CALIFORNIA 1977
77-10

AGRICULTURAL CROP ACREAGE ESTIMATES FOR SMALL LAND AREAS USING LANDSAT-- WHAT IS THE INFORMATION WORTH?

Galen F. Hart

Statistical Reporting Service
U. S. Department of Agriculture
Washington, D. C. 20250

1977

Abstract

The Statistical Reporting Service of the Department of Agriculture has been involved in exploratory research to determine if LANDSAT data can be converted to useful agricultural information. Research exploits the complete coverage or census data gathering capability of LANDSAT. Results show that, for a sizable agricultural producing area in Illinois, LANDSAT can be used as a secondary data source with a probability sample of ground data to provide estimates of crop acreage for counties with known statistical precision. Major problems in past and present investigative research such as signature extension, classification bias and cloud cover are eliminated by using statistical inference to convert pixel data through regression modelling to crop acreage estimates and domain theory to eliminate cloud cover bias. Projections indicate this improved estimating ability for small areas will be costly and the obvious question is raised--"What is the information worth?"

LANDSAT, A PASSIVE REMOTE DATA COLLECTOR

This paper discusses a statistical application of LANDSAT digital data. LANDSAT is a data collector and from a statistical point of view it offers desirable and undesirable features. First, some desirable features:

1. Complete or census like coverage data related to land cover is collected.

2. Data are current--there is the capability for near real time processing.
3. Data are consistent over both time and space.

To optimize use of this data collection means, these three important positive features should be exploited. Associated with the desirable are some undesirable features:

1. Resolution--"coarse" is a descriptive term that could be used. A resolution element (pixel) of slightly over one acre is hardly what one would describe as a highly discriminating unit of observation for detailed land use information.
2. Extracting land use information from reflectance data is, at best, a difficult task. A way of describing this statistically is that we have information concerning a population of interest, land use, imbedded in data from a population that we know to be different. There is no direct way to "scale" the covered population to the population of interest.
3. Nonresponse or weak response. Cloud cover is the same as nonresponse in a sample survey. Weak response could be the result of atmospheric variability or other conditions that result in reflectance values that do not distinguish different land uses.

There are a number of other desirable and undesirable features that could be mentioned but these are less important relative to the concerns of this paper.

AREA SAMPLING FRAME, AN ACTIVE GROUND ENUMERATION DATA COLLECTOR[1]

The Statistical Reporting Service (SRS) has a primary data collection methodology that has some features that are complementary to those of LANDSAT.

SRS uses a scientifically selected land area sample of about 16,000 enumeration units to estimate crop area and other production and economic items for the United States. The probability sample of the 16,000 enumeration units is from a frame of the total area of the United States excluding Alaska and Hawaii. Note that LANDSAT coverage and the population from which this sample of enumeration units is selected share a common base.

The entire land area of the United States is first partitioned or stratified by agricultural land use. Usually "county highway maps" are used in this operation. These maps meet certain basic requirements set forth by the Department of Transportation and can

be converted into the controlled map series of USGS on either 7-1/2 or 15 minute quadrangle maps.

Each partition or stratum is subdivided such that every acre has a known and unique probability of being selected as a unit of observation. Sampling units vary in size from a few city blocks in urban areas to several square miles in open grazing or ranching areas. A common size enumeration unit is about one square mile in major crop producing areas of the United States.

After a sample has been selected for a survey, photographic prints covering these areas are obtained. Prints are acquired from various sources but the primary supplier is the USDA Agricultural Stabilization and Conservation Service photographic laboratory in Salt Lake City. The scale used is about 8 inches to the square mile. This scale permits ground enumerators to locate field boundaries on the photography and identify the crop or land use for each parcel in the sampling unit. Crop or land use acreage is recorded by enumerators, expanded by the inverse of the probability of selection and summarized to state and national totals.

The major survey for acreage is in late May with a smaller subsample survey in late November to collect information for fall seeded crops. The sampling precision for major crop acreage estimates at the national level is on the order of 1.5 to 3.5 percent relative sampling error. For major crop acreage estimates at the state level, the relative sampling errors are on the order of 3 to 8 percent. The cost of the annual surveys, including the necessary increments for maintenance of the sampling frame, is about $3.5 million. Keep in mind these surveys also collect information on livestock and economic items as well as crop acreage and other land use.

