A NEW APPROACH TO

SMALL AREA CROP-ACREAGE ESTIMATION

Harold F. Huddleston
Statistical Reporting Service
U.S. Department of Agriculture

and

Robert M. Ray III
Center for Advanced Computation
University of Illinois

May 1976

ABSTRACT

This research was undertaken to assess the practicalities of improving the accuracy of crop acreages using remote sensing data. In obtaining state estimates, small area estimates are made which are partitioned into county estimates. The accuracy of the small area estimates is significantly improved leading to useful county estimates.

A NEW APPROACH TO
SMALL AREA CROP-ACREAGE ESTIMATION

INTRODUCTION

The basic approach has been to seek an integration of ILLIAC IV[1,2] and ARPA Network[3,4] software systems developed previously at CAC for more cost-efficient machine interpretation of LANDSAT data[5,6] with ground information systems. These systems have been implemented explicitly for interactive digitizing, storage, and retrieval of large quantities of crop-acreage information collected routinely by SRS in the course of the extensive field surveys associated with its on-going agricultural production estimation methodology. Our primary goal has been to determine the extent to which SRS ground survey samples may be employed successfully as ground-truth information for calibrating ILLIAC IV procedures for classification of LANDSAT multispectral scanner (MSS) imagery. Crop acreages are obtained by geographic regions corresponding to states, and groups of counties within states.

For this exploratory application of machine processing LANDSAT data, the state of Illinois was selected as the basic study area. All ground enumerated information was acquired during the Illinois 1975 growing season by SRS acting in collaboration with the Illinois Cooperative Crop Reporting Service. Digital data tapes for all 1975 late-summer, cloud-free LANDSAT imagery over Illinois were made available to the project by NASA's Office of Earth Observations Programs acting in cooperation with NASA's Ames Research Center.

In this paper we describe the overall methodology adopted for this investigation of the practicalities of LANDSAT imagery analysis for USDA crop-acreage estimation purposes and report research findings to date. We describe the general strategy pursued in developing a comprehensive LANDSAT imagery analysis system of the scale required for monitoring agricultural crop acreages over a geographic region of the scale of the state of Illinois. For a region corresponding to ten (10) western Illinois counties (a subset of the 102 counties of Illinois), we present preliminary crop-acreage estimation results derived from ILLIAC IV - ARPA Network analysis of LANDSAT data.

GROUND DATA COLLECTION, STORAGE, AND RETRIEVAL

In support of this research project, all crop-acreage information collected by SRS within the state of Illinois in the course of its 1975 crop and livestock surveys was retained and reformatted for use as ground-truth information for calibration of LANDSAT imagery analysis systems. These data contain complete descriptions of all agricultural and nonagricultural fields, i.e., areas of homogeneous land cover, for all operator tracts within each of 300 area segments of the SRS national survey sample that fall within the state of Illinois.

In accordance with SRS survey procedures, these 300 area segments had been selected earlier with respect to statistical sampling criteria and hence, while allocated heavily to agricultural terrains, may be considered randomly distributed throughout all land in the state. Each area segment corresponds to a geographic area of approximately one square mile. Each segment typically contains multiple operator tracts with numerous fields ranging in size from several-acre farmsteads, ponds, and forested areas to several-hundred-acre agricultural fields.

Following standard SRS survey practices, throughout the summer of 1975 ASCS aerial photographs (at a scale of 8" = 1 mile) were taken by survey enumerators to the location of each segment and used for delineation of all current field boundaries. Field boundaries for all tracts of all segments were monitored continually throughout the summer in conjunction with June, July, August, and September surveys conducted by SRS personnel. Field boundary changes from month to month were recorded using a color-coded marking system.

All crop-acreage data recorded by field enumerators on ASCS photos and interview forms were rechecked independently for consistency by personnel of the Illinois Cooperative Crop Reporting Service in Springfield. All crop-acreage data contained on survey forms were put into machine readable format. Output from this process consisted of a computer tape for which individual records represent crop-acreage information for all fields of all tracts in all segments for each of the four surveys conducted throughout the summer.

