STRATIFIED ACREAGE ESTIMATES IN THE ILLINOIS CROP-ACREAGE EXPERIMENT

RICHARD SIGMAN, CHAPMAN P. GLEASON, GEORGE A. HANUSCHAK, AND ROBERT R. STARBUCK
U. S. Department of Agriculture, Statistical Reporting Service

JUNE 1977

I. INTRODUCTION

The approach of the Statistical Reporting Service (SRS) for using LANDSAT remote sensor data is to use it as an auxiliary variable with existing operational ground surveys.[1] SRS objectives have been to investigate the use of LANDSAT data to improve crop-acreage estimates at several levels for which acreage statistics are needed; such as counties, groups of counties such as Crop Reporting Districts (CRD's), and entire states.

To determine the feasibility of these objectives, the Illinois crop-acreage experiment was established in 1975.[2] The experiment employs LANDSAT data for the state of Illinois and data from SRS's June Enumerative Survey (JES) for Illinois. The JES data was collected and edited by the Illinois Cooperative Crop Reporting Service. In addition the JES data was supplemented by monthly-updates conducted throughout the growing season and by low-altitude color-infrared photography for 202 of the 300 JES segments in Illinois.

This paper describes:

1. The statistical methodology for the auxiliary use of LANDSAT data to estimate crop acreages,

2. The procedure for designing the pixel classifier which is required by this methodology, and

3. Results obtained by applying this methodology for three LANDSAT frames in western Illinois.

Software systems have been developed jointly by SRS and the Center for Advanced Computation of the University of Illinois which implement the estimation methodology.[3]

The use of LANDSAT data as an auxiliary variable developed from a realization that using LANDSAT data as a survey variable produces biased estimates. The two major types of bias in using LANDSAT data as a survey variable are:

1. Mensuration biases due to the large pixel size of the LANDSAT data (57m x 79m), and

2. Classifier-related procedural biases due to different discrimination functions (linear or quadratic), training sets, prior probabilities, and classification categories used in the design of the classifier.

II. STATISTICAL THEORY AND METHODOLOGY

A. DIRECT EXPANSION ESTIMATION (GROUND DATA ONLY)

Aerial photography obtained from the Agricultural Stabilization and Conservation Service is photo-interpreted using the percent of cultivated land to define broad land-use strata. For example, the stratum definitions for Illinois are given in Table 1.

Within each stratum, the total area is divided into N_h area frame units. This collection of area frame units for all strata is called an area sampling frame. A simple random sample of n_h units is drawn within each stratum. The Statistical Reporting Service then conducts a survey in late May, known as the June Enumerative Survey (JES). In this general purpose survey, acres devoted to each crop or land use are recorded for each field in the sampled area frame units. Intensive training of field statisticians and inteviewers is conducted providing rigid controls to minimize non-sampling errors[4].

The scope of information collected on this survey is much broader than crop acreage alone. Items estimated from this survey include crop acres by intended utilization, grain storage on farms, livestock inventory by various weight categories, and agricultural labor and farm economic data.

Let h = 1, 2,..., L be the L land-use strata. For a specific crop (corn, for example) the estimate of total crop acreage for all purposes and the estimated variance of the total are as follows:

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

$\hat{Y}$ = Estimated total of corn acres for a state.

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

Then

$$\hat{Y} = \sum_{h=1}^{L} N_h \left(\sum_{j=1}^{n_h} y_{hj} \right) / n_h \qquad (1)$$

The estimated variance of the total is:

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

Note that we have not yet made use of an auxiliary variable such as classified LANDSAT pixels. The estimator in (1) is commonly called a direct expansion estimate, and we will denote this by $\hat{Y}_{DE}$.

As an example, for the state of Illinois in 1975, the direct expansion estimates were:

Corn $\hat{Y}_{DE}$ = 11,408,070 Acres
Relative Sampling Error = 2.4% = $\sqrt{v(\hat{Y})} / \hat{Y}$

Soybeans $\hat{Y}_{DE}$ = 8,569,209
Relative Sampling Error = 2.9% = $\sqrt{v(\hat{Y})} / \hat{Y}$

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

The regression estimator utilizes both ground data and classified LANDSAT pixels. The estimate of the total Y using this 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 $\bar{y}_h$ = the average corn acres per sample unit from the ground survey for the h^{th} land-use stratum

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

$\hat{b}_h$ = the estimated regression coefficient for the h^{th} land-use stratum when regressing ground-reported acres on classified pixels for the n_h sample units.

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

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

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

X_{hi} = number of pixels classified as corn in the i^{th} area frame unit of the h^{th} strata.

$\bar{x}_h$ = the average number of pixels of 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 (large sample) variance for the regression estimator is

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

where

r_h^2 = sample coefficient of determination between reported corn acres and classified corn pixels in the h^{th} land-use stratum.

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

Note that,

$$v(\hat{Y}_R) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h - 2} (1 - r_h^2)\, v(\hat{Y}) \qquad (2)$$

and so $\lim v(Y_R) = 0$ as $r_h^2 \to 1$ for fixed n_h. Thus a gain in lower variance properties is substantial if the coefficient of determination is large for most strata.

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

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

When LANDSAT passes do not cover the entire state on one date, it is necessary to work with analysis districts (domains) which are wholly contained within a LANDSAT scene or pass. In this study the analysis districts were collections of counties wholly contained in a LANDSAT pass. The

regression estimate for the i^{th} analysis district is

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

and the entire-state estimate is

$$\hat{Y}_R = \sum_{h=1}^{L_i} N_{hi} \bar{y}_{hi(reg)}.$$

When analysis districts are used, degrees of freedom for least squares regression by strata can become small. Under these circumstances it is necessary to pool strata, and the regression estimate for the i^{th} analysis district becomes:

$$\bar{y}^*_{ki(reg)} = \bar{y}^*_{ki} + \hat{b}^*_{ki} (\bar{X}^*_{ki} - \bar{x}^*_{ki})$$

for $k = 1, 2, \ldots, L^*_i$, and the entire-state estimate becomes

$$\hat{Y}_R = \sum_{k=1}^{L^*_i} N^*_{ki} \bar{y}^*_{ki(reg)},$$

where L^* = total number of pooled strata for the i^{th} analysis domain and N^*_{ki}, $\bar{X}^*_{ki}$, $\bar{x}^*_{ki}$, $\bar{y}^*_{ki}$ are adjusted for varying sizes of the sample units in each stratum. (Thus, h indexes individual stratum; whereas, k indexes pooled stratum. Consequently, the * notation is redundant and will not be used in the next section.)

C. COUNTY ESTIMATES USING A REGRESSION ESTIMATOR

Let $N_{k,c}$ = total number of area frame units in the k^{th} pooled strata for a set of C counties.

$\bar{X}_{k,c}$ = total number of pixels in the set of C counties classified as corn for the k^{th} pooled stratum divided by $N_{k,C}$.

Then an estimate based on the regression estimator of the total corn acreage for the C counties is:

$$\hat{Y}_{REG,c} = \sum_{k=1}^{L} N_{k,c} (\bar{y}_k + \hat{b}_k (\bar{X}_{k,c} - \bar{x}_k)) \qquad (4)$$

$$v(\hat{Y}_{REG,c}) = \sum_{k=1}^{L} N^2_{k,c} \frac{N_k - n_k}{n_k} S^2_{k,y} \frac{n_k - 1}{n_k - 2} \cdot$$

$$(1 - r^2_k) (I(C) + \frac{1}{n_k} + \frac{(\bar{X}_{k,c} - \bar{x}_k)^2}{\sum_{i=1} (x_{ki} - \bar{x}_k)^2})$$

where

I(C) = 1 if O(C) < total number of counties wholly contained in the analysis district
= 0 otherwise

O(C) is the cardinality of the set C.

$S^2_{k,y}$ = variance for the corn reported acreage for the k^{th} pooled stratum

$$= \sum_{j=1}^{n_k} (y_{k,i} - \bar{y}_k)^2 / (n_k - 1)$$

III. DESIGNING A CLASSIFIER

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 probability multiplied by the category's prior probability. If the prior probabilities used are correct for the population of pixels being classified, then the resulting Bayes classifier minimizes the posterior probability of misclassifying a pixel for a 0-1 loss function.[5]

In crop-acreage estimation, however, the objective is to minimize the variance of resulting acreage estimates. Since minimizing the posterior probability of misclassification does not necessarily achieve this objective, optimum acreage estimation may require the use of prior probabilities different than the optimum Bayes set.

For the case of multivariate normal signatures, the category likelihood functions are completely specified by the population means and covariances of the category signatures. Thus, the calculation of category discriminant functions involves the estimation of signature means and covariances and category prior probabilities.