With this as a brief description of the present SRS acreage estimating methodology, here are some of the associated desirable and undesirable features. As before, first the desirable features:

1. A sample is drawn from a complete coverage frame--no area of land is excluded from the sample selection process.
2. Data are current--there is about a three week time gap from the beginning of data collection to the completion of summarization.
3. Data collected are controlled for accuracy--within each selected sampling unit a tightly controlled complete accounting is obtained. There is a low potential for bias in estimating crop acreages.

4. A high response is obtained, or conversely, there is a low nonresponse. Generally, the nonresponse for a particular survey is less than 5 percent.

There are some undesirable features that align well with the desirable features of LANDSAT. These are:

1. A sampling methodology is used therefore there is an associated sampling error.
2. Sample size is too small to give sufficiently precise results for small land areas, counties or production area aggregations between the county and state levels, and for minor crops at the state level.

As was mentioned earlier, the frame from which the land area sample is selected is the geographic area of the United States. Therefore, the sample is a probability "ground truth" sample for satellite coverage. This is where the research task begins. The task being to utilize the complete coverage capability of the satellite in combination with probability ground data to improve crop acreage estimating ability for "small" land areas.

COMBINING THE DATA COLLECTORS[2,3]

The detail of individual steps in connecting the two data sources will not be discussed at length but are outlined as follows:

1. Registration--registering sampling unit locations to LANDSAT computer compatible tapes (CCT's).
2. Accurately locating field boundaries in CCT's.
3. Tagging individual pixels in CCT's with ground enumerated land use.
4. Establishing the relationship of reflectance values to land use for the sample units, commonly called "training the classifier."
5. Classifying all pixel data in a given area (i.e., an area within a LANDSAT scene or scenes on the same pass).
6. Combine ground and LANDSAT data with a regression estimator to generate a single estimate and compute its sampling variability.

Following are some points concerning these steps that have a bearing on the title of this paper.

1. Registering sample unit locations to LANDSAT CCT's is a time-consuming process. Presently registration, utilizing two

steps to locate data in CCT's at a half pixel accuracy, requires about 100 person hours per scene. This assumes approximately 40 sample units and 50 control points with a trained person performing the operation. The first step of registration should be eliminated with a product that will be coming from the EROS Data Center in Sioux Falls, South Dakota, in early 1978. However, the second step of registration will still be required and presently this operation takes about one-fourth of the two step registration time. Sample unit and stratum boundary registration information could be stored and reused for subsequent surveys.

2. The digitizing process to locate field boundaries requires about one person hour per sample unit or 40 person hours per scene. Field boundaries change from one survey period to the next so there is a limited carryover reduction in time for this operation. A hope here would be to use a scanner digitizing system that could potentially reduce the time required for this effort.
3. A satisfying reduction has already occurred in the cost of classification. A short time ago, soon after the launch of LANDSAT I, about $2,000 of computer time was used to classify a single LANDSAT scene. The present cost is about $700 and is likely to be reduced even further.
4. Classifier training is highly analyst interactive at present. Efforts are being made to reduce this input but it is unlikely that analyst interaction will be completely eliminated in the near future. Optimum classifier strategies for crop acreage estimation are also under continuing investigation.

Research hasn't progressed to the point where accurate operational cost information can be projected. However, at some unspecified time in the future, assuming optimistic progress associated with the points discussed above, it may be possible to generate estimates for major land uses for areas the size of a county for a few hundred dollars. The cost for a larger area, say the aggregate of ten to twenty counties, would be only slightly greater than for an individual county.