To make all crop-acreage data thus compiled convenient for LANDSAT imagery analysis purposes, all field, tract, and segment boundaries recorded on a complete set of area segment photos are presently being digitized jointly by personnel of CAC in Illinois and personnel of SRS in Washington. This task is being accomplished using graphics data tablet digitizing equipment connected via the ARPA Network to interactive DEC PDP-10 computers at Bolt, Beranek and Newman (BBN) in Boston. Data tablet digitizers at CAC are connected directly to the ARPA Network through CAC's own ANTS (ARPA Network Terminal System) computer facilities. SRS digitizing equipment has been linked to BBN computer systems via dial-up telephone line connection to ARPA Network node facilities at the National Bureau of Standards in Gaithersburg, Maryland and at Fort Belvoir, Virginia.

All "field" boundary digitizing is being accomplished using an interactive DEC PDP-10 data tablet software system developed at CAC explicitly for take-off of SRS crop-acreage data recorded on aerial photos.[7] This interactive data tablet software package was implemented as an extension of the EDITOR system -- a general PDP-10 LANDSAT imagery analysis system developed previously at CAC as an interactive ARPA Network interface to LANDSAT image interpretation procedures available on the ILLIAC IV computer at NASA's Ames Research Center.[6]

These additional procedures added to the EDITOR system for digitizing SRS crop-acreage data also include provisions for on-line geographic registration of all field boundaries digitized with respect to USGS quadrangle map coordinates. This task is done simply by mounting simultaneously both photo and quad map on the active surface of the data tablet (36" x 48") and digitizing points of geographic correspondence visible within both the photo and quad map.

After digitization and geographic registration of all segment, tract, and field boundaries delineated on any one photo, an areal-network mask is determined by the software system for the segment digitized. This segment network mask is stored as a DEC-10 disk file in terms of a list of network nodes and links representing respectively digitized field corners and boundaries.

Immediately following digitization and registration of all crop-acreage data on any photo, two-line plotter displays are produced using a drum plotter at CAC to provide a hard-copy record of the segment mask thus created. One of these displays is plotted at the exact scale of the photo itself and hence, by overlaying photo and plot, the correctness of all digitized boundaries may be conveniently checked (see Figure 1). Another display is plotted at the scale and cartographic projection of the USGS quad map and by overlaying this plot and quad map the accuracy of geographic registration may be verified.

LANDSAT IMAGERY SELECTION AND PREPROCESSING

Assuming ideal meteorological conditions, only eleven (11) frames of LANDSAT imagery are required for complete coverage of the entire state. Given prevailing conditions, however, a total of sixteen (16) frames of imagery acquired between the dates of 16 July and 7 September was deemed necessary to obtain cloud-free coverage of all of the 102 counties within Illinois. Digital data tapes and positive film imagery (both at 1:1,000,000 and 1:500,000) were obtained for each of these sixteen (16) scenes.

Having obtained a complete set of image files (LANDSAT frames and pseudo-frame) such that each county is completely contained in cloud-free fashion within at least one image file, the complete set of 102 counties is to be subdivided among nonoverlapping subsets of contiguous counties, one group of counties per each image file. These groups of counties are to be designated for project purposes as LANDSAT imagery analysis districts. All subsequent data management and machine processing of LANDSAT data is then to be structured in terms of the geographic regions corresponding to these analysis districts.[8]

Once a comprehensive set of analysis districts has been established and their corresponding LANDSAT image files created, the digital image data for each district is being geometrically corrected and geographically registered to USGS topo maps existing for the state.

Finally, all image files are being geographically registered to the SRS area segments available (and hence simultaneously also to the USGS map control already associated with all ground-truth).

DATA ANALYSIS SYSTEMS

Procedures are now operational on the ILLIAC IV for both multivariate cluster analysis and maximum-likelihood statistical classification of LANDSAT image samples.

To date results are available for only one analysis district of 10 counties in western Illinois. A maximum likelihood quadratic classifier using the prior probabilities for 10 land cover categories was used for classifying each pixel into one of the 10 categories for the entire analysis district. The prior probabilities for each category were calculated from the ground enumerated data in the 10 counties, but, in some cases, they could also have been calculated from historical acreage data.

