M. M. Hixson
B. J. Davis
M. E. Bauer
Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, IN 47907

Sampling Landsat Classifications for Crop Area Estimation*

The precision of the wheat area estimates increased as the segment (cluster) size decreased and the number of segments was increased.

Abstract: *The objective of this investigation was to evaluate the effect of several sampling plans on the precision and bias of crop area estimates made by sampling classifications of Landsat MSS data. Full-frame classifications of wheat and non-wheat for 80 counties in Kansas were repetitively sampled to simulate alternative sampling plans. Four sampling plans involving different numbers of samples and different size sampling units were evaluated. The precision of the wheat area estimates increased as the segment (cluster) size decreased and the number of segments was increased. Although the average bias associated with the various sampling schemes was not significantly different, the maximum absolute deviation was directly related to sampling unit size.*

Introduction

Accurate and timely crop production information is essential for planning the production, storage, transportation, and processing of grain crops; making marketing decisions; and determining national agricultural policies. Although most countries of the world gather crop production data, relatively few countries have reliable inventory systems. The synoptic view of the Earth provided by satellite remote sensing, along with computer processing of the data, provides the opportunity to identify and estimate the area of crops.

The most comprehensive investigation of the use of Landsat MSS data for crop surveys has been the Large Area Crop Inventory Experiment (LACIE) (MacDonald and Hall, 1980). The purpose of LACIE was to assimilate current remote sensing technology into an experimental system and evaluate its potential for determining the production of wheat in various regions of the world. In LACIE, area estimates were made from classifications of Landsat MSS data. Yield was estimated for fairly broad geographic regions using statistical regression models developed from historical weather and wheat yield data.

For the area estimation phase of LACIE, 5 by 6 nautical mile samples were drawn to represent about 2 percent of the agricultural land area to be inventoried. Segments (clusters) were allocated to political units (counties) based on a function of county size and the historical area of wheat. The sample segments were used both for training the classifier and for area estimation. The LACIE method was generally successful in obtaining unbiased and precise area estimates. Bias refers to the average size of deviations from the true parameter, while precision refers to the average size of deviations from the mean of all estimates of the parameter obtained through repeated applications of the sampling procedure (Cochran, 1963). Six hundred segments were selected in the United States, and 1900 in the Soviet Union, to achieve a sampling error of 2 percent.

An alternative to the LACIE sampling plan for obtaining crop area estimates had been utilized earlier in an investigation at LARS (Bauer *et al.*, 1978). A systematic sample of pixels (point samples) spread throughout a Landsat full-frame was

* Purdue University Agricultural Experiment Station Journal Paper No. 7963.

0099-1112/81/4709-1343$02.25/0

classified and used to make area estimates, while training data were obtained separately. The classifications were performed on a county basis using every other line and every other column of Landsat data. Training statistics were developed using photointerpretation from single date aerial photography taken along several flightlines dispersed throughout the state and were extended to counties lacking reference data but known to have similar land use, crops, and soils. The pixel sampling approach was demonstrated to have the capability to produce unbiased and precise area estimates for small (e.g., county) as well as large (e.g., state) geographic areas (Bauer *et al.*, 1978).

The bias and precision of crop area estimates made using remotely sensed data are affected both by measurement (non-sampling) error and sampling error. Measurement error is a function of the spectral features measured, the sensor employed, the timing of the crop observations, and the data analysis methodology utilized. Studies of measurement error due to these sources have been carried out; however, the effect on bias and precision of a particular plan for sampling classifications of Landsat MSS data has not been extensively researched.

Objectives

The overall objective of this investigation was to evaluate the effect of several sampling alternatives on the accuracy of crop area estimates made from classification of Landsat MSS data. The specific objective was to assess the precision and the bias associated with alternative sampling schemes involving different numbers of several sampling unit sizes.

Experimental Approach

Ideally, a study of bias and precision of a sampling scheme would be conducted by sampling repetitively from the population of interest. In this case, however, the population of interest is the true distribution of crops in a state (or other region), and this truth is not generally known for large regions. An alternative approach to actually sampling the population is to simulate its occurrence. Simulated data are used instead of truth, and they are repetitively sampled to determine a variance. Estimates of the bias are made by comparison with the mean of the simulated data.

The approach taken in this study was a combination of repetitive sampling and simulation. Full-frame classifications of Landsat MSS data for a portion of Kansas into wheat and non-wheat were used as simulated ground truth (Bauer *et al.*, 1978). Data for 80 counties comprising seven crop reporting districts were available and were used in the study. The study region will be referred to as the pseudo-state. The full-frame classifications were considered to be essentially the same as ground truth because the estimates of wheat area did not differ significantly from the USDA/ESCS estimates at the pseudo-state level and were sampled repetitively to simulate eight replications of each of four alternative sampling plans.