Designing the classifier for this experiment consisted of the following steps:

1. Identification of classification categories.

2. Calculation of signature means and covariances and category prior probabilities from a training set of labeled pixels (called "training the classifier").

3. Measurement of classifier performance on a test set of labeled pixels (called "testing the classifier").

4. Heuristic optimization of the classifier by repeating steps 1 through 3 for different numbers of categories and/or different prior probabilities, and then proceeding to step 5 for the "optimized" classifier.

5. Estimation of classifier performance in classifying the entire pixel population.

Because of the availability of ground data, which supplied the location and cover type of

agricultural fields, supervised identification of classification categories was possible. A classification category was created for each cover type in which the number of training pixels exceeded a specified threshold, usually 100 pixels. In addition, a classification category for surface water was created using pixels from rivers, lakes, and ponds.

A classifier was heuristically optimized through a series of classification trials using field-interior pixels to train and all segment-interior pixels to test. The various trials used different combinations of the number of categories and the method of computing prior probabilities. These classification trials, along with additional details on the classifier design procedure, are described in the next section.

IV. ANALYSIS RESULTS FOR WESTERN ILLINOIS

The purpose of the Illinois crop-acreage experiment is to investigate the effectiveness of LANDSAT data to serve as an auxiliary variable for crop acreage estimates. In the analysis of the LANDSAT pass covering western Illinois, referred to simply as the "Western Pass", this investigation had three major objectives. These were:

1. To investigate the influence or lack of influence of various factors, both methodological and geographical, on classifier performance.

2. To compute LANDSAT-based regression estimates for crop acreages in all counties wholly contained in the Western Pass and for the Western Crop Reporting District (CRD) and then compare the precisions of these estimates to JES direct expansion estimates for these areas.

3. To compute crop-acreage regression estimates plus the relative sampling errors of these estimates for the twenty-nine individual counties wholly contained within the Western Pass.

A. CLASSIFIER PERFORMANCE STUDY

The following factors were investigated for their influence or lack of influence on classifier performance:

1. Scene Domain. The northwest Illinois LANDSAT scene, denoted W1 (scene 2194-16035, August 4, 1975), and the west-central scene, denoted W2 (scene 2194-16042, August 4, 1975) were first analyzed separately and then collectively within the Western Pass joined-scene, denoted W123. The southern scene denoted W3 was not analyzed individually since only four segments were on this scene.

2. Number of Classification Categories. This factor investigated the influence of intra-crop clustering to create multiple categories per crop (MCPC) versus straight supervised clustering with a single category per crop (SCPC). The SCPC set of categories consisted of seven categories for W2 and ten categories for W1 and W123. The MCPC set of categories consisted of fifteen categories and was developed by clustering the ten-category SCPC set of covers. This resulted in three categories for alfalfa--cut, uncut, and dried; two categories for hay; and two categories for oat stubble.

3. Prior Probabilities. This factor investigated the effect on classifier performance of using "different prior probabilities" for the classification categories. Strictly speaking, there is only one correct set of prior probabilities for a given geographical region. Using "different prior probabilities" actually means using different weighting factors for the likelihood probabilities in the class discriminant functions. The two sets of prior probabilities which were studied were using priors proportional to expanded reported acres, denoted PER, and using equal priors, denoted EP.

4. Training/test data sets. This factor investigated the data sets on which the classifier was trained and tested. The following methods were employed to allocate the LANDSAT data associated with JES segments between the training and test data sets:

a. Resubstitution, in which all of the segment data, denoted NB for "not background", was used to both train and test the classifier.

b. Sample partition, in which the classifier was trained on a 50% sample of segment fields, denoted FLDS, and then tested on all of the segment data.

c. Jackknifing, denoted JK, in which the training set was 3/4 of the data and the test set was the remaining 1/4. This allocation was repeated four times so that the union of the four test sets was the entire collection of segment data.

The jackknifing technique used was that referred to by Toussaint as the Pi-method.[6] Thus, four separate estimates of classifier performance were obtained and then averaged to yield the jackknife estimate.

There are two reasons why the training/test factor was of interest. The first reason was the desire to minimize the work involved with evaluating a classifier. The resubstitution and sample partition methods are easy to perform but are known to produce biased evaluations of the classifier in small samples. On the other hand, the jackknife is known to give a less biased evaluation but also involves substantially more work to perform. Consequently, if in this investigation the three methods give similar results, then in future experiments of the same size or larger the much-easier-to-apply resustitution and sample partition methods will be compared. If there is no difference between the resubstitution and sample partition methods then

these will be used and jackknifing will not be investigated.

The second reason for investigating this factor was to study the sensitivity of the classifier to the selection of the training data. This was the purpose of performing sample partition and then comparing the results with those from the other two methods of classifier evaluation.

5. Strata poolings. Table 2 shows the distribution of JES segments by stratum for W1, W2, and W123. As can be seen, a number of strata have zero or very few segments in them. Thus, it was necessary to pool a number of strata together and then compute $\bar{y}_{h(reg)}$ on the pooled strata. Three different strata poolings were tried and are denoted by the pooled strata given in Table 2.

The purpose of the classifier performance study was to investigate the influence of the above factors on classifier performance. Traditionally, the performance of a classifier has been measured in terms of its confusion matrix of percents correct and commission error rates. However, if a classifier is being used to estimate crop acreages, then it should be evaluated in terms of how well it does exactly that. Thus, the classification objective is to minimize the variance of the resulting regression estimates, and as shown in equation (2) this is accomplished by maximizing the r_h^2's (r-squares). Hence, to compare classifier performance on the same stratum, the respective r-squares were compared. For multi-strata regions, classifier performances were compared in terms of the relative efficiencies (equation (3)) of the resulting estimates. Two types of relative efficiency were calculated. The first type, denoted RE1, was calculated with respect to the direct expansion estimator which uses the same poolings as the regression estimator. RE1 measures the gain in terms of lower variance, of the regression estimate over the pooled JES direct expansion estimate. Of course this doesn't take into account the strata in the direct expansion estimate. However, a second type of relative efficiency, denoted RE2, was calculated with respect to direct expansion over the 11-12-20-30 pooling. Thus RE2 measures the gain, in terms of increased precision, of the regression estimate over the unpooled JES direct expansion estimate.

Counting the different strata poolings as separate trials, thirty-four separate classification trials were performed in the classification performance study. Even this, however, is far short of the number of trials required for a complete factorial analysis. Nevertheless, the influence of each factor on classifier performance can be determined but only on a subset of the levels of other factors. The factor levels for the different trials are summarized in Table 3.

Table 4 compares the r-squares and percents correct for corn in twenty-seven of the classification trials. The MCPC and JK trials are not included in this table. Items of note in this table are:

a. Percents correct are greater for PER priors than for equal priors, but for r-square the opposite is true.

b. Training on a 50% sample of fields yields r-squares very close to those for training on NB.

c. r-square is very small in stratum 20.

d. The r-squares in W1 are generally larger than the corresponding r-squares in W2. W123 is in-between but closer to W2 than W1.

Table 5 presents the relative efficiencies for corn for the same twenty-seven trials. As expected, RE1 and RE2 have the same rankings across factor levels as noted for r-square in Table 4. An interaction between domain location and the optimum strata pooling can be noted. In W1 and W123 the 11-12-20-30 pooling is optimum for RE2, but in W2 the 10-50 pooling is best.

A possible explanation of the effect of domain location on classifier performance is that scenes W1 and W2 are markedly different agriculturally. These differences are exhibited in Table 6 which indicates the amount of land in W1, W2, and W123 devoted to various levels of agricultural activity.

Tables 7 and 8 present results for soybeans for twenty-seven of the classification trials. Unlike corn, the effect of different priors on the classification results for soybeans is very slight, with PER being slightly better than EP. Again, an interaction between location and the optimum strata pooling for RE2 is exhibited, and the nature of this interaction is different from that observed for corn.

Table 9 presents the results of trial JK in which jackknife training and testing is used. Table 10 compares the results of this trial to the corresponding resubstitution trial (Trial W123.2). The jackknife and resubstitution r-square values are quite similar, the major dissimilarities being for those cover types which have large coefficients of variation and small r-squares in Table 9. This suggests that for sufficiently large sample sizes, the resubstitution method will yield r-square values whose biases are acceptably small.

Table 11 compares MCPC versus SCPC. For corn, MCPC is superior; whereas for soybeans an interaction with type of priors can be noted. For the soybeans EP case, SCPC is better. On the other hand, for soybeans PER the MCPC method is superior.

Finally, Table 12 compares classifier performance for all covers and two different priors. Items of note are the low r-square and RE1 values for minor crops and the fact that no single type of prior probability, neither EP nor PER, is optimum for every cover.

