PRECISION OF CROP-AREA ESTIMATES

George A. Hanuschak
Statistical Research Division
Economics, Statistics, & Cooperatives Service
U.S. Department of Agriculture
Washington, D.C. 20250

Invited Paper at the
Thirteenth International
Symposium of Remote Sensing
of Environment
April 1979
Ann Arbor, Michigan

PRECISION OF CROP-AREA ESTIMATES

GEORGE A. HANUSCHAK

Statistical Research Division
Economics, Statistics, and Cooperatives Service
U.S. Department of Agriculture
Washington, D.C.

1. INTRODUCTION

The utility of LANDSAT data in developing crop-area estimates has been demonstrated by several investigators. The major issue in evaluating crop-area estimates is how to measure the precision and accuracy of the estimates. This paper describes the methods used by the Economics, Statistics, and Cooperatives Service (ESCS) of the U.S. Department of Agriculture (USDA) for evaluating crop-area estimates.

Annually in late May and early June ESCS conducts a nationwide agricultural survey, referred to as the June Enumerative Survey (JES), consisting of interviews with farm operators in randomly sampled areas of land called segments. Segments enter the JES by selection through stratified random sampling. The strata are land use categories determined by visual interpretation of aerial photography or LANDSAT imagery and delineated on county highway maps. The JES segments are typically 2.59 square kilometers in size. Strict survey quality control methods are used prior to, during, and after the data collection period to minimize nonsampling errors at the elementary sample unit (segment) level. Methods such as training of statisticians and interviewers prior to each JES, use of aerial photographs during the interview with the farm operators, reinterviews by supervisory interviewers, follow-up survey interviews, data editing-manual and machine, current aerial photography for comparison, etc. are used to insure data quality. The relative sampling errors for major crops at the national and regional level are on the order of 1-3 percent. At the state level they are on the order of 2-10 percent.

Any use of LANDSAT data by ESCS must be an improvement over the extremely efficient JES. The Statistical Research Division (SRD) of ESCS has developed techniques using LANDSAT and JES data together that produce lower sampling errors than the JES alone. The basic method used by the SRD is a simple application of regression estimation as described in William Cochran's, Sampling Techniques. This technique has been applied in Illinois (1975), Kansas (1976), and Iowa (1978). The Iowa project was completed during the 1978 crop year and in time for the regression estimates to be input to the official USDA Crop Reporting Board's Annual Crop Summary for Iowa released on January 16, 1979.

2. STATISTICAL METHODOLOGY

2.1 Direct Expansion Estimation (Ground Data Only)

Aerial photography obtained from the Agricultural Stabilization and Conservation Service is visually interpreted using the percent of cultivated land to define broad land-use strata. 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. ESCS then conducts a survey in late May, known as the JES. In this general purpose survey area devoted to each crop or land use is recorded for each field in the sampled area frame units. The scope of information collected on this survey is much broader than crop-area alone. Items estimated from this survey include crop-area by intended utilization, grain storage on farms, livestock inventory by various weight categories, and agricultural labor and farm economic data. Intensive training of field statisticians and interviewers is conducted providing rigid controls to minimize nonsampling errors.

The form of an estimated state total for a crop from a stratified random sample is as follows:

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

Let Y = Total corn area for a state (Iowa, for example).

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

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

Then,

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

The estimated variance of the total is:

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

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

As an example, for the state of Iowa in 1978, the direct expansion estimates were·

Corn $\hat{Y}_{DE}$ = 5,525,807 Hectares

Relative Samping Error = $\sqrt{v(\hat{Y})} / \hat{Y}$ = 2.4%

Soybeans $\hat{Y}_{DE}$ = 3,205,320 Hectares

Relative Sampling Error = $\sqrt{v(\hat{Y})} / \hat{Y}$ = 3.9%

2.2 Regression Estimation (Ground Data and Computer Classified LANDSAT Data)

By means of a regression estimator both ground data and classified LANDSAT data can be utilized to estimate crop hectarage.[2] (Regression estimators are discussed in most sampling texts, e.g. Cochran[1]) The estimate of Y using the separate form of the regression estimator is

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

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.