The field acres in each crop or land use type were "expanded" to correct for varying probabilities of selecting segments. Then a sample of fields was selected independently for each crop so that each acre (or pixel) had an equal chance of being selected for cover types with 80 or more fields. That is, the probability of a field being selected was proportional to its expanded acres. The selection was made from a listing of fields ordered by segment numbers to help insure that fields would be spread over the entire LANDSAT image. The number of fields selected for calculating mean vectors and covariance matrices are given in Table 1.

The pixels for all the selected fields of a crop or cover type were combined and treated as one large field for analysis purposes; however, only the nonborder pixels were used in calculating the mean vector and covariance matrix. The data vector for each pixel represents the radiometric readings corresponding to the four sensors on LANDSAT.

Table 1. Number of Training Fields by Cover Type

Crop or Cover Type	Number Fields	Nonborder Pixels*
Corn	50	1648
Soybeans	50	1107
Perm. pasture	25	297
Dense woods	40	453
Hay	16	153
Wasteland	8	492
Alfalfa	40	183
Wheat stubble	27	86
Water (Farm ponds & lakes)	17	73
Crop pasture	21	119

* Pixels which do not touch the field boundaries

ESTIMATION PROCEDURE

Following ILLIAC IV classification of all LANDSAT pixels contained within the counties making up a particular analysis district, classification results for each crop type was aggregated to obtain individual totals for all segments sampled within the district. Also, a classified pixel total for each crop type was determined for the entire analysis district itself as well as each county.

An estimator[9] of the total acreage for a particular crop in a particular analysis district and its sampling error may then be computed as follows. The total acreage may be estimated as

$$\hat{Y}_i = N_i(\bar{y}_i - \hat{B}_i(\bar{x}_i - \bar{X}_i))$$

and the variance for a large sample of segments is:

$$V(\hat{Y}_i) = N_i^2 \; V(\bar{y}_i)(1 - r_i^2)\left(\frac{n_i - 1}{n_i - 2} \right) \quad .$$

For the individual analysis districts, the normal approximation for small samples is used, that is $V(\hat{Y}_i)$ for large samples multiplied by $(1 + \frac{1}{n_i - 3})$.

Where $\hat{Y}_i$ = total acres of the crop within all area segments contained within the i^{th} analysis district

N_i = total number of all segments contained within the i^{th} analysis district (known from sampling frame)

n_i = the number of area segments sampled in the i^{th} district

$\bar{y}_i$ = average number of acres of the crop reported per area segment for all n_i area segments sampled in the i^{th} district

$\bar{x}_i$ = average number of pixels classified into the crop per area segment for all n_i area segments sampled in the i^{th} district

$\bar{X}_i$ = average number of pixels classified into the crop per segment over all possible segments for the i^{th} district

$\hat{B}_i$ = the regression coefficient between y_{ij} and x_{ij} based on the n_i area segments sampled in the i^{th} district

y_{ij} = number of acres of the crop enumerated for the j^{th} segment sampled in the i^{th} district

x_{ij} = number of pixels classified into the crop for the j^{th} segment sampled in the i^{th} district

$$V(\bar{y}_i) = \frac{\sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{j=1}^{n_i} y_{ij}\right)^2}{n_i}}{n_i(n_i - 1)}$$

r_i^2 = correlation coefficient squared between y_{ij} and x_{ij} for the i^{th} district.

The formulas given are appropriate for a simple random sample within each analysis district. However, the SRS surveys are stratified by land use categories which require that item totals, sums of squares, and sums of cross products be weighted and combined (or pooled) in order to obtain the equivalent of a simple random sample over the entire analysis district.

The estimates and their errors are based on the 33 segments falling in the 10 western Illinois counties comprising the first analysis district corresponding to LANDSAT image ID#2194-16042 of August 4, 1975. The estimates are shown in Table 2 and their sampling errors squared in Table 3 for eight agricultural land use categories. The LANDSAT window containing the 10 counties included 4,887,960 pixels and required less than 80 seconds for classification on the ILLIAC IV. If the analysis is confined to a single district, the subscript i may be dropped from all variables.