B. Large-area Estimates

The relative efficiencies obtained in the classification trials indicated that the auxiliary use of LANDSAT data can reduce the variance of acreage estimates for corn and soybeans. Consequently, the regression estimates for these crops were calculated for the nine-county Western Crop Reporting District (CRD) and for the entire twenty-nine county region contained in the Western Pass. These large-area estimates were then compared to the corresponding direct expansion estimates and to estimates based on the Illinois State Farm Census.

The Western CRD is completely contained in scene W2 and occupies about half of the W2 land area. Regression estimates for the CRD were calculated by first classifying all pixels in W123 with the classifier from classification trial W123.2; i.e., EP, SCPC with ten crops, and training on NB in W1 + W2. The classification results for only those pixels in the Western CRD were then used with a 10-50 strata pooling to compute the $X_{k,c}$ values for equation (4).

Table 13 compares the regression and direct expansion estimates for corn and soybeans in the Western CRD. For each crop the difference between the regression estimate and the direct expansion estimate is less than the standard error of either estimate. For corn the regression estimate C.V. is 54% of the C.V. for direct expansion. For soybeans, however, the regression estimate C.V. is 81% of the direct expansion C.V. Thus, the gain, in terms of lower variance, of the regression estimator over direct expansion is smaller for soybeans than for corn. One reason for this is the fact that an EP classifier was used. The classification trials indicate that EP is optimal for corn but sub-optimal for soybeans.

Table 13 also compares the direct expansion estimates for the Western CRD with acreage estimates based on the Illinois State Farm Census. For each crop the difference between the two estimates exceeds 1.5 times the standard error of the direct expansion estimate. The two estimates, however, measure different quantities--the direct expansion estimate measures standing acres, whereas the State Farm Census measure acres harvested.

Table 14 lists acreage estimates for the entire twenty-nine county region contained in the Western Pass. These estimates were computed using the same classifier as that used for the Western CRD.

C. County Estimates

Regression estimates for corn and soybeans were calculated for the twenty-nine individual counties in joined-scene W123. These are listed in Table 15 and were also computed with the same classifier as that used for the CRD estimates. With two exceptions the C.V.'s for corn ranged between 15 and 20% on a county-by-county basis in northwest Illinois. The exceptions were Jo Davies county (34% C.V.), which is almost entirely stratum 20, and Peoria county (24% C.V.), which is largely urban.

The high C.V.'s in stratum 20 are to be expected due to the very nature of this stratum. Basically, stratum 20 is a "catch-all" stratum in which areas of highly heterogeneous land use are placed.

In west-central Illinois the C.V.'s for corn ranged as high as 33% on a county-by-county basis. Counties with the largest C.V.'s were located on the Mississippi or Illinois rivers.

The C.V.'s for soybeans were considerably larger than those for corn. One reason for this, as was also the case for the CRD estimates, is that the EP classifier is sub-optimal for soybeans.

V. SUMMARY

In order to investigate the effectiveness of LANDSAT data as an auxiliary variable for crop acreage estimates, three LANDSAT frames from an August 4, 1975 satellite pass over western Illinois were analyzed. It was observed that the pixel classifier used in the crop-acreage methodology was influcenced by a number of factors, both methodological and geographical.

Large-area corn and soybean acreage estimates were calculated using LANDSAT data as an auxiliary variable for both a twenty-nine county area and a nine-county Crop Reporting District. Significant increases in precision over ground survey estimates were demonstrated.

It was also shown that small-area crop-acreage estimates for individual counties with measurable precision are technically feasible. However, the large coefficients of variation of some of these estimates may make them unsuitable for operational publications.

ACKNOWLEDGEMENTS

The authors wish to acknowledge several individuals for their contribution to this paper and project: Martin Ozga and Walt Donovan under the direction of Robert M. Ray provided heroic programming and systems development in support of our efforts, Richard D. Allen of the Illinois Cooperative Crop Reporting Service for his management of the ground data collection in Illinois, Harold F. Huddleston for his statistical leadership and guidance, and William H. Wigton and fellow members of the New Techniques Section for their efforts and support in this large scale project.

REFERENCES

1. Von Steen, Donald H. and Wigton, William H., "Crop Identification and Acreage Measurement Utilizing LANDSAT Imagery", Statistical Reporting Service, U. S. Department of Agriculture, Washington, D.C., March 1976.

2. Ray, Robert M., III and Huddleston, Harold F., "Illinois Crop-Acreage Estimation Experiment", Proceedings of the 1976 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana.

3. Ozga, Martin; Donovan, Walter E.; and Gleason, Chapman P., "An Interactive System for Agricultural Acreage Estimates Using LANDSAT Data", Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana.

4. 1975 June and December Enumerative Supervising and Editing Manual, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.

5. Anderson, T.W., An Introduction to Multivariate Statistical Analysis, New York: Wiley, 1958.

6. Toussaint, Godfried T., "Bibliography on Estimation of Misclassification", IEEE Transactions on Information Theory, Vol. IT-20, No. 4, (July, 1974), pages 472-79.

Table 1. Stratum numbers and definitions

stratum #	stratum description	sub-stratum #	sub-stratum description
10	intensive agriculture	11	75%+ cultivated
		12	50% - 75% cultivated
50	non-intensive agriculture	20	15% - 49% cultivated
		31\ 32 :urban 33/	:non-agricultural
		40	range land : (# 30)
		61	proposed water :
		62	water /

Table 2. Sample Sizes within Strata and Strata Poolings

original stratum #	# segments* W1	W2	W123	pooled stratum # 0	10-50	11-12-20-30
11	30	16	44	0	10	11
12	6	10	16	0	10	12
20	5	11	17	0	50	20
31	2	1	3	0	50	30
32	1	0	1	0	50	30
33	0	0	0	0	50	30
40	0	1	1	0	50	30
61	0	1	1	0	50	30

*W1 and W2 entries are on an entire scene basis. W123 entries are for the counties wholly contained in W1+W2+W3.

Table 3. Summary of Classifier Performance Study

	factor										
	domain			categories		priors		train/test			strata
trial	W1	W2	W123	SCPC	MCPC	EP	PER	NB	FLDS	JK	poolings
W1.1	X			X/10			X	X			all 3
W1.2	X			X/10			X		X		poolings
W1.3	X			X/10		X		X			
W1.4	X			X/10		X			X		.
W2.1		X		X/7		X			X		
W2.2		X		X/7			X	X			.
W2.3		X		X/7		X		X			
W123.1			X	X/10			X	X			.
W123.2			X	X/10		X		X			
W123.3			X		X/15	X			X		all 3
W123.4			X		X/15	X		X			poolings
JK			X	X/10		X				X	pooling 0

Table 4. Sample coefficients of determination (r-squares) and percents correct for corn in SCPC classifications

analysis district	train/ test	priors	stratum r-square 0	10-50 10	10-50 50	11-12-20-30 11	11-12-20-30 12	11-12-20-30 20	11-12-20-30 30	% correct*
W1	NB	EP	.83	.80	.36	.86	.62	.09	1.00	54
		PER	.64	.56	.50	.65	.60	.06	.95	88
	FLDS	EP	.84	.82	.31	.89	.57	.15	1.00	57
		PER	.70	.62	.51	.72	.56	.07	.97	84
W2	NB	EP	.63	.66	.19	.66	.71	.06	.28	51
		PER	.41	.55	.15	.72	.48	.25	.00	85
	FLDS	EP	.69	.74	.30	.82	.58	.12	.53	54
W123	NB	EP	.70	.72	.21	.78	.54	.00	.58	52
		PER	.52	.56	.18	.67	.57	.00	.20	86

*Based on all segment interior pixels, including field boundaries.

