CROP AREA ESTIMATION USING GROUND-GATHERED AND SAMPLED LANDSAT DATA

James Weldon Mergerson

1,1	1,2	1,3	1,4				
2,1	2,2	2,3	2,4				
3,1	3,2	3,3	3,4				
4,1	4,2	4,3	4,4				
				5,5			
					6,6		
						7,7	
							8,8

United States Department of Agriculture

Economics and Statistics Service

Statistical Research Division

May 1981

ESS Staff Report No. AGESS810408

CROP AREA ESTIMATION USING GROUND-GATHERED AND SAMPLED LANDSAT DATA. BY JAMES WELDON MERGERSON, Statistical Research Division, Economics and Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250, ESS Staff Report No. AGESS810408, May 1981.

ABSTRACT

The report describes a comparative study in which a procedure using ground gathered data and classified LANDSAT data for estimating crop area was compared to procedures using ground gathered data and sampled LANDSAT data. Data from parts of Iowa and Missouri were used. Unitemporal data were used in Iowa and multitemporal data were used in Missouri. Results indicate that some sampling schemes can be used without any significant difference in the crop area estimates, but with a large reduction in cost for corn, soybeans, and winter wheat.

Key words: LANDSAT, regression, sampling, Iowa, Missouri.

****
* This paper was prepared for limited distribution *
* to the research community outside the *
* U.S. Department of Agriculture *
****

ACKNOWLEDGMENTS

The author would like to thank Paul Cook and Richard Sigman for helpful comments, observations and suggestions, Martin Ozga for his exceptional programming, and Sandy Stutson for an excellent registration of the Iowa scene.

CONTENTS

CROP AREA ESTIMATION USING GROUND-GATHERED AND SAMPLED LANDSAT DATA

James W. Mergerson
Statistical Research Division
Economics and Statistics Service
U.S. Department of Agriculture
Washington, D.C. 20250

INTRODUCTION

The Economics and Statistics Service (ESS) is investigating the operational use of LANDSAT data. The ESS approach is to use classified LANDSAT pixels as an auxiliary variable with existing operational June Enumerative Survey (JES) ground gathered data to improve the precision of crop area estimates. A regression estimator which utilizes both ground gathered JES data and classified LANDSAT pixels is used. All pixels in an analysis district are classified. An analysis district is a group of counties and sub-counties wholly or partially contained in a LANDSAT scene. The purpose of this study was to investigate the use of sampled LANDSAT data in conjunction with JES data to estimate crop areas.

This report describes a study in which various sampling schemes were compared with the current approach. Some of these schemes produced results that were not significantly different from the current approach.

This report, intended for those with some knowledge of Remote Sensing Applications, will be useful to researchers considering the use of sampled LANDSAT data in estimating crop areas.

REGRESSION ESTIMATION (GROUND DATA AND SAMPLED LANDSAT DATA)

The regression estimator utilizes ground gathered JES data, sampled classified and classified sampled LANDSAT pixels. The estimate of the total crop area for a given crop 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 reported crop area of a given crop per segment from the ground survey for the h-th land-use stratum

$\hat{b}_h$ - the estimated regression coefficient for a given crop in the h-th land-use stratum when regressing the ground data reported area on the number of corresponding sampled classified pixels

$\bar{x}'_h$ - the average number of classified sampled landsat pixels, classified as a given crop, per frame unit in the h-th land-use stratum

x_{hj} - the number of sampled classified pixels classified as a given crop in the j-th segment of the h-th stratum

$\bar{x}_h$ - the average number of sampled classified pixels classified as a given crop per segment in the h-th land-use stratum

y_{hj} - the total area of a given crop in the j-th segment in the h-th stratum

n_h - the number of segments selected in the h-th stratum

N_h - the number of area frame units in the h-th stratum

The estimated variance of the regression estimator is:

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

where

r_h^2 - the sample coefficient of determination between reported area for a given crop and sampled classified pixels classified as the given crop in the h-th land-use stratum

LANDSAT DATA

The data sets used in the study were selected from two states, Iowa and Missouri. The Iowa data set was a unitemporal data set. The Missouri data set was a multitemporal data set created from two dates of coverage. Information on location, cloud cover and dates is summarized in Table 1.

TABLE 1: LANDSAT IMAGERY

Path	Row	Date	Cloud Cover	Scene ID	State
28	30	9/02/80	0%	22050-16145	Iowa
28	32	5/14/79	30%	30435-16165	Missouri
28	32	8/03/79	10%	21654-16100	Missouri

ANALYSIS

Twenty sampling schemes were used. All schemes involved taking systematic samples starting in row one and column one. These schemes are listed in Table 2 and some sampling schemes are illustrated in Table 3.