Two types of samples were considered: cluster (segment) sampling and point (pixel) sampling of full-frames. Four sampling schemes were selected for testing: (1) 75 5 by 6 nm samples, (2) 137 4 by 4 nm samples, (3) 560 2 by 2 nm samples, and (4) 427,587 pixel samples. Each of these four alternative sampling plans was designed to contain about the same number of pixels, which kept each scheme at a fixed cost with our analysis and classification methodology.

Procedures similar to those followed in LACIE were used to determine the allocation (number) of samples, location (geographic placement) of segments, and the aggregated area estimate of wheat (Hallum *et al.*, 1978; LACIE Staff, 1974). These procedures are described in the remainder of this section.

Sample Segment Allocation

Based on a total of 84 sample segments allocated to the state of Kansas in LACIE, 75 sample segments were needed to sample the pseudo-state region at the same rate. For each sampling plan, the number of segments per county was computed. The threshold value, t_k^*, for each county was computed based on the total number of acres in the county and the standard deviation of the proportion of wheat in that county. For county k,

$$t_k^* = A_k \sqrt{p_k (1 - p_k)},$$

where A_k is the total land area in county k, and p_k is the historical proportion of wheat in county k. The proportional number of sample segments allotted to each county was computed employing the equation

$$N_k = \frac{84\, t_k^*}{\sum_{k=1}^{n} t_k^*}$$

where t_k^* is as defined above and n is the number of counties in the pseudo-state. The number of sample segments allotted to each crop reporting district (CRD) was computed similarly.

The sampling method used for each county was then determined by the following procedure:

- Stratified sample segment—all counties with $N_k > 0.6$ will have at least one sample segment; the actual number of segments is the rounded value of N_k.
- No sample segments allotted if $N_k < 0.3$.
- The counties in the CRD with $0.3 < N_k < 0.6$ were allotted the remaining number of segments determined for the CRD by sampling with probability proportional to size (pps).

Allocations strictly according to the LACIE procedure produced county allocations which did not add to the total number allocated for the crop reporting district, so the rules presented above were slight modifications of the LACIE rule to achieve consistency. Two counties received two sample segments, seven counties received no sample segments, and the remainder of the counties received one segment in the 5 by 6 nm segment allocation. The criteria were generalized for other segment sizes.

SAMPLE SEGMENT LOCATION

The selection of sample segments was computer-implemented. This allowed a large number of segments to be chosen with little personnel time and also facilitated choice of any segment size or number of segments. The greater number of samples which could be taken through automated selection permitted statistical tests of precision. The description of the procedure which was implemented follows.

A grid, spaced 6 nm in the east-west direction and 5 nm in the north-south direction, was defined to cover the study region. To select a sample for a given county, the number of segments whose centers were inside the county boundaries and which did not fall entirely in the defined nonagricultural areas was determined, and a sample was randomly selected from these.

The location of sample segments differed in two respects from the location of LACIE segments: first, in the definition of nonagricultural areas, and second, in the number of segments permitted in a window or extended rectangle about a given segment.

Nonagricultural areas of at least 2 by 2 miles in size were excluded from consideration as sample segments. The boundaries of urban areas, federal lands, reservoirs and other nonagricultural areas appearing on county maps prepared by the State Highway Commission of Kansas, Department of Planning and Development, were found using a coordinate digitizer. The boundary definitions of nonagricultural areas were somewhat more crude than those defined by LACIE. The reasons for this include (1) constraints of time (including computer time) and resources (including detailed maps) and (2) the belief that only major nonagricultural areas needed to be excluded because experience has indicated that, even when few nonagricultural areas are excluded, estimates of high accuracy can be obtained (Bauer *et al.*, 1978). This constraint that a sample segment not fall within a nonagricultural area was ignored with the pixel sampling method due to excessively high costs of computer checking for each of the nearly four million samples.

The selected segment was then checked against a set of system constraints. The first constraint was that a new 5 by 6 nm segment was discarded if there was already a sample segment selected within a 10.5 by 12 nm rectangle centered about the new segment. A second constraint defined two extended rectangles: one, running in the east-west direction, was 10.5 by 80 nm, and the other, running north-south, was 100 by 12 nm. Only four sample segments were permitted to fall in the east-west extended rectangle, and no more than eight sample segments were permitted to fall in the north-south extended rectangle. If the new segments caused either of these constraints to fail, it was discarded and a new random draw was made.

The constraints concerning the number of segments permitted in a given size rectangle centered about the sample segment and its east-west and north-south extensions to 80 and 100 nm, respectively, were adjusted by number and size of the rectangle to be relatively consistent with the constraints for the LACIE 5 by 6 nm segments. It was not feasible to use this type of constraint for the pixel selection procedure.