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

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

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

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.

$$\bar{x}_h = \sum_{j=1}^{n_h} x_{hj}/n_h$$

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

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.

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

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

Note that,

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

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.

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

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

where

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

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

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

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

and

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

The approximate variance of the combined regression estimator and the expression for $\hat{b}_c$ are given in Cochran,[1] [pp 202-203] and applied by Craig et. al.[3] in a 1976 Kansas LANDSAT study.

Since a LANDSAT pass does not cover an 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:

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

One problem associated with the use of LANDSAT data for crop-area estimation is cloud covered areas on the imagery. In essence, this becomes a non-response domain or post-stratum for LANDSAT studies . In the case of ESCS, the problem does not prohibit inferences at the state level since the random sample of JES ground data is available. The direct expansion estimate of the JES sample segment data is used for the cloud covered post-stratum area.[3]

All of the above formulas refer to sample estimates and their respective precision. The major item of interest in evaluating crop-area estimates, however, is accuracy. Cochran states, "Accuracy refers to the size of deviations from the true mean u, whereas precision refers to the size of deviations from the mean m obtained by repeated application of the sampling procedure."

In complex large scale applications such as crop production surveys, there is usually only one practical method in controlling the accuracy of the forecasts or estimates. That method is

to design a sound probability sample for the characteristics of interest (crop-area, in this case) where there is strict control over measurement error at the elementary sample unit level. Statistical formulas for precision of unbiased estimators (such as direct expansion) will then relate to accuracy if the characteristics of interest have been properly defined and the measurement errors are insignificant. An alternative is to know the true population parameter to measure the deviation of the forecast or estimate against. Seldom does this alternative exist in a complex application and even if known, cannot provide as estimate of the distribution of bias or measurement error.

ESCS's Statistics Unit fixes the precision of crop-area estimates and then minimizes the nonsampling errors by using previously mentioned strict survey quality control methods. Farm operators are interviewed by well trained personnel with an aerial photograph. All field boundaries are drawn onto the aerial photograph and the crop or land use type and area is recorded on a questionnaire. The data is then carefully reviewed and edited both manually and by computer processing. It is in this fashion that ESCS's Statistics Unit minimizes nonsampling errors.

Several special studies that compared farmer reported crop-area to digitized crop-area from current photo interpreted color infrared aerial photography have shown no significant differences in the estimates. The differences were .4 percent for wheat in Kansas in 1976 and less than 1 percent for corn and soybeans for a 29 county area in Western Illinois.

3. 1978 IOWA LANDSAT PROJECT

Twelve LANDSAT scenes were required to virtually cover the state of Iowa (See Figure 1). The dates of the LANDSAT data used ranged from August 6, 1978 to September 4, 1978. Median delivery time for LANDSAT products to ESCS from NASA's Goddard Space Flight Center was 49 days. Median time for ESCS to register and analyze the LANDSAT data was 30 days. The state was divided into ten post strata (analysis districts) as seen in Figure 2. As described in previous papers,[2,5] a modified supervised[6] approach is used for classification of the LANDSAT data into cover types. The algorithm used is the Gaussian maximum likelihood classifier. Within a known cover type clustering is used to minimize the chances of having multi-modal data.

As seen in the variance formula for the regression estimate when the $\hat{R}_h^2$ for h=1,2,...,L are at a maximum value the variance of the estimator is at a minimum. What then is the relationship between R_h^2 and the traditonal percent correct classification measures which are commonly used as success criteria in remote sensing projects? In the Iowa project percent correct was measured using the following 2 data sets: 1.) all pixels for a cover type (including field boundary pixels) and 2.) only field interior pixels. Tabel I shows the percent correct measures and $\hat{R}_h^2$ for the ten analysis districts in Iowa. There is no obvious relationship and this is a research topic that warrants further investigation. The only relationship that is presently obvious is that if the classification matrix approaches perfection then R_h^2 approaches 1 and two become the same criterion. However, this is rarely the case in large scale crop classifications.