TABLE 2: SAMPLING SCHEMES

[1,1]	[1,2]	[1,3]	[1,4]	[5,5]
[2,1]	[2,2]	[2,3]	[2,4]	[6,6]
[3,1]	[3,2]	[3,3]	[3,4]	[7,7]
[4,1]	[4,2]	[4,3]	[4,4]	[8,8]

[a,b] a - Row Increment
b - Column Increment

TABLE 3: SAMPLING SCHEME ILLUSTRATIONS

Sampling Scheme [2,3]

(x)	X	X	(x)	X	X	(x)	X	X
X	X	X	X	X	X	X	X	X
(x)	X	X	(x)	X	X	(x)	X	X
X	X	X	X	X	X	X	X	X
(x)	X	X	(x)	X	X	(x)	X	X
X	X	X	X	X	X	X	X	X
(x)	X	X	(x)	X	X	(x)	X	X
X	X	X	X	X	X	X	X	X

Sampling Scheme [2,2]

(x)	X	(x)	X	(x)	X
X	X	X	X	X	X
(x)	X	(x)	X	(x)	X
X	X	X	X	X	X
(x)	X	(x)	X	(x)	X
X	X	X	X	X	X

Six treatment combinations were applied to each sampling scheme. The six treatment combinations are illustrated in Table 4.

TABLE 4: TREATMENT COMBINATIONS

	A	B	C	D	E	F
STATE	I	I	I	M	M	M
YEAR	0	0	0	9	9	9
CROP	C	S	S	C	S	W
PROBS	E	E	P	P	P	P

I - IOWA M - MISSOURI
0 - 1980 9 - 1979
C - CORN S - SOYBEANS W - WINTER WHEAT
E - EQUAL PROBABILITIES
P - PROBABILITIES PROPORTIONAL TO UNEXPANDED REPORTED ACRES

For each sampling scheme and treatment combination, a small scale analysis was performed. A regression estimator with JES data as the dependent variable and sampled classified LANDSAT pixels as the independent variable was used. For the purpose of estimating crop areas, ESS's evaluation criteria is not the percent of pixels classified correctly, but is how precisely the crop area is estimated for the area of interest. Maximization of the R-square values minimizes the variances of the regression estimates. Thus, the major criterion used to compare the various sampling schemes was the respective R-squares. Another factor was cost.

The [1,1] scheme utilizes all of the LANDSAT data in the analysis district and is the scheme that is currently used. The other sampling schemes were considered to be alternative candidate schemes. For each alternative candidate scheme and treatment combination a transformed absolute difference was computed relative to the [1,1] scheme and summed over each alternative candidate scheme. The alternative candidate schemes were then ranked in increasing order of the differences.

A t-test was used to determine if there was a significant difference between the R-square obtained using the [1,1] scheme, and the R-square for each of the alternative candidate sampling schemes for the various treatment combinations. The t-test is outlined in Table 5. If the difference between the two R-squares was significant that sampling scheme was eliminated from further consideration as an alternative candidate sampling scheme. Of the remaining schemes the ones with a higher rank were then also eliminated.

After these elimination processes were completed five alternative candidate schemes were left. Large scale estimation was then performed for all six remaining schemes for a randomly selected treatment combination. Results are displayed in Table 6.

Cost figures are not available for the phases directly associated with obtaining large scale estimates. As a means of reducing the cost of this study, a full frame classification was performed using all pixels and the classified pixels were then sampled using the remaining sampling schemes. One would expect a lower cost for a smaller amount of data.

TABLE 5: T-TEST

To test for a significant difference between two R-square values, the t distribution was used, where

$$t = (r1 - r2)\sqrt{[(n-3)(1+r0)] / 2D}$$

and D is

$$\begin{vmatrix} 1 & r1 & r2 \\ r1 & 1 & r0 \\ r2 & r0 & 1 \end{vmatrix}$$

with n-3 degrees of freedom.

r1 - coefficient of correlation between the reported area and the number of pixels classified into a given cover using all classified pixels in the segments

r2 - coefficient of correlation between the reported area and the number of sampled classified pixels for a given alternative sampling scheme

r0 - coefficient of correlation between all pixels classified into a given cover type and the number of sampled classified pixels for a given alternative sampling scheme

n - number of segments

TABLE 6: LARGE SCALE ESTIMATES

SAMPLING SCHEME	$\hat{Y}$ (10,000 hectares)	STANDARD DEVIATION
[1,1]	25.2	1.4
[1,2]	25.6	1.4
[1,3]	25.7	1.5
[2,1]	25.0	1.3
[2,2]	25.2	1.3
[2,3]	25.0	1.4

SUMMARY AND CONCLUSIONS

In obtaining crop area estimates using ground data and classified LANDSAT pixels, the current ESS approach is to classify all the pixels. The use of sampled classified LANDSAT data was investigated. Results indicate that some sampling schemes may be used without resulting in a significant difference in the R-square values, the estimate or the variance.