Table 2. Estimates of Agricultural Cover Types

Crop or Cover Type	Reported Acres July 27	Regression Estimate	Pixel Count x 1.114
	---	(000 acres)	---
Corn	1286	1390	2105
Soybeans	631	701	610
Perm. pasture	533	434	678
Hay	179	154	104
Alfalfa	69	71	14
Wheat stubble	39	39	0.3
Water (Farm ponds & lakes)	28	32	10
Crop pasture	45	45	0

Table 3. Variances of Estimates of Agricultural Cover Types for 10-County Analysis District

Crop or Cover Type	Variance Reported (10^6acres^2)	Variance Regression Estimate (10^6acres^2)	Information Gain or Loss (1) ÷ (2)
	(1)	(2)	(3)
Corn	17202	2459	7.00
Soybeans	5880	847	6.94
Perm. pasture	4489	1096	4.09
Hay	630	376	1.67
Alfalfa	155	135	1.14
Wheat stubble	66	70	.94
Water (Farm ponds & lakes)	30	11	2.71
Crop pasture	88	94	.94

For an individual county within one district the same type of regression estimator as used for the entire analysis district can be employed if the counties are similar with respect to the individual crops being estimated. However, the following symbols are redefined to denote county means and totals for individual crops where the subscript i now refers to a county.

$\hat{Y}_i$ = the total acreage in a given crop in the i^{th} county

N_i = the total number of sampling units (segments) in the i^{th} county

$\bar{X}_i$ = average number of pixels classified into the crop per segment over all possible segments for the i^{th} county

but $\bar{y}$, $\bar{X}$, and $\hat{B}$ are as defined previously for the district with the subscript i dropped; that is, they are the means and regression coefficient obtained for the 10-county analysis district. This partition regression estimate is used to obtain county acreage totals even though $\bar{y}$, $\bar{x}$, and $\hat{B}$ do not relate to the individual counties but the group of 10 counties which were sampled. This implies the same relationship (i.e., regression model) between segment crop acres and classified pixels holds over all counties in the analysis district, and all pixels in each county have been classified into crop types (i.e., $\bar{X}_i$ can be determined for each county).

The estimates of corn and soybean acreages are shown for individual counties in Table 4. A gray-scale film display of the classified results for a county can also be produced to show the relative density and distribution of the two crops within the county (see Figure 2). Figure 3 shows a false color enhanced image made from the LANDSAT digital tapes.

The sampling errors were derived for the individual counties based on the regression error for the analysis district. The expected error at the county level is approximately

$$Sy \cdot x = \sqrt{1 + \frac{1}{n} + (X_i - \bar{X})^2 \div \Sigma x^2}$$

where an independent prediction for a single X value in the county is made and n = the number of segments in the analysis district.

For example, the regression estimate and expected sampling error for corn in McDonough County were computed as follows:

$$\hat{Y} = 669(206.11 - .73545(279.88 - 292.82)) = 669(215.43) = 144{,}256 \text{ acres}$$

$$\sqrt{V(\hat{Y})} = 669(45.658 \sqrt{1 + \frac{1}{33} + \frac{167.44}{27{,}450})} = 669(47.32) = 31{,}657 \text{ acres}$$

If the X value corresponding to the prediction is considered as representing the 669 segments in the county, then the error would be much less providing the between county variance component was small and dropped; that is:

$$S\bar{y} \cdot x = \sqrt{\frac{1}{n} + (\bar{X}_i - \bar{X})^2 \div \Sigma x^2}$$

or $\sqrt{V(\hat{Y})} = 669(45.658) \sqrt{\frac{1}{33} + \frac{167.44}{27{,}450}} = 669(8.712) = 5{,}828$ acres .

Based on the counties in this analysis district, the latter expression for the county error terms appears to be appropriate.