The direct expansion, LANDSAT based regression, and pixel count crop-area estimates for the state and ten-analysis districts are presented in Tables II and III.

There were substantial improvements in the precision of the regression estimates versus the direct expansion estimates. This was the first time ESCS researchers were able to receive and analyze LANDSAT data in time to be used for a regularly scheduled crop production report.

4. SUMMARY

Precision and "controlled accuracy" are the major criteria used by ESCS for evaluating crop-area estimates. Regression estimates which utilize both LANDSAT data and ESCS's JES data were substantially more precise than the direct expansion estimates (ground data only)

for the 1978 Iowa LANDSAT Project. The regression estimates were input to the USDA Crop Reporting Board's Annual Crop Summary for Iowa released on January 16, 1979. The repeatability of such efforts, however is highly dependent on rapid LANDSAT data delivery to ESCS and cloud free coverage of the analysis areas.

5. ACKNOWLEDGMENTS

The author wishes to thank the following organizations and individuals for their deeply appreciated contributions to this paper: Fellow members (current and former) of ESCS's New Techniques Section for their development of the regression estimator methodology over the last several years and their heroic efforts on the Iowa LANDSAT Project; Charles Caudill, Galen Hart, Harold Huddleston and William Wigton of ESCS for their management and technical support; Charles E. Miller of the New Techniques Section for the percent correct and pixel count estimates; and Tricia Brookman, ESCS for her fine typing efforts.

6. REFERENCES

1. Cochran, William G., Sampling Techniques (Third Edition), John Wiley and Sons, 1977.

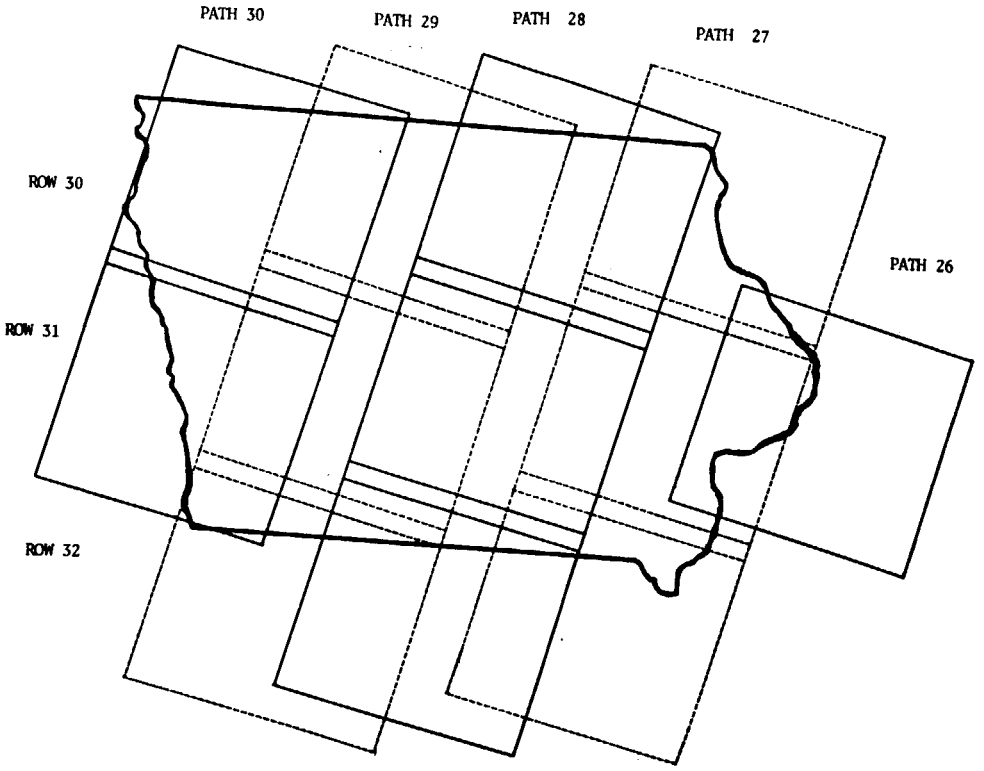
2. Gleason C., Starbuck R., Sigman R., Hanuschak G., Craig M., Cook P., and Allen R., "The Auxiliary Use of LANDSAT Data in Estimating Crop Acreages: Results of the 1975 Illinois Crop Acreage Experiment," Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C., October 1977.

