80-2

FILE COPY

CROP AREA ESTIMATES FROM LANDSAT AND GROUND SURVEY DATA*

by

Richard Sigman
ESCS/U.S. Department of Agriculture

1980

I. INTRODUCTION

The Economics, Statistics, and Cooperatives Service (ESCS) of the U.S. Department of Agriculture is presently conducting research in possible uses of LANDSAT satellite data in agricultural surveys. This research is in the following areas:

1. improvement of crop-hectarage estimates for multi-county areas, such as Crop Reporting Districts and states,

2. development of small-area crop-hectarage estimates for individual counties, and

3. photo-interpretive use of LANDSAT imagery in developing area sampling frames.

This paper briefly describes ESCS's statistical methodology and some recent applications in using LANDSAT data to improve crop-hectarage estimates for multi-county areas. Cardenas, et al [1] discuss ESCS's research in developing small-area estimates from LANDSAT data; whereas, Hanuschak and Morrissey [2] describe ESCS's use of LANDSAT imagery in developing area sampling frames.

II. DATA SOURCES

A. GROUND-SURVEY DATA

As a part of its operational program, ESCS conducts in late May an annual nationwide agricultural survey called the June Enumerative Survey (JES). The JES sample units, called segments, are well-defined areas of land, typically

*Presented at Canadian Federal-Provincial Committee on Agricultural Statistics; March 11-12, 1980; Ottawa, Ontario.

one-square mile in size. Two levels of stratification are employed. The first-level strata are the individual states. Secondary strata are areas of land within a state which have similar patterns of land use. Defined in terms of the percent of land under cultivation, these secondary strata are determined by photo-interpretation of aerial photography. Stratum definitions in the state of Illinois, for example, are given in Table 1.

During the JES interviews, the hectares devoted to each crop or land use are recorded for each field in the sample units. The scope of information collected by the JES, however, is much broader than crop hectarage alone. Estimated items include crop hectares by intended utilization, grain storage on farms, livestock inventory by various weight categories, and agricultural labor and farm economic data. The ground data used in the studies reported here have been derived from special tabulations in conjunction with the JES and include information to update the data to near-date of the LANDSAT acquisition.

B. LANDSAT DATA

The basic element of LANDSAT data is the set of measurements by the satellite's multispectral scanner (MSS) of a .4 hectare area of the earth's surface. The MSS measures the amount of radiant energy reflected and/or emitted from the earth's surface in four different regions (bands) of the electromagnetic spectrum--green, red, and two near-infrared regions.

The individual .4 hectare MSS resolution areas, referred to as pixels, are arrayed along east-west running rows within the 185 kilometer wide north-to-south pass of the LANDSAT satellite. A given point on the earth's surface is imaged once every eighteen days by the same LANDSAT satellite and once every nine days by either one of two satellites. Satellite passes which are adjacent on the earth's surface are at least one day apart with respect to their dates of imagery.

III. STATISTICAL METHODOLOGY

ESCS's approach for using LANDSAT data is to use it as an auxiliary variable with existing operational ground surveys [3]. The information from these surveys is actually used twice in the ESCS procedure for computing LANDSAT-based crop-hectarage estimates. The ground-survey data is used (1) as "ground-truth" for developing a set of discrimination functions for the LANDSAT data, and (2) as the primary survey variable for estimating crop-hectarage.

A. DIRECT EXPANSION ESTIMATION (GROUND DATA ONLY)

The estimation procedure presented here is for a given state. National totals are then obtained by appropriately combining state totals.

Let $h = 1,2,\ldots,L$ be L land-use strata. Within each stratum, the total area is divided into N_h area-frame units from which a simple random sample of n_h units is drawn. Using only JES data for the L strata, an estimate of total hectares of a particular crop (corn, for example) can be computed by direct expansion as follows:

Let Y = Total corn hectares for a state (Illinois, for example).

$\hat{Y}$ = Estimated total corn hectares for the state.

y_{hj} = Total corn hectares in j^{th} sample unit in the h^{th} stratum.