Table 4. County Estimates of Corn and Soybeans and Preliminary Assessor Census Data

County	Corn		Soybeans	
	Regression Estimate	Assessor Census	Regression Estimate	Assessor Census
	(000)	(000)	(000)	(000)
Adams	166	121	81	112
Brown	70	44	27	28
Cass	84	75	63	51
Fulton	214	155	112	95
Hancock	178	150	78	120
Henderson	94	100	42	40
Knox	186	169	92	65
McDonough	144	157	92	93
Schuyler	106	68	39	57
Warren	148	157	75	78
10-County Total	1390	1196	701	739

$\bar{X}_i$ = average number of pixels classified into crop per segment for i^{th} county,

$\bar{X}$ = average number of pixels classified into crop per segment for the district, and

Σx^2= the corrected sums of squares derived from the 33 segments in the district.

Using the first expression for the error, the relative error, i.e., coefficient of variation of the total acres, at the county level is approximately 23 percent for corn and 27 percent for soybeans. A comparison of the regression estimate with the Assessor Census Data which is not subject to sampling error also gives a measure of the average error at the county level. The average error at the county level, ignoring the sign, based on this comparison is approximately 25 percent for corn and 19 percent for soybeans. In general, this type of comparison is available for only a few states. However, it should be pointed out that there are some conceptual differences between the regression estimate and the Assessor Census Data given in Table 4. The regression estimate is the standing acreage at a given point in time, i.e., about August 1, while the Assessor Census relates to the acreage harvested for all purposes during the crop year. For corn and soybeans in Illinois, this difference is not considered important by the authors.

If the relationship of reported acres to pixels for all counties are not similar for individual crops, a different model may need to be considered which will include a county "effect." Also a ratio estimate might be considered as the basis for partitioning the area estimate to the individual counties.

The ratio (or fraction) for a given crop in a county would be the total classified pixels in the county divided by the total classified pixels in the district. For this group of counties in Illinois, the ratio estimates are almost identical with the regression estimates shown in Table 4.

SUMMARY

The use of LANDSAT data in conjunction with the SRS area sample data appears promising as a technique to obtain small area and county acreage estimates on an annual basis. The degree of precision of these estimates may vary considerable because the size of the current SRS area sample varies considerable from state to state. In addition, the current sample is not designed to provide small area estimates and is likely to be useful only for crops which have large acreages in a given area. Nevertheless, it seems likely that both the quantity and quality of small area data can be significantly improved by the joint use of LANDSAT and SRS information. Many of the software and computer facilities are now available for developing annual statistics by small geographic areas. Also, the modifications in the current SRS system appears feasible in terms of data collection requirements.

The successful use of the LANDSAT images for small area estimates does require certain conditions be met, namely:

(1) excellent quality, cloud-free LANDSAT imagery

(2) good geographic registration of ground data and small areas to LANDSAT imagery

(3) estimate of mean vectors and covariance matrices for each crop for each LANDSAT image used (i.e., the small area)

(4) prior probabilities for each crop type, i.e., approximate fraction of land devoted to crop

(5) an adequate number of ground segments for each LANDSAT image (100nm x 100nm) to compute the regression coefficient, correlation coefficient and means used in estimator formulas

(6) the small area estimate obtained through the use of current year crop acreages from the SRS area sample with classified crop data from LANDSAT to remove the bias from the LANDSAT data and reduce the sampling error to produce more accurate area estimates

(7) the small area estimate can then be partitioned to individual counties within the area based on a linear model appropriate for individual counties.

REFERENCES

1. D. L. Slotnick, "The Fastest Computer," Scientific American, Vol. 224, No. 2, February 1971, pp. 76-87.

2. W. J. Bouknight, S. A. Denenberg, D. E. McIntyre, J. M. Randall, A. H. Sameh, and D. L. Slotnick, "The ILLIAC IV System," Proceedings of the IEEE, Vol. 60, No. 4, April 1972, pp. 369-388.

3. L. G. Roberts and B. D. Wessler, "Computer Network Development to Achieve Resource Sharing," 1970 Spring Joint Computer Conference, AFIPS Conference Proceedings, Spartan, 1970, pp. 543-549.

4. L. G. Roberts, "Network Rationale: A 5-Year Reevaluation," COMPCON 73 Proceedings, Seventh Annual IEEE Computer Society International Conference, March 1973, pp. 3-5.