3. Craig M., Sigman R., and Cardenas M., "Area Estimates by LANDSAT: Kansas 1976 Winter Wheat, "Economis, Statistics, and Cooperatives Service, U.S. Department of Agriculture, Washington, D.C., August 1978.

4. Hanuschak, G., "LANDSAT Estimation with Cloud Cover," Proceedings of the 1976 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana

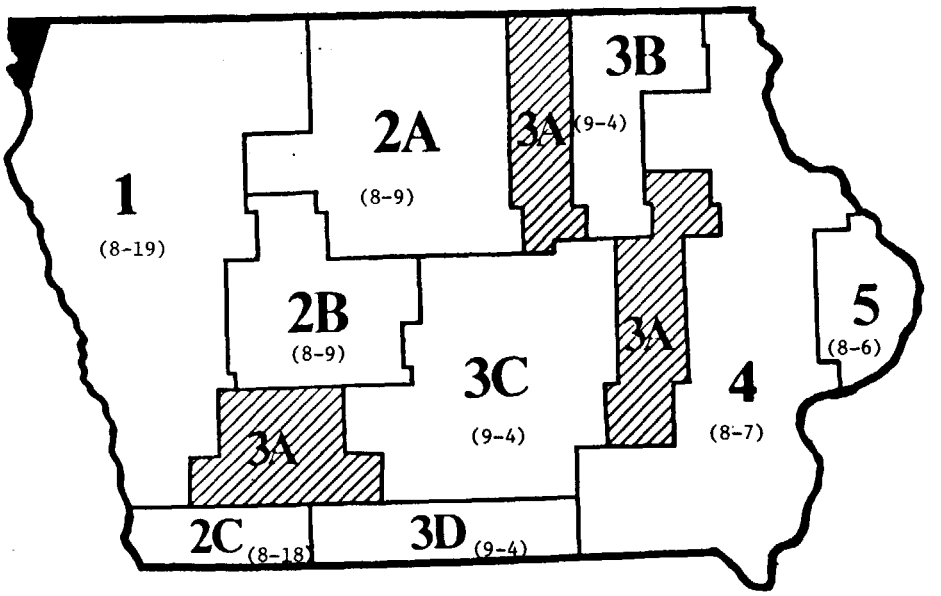
5. Sigman R., Hanuschak G., Craig M., Cook P., Cardenas M., "The Use of Regression Estimation with LANDSAT and Probability Ground Sample Data," 1978 American Statistical Association Annual Meeting, San Diego, California.

6. Fleming M. D., Berkebile J. S., Hoffer R. M., "Computer Aided Analysis of LANDSAT-1 MSS Data: A Comparison of Three Approaches Including a Modified Clustering Approach," Proceedings of the 1975 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana.



Path	Row	Date	Percent Iowa Cloud-Cover	Scene ID
30	30	August 19	0	30167-16274
	31	August 19	0	30167-16280
29	30	August 9	0	21295-16013
	31	August 9	40	21295-16020
	32	August 18	0	30166-16224
28	30	September 4	60	30183-16162
	31	September 4	0	30183-16164
	32	September 4	0	30183-16171
27	30	August 7	10	21293-15500
	31	August 7	15	21293-15502
	32	August 7	10	21293-15505
26	31	August 6	0	21292-15444

Figure 1. LANDSAT Images and Dates.



LANDSAT Data not analyzed.

Cloud covered.

Figure 2. Analysis Districts

TABLE I. CLASSIFICATION PERCENTS CORRECT BY ANALYSIS DISTRICT

Analysis District	Corn			Soybeans		
	% Correct Using All Pixels	% Correct Using Interior Pixels	Range of r^2 1/	% Correct Using All Pixels	% Correct Using Interior Pixels	Range of r^2 1/
1	72.13	79.98	.57-.92	67.34	76.37	.58-.88
2A	81.46	87.03	.71	71.21	79.43	.74
2B	79.59	90.39	.78-.94	71.36	85.14	.74-.98
2C	50.55	63.17	.30	59.63	74.31	.80
3B	77.58	77.41	.38	37.49	44.15	.79
3C	56.57	65.24	.34-.40	59.37	68.97	.77
3D	33.71	51.27	.07	54.93	70.47	.89
4	56.68	60.94	.65-.71	26.52	29.36	.45-.83
5	50.00	54.35	.75	45.23	75.51	.86

1/ Range by land use strata.

TABLE II. 1978 IOWA CORN RESULTS (PLANTED HECTARES)

Analysis District	Classified Pixels 1/	$\hat{Y}_{DE}$ Direct Expansion Estimate	C.V. $\hat{Y}_{DE}$	$\hat{Y}_{R}$ LANDSAT Regression Estimate	C.V. $\hat{Y}_{R}$	Range of r^2 for h=1, . ., L	Relative Efficiency
1	1,306,217	1,462,074	3.48	1,460,234	2.20	.57-.92	2.51
2A	923,626	828,772	4.47	818,892	2.50	.71	3.28
2B	463,957	332,050	11.50	454,252	3.40	.78-.94	5.98
2C	124,767	106,036	10.98	109,959	9.50	.30	1.24
*3A	-	657,462	4.36	-	-	-	-
3B	345,293	276,112	10.05	268,022	8.47	.38	1.49
3C	589,898	550,581	7.46	542,081	6.02	.34-.40	1.58
3D	58,843	83,658	17.76	82,798	18.65	.07	0.93
4	1,058,692	1,029,688	6.72	896,084	4.47	.65-.71	2.99
5	132,166	148,148	11.10	149,820	6.03	.75	3.32
State	5,660,921	JES= 5,525,807	2.3	5,439,604	1.5	.07-.94	2.43

*LANDSAT data not available

1/ converted to hectares

TABLE III. 1978 IOWA SOYBEANS RESULTS (PLANTED HECTARES)

Analysis District	Classified Pixels 1/	$\hat{Y}_{DE}$ Direct Expansion Estimate	C.V. $\hat{Y}_{DE}$	$\hat{Y}_{R}$ LANDSAT Regression Estimate	C.V. $\hat{Y}_{R}$	Range of r^2 for h=1, . .,L	Relative Efficiency
1	760,215	747,759	8.11	781,566	4.04	.58-.88	3.70
2A	650,382	655,049	6.75	675,293	3.42	.74	3.68
2B	244,275	256,944	12.91	255,540	6.11	.74-.98	4.55
2C	93,828	95,196	24.97	97,497	11.67	.80	4.37
*3A	-	401,671	9.20	-	-	-	-
3B	84,102	86,550	28.00	125,300	9.37	.79	4.26
3C	369,662	328,662	14.51	338,363	7.06	.77	3.98
3D	78,841	82,633	32.55	95,933	10.20	.89	7.59
4	343,162	441,032	12.68	424,782	7.97	.45-.83	2.73
5	34,575	47,060	29.20	48,580	12.53	.86	5.10
State	3,060,122	3,205,320	3.91	3,244,525	2.50	.45-.98	2.38

*LANDSAT data not available

1/ Converted to hectares