Then

$$\hat{Y}_{DE} = \sum_{h=1}^{L} N_h \bar{y}_h \tag{1}$$

where $\bar{y}_h$ = the average corn hectares per sample unit from the ground survey for the h^{th} land-use stratum

$$= \sum_{j=1}^{n_h} y_{hj} / n_h$$

The estimated variance of the estimate is:

$$v(\hat{Y}_{DE}) = \sum_{h=1}^{L} v_h(\hat{Y}_{DE})$$

$$= \sum_{h=1}^{L} \frac{N_h^2}{n_h(n_h - 1)} \frac{N_h - n_h}{N_h} \cdot \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2$$

Note that we have not yet made use of an auxiliary variable such as classified LANDSAT pixels. For major crops the JES provides state-level estimates with relative sampling errors on the order of 3 to 8 percent.

B. REGRESSION ESTIMATION (GROUND DATA AND CLASSIFIED LANDSAT DATA)

ESCS extracts information from LANDSAT data by classifying individual pixels as to probable crop type. This classification is performed by a collection of discriminant functions which are defined over the MSS measurement space. (Pixel classification is explained in more detail in the next section.)

By means of a regression estimator both ground data and classified LANDSAT data can be utilized to estimate crop hectarage. (Regression estimators are discussed in most sampling texts, e.g. Cochran [4].) The estimate of Y using the <u>separate</u> form of the regression estimator is

$$\hat{Y}_R = \sum_{h=1}^{L} N_h \cdot \bar{y}_{h(reg)}$$

where

$$\bar{y}_{h(reg)} = \bar{y}_h + \hat{b}_h \cdot (\bar{X}_h - \bar{x}_h)$$

and $\hat{b}_h$ = the estimated regression coefficient for the hth land-use stratum when regressing ground-reported hectares on classified pixels for the n_h segments.

$$= \frac{\sum_{j=1}^{n_h}(x_{hj} - \bar{x}_h)(y_{hj} - \bar{y}_h)}{\sum_{j=1}^{n_h}(x_{hj} - \bar{x}_h)^2}$$

$\bar{X}_h$ = the average number of pixels classified as corn per frame unit for all frame units in the hth land-use stratum. Thus whole LANDSAT scenes must be classified to calculate X_h. Note that this is the mean for the population and not the sample.

$$= \sum_{i=1}^{N_h} X_{hi}/N_h$$

where X_{hi} = number of pixels classified as corn in the ith area-frame unit of the hth stratum.

$\bar{x}_h$ = the average number of pixels classified as corn per sample unit in the hth land-use stratum

$$= \sum_{j=1}^{n_h} x_{hj}/n_h \ .$$

x_{hj} = number of pixels classified as corn in the jth sample unit in the hth strata.

The estimated (approximate) variance for the separate regression estimator is

$$v(\hat{Y}_R) = \sum_{h=1}^{L} \frac{N_h^2}{n_h} \frac{N_h - n_h}{N_h} \cdot \sum_{j=1}^{n_h}(y_{hj} - \bar{y}_h)^2 . \frac{1 - R_h^2}{n_h - 2}$$

where $\hat{R}_h^2$ is an estimate of

R_h^2 = population coefficient of determination between reported corn hectares and classified corn pixels in the hth land-use stratum.

$$= \frac{[\sum_{i=1}^{N_h} (Y_{hi} - Y_h)(X_{hi} - X_h)]^2}{[\sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2] [\sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2]}$$

Note that,

$$v(\hat{Y}_R) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h - 2} (1 - \hat{R}_h^2) v_h(\hat{Y}_{DE}) \qquad (2)$$

and so $\lim v(\hat{Y}_R) = 0$ as $\hat{R}_h^2 \longrightarrow 1$ for fixed n_h. Thus a substantially lower variance is obtained if the coefficient of determination is close to 1 for most strata. (Methods for estimating R_h^2 are discussed in the next section).

The estimate of Y using the combined form of the regression estimator is

$$\hat{Y}_R = N \bar{y}(reg)$$

where $N = \sum_{h=1}^{L} N_h$

$$\bar{y}(reg) = \bar{y} + b_c (\bar{X} - \bar{x})$$

$$\bar{X} = (\sum_{h=1}^{L} \sum_{i=1}^{N_h} X_{hi})/N$$

$$\bar{x} = (\sum_{h=1}^{L} N_h \bar{x}_h)/N$$

and $\bar{y} = (\sum_{h=1}^{L} N_h \bar{y}_h)/N$.