Table 5. Relative efficiencies for corn in SCPC classifications

analysis district	train/ test	priors	RE1 pooling 0	RE1 pooling 10-50	RE2 pooling 0	RE2 pooling 10-50	RE2 pooling 11-12-20-30
W1	NB	EP	5.69	3.95	3.03	3.78	4.25
		PER	2.74	2.15	1.46	2.06	2.46
	FLDS	EP	5.97	4.20	3.18	4.02	4.58
		PER	3.26	2.44	1.74	2.33	2.77
W2	NB	EP	2.66	1.68	1.61	1.76	1.27
		PER	1.65	1.47	1.00	1.54	1.15
	FLDS	EP	3.16	2.03	1.91	2.13	1.67
W123	NB	EP	3.34	2.23	1.73	2.00	2.23
		PER	2.08	1.74	1.07	1.56	1.81

Table 6. Distribution of population segments by stratum within analysis districts

stratum	% of population segments in analysis district contained in each stratum W1	W2	W123
11	53.7	32.5	39.8
12	13.0	16.6	15.7
20	10.9	30.8	23.4
31	11.4	8.6	9.7
32	9.4	5.5	7.2
33	1.0	1.8	1.4
40	.5	3.1	2.0
61	.2	1.1	.8
	100.0	100.0	100.0

Table 7. Sample coefficients of determination (r-squares) and percents correct for soybeans in SCPC classifications

analysis district	train/ test	priors	stratum r-square 0	10-50 10	10-50 50	11-12-20-30 11	11-12-20-30 12	11-12-20-30 20	11-12-20-30 30	% correct*
W1	NB	EP	.81	.82	.83	.82	.70	.98	.98	72
		PER	.82	.83	.83	.83	.72	.98	.98	74
	FLDS	EP	.81	.82	.84	.82	.75	.99	.98	71
		PER	.82	.82	.84	.82	.72	.97	.98	74
W2	NB	EP	.62	.60	.49	.73	.31	.63	.55	65
		PER	.63	.62	.49	.73	.38	.58	.55	63
	FLDS	EP	.63	.61	.51	.73	.34	.63	.02	65
W123	NB	EP	.67	.69	.49	.77	.44	.57	.56	63
		PER	.74	.74	.50	.78	.62	.55	.66	67

*Based on all segment interior pixels, including field boundaries.

Table 8. Relative efficiencies for soybeans in SCPC classifications

analysis district	train/ test	priors	RE1 pooling 0	RE1 pooling 10-50	RE2 pooling 0	RE2 pooling 10-50	RE2 pooling 11-12-20-30
W1	NB	EP	5.25	5.26	4.73	4.81	5.56
		PER	5.42	5.43	4.89	4.97	5.76
	FLDS	EP	5.20	5.25	4.69	4.81	5.62
		PER	5.41	5.42	4.87	4.96	5.74
W2	NB	EP	2.53	2.10	2.26	2.18	1.97
		PER	2.63	2.15	2.34	2.23	1.97
	FLDS	EP	2.60	2.16	1.67	2.13	1.91
W123	NB	EP	2.99	2.56	2.84	2.60	2.52
		PER	3.32	2.78	3.15	2.82	2.91

Table 9. r-squares for jackknifed classification (W123, SCPC, EP, pooling 0)

cover	pooled-stratum-0 r-square jackknife group 1	2	3	4	Ave	S.E.	C.V. (%)
Alfalfa	.002	.001	.195	.078	.069	.09	132.7
Corn	.734	.814	.639	.680	.717	.07	10.5
Dense Woods	.097	.003	.030	.213	.086	.09	109.2
Hay	.017	.245	.042	.271	.144	.13	92.2
Oat Stubble	.000	.016	.119	.004	.035	.06	163.9
Oats	.119	.001	.069	.109	.094	.08	87.8
Permanent Pasture	.339	.304	.552	.269	.366	.13	34.8
Soybeans	.578	.745	.843	.520	.671	.15	22.2
Wasteland	.847	.732	.062	.248	.472	.38	79.9

Table 10. Comparison of jackknifed and resubstitution r-squares (W123, SCPC, EP, Pooling 0)

cover	train/test JK	NB
Alfalfa	.069	.09
Corn	.717	.70
Dense Woods	.086	.01
Hay	.144	.25
Oat Stubble	.035	.06
Oats	.094	.15
Permanent Pasture	.366	.36
Soybeans	.671	.67
Wasteland	.472	.81

Table 11. Relative efficiencies for corn and soybeans in W123 classifications

cover	priors	categories	train/ test	RE1 pooling 0	RE1 pooling 10-50	RE2 pooling 0	RE2 pooling 10-50	RE2 pooling 11-12-20-30
Corn	EP	SCPC/10	NB	3.34	2.23	2.00	1.73	2.23
		MCPC/15	FLDS	3.90	2.54	2.02	2.28	2.48
	PER	SCPC/10	NB	2.08	1.74	1.07	1.56	1.81
		MCPC/15	FLDS	2.32	1.86	1.20	1.67	1.91
Soybeans	EP	SCPC/10	NB	2.99	2.56	2.84	2.60	2.52
		MCPC/15	FLDS	2.61	2.29	2.48	2.33	2.31
	PER	SCPC/10	NB	3.32	2.78	3.15	2.82	2.91
		MCPC/15	FLDS	3.39	2.84	3.22	2.89	2.97

Table 12. r-squares and relative efficiencies for all covers (W123, MCPC, FLDS, Pooling 0)

Cover	r-square		RE1	
	EP	PER	EP	PER
Water	.89	.84	8.70	6.23
Waste	.78	.82	4.47	5.45
Soybeans	.62	.71	2.61	3.39
Corn	.75	.57	3.90	2.32
Permanent Pasture	.32	.35	1.44	1.51
Woods	.02	.24	1.01	1.31
Alfalfa	.05	.13	1.04	1.13
Hay	.20	.10	1.24	1.10
Oats	.14	.05	1.15	1.04
Oat Stubble	.01	.03	1.00	1.02

Table 13. Estimated acres of corn and soybeans in the Western CRD

Estimator	Corn		Soybeans	
	Acres	C.V.	Acres	C.V.
Direct Expansion	1,316,000	8.5%	562,000	13.1%
Regression	1,269,000	4.6%	574,100	10.6%
Farm Census	1,121,000		688,700	

Table 14. Estimated acres of corn and soybeans in Western Pass 29-county region

Estimator	Corn		Soybeans	
	Acres	C.V.	Acres	C.V.
Direct Expansion	4,110,150	3.6%	1,539,200	7.7%
Regression	4,125,400	2.5%	1,681,800	5.2%
Farm Census	3,653,800		1,707,400	

Table 15. Regression estimates for corn and soybeans in individual counties in Western Pass

County	Corn		Soybeans	
	Acres	C.V.	Acres	C.V.
Adams	166,600	24.0%	83,600	35.3%
Brown	53,700	33.4	24,300	50.7
Bureau	254,000	18.7	110,600	33.4
Calhoun	56,700	25.1	23,3000	39.9
Carroll	126,500	17.5	57,200	29.6
Cass	91,700	20.3	54,100	25.5
Fulton	172,100	29.0	91,400	37.8
Greene	136,800	19.2	76,000	24.8
Hancock	190,500	19.3	74,800	36.2
Henderson	104,000	17.3	37,100	36.4
Henry	276,800	17.2	79,400	46.6
Jersey	85,700	21.6	48,900	27.0
Jodaviess	108,300	34.1	27,100	94.2
Knox	174,100	19.5	79,600	31.6
Mason	129,100	21.3	76,100	27.9
McDonough	162,500	17.4	82,500	26.3
Mercer	139,800	18.7	43,900	43.4
Morgan	147,200	17.6	93,700	20.9
Ogle	223,000	19.0	51,500	64.2
Peoria	124,000	24.0	65,300	32.6
Pike	160,100	25.7	78,300	37.3
Rock Island	107,000	18.7	27,500	52.7
Schuyler	84,000	29.0	36,650	46.2
Scott	61,100	19.9	31,500	28.6
Stark	92,000	18.2	40,600	32.1
Stephenson	172,100	18.6	30,600	81.8
Warren	161,800	16.5	64,100	32.2
Whiteside	242,800	16.2	62,400	49.0
Winnebago	121,500	21.5	29,600	68.0

ILLINOIS CROP-ACREAGE ESTIMATION EXPERIMENT*

Robert M. Ray III and Harold F. Huddleston

Center for Advanced Computation
University of Illinois at Urbana-Champaign
Urbana, Illinois

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

JUNE 1976

I. ABSTRACT

This paper describes remote-sensing data analysis research conducted collaboratively during the last year by personnel of the Center for Advanced Computation (CAC) of the University of Illinois at Urbana-Champaign and the Statistical Reporting Service (SRS) of the U. S. Department of Agriculture. The research reported has been undertaken by CAC and SRS to assess the practicalities of existing high-volume earth observations data acquisition, processing, and communication technologies such as LANDSAT, the ILLIAC IV parallel computer, and the ARPA Network as mechanisms for improving the accuracy of USDA annual estimates of agricultural crop acreages for geographic regions corresponding to U. S. states.

II. INTRODUCTION

Throughout this research, our 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 geographic information systems 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 ongoing 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 for regions corresponding to U. S. states.

For this exploratory application of machine processing of LANDSAT data, the state of Illinois was selected as the basic study area. All ground-truth 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. Assuming the practicality of similar analyses for LANDSAT imagery covering the entire state, we discuss a procedure for evaluating statistically the information to be gained by estimating state crop-acreage totals from LANDSAT imagery classification results where SRS sample survey data are used as ground-truth information for classification training as opposed to estimating state crop-acreage totals directly from SRS survey data alone.