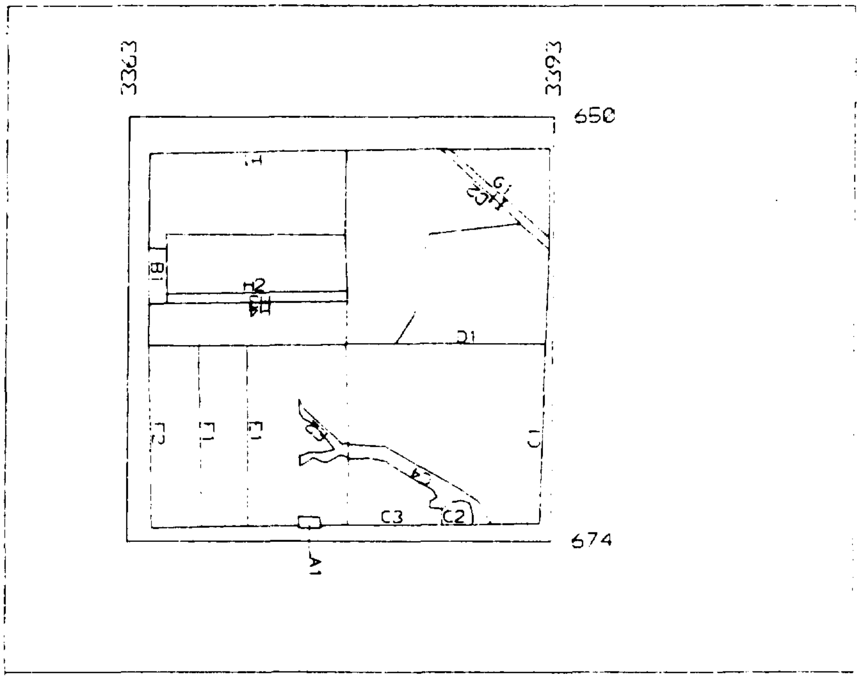
5. Robert M. Ray III, John D. Thomas, Walter E. Donovan, and Philip H. Swain, "Implementation of ILLIAC IV Algorithms for Multispectral Image Interpretation," CAC Document No. 112, (June 1974), Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801.

6. Robert M. Ray III, Martin Ozga, Walter E. Donovan, John D. Thomas, and Marvin L. Graham, "EDITOR: An Interactive Interface to ILLIAC IV - ARPA Network Multispectral Image Processing Systems," CAC Document No. 114, (June 1975), Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801.

7. Walt Donovan and Martin Ozga, "Retrieval of LANDSAT Image Samples by Digitized Polygonal Windows and Associated Ground Data Information," CAC Technical Memorandum No. 57, (August 1975), Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801.

8. Robert M. Ray III and Harold F. Huddleston, "Illinois Crop-Acreage Estimation Experiment," Presented at the Third Symposium on Machine Processing of Remotely Sensed Data, Purdue University, June 1976.

9. W. G. Cochran, "Sampling Techniques," (2nd Ed.) (1963), Wiley and Sons, pp. 193-200.



```
SEGMENT 5159. PART 1 OF 1.  18 FIELDS  92 EDGES  75 VERTICES
TOTAL AREA       653.3 ACRES. SCALE IS 1.  11891
PLOT FILE <ILLINOIS>5159.SEG;5.22-MAR-76 05.00.25
USER IS   replot of segments using local calibration
CAL  FILE <ILLINOIS>5159.CAL;1.25-MAR-76 13.48.49
```

Figure 1. Example USDA/SRS Area Segment Mask Plotted at Scale of Photo Digitized. (Shown reduced here.)

Figure 2. Distribution of Classified Corn and Soybean Pixels for McDonough County.

Legend: Off-white = Corn
Yellow = Soybean
Green = Pasture & woods
Blue = Water

Black = Wasteland (include two cities at center and in upper right hand quarter)
Brown = Wheat Stubble, Hay, Alfalfa

Figure 3. False Color Enhanced Image of McDonough County.