Now, for some research results for Illinois.[4] First the formulation. For the area sample direct expansion estimate and its variance:

Let $\hat{Y}$ = direct expansion acreage estimate of a crop using SRS land area sampling unit data

y_{hi} = acreage of a crop in i^{th} sampling unit of the h^{th} stratum (partition)

$v(\hat{Y})$ = sampling variability of the estimate $\hat{Y}$

r.s.e.$(\hat{Y})$ = relative sampling error of the estimate $\hat{Y}$

Then

$$\hat{Y} = \sum_h \left[\frac{N_h}{n_h} \left(\sum_i y_{hi} \right) \right]$$

$$v(\hat{Y}) = \sum_h \left\{ \frac{N_h}{n_h} \cdot \frac{N_h - n_h}{n_h - 1} \cdot \left[\sum_i y_{hi}^2 - \frac{\left(\sum_i y_{hi} \right)^2}{n_h} \right] \right\}$$

$$\text{r.s.e.}(\hat{Y}) = \frac{\sqrt{v(\hat{Y})}}{\hat{Y}}$$

For the regression estimator and its variance:

Let $\hat{Y}_R$ = regression acreage estimate of a crop

$\hat{\bar{Y}}_R$ = regression estimate of average crop acres per land area sampling unit using SRS land area sampling unit and classified LANDSAT data

$\bar{y}_h$ = average enumerated crop acres per sampling unit in the h^{th} stratum

$\bar{x}_h$ = average LANDSAT pixels classified as a crop per sample unit in the h^{th} stratum

$\bar{\bar{X}}_h$ = average LANDSAT pixels classified as a crop per sample unit for the entire land area in the h^{th} stratum (a population mean, not a sample mean, for the stratum)

$\hat{b}_h$ = regression coefficient for the h^{th} stratum when regressing enumerated crop acres on classified pixels for the sampled units

r_h^2 = coefficient of determination between enumerated crop acreage and classified crop pixels for sampled units in the h^{th} stratum

$v(\hat{Y}_R)$ = sampling variability of the estimate $\hat{Y}_R$

$r.s.e.(\hat{Y}_R)$ = relative sampling error of the estimate $\hat{Y}_R$

Then

$$\hat{Y}_R = N \hat{\bar{Y}}_R = N \left[\bar{y}_h - \hat{b}_h (\bar{x}_h - \bar{X}_h) \right]$$

$$v(\hat{Y}_R) = \sum_h \left[\frac{n_h - 1}{n_h - 2} \cdot (1 - r_h^2) \cdot v(\hat{Y}_h) \right]$$

$$r.s.e.(\hat{Y}_R) = \frac{\sqrt{v(\hat{Y}_R)}}{\hat{Y}_R}$$

Note that if the correlation, expressed by r_h^2, between pixels and acreage is high enough a significant reduction would be expected from the sampling variability of the land area sample alone.

Following are results for 29 Illinois counties covered on a single LANDSAT pass on August 4, 1975 by three scenes. Ground enumerated data from 83 sample units were matched to this date.

Also available for comparison are data from the "Illinois State Farm Census." This is a post growing season accounting of specified crop and livestock items obtained as an adjunct to a state tax accounting. The "census" is not a controlled accounting

and adjustments are made for consistency but these data provide an independent comparison for level. The cost of this program is about $120,000.

Correlations and percent correct classification, as shown in Table 1, are not particularly dramatic but the correlations are high enough to significantly reduce relative sampling errors, as shown in Table 2. Note the seeming lack of relation between correlation and percent correct classification.

Table 1.--Relationship of ground acreage data to LANDSAT pixel data, Illinois western pass, August 4, 1975, 29 counties

Scene*	Crop	r^2	Correct classification percent
North	Corn	.83	54
	Soybeans	.81	72
Central	Corn	.63	51
	Soybeans	.62	65
Pass	Corn	.70	52
	Soybeans	.67	63

*Scene "South" contained only 3 of the 29 counties and was excluded because of insufficient data for meaningful comparison. All three scenes were aggregated for "Pass."

Only two of ten land use categories are shown in Table 2. Corn and soybeans are major crops in this 29 county area. Results were generally less impressive for the remaining eight categories not shown in the tables. Note, in Table 2, that the August 4, 1975 estimates for the direct expansion and ratio estimators are not estimating the same population as the farm census. The direct expansion and ratio estimates are "standing" acres while the state farm census represents a harvested for grain estimate. The relative sampling error for the ratio estimator is significantly lower than for the direct expansion estimator. If the

land area sampling frame alone was used, the sample size would have to be more than doubled to achieve this result.