III. 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 non-agricultural fields, i.e., areas of homogeneous land cover, for all ownership tracts within each of 300 area segments of the SRS national survey sample that fall within the state of Illinois.

*This research was supported in part by the U. S. Department of Agriculture through USDA Research Agreement No. 12-18-04/-8-1794-X, and in part by the National Aeronautics and Space Administration through NASA Grant NGR 14-005-202.

In accordance with SRS survey procedures, these 300 area segments had been selected earlier with respect to strict 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 ownership 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 enuerators on ASCS photos and interview forms were rechecked independently for consistency by personnel of the central offices 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.

As still another source of ground-truth information, low-altitude infrared photography (at a scale of approximately 5" = 1 mile) was obtained commercially for a subsample of 202 area segments. This current aerial photography provided directly an accurate, high-resolution picture of the agricultural crops and land uses actually existing in late summer for the 202 segments covered. Hence, it was possible to check the degree of accuracy with which 1975 field boundaries had been delineated on the older ASCS photos. For those segments for which summer 1975 photography had been obtained, field boundaries (and changes) were redrawn directly on the current photography making reference both to data recorded by survey enumerators on ASCS photos and to features visible directly in the current infrared photos themselves. This task was also done in Springfield by SRS personnel. A quantitative evaluation of the relative advantages for LANDSAT imagery classification objectives of this current photography and the older ASCS photography is currently being conducted by SRS.

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 (202 current infrareds and 98 ASCS 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 agricultural field boundary digitizing is being accomplished using an interactive DEC PDP-10 data tablet software system developed at CAC explicitly for takeoff 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. The other 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. (See Figures 1-2.)

IV. LANDSAT IMAGERY SELECTION AND PREPROCESSING

All LANDSAT imagery collected over Illinois during the summer of 1975 was acquired from NASA in the form of 70 mm film transparencies and evaluated by SRS and CAC with regard to project objectives. 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.

Since one of the goals of our project is to obtain crop-acreage estimates for the entire state of Illinois, and since counties represent smaller geographic units more convenient for estimation of state-wide crop-acreages in terms of regional subtotals, it was decided to preprocess and reformat all LANDSAT imagery acquired from NASA into a set of image files such that each of the 102 counties of Illinois was contained wholly and cloud-free within at least one image file. To accomplish this objective, the following strategy has been adopted.

Despite the vertical overlap of approximately fifteen (15) miles between successive LANDSAT frames of the same orbit, in many cases individual counties falling on north and south frame boundaries are not wholly contained in any one frame. Hence, in numerous instances it is necessary to compile pseudo-frames of LANDSAT digital imagery by concatenating data records of a top portion of one frame to data records of a bottom portion of another frame of the same orbit. Such pseudo-frames are to be compiled wherever they are necessary to achieve continuous cloud-free imagery for a particular county.

Due to the size of counties in Illinois, the horizontal overlap of approximately fifty miles between frames of successive orbits is sufficient to insure that no county fails to lie wholly within the swath of at least one orbit. Thus fortuitously, the considerably more difficult problem of splicing LANDSAT imagery horizontally across orbits does not arise.

Having obtained a complete set of image files (LANDSAT frames and pseudo-frames) 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. Inspection of the imagery available suggests that for 1975 Illinois LANDSAT imagery only fourteen (14) such analysis districts -- ranging in size from as few as two or three counties to as many as a dozen -- are required to provide integral-county, cloud-free LANDSAT coverage for the entire state.

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. This task is being performed using an image skew transformation procedure developed at CAC for efficient de-skewing and rotation of LANDSAT digital data to map orientation[8] in conjunction with other systems developed at CAC for precision geographic registration of LANDSAT imagery.[9]

Finally, all image files are being geographically registered to the SRS ground-truth data available (and hence simultaneously also to the USGS map control already associated with all ground-truth). This step is being accomplished in the following manner.

First, with respect to digital image calibration information available, all SRS area segments are located approximately in terms of digital image file row and column coordinates. Gray-scale displays for windows of LANDSAT data known to contain all SRS area segments are produced using a conventional line printer and over-printing techniques. Then digitized SRS segment masks (described above) are again plotted this time at the exact scale of the line-printer LANDSAT imagery displays. Following manual overlay and visual correlation of line-printer and plotter displays on a light table, overlay positions of maximum geographic correspondence between LANDSAT image pixels and SRS segment masks are recorded. (See Figures 3-4.)

V. LANDSAT DATA ANALYSIS SYSTEMS

As the SRS ground-truth data is digitized and LANDSAT imagery preprocessed for each analysis district, LANDSAT data is being analyzed collaboratively by SRS and CAC using a common set of computer facilities available via the ARPA Network. For small-scale analyses of SRS area segment data the EDITOR software system at BBN is used. For specific large-scale LANDSAT image analysis functions, the ILLIAC IV at NASA's Ames Research Center is employed also via the ARPA Network but addressed conveniently through the EDITOR system at BBN.

All ILLIAC IV - ARPA Network image analysis systems implemented to date have been designed and developed in close collaboration with personnel of the Laboratory for Applications of Remote Sensing (LARS) of Purdue University. Hence all software procedures implemented specifically for machine interpretation of LANDSAT data follow closely multispectral image interpretation methods researched previously at LARS.[10,11]

Specifically, ILLIAC IV procedures are now operational for both multivariate cluster analysis and maximum-likelihood statistical classification of LANDSAT image samples. The speed of the ILLIAC IV with respect to these two image interpretation procedures has proven to be generally two orders of magnitude faster than execution times observed for the same processing task using the IBM 360/67 computer at LARS.[5]

Such LANDSAT imagery interpretation capabilities available via ILLIAC IV batch processing, together with the availability for classifier training operations of the interactive image processing software of the EDITOR system, suggest that operational crop-acreage monitoring via digital processing of orbital remote-sensing imagery may indeed be practical. Our project has been undertaken to assess more exactly the potentialities existing in this area. In the next section, we present preliminary results of one LANDSAT imagery analysis experiment using SRS data available for a single analysis district consisting of ten (10) counties in western Illinois.

VI. EVALUATION PROCEDURE

Of central importance to our experiment is the evaluation of the extent to which regional crop-acreage estimates may be improved by estimation with respect to LANDSAT imagery classification methods as opposed to estimation directly from SRS survey data alone. Statistical regression techniques may be used to obtain estimates of the total acreage of each crop type for each analysis district.[12] Following estimation of crop-acreages for all analysis districts separately, state-wide acreage estimates for each crop may easily be obtained by simply summing individual district estimates. Such estimates may be determined both with and without use of LANDSAT classification results and a measure of the value of the LANDSAT data may be computed.

Following ILLIAC IV classification of all LANDSAT pixels contained within the counties making up a particular analysis district, classification results for each crop type will be aggregated to obtain individual totals for all segments sampled within the district. Also, acreage totals for each crop type will be determined for the entire analysis district itself.

An estimator 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

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

and the variance for a large sample of segments is:

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

For the individual analysis districts, the normal approximation for small samples is used, that is

$V(\hat{Y}_i)$ for large samples multiplied by $\left(1 + \frac{1}{n_i - 3}\right)$.

Where $N_i \hat{\bar{Y}}_i$ = 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)

$\hat{\bar{Y}}_i$ = a regression estimate of the average number of acres of the crop per area segment for the i^{th} district

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 in order to obtain the equivalent of a simple random sample over the entire analysis district.

An estimate of the total state-wide acreage for each crop may be obtained by making use of the additive property of the estimator and its sampling error over all districts. The estimators are

$$\hat{Y} = \sum_{i=1}^{f} \hat{Y}_i$$

$$V(\hat{Y}) = \sum_{i=1}^{f} V(\hat{Y}_i) .$$

A measure of the gain in relative efficiency of estimation of state-wide acreage obtained by using machine-interpreted LANDSAT data may be computed

$$RE = \frac{\sum_{i=1}^{f} N_i^2 V(\bar{y}_i)}{\sum_{i=1}^{f} N_i^2 V(\bar{y}_i)(1 - r_i^2)(\frac{n_i - 1}{n_i - 3})}$$

where the value of RE is expected to be greater than 1.0. A value of RE less than 1.0 would indicate that information had been lost through use of LANDSAT data. A value of 5.0 would indicate the regression estimator using LANDSAT classification results is equivalent to increasing the number of area segments by five times if costs of acquiring the LANDSAT data are equal to the costs of collecting the area segment data. For the single analysis district analyzed, the information gain or loss is:

$$\frac{1}{(1 - r_h^2)} \cdot \frac{n_h - 3}{n_h - 1}$$

where n_h is the number of area segments sampled. These values for the first analysis district are shown in the last column of table 5.