The approximate variance of the combined regression estimator and the expression for $\hat{b}_c$ are given in Cochran [4, pp 202-203].

When a LANDSAT pass does not cover the entire state on one date, it is necessary to partition the state into analysis areas which are wholly contained within the individual passes. The estimation procedure described above is carried out in each analysis area, and then analysis-area-level estimates as well as variances are combined to the state level by treating the analysis areas as post-strata.

The relative efficiency of the regression estimator compared to the direct expansion estimator will be defined as the ratio of the respective variances:

$$\text{R.E.} = v(\hat{Y}_{DE}) \;/\; v(\hat{Y}_R) \tag{3}$$

The auxiliary variables described above, i.e.

$$x_{hj} = \sum_k c(z_{hjk}) \text{ and } X_{hj} = \sum_k c(Z_{hik}) \tag{4}$$

where the variable z_{hjk} (Z_{hik}) is the MSS data for the k^{th} pixel of the j^{th} sample unit (i^{th} area-frame unit) in the h^{th} stratum and the function c(z) is 1 if radiometric measurement z is classified as the crop of interest and 0 otherwise, are probably not optimum in the sense of producing the estimate of Y with smallest possible variance. Alternate approaches which are being investigated are

1. using a multiple regression estimator, where the set of auxiliary variables includes not only the quantities in equation (4) but also the classification results into cover types other than the crop of interest (discussed in [5]); and

2. changing c(z) in equation (4) to the posterior probability that a pixel having radiometric measurement z is from the crop of interest. The posterior probability function can be estimated by approximating it with a linear combination of basis functions with the coefficients estimated by least squares

(suggested by Fuller [6]) or by assuming a logistic form for the posterior probability and then estimating unknown parameters by maximum likelihood.

C. PIXEL CLASSIFICATION

The pixel classifier is a set of discriminant functions corresponding one-to-one with a set of classification categories. Each discriminant function consists of the category's likelihood multiplied by the category's prior probability. If the prior probabilities used are correct for the population of pixels being classified, then the resulting set of discriminant functions, called a Bayes classifier, minimizes the over-all probability of misclassifying a pixel.

In crop-hectarage estimation, however, the objective is to minimize the variance of resulting hectarage estimates. Since minimizing the over-all probability of misclassification does not necessarily achieve this objective, optimum hectarage estimation may require the use of prior probabilities different from the optimum Bayes set. (Strictly speaking, there is only one correct set of prior probabilities for a given geographical region, i.e. the actual probabilities of occurrence for the various cover types. Using "different prior probabilities" actually means using different weighting factors for the category likelihoods in computing the category discriminant functions.) We have investigated two types of "prior probabilities": equal probabilities and probabilities proportional to direct-expanded hectarage, i.e. the $\hat{Y}_{de}$'s. Equal prior probabilities have yielded more precise crop-hectarage estimates (compared to using probabilities proportional to direct-expanded hectares) in most cases for corn and for wheat and in some cases for soybeans.

Since the type of ground cover in every JES field is known as a result of JES enumeration, the pixels lying inside JES fields are of known cover type.

These pixels, called field-interior pixels, determine the cover types for which classification categories are created. In addition, pixels are selected from rivers, lakes, and ponds to determine classification categories for surface water.

The field-interior pixels for a given cover type are extracted from the LANDSAT data, and the corresponding MSS data are clustered in MSS measurement space. A classification category is then associated with each cluster which has more than some specified number of pixels (usually 100 pixels).

Category likelihoods are computed by assuming that the MSS data in a given category follow a multivariate normal distribution. Thus, the calculation of category discriminant functions involves the estimation of category means, covariances, and prior probabilities. Once this has been done, all the JES *segment-interior* pixels (includes field-boundary pixels) can be classified and the *sample* coefficient of determination

$$r_h^2 = \frac{\left[\sum_{j=1}^{n_h}(y_{hj} - \bar{y}_h)(x_{hj} - \bar{x}_h)\right]^2}{\left[\sum_{j=1}^{n_h}(y_{hj} - \bar{y}_h)^2\right]\left[\sum_{j=1}^{n_h}(x_{hj} - \bar{x}_h)^2\right]}$$

calculated. In small samples, however, r_h^2 can have a large positive bias as an estimate of R_h^2 because much of the same data is used to both develop the sample discriminant functions and to compute r_h^2. Less biased estimates for R_h^2 can be obtained by many of the same methods used to estimate error rates in discriminant analysis; e.g., jackknifing, sample partition, etc. We have found, however, that in moderate size samples, e.g. $n_h = 84$, that the difference between r_h^2 and a jackknifed estimate of R_h^2 is acceptably small so as to not warrant the additional labor involved in performing the jackknife calculations [7, 8].