Table 2.--Estimated acres* and precision, Illinois western pass, August 4, 1975, 29 counties

Estimator	Corn		Soybeans	
	Acres	r.s.e.	Acres	r.s.e.
Direct expansion (area sample)	4,110,200	3.6%	1,539,200	7.7%
Regression (area sample and LANDSAT)	4,126,400	2.5%	1,681,800	5.2%
Farm census	3,653,800	N.A.	1,707,400	N.A.

*Acres rounded to hundreds but zeros entered to retain magnitude.

Table 3 provides some results for selected counties. The relative sampling errors are substantial however, without LANDSAT data it would not be possible to generate a probability estimate at the county level at all.

These Illinois scenes are cloud free and this paper will not discuss the procedures for handling cloud cover.[5] It should be mentioned however that significant crop acreage differences between cloud cover and cloud free domains have been found and reported in associated research.[5]

Table 3.--Regression estimated acres* and precision for selected counties, Illinois western pass, August 4, 1975, 29 counties

County	Corn		Soybeans	
	Acres	r.s.e.($\hat{Y}_R$)	Acres	r.s.e.($\hat{Y}_R$)
		Percent		Percent
Adams	166,600	24.0	83,600	35.3
Bureau	254,000	18.7	110,600	33.4
Carroll	126,500	17.5	57,200	29.6
Cass	91,700	20.3	54,100	25.5
Greene	136,800	19.2	76,000	24.8
McDonough	162,500	17.4	82,500	26.3
Ogle	223,000	19.0	51,500	64.2
Peoria	124,000	24.0	65,300	32.6

*Acres rounded to hundreds but zeros entered to retain magnitude.

SUMMARY

Reviewing--the major features of this research application are:

1. LANDSAT data are used as a secondary, not a primary, data source. Probability ground data are required.
2. The impact of problems such as classification accuracy, signature extension and cloud cover (or more generally non-response) are eliminated or minimized. Statistical inference permits one to circumvent these problems--not solve them.
3. The advantages to be gained, over the present SRS data collection system, are reduced sampling variability and the ability to generate reasonably precise estimates for small land areas. These small areas do not have to correspond to the political boundaries of counties and states.

Please note, and this is of particular importance, that classification accuracy, measured by "percent correct," need not be high to achieve correlations that are high enough to improve estimating ability. Also, classification accuracy can be sacrificed for improved correlation. This seems contrary to what one might think but there is theoretical as well as empirical evidence this is in fact a true condition.

A major area that hasn't been explored, but in which gains could be expected, is in the use of temporal data. The proper use of temporal data could result in greater improvements in correlation than improvements to be made through classification techniques and other refinements.

THE QUESTION

The research discussed in this paper has not progressed to the point where a positive recommendation can be made to the agency to become involved in a long term commitment to utilize satellite remotely sensed data. Information for making a recommendation should be available within the next few months. If the recommendation would be positive, a likely course of action would be that the agency establish, on a trial basis, a program whereby small area statistics would be offered users on a cost reimbursable basis. Public money would be necessary to support the trial. If, after a trial period, it did not appear that the service was cost effective, either through direct reimbursement by users or through public support, the program would be withdrawn.

So, one Federal Government agency has the question before it right now. Will a user public, yet to be fully identified, pay a significant amount more for a better product?

References

[1]"Scope and Methods of the Statistical Reporting Serivce," Miscellaneous Publication No. 1308, USDA, SRS, July 1975.

[2]Von Steen, Donald and Wigton, William, "Crop Identification and Acreage Measurement Utilizing LANDSAT Imagery," USDA, SRS.

[3]Ray, Robert M., III and Huddleston, Harold F., "Illinois Crop-Acreage Estimation Experiment," Purdue: LARS Symposium Proceedings, Machine Processing of Remotely Sensed Data, June 1976, IEEE Catalog No. 76CH1103-1 MPRSD.

[4]Gleason, C. P., Hanuschak, G. A., Starbuck, R. R., and Sigman, R. S., "Stratified Acreage Estimation in the Illinois Crop-Acreage Experiment," USDA, SRS.

[5]Hanuschak, George, "LANDSAT Estimation With Cloud Cover," Purdue: LARS Symposium Proceedings, Machine Processing of Remotely Sensed Data," June 1976, IEEE Catalog No. 76CH1103-1 MPRSD.