To date limited 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 were calculated from the ground enumerated data in the 10 counties.

Table 1. Prior Probabilities for Land Cover Categories

Crop or land use	Prior probabilities (Survey land use July 27, 1975)
Corn	.3282 (.3097*)
Soybeans	.1602 (.2297*)
Permanent pasture	.1392
Dense woods	.0935
Alfalfa hay	.0180
Other hay	.0467
Wheat stubble	.0101
Crop pasture	.0118
Water (farm ponds & lakes)	.0074
Wasteland (no agri. prod.)	.1148
Other crops & land uses (Training data not available)	.0701

*Based on 1974 crop year data from Illinois Assessor Census

A sample of fields was selected from the segments falling in the LANDSAT image. The acres in each crop or land use type was "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 tables 2 and 3.

Table 2. Number of Sample Fields by Cover Type

Crop or cover type	Number fields	Acres/ field	Total pixels	Nonborder pixels
Corn	425	23.3	9026	5604
Soybeans	215	22.5	4502	2712
Perm. pasture	163	19.7	2780	1289
Dense woods	144	16.8	2147	784
Hay	83	11.8	1069	477
Wasteland	274	8.7	2087	920
Alfalfa	40	11.0	423	183
Wheat stubble	27	11.2	259	86
Water	17	12.1	190	73
Crop pasture	21	13.3	280	119

Table 3. 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	17	73
Crop pasture	21	119

The pixels for all the selected fields 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. (It is planned to investigate the use of all fields and all pixels in developing mean vectors and covariance matrices as well as using equal prior probabilities in the classification.)

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 4 and their sampling errors squared in table 5 for eight agricultural land use categories. The window containing the 10 counties included 4,887,960

pixels and required less than 80 seconds for classification on the ILLIAC IV.

Table 4. 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	28	32	10
Crop pasture	45	45	0

Table 5. Variances of Estimates of Agricultural Cover Types for 10-County Analysis District

Crop or cover type	Variance reported (10^6 acres^2)	Variance regression estimate (10^6 acres^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	30	11	2.71
Crop pasture	88	94	.94

VII. SUMMARY

The four dimensional distributions by cover types exhibited considerable overlap except for water. In general, the distributions were unimodal or where several modes were present, one mode was much higher than the other. Soybeans had two distinct modes (one major and one minor), but no factor could be isolated as a basis for grouping fields into two categories. Based on the principal component analysis, bands 6 and 7 (IR) explain practically all the variation for most cover types, but for several cover types bands 4 and 5 (visible) were equally important. The quality of the data for band 7 appeared to be superior to the other bands which had discontinuities or gaps in the data values.

The LANDSAT data for major crops or cover types (i.e., large acreages) showed significant improvements can be expected in the estimates, but minor cover types showed little improvement or even loss of information unless their distributions were reasonably separated (i.e., in the measurement space) from the distributions of the major cover types.

These results for the first LANDSAT image are quite encouraging. Assuming LANDSAT digital tapes and near real-time processing of the ground and classification data, acreage estimates of spring planted crops could have been significantly improved for the area of this LANDSAT image by September 1. The authors believe that similar results can probably be achieved in other areas if the same conditions can be met; namely:

(1) excellent quality, cloud-free LANDSAT imagery

(2) good geographic registration of ground segments to LANDSAT imagery

(3) mean vector and covariance matrices for each crop for each LANDSAT frame (i.e., localized classifiers)

(4) prior probabilities for each LANDSAT frame (i.e., localized priors)

(5) sufficient ground data for each crop for classifier training

(6) an adequate number of ground segments for each LANDSAT frame to compute the regression and correlation coefficients for each crop.

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 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. Walt Donovan, "Oblique Transformation of ERTS Images to Approximate North-South Orientation," CAC Technical Memorandum No. 38, (November 1974), Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801.

9. Walter E. Donovan, Martin Ozga, and Robert M. Ray III, "Compilation and Geographic Registration of ERTS Multitemporal Imagery," CAC Technical Memorandum No. 52, (May 1975), Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801.

10. "Remote Multispectral Sensing in Agriculture," Research Bulletin 873, (December 1970), Laboratory for Agricultural Remote Sensing, Purdue University, West Lafayette, Indiana 47907.

11. P. H. Swain, "Pattern Recognition: A Basis for Remote Sensing Data Analysis," LARS Information Note 11572, (1972), Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana 47907.

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

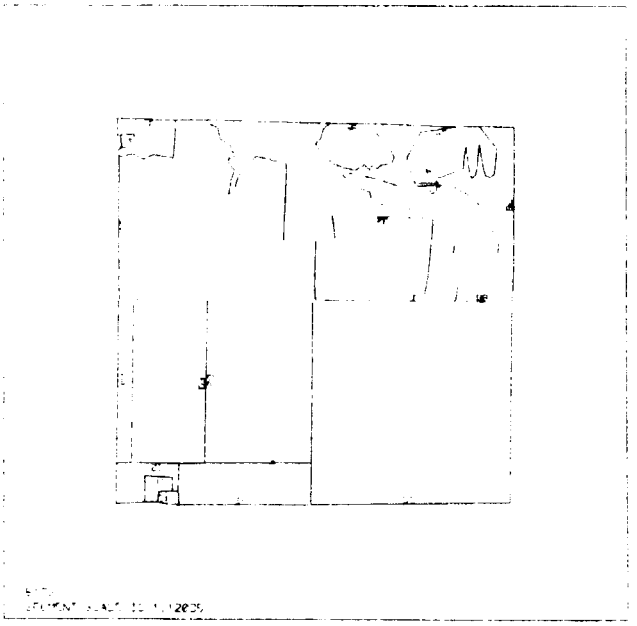


Figure 1. Example USDA/SRS Area Segment Mask Plotted at Scale of Photo Digitized. (Shown reduced here.)

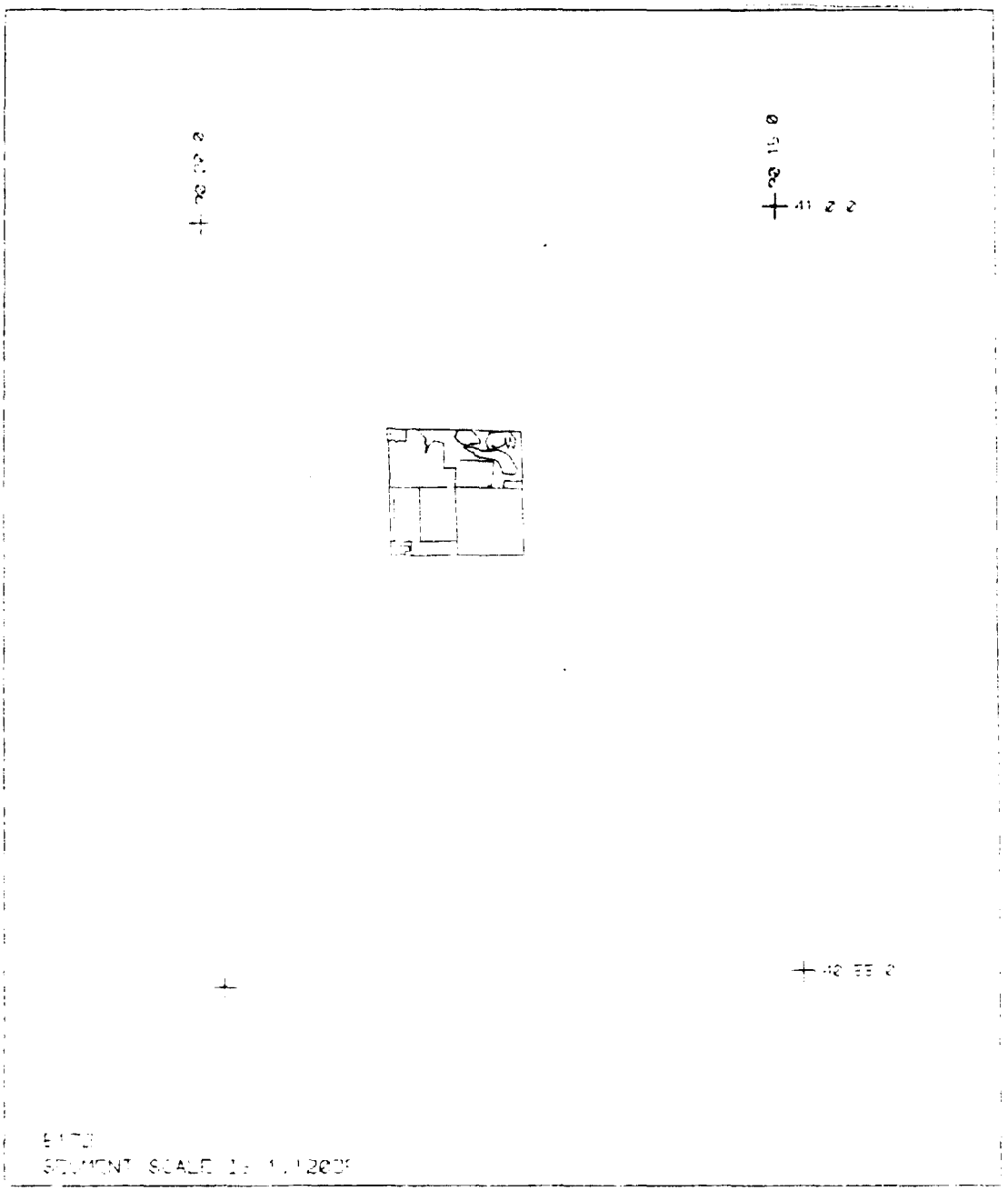
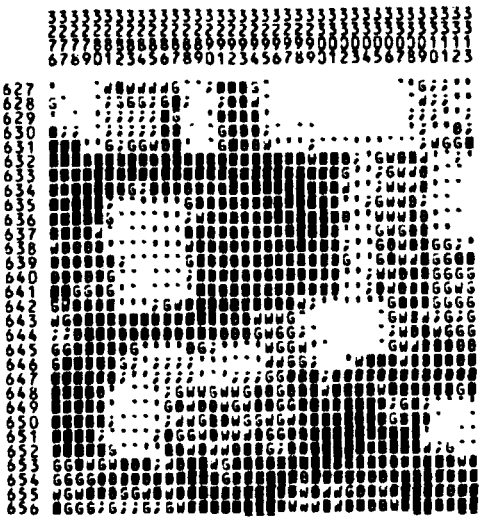