For future operational projects involving the use of LANDSAT data in obtaining crop area estimates, use of the [2,2] sampling scheme is strongly recommended. This would greatly reduce the cost of full frame classification without any significant difference in the precision of the estimate. This scheme is also intuitively appealing since one would expect adjacent pixels to have a high probability of being similar.

The use of sampling can be implemented in future operational projects quite easily. The additional steps required in the small scale phase could be performed in about one half hour. After classifying the segment data, the resulting classified file could then be partially tabulated by sampling scheme in addition to complete tabulation. Small scale regression estimation could then be performed using each tabulated file. A test could then be performed to determine if there is a significant difference between the two R-squares and a decision could then be made as to whether or not to use sampling for large scale estimation.

In creating the Illiac LANDSAT data tape, an additional tape containing only the sampled data for a selected sampling scheme could also be created. Both tapes would be sent for possible processing on the Illiac or whatever computer that may be used in the future. If one elects to use sampled LANDSAT data for large scale estimation the tape containing only the sampled data would be used as an input file for large scale classification and aggregation. The reduction in cost for this phase would be approximately proportional to the reduction in the number of pixels.

REFERENCES

1. Kish, Leslie *Survey Sampling*, John Wiley and Sons, 1965.

2. Sterling, T., S. Pollack, *Introduction to Statistical Data Processing*, Prentice-Hall, Inc., 1968.

APPENDIX A

TABLE A1: TREATMENT COMBINATION A RESULTS

SAMPLING SCHEME	B0	B1	$\bar{X}$	R-square
(1,1)	3.9	0.3	376.3	0.74
(1,2)	-5.0	0.7	188.8	0.77
(1,3)	6.4	0.9	126.5	0.69
(1,4)	-4.9	1.4	95.1	0.79
(2,1)	6.5	0.6	187.4	0.75
(2,2)	1.4	1.3	94.4	0.75
(2,3)	9.0	1.8	62.7	0.76
(2,4)	6.2	2.5	47.7	0.75
(3,1)	9.4	0.9	124.2	0.71
(3,2)	-4.5	2.1	62.2	0.77
(3,3)	-5.2	3.0	42.4	0.71
(3,4)	0.8	3.9	31.8	0.67
(4,1)	18.9	1.1	95.5	0.60 *
(4,2)	8.8	2.4	48.3	0.63
(4,3)	19.5	3.3	32.1	0.60
(4,4)	1.2	5.0	24.7	0.67
(5,5)	11.4	7.1	15.9	0.62
(6,6)	38.6	8.6	9.9	0.57
(7,7)	38.1	10.4	8.2	0.68
(8,8)	75.2	7.9	6.1	0.27 **

* $P < 0.05$

** $P < 0.01$

TABLE A2: TREATMENT COMBINATION B RESULTS

SAMPLING SCHEME	B0	B1	$\bar{X}$	R-square
(1,1)	-21.8	0.3	344.3	0.65
(1,2)	-20.4	0.6	170.8	0.62
(1,3)	-19.6	1.0	113.7	0.60
(1,4)	-22.6	1.3	85.1	0.62
(2,1)	-21.4	0.6	172.9	0.67
(2,2)	-18.5	1.3	85.6	0.63
(2,3)	-13.8	1.8	58.0	0.60
(2,4)	-20.6	2.6	42.6	0.67
(3,1)	-26.2	1.0	115.8	0.63
(3,2)	-24.4	2.0	57.0	0.56
(3,3)	-20.3	3.0	37.2	0.57
(3,4)	-28.2	4.2	28.2	0.53
(4,1)	-16.5	1.2	85.6	0.59
(4,2)	-11.0	2.4	42.3	0.54
(4,3)	-1.4	3.2	28.2	0.51
(4,4)	-17.6	5.2	20.8	0.61
(5,5)	28.4	4.7	13.1	0.19 **
(6,6)	0.9	9.4	9.5	0.66
(7,7)	12.9	11.8	6.5	0.56
(8,8)	47.6	8.1	5.2	0.12 **

