The Use of Regression Estimation With LANDSAT and Probability Ground Sample Data

Richard S. Sigman, George A. Hanuschak, Michael E. Craig, Paul W. Cook, and Manuel Cardenas
U. S. Department of Agriculture

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 discusses some recent applications in using LANDSAT data to improve crop-hectarage estimates for multi-county areas. ESCS's research in developing small-area estimates from LANDSAT data is discussed in another paper at this conference [1]. 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 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 visual interpretation of aerial photography. Stratum definitions in the state of Illinois, for example, are given in Table 1.

Table 1. Stratum numbers and definitions

stratum description	sub-stratum 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 : (30) 61 proposed water : 62 water

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, called an individual signature, 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 various regions (bands) of the electromagnetic spectrum. The LANDSAT II and LANDSAT III MSS's have four and five bands, respectively.

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 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 data acquired from 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 discriminant 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),

$\hat{Y}_{DE}$ = Direct expansion estimate of total corn hectares in the state.

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

Then

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

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 2 to 8 percent.

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

ESCS extracts information from LANDSAT data by classifying individual signatures 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 separate form of the regression estimator is

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

where

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

and $\hat{b}_h$ = the estimated regression coefficient for the h^{th} 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 h^{th} 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 i^{th} area-frame unit of the h^{th} stratum.

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

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

x_{hj} = number of pixels classified as corn in the j^{th} sample unit in the h^{th} strata.

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

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

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 h^{th} land-use stratum.

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

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 $\hat{R}_h^2 \to 1$ 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 $N = \sum_{h=1}^{L} N_h$

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

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

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

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 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:

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

The auxiliary variables described above, i.e.

$$x_{hj} = \sum_{k} c(z_{hjk}) \text{ and } X_{hj} = \sum_{k} c(Z_{hik}) \qquad (4)$$

where the variable z_{hjk} (Z_{hik}) is the signature of 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 signature z is classified as the crop of interest and 0 otherwise. These auxiliary variables 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 with signature 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 Y_{DE}. The results of this investigation are discussed in the next section.

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 signatures 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 signatures in a given category follow a multivariate normal distribution. Thus the calculation of category discriminant functions involves the estimation by category of signature means and covariances and prior probabilities. Once this has been done, all the JES segment-interior pixels (field-boundary pixels included) can be classified and the sample coefficient of determination

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

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 U.S. over the past several years. Major demonstration efforts have been conducted in Illinois, Kansas, and Kings County, California. All of these studies have been performed in a purely research mode, and except for the 1977 study effort in Kings County, California, none of these demonstration projects have produced timely crop hectarage estimates. Also, this methodology is not yet demonstrably cost effective. In 1978, however, ESCS expects to complete LANDSAT crop-hectarage estimates in time for input to USDA final season estimates for Iowa.

A. 1975 ILLINOIS STUDY [7,8]

1975 LANDSAT data for the entire state of Illinois was used to estimate crop hectarages for Illinois spring-seeded crops at county and multi-county levels. Requiring three LANDSAT passes to completely image the state, the dates of imagery of the analyzed LANDSAT data ranged from July 16 to September 7. On account of the different dates of analyzed LANDSAT data, the state was partitioned into six analysis areas,

The distribution of the 300 Illinois JES segments into the six areas ranged from 30 to 84 segments per analysis region.

The separate form of the regression estimator was used in Illinois. Cover types for which classification categories were created were corn, soybeans, alfalfa, other hays, permanent pasture, wheat stubble, oats and oat stubble, dense woodlands, water, and other non-agricultural land (called waste). Only for corn, soybeans, water, and waste, however, did the use of LANDSAT data result in significant increases in precision (relative to using JES data alone) of analysis-area crop-hectarage estimates. For the analysis-area estimates, the regression estimate relative efficiencies for corn ranged from 1.3 to 6.3; for soybeans, from 1.1 to 5.8.

One of the major factors determining the ability of LANDSAT data to improve crop-hectarage estimates was the acquisition data of the LANDSAT imagery. Best results were obtained for August 3 and 4, when corn was nearly 100% silked. In the calculation of category discriminant functions, it was observed that using equal prior probabilities yielded more precise crop-hectarage estimates (compared to using probabilities proportional to direct expanded hectares) in most cases for corn and in some cases for soybeans.

B. 1976 KANSAS STUDY

The objective of this study was to estimate winter wheat hectarages for Kansas using 1976 LANDSAT data. In order to completely image the state, six LANDSAT passes are required. The easternmost pass, covering only four counties, was not analyzed because of insufficient JES data to estimate the required parameters. Also, the central pass was almost completely cloud covered during April, May, and June, causing loss of LANDSAT acquisitions for some major wheat-producing counties. Acquired from April 1 to May 6, usable LANDSAT data covered 87 of the 105 Kansas counties. [9]

A 40% subsample of segments from the Kansas JES was used in the LANDSAT analysis. The number of segments in the subsample ranged from 11 to 35 per pass. The combined form of the regression estimator was used because of the small number of segments from the subsample within each stratum in a pass. Since only winter wheat estimates were of interest, classification categories were created only for wheat and 'other'. The 'other' cover type was a catch-all name for anything (crop, waste, pasture, etc) not labelled as winter wheat by the USDA enumerators.

Sample coefficients of determination between classification results and ground truth were high, ranging from .60 to .92. Relative efficiencies (with respect to the subsample) ranged from 3.1 to 13.0, with the exception of the central pass. This pass was mostly cloud covered and analysis was done for only 7 counties using 11 segments. The resulting relative efficiency was slightly less than one.

C. 1977 CALIFORNIA STUDY

In both 1976 and 1977, crop-hectarage estimates using LANDSAT data were calculated for Kings County, California. In 1977, timeliness of the estimates was a primary objective. This goal was successfully achieved: using July 7 LANDSAT data, the analysis was completed on August 15, 1977.

Kings County is several times larger in size than a typical Illinois or Kansas County. In 1977, sixty JES segments were allocated to the county. From these a random sub-sample of fifteen segments was selected for use in the LANDSAT study.

Major crops were cotton, barley, wheat, and alfalfa. For these crops all r_h^2 values exceeded 0.80 and regression estimator relative efficiencies(with respect to sub-sample direct expansion) ranged from 5.2 to 28.0.

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 Measurement 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; "Stratified Acreage Estimates in the Illinois Crop-Acreage Experiment," Proceedings of the 1977 Symposium on Machine Processing of Sensed Data, Purdue University, West Lafayette.

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. Hanuschak, George A.; "The Effect of the LANDSAT Cloud Cover Domain on Winter Wheat Acreage Estimation in Kansas During 1976," Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana.