. The alternative hypothesis was that TM produced smaller absolute residuals than SPOT. The mean value of the difference over 181 observations was 17.376, the standard error was 3.224 and the t-statistic was 5.389. The null hypothesis of no significant difference was rejected.

CONCLUSION: Given both the statistical analysis and operational considerations TM data is preferred to SPOT for crop area estimation. The TM data produces the most statistically accurate regression estimates and requires less analysis time than SPOT.

APPENDIX A: REGRESSION ESTIMATOR (GROUND DATA AND CLASSIFIED SATELLITE DATA) FOR A STRATUM IN AN ANALYSIS DISTRICT AND THE ESTIMATE OF THE VARIANCE FOR THE STRATUM

$\hat{y} = N [\bar{y} + b (\bar{X} - \bar{x})]$, where :

N = The number of sampling units in the stratum.

$\bar{y}$ = The June Enumerative sample average reported crop acres in the stratum.

b = The estimated regression coefficient for the stratum.

$\bar{X}$ = The average classified crop pixels for the stratum, total classified pixels divided by total units.

$\bar{x}$ = The average sample classified crop pixels for the stratum, sample classified pixels / sample units.

$$\hat{\sigma}^2 = N^2 * (1 - n/N) * \left[\frac{\Sigma (y_i - \bar{y})}{(n - 2)} \right] * (1 - R^2) * \left[1 + \frac{1}{(n - 3)} \right]$$

APPENDIX B: APPROXIMATE LOCATION OF SENSOR SCENES AND ASSOCIATED DATES IN RELATION TO EIGHT COUNTIES IN WESTERN KANSAS.