IV. RECENT APPLICATIONS

ESCS has applied the methodology described above in a number of different areas in the United States over the past several years. Major demonstration efforts have been conducted for entire states--Illinois, Kansas, Iowa, and Arizona--and for various sub-state areas--Kings county, California; eastern Arkansas; and Spink county, South Dakota [7, 8, 9, 10, 11, 12]. Results from these studies are summarized in Table 2.

The majority of these LANDSAT studies by ESCS have been completely research oriented. Nevertheless, timely LANDSAT regression estimates for use by operational elements of ESCS were obtained in 1978 in Iowa and again in 1979 in Arizona. In both of these studies ESCS completed all processing of LANDSAT data and calculation of regression estimates by mid-December of the crop year. Consequently, in 1978 and 1979 LANDSAT-based state level estimates were input to USDA's Crop Reporting Board for use in compiling the annual crop summaries for Iowa and Arizona, respectively. In addition, the regression estimates for individual LANDSAT analysis districts were supplied to ESCS's Iowa and Arizona state statistical offices for developing sub-state crop area estimates. The LANDSAT-based regression estimates were not the sole source of data in determining state and sub-state estimates, however.

ESCS's Iowa and Arizona LANDSAT studies also demonstrated, however, that a number of difficulties accompany attempts to obtain timely results from LANDSAT. Chief among these are unusable LANDSAT acquisitions as a result of clouds and the delayed delivery of LANDSAT data tapes. During the Iowa and Arizona projects, both LANDSAT's II and III were in operation. Nevertheless, because of clouds 13 out of 99 Iowa counties had no usable LANDSAT data. For these 13 counties crop-area estimates were calculated soley from ground data. With only LANDSAT

III presently operating, the problem of lack of usable LANDSAT data will probably worsen.

For the Iowa study data delivery time for a LANDSAT tape (that is, time from satellite overpass to receiving of data by ESCS) ranged from 4 to 13 weeks with a median delivery time of 7 weeks. For the 1979 Arizona study, three LANDSAT tapes were not available until February 1980, several months past the last date of usefulness of this data for developing timely crop-year regression estimates. (Ground-data only were used to develop crop area estimates for this three scene area; whereas, LANDSAT data covering 10 scenes were used to develop regression estimates for the remainder of Arizona.) The value of LANDSAT-based regression estimates for spring-seeded crops would increase if they were available earlier than December 31, such as by December 1 or even by November 1. For this to occur, however, requires that the delivery time for LANDSAT data tapes be greatly shortened, such as reciving data two to four weeks after acquisition.

V. REFERENCES

1. Cardenas, Manuel; Blanchard, Mark M.; Craig, Michael E.; "Small Area Estimators: County Crop Acreage Estimates Using LANDSAT Data," contributed paper, 1978 annual ASA meeting, San Diego, California.

2. Hanuschak, George A. and Morrissey, Kathleen M., "Pilot Study of the Potential Contributions of LANDSAT Data in the Construction of Area Sampling Frames," Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C., October 1977.

3. Von Steen, Donald H. and Wigton, William H., "Crop Identification and Acreage Measuremtnt Utilizing LANDSAT Imagery," Statistical Reporting Service, United States Department of Agriculture, Washington, D.C., March 1976.

4. Cochran, William G., Sampling Techniques, (2nd Ed.), John Wiley & Sons, 1963.

5. Hanuschak, George A. and Cardenas, Manuel, "Multiple Regression Estimation Using Classified LANDSAT Data," Economics, Statistics, and Cooperative Service, U.S. Department of Agriculture, Washington, D.C., April 1978.

6. Fuller, Wayne, personal communication to William Wigton, December 1977.

7. Sigman, Richard S.; Gleason, Chapman P.; Hanuschak, George A.; and Starbuck, Robert A.; "Stratified Acreage Estimates in the Illinois Crop-Acreage Experiment," Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana.

8. Gleason, Chapman; Starbuck, Robert R.; Sigman, Richard S.; Hanuschak, George A.; Craig, Michael E.; Cook, Paul W.; and Allen, Richard D.; "The Auxiliary Use of LANDSAT Data in Estimating Crop Acreages: Results of the 1975 Illinois Crop-Acreage Experiment," Statistical Reporting Service, U.S. Department of Agriculture, Washington D.C., October 1977.

9. Craig, Michael E.; Sigman, Richard S.; and Cardenas, Manuel; "Area Estimates by LANDSAT: Kansas 1976 Winter Wheat," Economics, Statistics, and Cooperatives Service; U.S. Department of Agriculture, Washington D.C., August 1978.

10. Hanuschak, George; Sigman, Richard; Craig, Michael; Ozga, Martin; Luebbe, Raymond; Cook, Paul; Kleweno, David; and Miller, Charles; Economics, Statistics, and Cooperatives Service; U.S. Department of Agriculture, Washington D.C., Technical Bulletin No. 1609, August 1979.

11. Craig, Michael E. and Miller, Charles E.; "Area Estimates by LANDSAT: Arizona 1979," Economics, Statistics, and Cooperatives Service; U.S. Department of Agriculture, Washington D.C., January 1980.

12. Cook, Paul W.; "Sunflower Acreage Estimation in South Dakota," Economics, Statistics, and Cooperatives Service; U.S. Department of Agriculture, Washington D.C., in progress.

Table 1. Illinois stratum definitions .

stratum		sub-stratum	
#	description	#	description
10	intensive agriculture	11	75%+ cultivated
		12	50% - 75% cultivated
50	non-intensive agriculture	20	15% - 49% cultivated
		31\	\
		32	:urban :non-
		33/	:cultivated
		40	range land :
		61	proposed water :
		62	water /

ENTIRE-STATE STUDIES

AREA	IMAGE DATES	SEGMENTS/ ANALYSIS DISTRICT	MAJOR CROPS	$\hat{R}^2_H$'S	ANALYSIS DISTRICT RELATIVE EFFICIENCIES
ILLINOIS 1975	JULY 16 - SEP 7	30 - 84	CORN, SOYBEANS	.05 - .86 .22 - .98	1.9 - 6.3 1.1 - 5.8
KANSAS 1976	APR 1 - MAY 6	11 - 35 (SUB-SAMPLE)	WINTER WHEAT	.60 - .92	3.1 - 13.0 (WRT SUB-SAMPLE)
IOWA 1978	AUG 6 - SEP 4	9 - 80	CORN, SOYBEANS	.07 - .94 .45 - .98	1.0 - 6.0 2.7 - 7.6
ARIZONA 1979	JULY 2 - JULY 26	9 - 52	COTTON, ALFALFA	.53 - .84 0.0 - .66	2.0 - 6.1 1.6 - 2.9

SUB-STATE STUDIES

AREA	IMAGE DATES	SEGMENTS/ ANALYSIS DISTRICT	MAJOR CROPS	R_H^2'S	ANALYSIS DISTRICT RELATIVE EFFICIENCIES
KINGS CNTY CALIFORNIA 1977	AUG 15	15 (SUB-SAMPLE)	COTTON, BARLEY, WHEAT	≥.80	5.2 - 28.0 (WRT SUB-SAMPLE)
EASTERN ARKANSAS 1978	JUNE 30	37, 42	RICE,	.02 - .81	2.5, 5.1
			SOYBEANS	.62 - .72	2.3, 2.6
			COTTON	.36 - .65	1.5, 1.5
SOUTH DAKOTA SPINK CNTY 1979	AUG 25	53 (¼ SECTIONS)	SUNFLOWERS	.53, .93	13.8
SNAKE RIVER VALLEY IDAHO 1978	JULY 18 JULY 26	56 82	WHEAT	.25-.85	3.3, 5.0
			POTATOES	.06-.80	5.6, 1.2
			BARLEY	.06-.67	1.7, 1.5
			ALFALFA	.23-.76	2.2, 1.9