Figure 2. Example Area Segment Mask Plotted at Scale of USGS 15' Quad and Showing 5' Quad Ticks.



EQUIPROBABLE SCALING USED.

Figure 3. Gray-scale Line-printer Display of Digital LANDSAT Data (IR Band 4) Corresponding to a Specific USDA/SRS Area Segment.

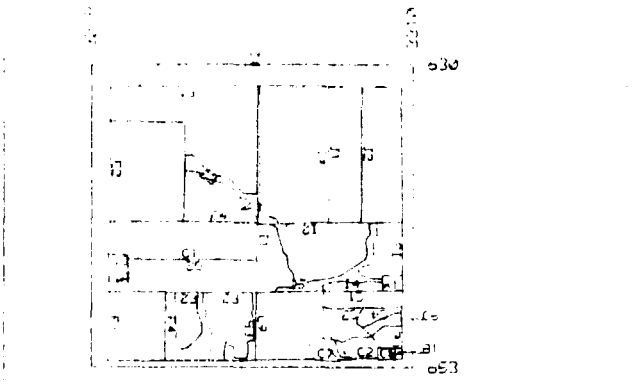


Figure 4. Plotter Display of Same Area Segment Scaled Horizontally and Vertically to Overlay Line-printer Display for Precision Local Registration of Ground Truth Data and LANDSAT Imagery.

USE OF LANDSAT TECHNOLOGY BY STATISTICAL REPORTING SERVICE

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

JUNE 1976

I. ABSTRACT

This paper describes an area sampling frame and defines the sampling error and bias of an estimate. LANDSAT data is explained in the Statistical Reporting Service framework and the essential components of computer classification are delineated. A procedure is presented that utilizes satellite data to improve an estimator with 3 percent sampling error.

II. AREA SAMPLING FRAME

The area sampling fr
tical Reporting Service (
estimates at both state a
addition, the use of the
cial for our application
improve these crop acreag
it is essential to spend
its function and use in g

accuracy
bias
random sample
estimators
coefficient of variation

The concepts of area
simple:

1. Divide the total
N small contiguou
without any overl
2. Select a random s
3. Obtain the desire
units of the popu
sample blocks.
4. Estimate the population totals by multiplying the sample totals by $\frac{N}{n}$. 1/

This procedure, as outlined above, is used for crop acreage, livestock, and other farm data estimation, and is a dependable method. The use of random numbers in selecting a sample from the universe accomplishes two things:

1. It gives a basis for making inference about the total production of all farms in the U.S.
2. It provides a basis for the computation of sampling errors which will be discussed in the following section.

III. CALCULATION OF THE ACCURACY OF AN ESTIMATE

To determine the accuracy of any estimate, one requires the population target value or the actual number which is being estimated. Of course,
ere available, it would not
the estimate of the target
becomes mandatory to use
evaluate an estimate. An
illustrates the use of an
nd sampling theory.

rest is divided into 30,000
nd a random sample of 300 is
btained and an estimate of
ed by multiplying the total
$\frac{000}{0}$ = 100. If another 300
lected, the estimate would
If the estimates do not
om one sample of size 300 to
te is fairly stable. However,
y considerably, then we would
timator has a large variance
The variation of the estimate
00 to other samples of 300
manner is sampling error.
l sampling errors are most de-
here is another criterion that is also important--the element of bias.

Bias

If there is a difference between the center of the distribution that defines sampling error and the true value being estimated, this difference is defined as bias.

Whether or not the true value being estimated is at the center of the sampling error distribution is controlled by:

1. The completeness of the sampling frame.

2. The importance of giving every element in the population a known positive chance of selection.

3. The use of high quality control standards of enumeration and other nonsampling errors.

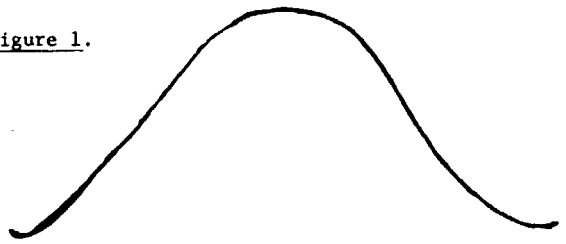
4. Technical properties of the estimators.

If the estimator that is being generated by selecting 300 segments is centered around the true value and the variation is small, then our one estimate is an accurate one--one that is close to the true value.

Often, one cannot tell about sampling errors unless other samples of 300 segments are selected and enumerated. However, with proper sampling techniques the variation can be measured with only one sample. The segment to segment variation is used to calculate the sample to sample variation. In essence, sample to sample variation is estimated with only one sample.

From one sample, then, the sampling distribution is estimated.

Figure 1.



Let us assume that the distribution looks like the distribution curve illustrated in Figure 1. We do not know where in the distribution our sample point lies. We only know that it was drawn from this distribution at random. We know, also, from the sampling procedure and the estimating formula that the statistic is unbiased. We have a better estimate if it comes from a distribution curve such as Figure 2, than from a curve such as Figure 3, because the values are clustered closer to the center.

Figure 2.

Figure 3.

If we improve the currnet estimates from the area frame with LANDSAT, then we must alter the distribution of the possible estimates by reducing the spread.

IV. APPLICATION OF LANDSAT CLASSIFICATION

Description of LANDSAT Data

The satellite data used in this report is LANDSAT Multi-Spectral Scanner (MSS) data and is described in Section 3 of data User's Handbook. 2/

The MSS is a passive electro-optical system that can record radiant energy from the scene being sensed. All energy coming to earth from the sun is either reflected, scattered, or absorbed, and subsequently, emitted by objects on earth. 3/ The total radiance from an object is composed of reflected radiance forms, a dominant portion of the total radiance from an object at shorter wavelengths of the electromagnetic spectrum, while the emissive radiance becomes greater at the longer wavelengths. The combination of these two sources of energy would represent the total spectral response of the object. This, then, is the "spectral signature" of an object and it is the differences between such signatures which allows the classification of objects using the statistical techniques about to be discussed.

Classification Techniques

Let us suppose that we wish to classify a LANDSAT frame. The way this is done in the computer is by use of discriminant functions. Computers must differentiate between crops on the basis of reflected energy. To start, we must have two or more crops and a sample of individual pixels for each. The problem is to set up a rule using the sample pixels for each crop, which will enable us to allot some unknown crop pixel outside the sample to the correct crop type given only the amount of reflected energy of that pixel.

This can be formulated statistically, but let me introduce some notation.

If all data in a LANDSAT frame were plotted in a scatter diagram it might appear as Figure 4.