** $p < 0.01$

TABLE A3: TREATMENT COMBINATION C RESULTS

SAMPLING SCHEME	B0	B1	$\bar{X}$	R-square
(1,1)	-21.1	0.4	302.6	0.79
(1,2)	-21.8	0.7	149.9	0.79
(1,3)	-20.5	1.1	99.3	0.78
(1,4)	-15.9	1.4	74.0	0.76
(2,1)	-21.3	0.7	151.5	0.79
(2,2)	-20.8	1.5	75.3	0.78
(2,3)	-16.5	2.1	50.4	0.77
(2,4)	-17.9	2.9	37.1	0.77
(3,1)	-30.6	1.2	102.3	0.82
(3,2)	-34.1	2.5	50.3	0.81
(3,3)	-28.9	3.6	32.8	0.80
(3,4)	-29.6	4.9	24.2	0.74
(4,1)	-18.0	1.5	73.8	0.77
(4,2)	-17.1	2.9	36.8	0.71
(4,3)	-9.6	4.1	24.0	0.73
(4,4)	-19.8	6.0	18.4	0.72
(5,5)	11.0	6.7	11.8	0.33 **
(6,6)	-3.4	11.3	8.2	0.80
(7,7)	17.8	12.7	5.7	0.61 *
(8,8)	23.0	14.3	4.7	0.26 **

* $p < 0.05$

** $p < 0.01$

TABLE A4: TREATMENT COMBINATION D RESULTS

SAMPLING SCHEME	B0	B1	$\bar{X}$	R-square
(1,1)	-1.1	0.4	45.8	0.79
(1,2)	-1.3	0.7	23.3	0.78
(1,3)	-1.1	1.1	15.2	0.81
(1,4)	-0.5	1.4	11.6	0.75
(2,1)	-1.0	0.7	23.2	0.78
(2,2)	-1.2	1.4	11.9	0.76
(2,3)	-0.9	2.2	7.8	0.79
(2,4)	-0.2	2.7	6.1	0.71 *
(3,1)	-1.4	1.1	15.4	0.77
(3,2)	-1.9	2.3	7.8	0.75
(3,3)	-1.1	3.4	5.0	0.74
(3,4)	0.2	4.3	3.7	0.70
(4,1)	-0.5	3.6	11.1	0.77
(4,2)	0.2	2.7	5.8	0.72 *
(4,3)	-0.4	4.3	3.8	0.77
(4,4)	1.3	5.2	2.8	0.66 *
(5,5)	2.9	7.4	1.8	0.55 **
(6,6)	5.3	8.0	1.3	0.36 **
(7,7)	2.7	11.2	1.2	0.47 **
(8,8)	9.3	8.8	0.8	0.25 **

* $p < 0.05$

** $p < 0.01$

TABLE A5: TREATMENT COMBINATION E RESULTS

SAMPLING SCHEME	B0	B1	$\bar{X}$	R-square
(1,1)	-2.4	0.3	119.8	0.86
(1,2)	-2.2	0.6	59.5	0.87
(1,3)	-2.8	0.9	40.3	0.86
(1,4)	-1.2	1.1	30.0	0.90
(2,1)	-2.7	0.6	59.9	0.85
(2,2)	-2.1	1.2	29.6	0.86
(2,3)	-2.3	1.8	19.8	0.85
(2,4)	-0.1	2.2	14.9	0.87
(3,1)	-1.0	0.8	40.1	0.83
(3,2)	-1.5	1.7	20.0	0.84
(3,3)	-1.0	2.5	13.2	0.82 *
(3,4)	-1.2	3.3	10.2	0.86
(4,1)	-2.2	1.1	30.4	0.83
(4,2)	-0.8	2.2	15.1	0.81
(4,3)	-3.1	3.5	10.0	0.85
(4,4)	-0.2	4.2	7.7	0.85
(5,5)	2.5	6.3	4.8	0.80
(6,6)	3.1	10.0	2.9	0.75 *
(7,7)	7.8	11.2	2.2	0.69 **
(8,8)	4.7	12.3	2.3	0.65 **

* $p < 0.05$

** $p < 0.01$

TABLE A6: TREATMENT COMBINATION F RESULTS

SAMPLING SCHEME	B0	B1	$\bar{X}$	R-square
(1,1)	0.54	0.35	10.8	0.62
(1,2)	0.72	0.68	5.4	0.64
(1,3)	0.52	1.06	3.6	0.65
(1,4)	1.15	1.18	2.7	0.55
(2,1)	0.61	0.69	5.4	0.63
(2,2)	0.92	1.28	2.7	0.62
(2,3)	0.90	1.82	1.9	0.63
(2,4)	1.20	2.22	1.4	0.51
(3,1)	1.04	0.92	3.6	0.57
(3,2)	1.52	1.59	1.8	0.55
(3,3)	1.55	2.38	1.2	0.47
(3,4)	1.98	2.81	0.8	0.49
(4,1)	1.11	1.23	2.6	0.57
(4,2)	1.39	2.19	1.4	0.57
(4,3)	1.45	3.01	1.0	0.48
(4,4)	1.81	3.84	0.7	0.47
(5,5)	1.93	6.19	0.4	0.37
(6,6)	2.29	6.23	0.3	0.36 *
(7,7)	3.49	5.81	0.2	0.17 **
(8,8)	4.02	11.36	<.1	0.10 **

* $p < 0.05$

** $p < 0.01$

TABLE A7: R-SQUARES

SAMPLING SCHEME	TREATMENT COMBINATIONS A	B	C	D	E	F
(1,1)	0.74	0.65	0.79	0.79	0.86	0.62
(1,2)	0.77	0.62	0.79	0.78	0.87	0.64
(1,3)	0.69	0.60	0.78	0.81	0.86	0.65
(1,4)	0.79	0.62	0.76	0.75	0.90	0.55
(2,1)	0.75	0.67	0.79	0.78	0.85	0.63
(2,2)	0.75	0.63	0.78	0.76	0.86	0.62
(2,3)	0.76	0.60	0.77	0.79	0.85	0.63
(2,4)	0.75	0.67	0.77	0.71	0.87	0.51
(3,1)	0.71	0.63	0.82	0.77	0.83	0.57
(3,2)	0.77	0.56	0.81	0.75	0.84	0.55
(3,3)	0.71	0.57	0.80	0.74	0.82	0.47
(3,4)	0.67	0.53	0.74	0.70	0.86	0.49
(4,1)	0.60	0.59	0.77	0.77	0.83	0.57
(4,2)	0.63	0.54	0.71	0.72	0.81	0.57
(4,3)	0.60	0.51	0.73	0.77	0.85	0.48
(4,4)	0.67	0.61	0.72	0.66	0.85	0.47
(5,5)	0.62	0.19	0.33	0.55	0.80	0.37
(6,6)	0.57	0.66	0.80	0.36	0.75	0.36
(7,7)	0.68	0.56	0.61	0.47	0.69	0.17
(8,8)	0.27	0.12	0.26	0.25	0.65	0.10

TABLE A8: TRANSFORMED DIFFERENCES (TD)

SAMPLING SCHEME	TREATMENT COMBINATIONS A	B	C	D	E	F
(1,1)	-	-	-	-	-	-
(1,2)	-3	3	0	1	-1	-2
(1,3)	5	5	1	-2	0	-3
(1,4)	-5	3	3	4	-4	7
(2,1)	-1	-2	0	1	1	-1
(2,2)	-1	2	1	3	0	0
(2,3)	-2	5	2	0	1	-1
(2,4)	-1	-2	2	8	-1	11
(3,1)	3	2	-3	2	3	5
(3,2)	-3	9	-2	4	2	7
(3,3)	3	8	-1	5	4	15
(3,4)	7	12	5	9	0	13
(4,1)	14	6	2	2	3	5
(4,2)	11	11	8	7	5	5
(4,3)	14	14	6	2	1	14
(4,4)	7	4	7	13	1	15
(5,5)	12	46	46	24	6	25
(6,6)	17	-1	-1	43	11	26
(7,7)	6	9	18	32	17	45
(8,8)	47	53	53	54	21	52

TABLE A9: SAMPLING SCHEME ELIMINATION

SAMPLING SCHEME	Σ|TD|	RANK	ELIMINATION CODE
(1,2)	10	3	0
(1,3)	16	5	0
(1,4)	26	8	2
(2,1)	6	1	0
(2,2)	7	2	0
(2,3)	11	4	0
(2,4)	25	7	1
(3,1)	18	6	1
(3,2)	27	9	2
(3,3)	36	11	1
(3,4)	46	12	2
(4,1)	32	10	1
(4,2)	47	13	1
(4,3)	51	15	2
(4,4)	47	13	1
(5,5)	159	18	1
(6,6)	99	16	1
(7,7)	127	17	1
(8,8)	280	19	1

0 NOT ELIMINATED

1 FIRST GROUP ELIMINATED

2 SECOND GROUP ELIMINATED