Figure 4. Scatter Diagram of All Values in One LANDSAT Frame for Three Crops. C-Corn, S-Soybeans, W-Water

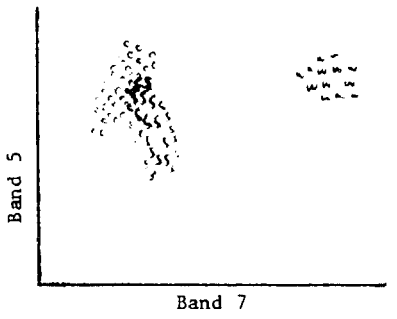
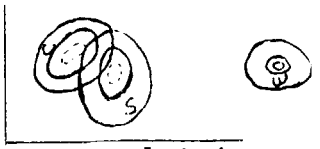


Figure 5 shows confidence limits for above data.

Figure 5. Confidence Limits for Data in Figure 4



If we study Figure 5, it does not take long to make some observations:

1. The location of the center of these concentric circles directly effects the type of rule to be followed.
2. The data looks quite elliptical (often this is not the case for actual data).
3. The spread of the data varies considerably for the crops. Soybeans has wide variability for example.
4. It will be impossible to tell with certainty which crops we have, if the reflected energy comes from the overlap region of corn with soybeans, because both are possible.
5. It would be ideal if the data for each crop were as far apart as water from corn if the spread were as small as the spread for water and elliptical in form and there were no areas of overlap.

However, it appears that these items are not under our control. The sensor (bands and bands width) determines the location of the centers of the spread of points.

The spread of the data and its contour are determined by factors such as soil conditions, varieties of crops, amount of fertilizer used, planting dates, atmospheric conditions, NASA preprocessing, and many more things.

As far as the overlap areas, where mislabeling or misclassification is inevitable, nature herself is the problem. Some items that we would like to be able to tell apart reflect solar energy similarly. We can not change the nature of things but simply to estimate what is there.

The best we can hope for is to estimate from a sample the scatterdiagram of the population and this we know how to do if we treat it like anything else that we estimate.

We want a valid statistical estimate that requires a random sample from the population of interest. This requires that all parts of the population of interest must have a chance of selection and the size must be large enough to adequately represent the population. If the population structure is as complicated as water in Figure 4 or if estimates are needed that are quite accurate, as in corn and soybeans, then, a fairly substantial sample size is required.

The area sampling frame is ideal because a valid statistical estimate can be made for the LANDSAT frame. In addition a random sample of all possible segments is available and reflected energy for the crop types can be determined for the sample fields inside the segments. These signatures are estimated for the scene they are in, so it is valid to use these values for computer training of the discriminant functions. After population scatterdiagrams have been estimated, rules are set up to allot pixels with known energy readings but unknown crop labels to crop categories. Rules are simple; they amount to drawing lines that partition the space. Figure 6 shows an example of this

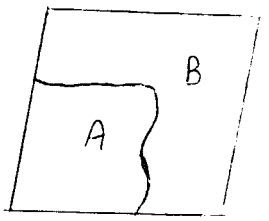
Figure 6. Partioned Space Showing Population Scatterdiagram

The rest is simple. All pixels that need crop labels should be plotted on the partitioned space. If they fall in partition one, give it a label of corn, even though some soybeans will creep in; obviously, we will do well with water.

Incidently, it turns out that the location, size and shape of these population scatterdiagrams shift relative to each other in different scenes and even different parts of the same scene. Hence, LANDSAT scene to label pixels from another locale is hazardous.

There are two cases, both are quite different. One is reasonable, and the other is not. Let us divide an image into two parts. Figure 7 shows a possible division of a LANDSAT scene.

Figure 7. LANDSAT Frame Divided Into Two Parts.



Let us imagine that we have divided Section A into 600 small parts. We then draw a random sample of 60 parts from the 600. This may or may not be truly representative. If it is, then, the reflective and emitted energy (the signature) from these 60 segments adequately represents the reflected energy in all of Section A. We do not consider the use of the signature in the sample of 60 segments to classify the 600, a signature extension. This is simply a valid statistical inference. It is a commonly misunderstood notion that one does not have to sample from the population of interest to make an inference, for that population.

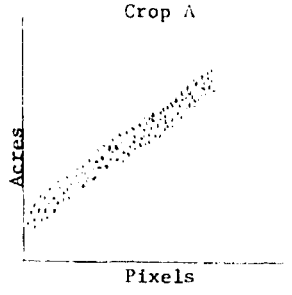
Should we wish to classify crops in Section B, it would be necessary to divide Section B into segments and draw a random sample from these segments as representative for signatures in Section B. One must sample the population of interest or the inference will be erroneous.

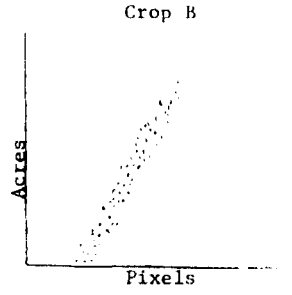
Model Utilizing LANDSAT

In order to make use of LANDSAT to reduce the sampling variation we shall first estimate the linear relationship between classified pixels for a crop and acres of the crop.

Figure 8 illustrates this relationship.

Figure 8. Population Relationship Between Classification Results and Reported Acres of the Same Crop for One LANDSAT Scene





Again, these relationships are population relationships that we do not know, so we wish to estimate them from a sample.

Our area frame sample segments can be used to estimate this relationship. The sample observations for Crop A are shown in Figure 9 and Figure 10.

Figure 9. Sample Data Points for Crop A Showing Relationship Between Pixels and Acres

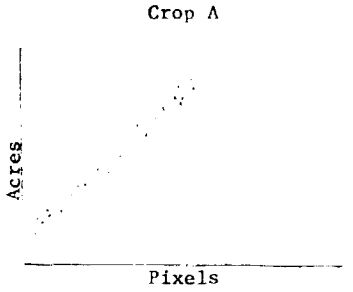


Figure 10. Estimated Population Linear Relationship Based on Sample Data in Figure 9

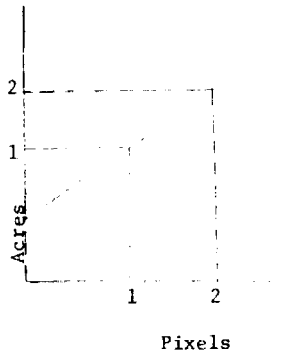


Figure 10 illustrates the relationship that is needed in order to use LANDSAT results.

This is on a per segment relationship. Therefore, we can locate a segment in LANDSAT, classify the segment and count the pixels of Crop A. If the pixels for Crop A turn out to be at point 1 then we read the corresponding value on the y-axis. If on the other hand, the classified pixels for the segment turn out to be at point 2 then we read that value on the y-axis.

This procedure could be completed for each segment in the population and we could sum up all the segments to get an estimate using satellite information across the whole area. However, all this is unnecessary.

Since we know N, the total number of segments in the LANDSAT frame, we can classify every pixel in the frame and divide the total number of pixels in Crop A by the number of segments in the frame. This then would equal the average number of pixels in Crop A for the average segment.

Also, we know total number of pixels of Crop A in sample segments (n). With this information we can adjust the direct expansion estimate for the difference between the pixels in Crop A for the

sample (n) versus the total of the population (N). That is, the difference between the mean number of pixels for the sample (n) and the mean number of pixels for the population (N) parts is a measure of how unrepresentative the selected sample is.

Figure 10 illustrates how the adjustments would be made. Say a difference between the average pixels for Crop A for the sample is at point 1 and the average for the universe is at point 2. The adjustment in acres is made on the y-axis. The formula is:

$$\hat{Y}_{reg} = \bar{Y} + b\ (\bar{X}_{total} - \bar{x}_{sample})$$

$\hat{Y}_{reg}$ is the adjusted number of acres in the average segment. $\hat{Y}_{reg}$ is then multiplied by N to get an estimate for the total.

The variance for $\hat{Y}_{reg}$ is $\frac{n - 1}{n - 2}\ (1 - r^2)$ times the variance of the direct expansion. This regression model reduces the spread of the sampling error distribution by a factor of $(1 - r^2)$.

In summary, we have ground data for a properly selected statistical sample, as well as the computer classification for the same. Thus, the necessary information is available to adjust a full frame classification for all systematic errors. If there is a good linear relationship between ground data and what the computer classifies as being on the ground, the sampling error will be materially reduced as compared to not having remotely sensed data.

REFERENCES

1/ Houseman, Earl E., Area Frame Sampling in Agriculture, Washington, D.C., United States Department of Agriculture: 1975

2/ LANDSAT Multi-Spectral Scanner Data User's Handbook, Goddard Space Flight Center, Greenbelt, Maryland: 1971

3/ Baker, J. R. and Mikhaul, E. M., Geometric Analysis and Restitution of Digital Multispectral Scanner Data Arrays, LARS Information Note 052875

4/ Von Steen, Donald and Wigton, William, Crop Identification and Acreage Measurement Utilizing LANDSAT Imagery, Statistical Reporting Service, United States Department of Agriculture, Washington, D. C. 20250