AREA ESTIMATION PROCEDURE

Wheat area estimates were calculated for each replication for the counties and were aggregated to obtain estimates for the crop reporting districts and the pseudo-state. For crop reporting district j, the area estimate was computed by

$$A_j = A_{1j} + A_{2j} + A_{3j}$$

where A_{1j} is the estimate of the area in the counties within the CRD which had no segments allocated, A_{2j} is the estimate for those counties which were allocated segments with probability proportional to size, and A_{3j} is the estimate for counties allocated one or more segments.

The wheat areal proportion in the kth county estimated directly from the segments in that county is $\hat{p}_{jk}'$ where

$$\hat{p}_{jk}' = \frac{\text{no. of pixels classified as wheat}}{\text{total no. of pixels classified}}.$$

In order to adjust this proportion for the nonagricultural area in the county which was not in the sampling frame, $\hat{p}_{jk}$ was computed as

$$\begin{aligned}\hat{p}_{jk} &= \hat{p}_{jk}' (A_k - N_k) + 0\,(N_k)\\ &= \hat{p}_{jk}' (A_k - N_k)\end{aligned}$$

where $\hat{p}_{jk}'$ is as defined above; A_k is the total land area in the kth county; and N_k is the total nonagricultural land area in the kth county.

For the m_{3j} counties containing one or more segments, A_{3j} is simply the sum of the proportion of wheat in each county as estimated from the sample segments multipled by the area of the counties containing the segments: i.e.,

$$A_{3j} = \sum_{k=1}^{m_{3j}} \hat{p}_{jk} A_k$$

where $\hat{p}_{jk}$ is the wheat areal proportion in the kth county estimated from the segments and A_k is the total land area in the kth county.

For that set of m_{2j} counties in a crop reporting district to which segments were allocated with probability proportional to size, the area of wheat was estimated by employing the equation

$$A_{2j} = A_2 \frac{p_j}{m_{2j}} \sum_{k=1}^{m_{2j}} \frac{\hat{p}_{jk}}{p_{jk}}$$

where A_2 is the total land area of counties in group 2; p_j is the agricultural census proportion for all counties in that group; m_{2j} is the number of sample segments in this set of counties; $\hat{p}_{jk}$ is the Landsat estimate of wheat proportion in the kth county; and p_{jk} is the agricultural census wheat proportion for the kth county.

For the m_{1j} counties in the jth district which received no sample segments, the area estimate is

$$A_{1j} = \frac{A_{2j} + A_{3j}}{\sum_{k=1}^{m_{2j}} A_k p_{jk} + \sum_{k=1}^{m_{3j}} A_k p_{jk}} \sum_{k=1}^{m_{1j}} A_k p_{jk}$$

where A_{2j} and A_{3j} are as defined above; A_k is the total land area in county k; and p_{jk} is the agricultural census wheat proportion for county k.

RESULTS EVALUATION

For each sampling plan, eight replications were carried out. The 32 resulting area estimates were then compared. The first comparison was a qualitative one, plotting the pseudo-state area estimates by sampling plan. As the nonhomogeneous variances did not satisfy the requirements for classical statistical testing, two nonparametric tests, the Kruskal-Wallis rank sum test for one-way classifications and the sign test (Hollander and Wolfe, 1973), were carried out to determine if a bias were present using any of the methods.

The variability of the estimates made by each of the sampling plans was first examined qualitatively. In order to perform statistical testing, at least two sampling errors per plan were required. Thus, for each sampling plan, a standard deviation of the estimate was computed using four replications. These two standard deviations per plan were compared using a distribution-free multiple comparisons test based on the Kruskal-Wallis rank sums.

Results and Discussion

The effects of jointly varying sampling unit size and the number of samples on the accuracy of crop area estimates are illustrated in Figure 1 and summarized in Table 1.

PRECISION OF ESTIMATES

Figure 1 shows that the use of fewer, larger sample unit sizes results in a greater range and more variability in the estimates than the use of more, smaller samples. The standard deviations obtained range from 12,000 ha for pixel samples to 223,500 ha for 5 by 6 nm segments (Table 1). Coefficients of variation range from 0.2 percent for pixel samples to 4.0 percent for 5 by 6 nm segments. The variability associated with the pixel samples is, thus, nearly negligible, while the 4 percent variability associated with the 5 by 6 nm segments is not negligible.

These observations are supported by statistical results. A distribution-free multiple comparisons test based on the Kruskal-Wallis rank sums was performed (Hollander and Wolfe, 1973) to assess which pairs of sampling plans, if any, had significantly different standard deviations. At the 5 percent level of significance, the only sampling plans which had significantly different standard deviations were those using 5 by 6 nm and pixel samples.

BIAS OF ESTIMATES

The results presented in Figure 1 indicate that there may be some difference in the means of estimates made using the different sampling plans.

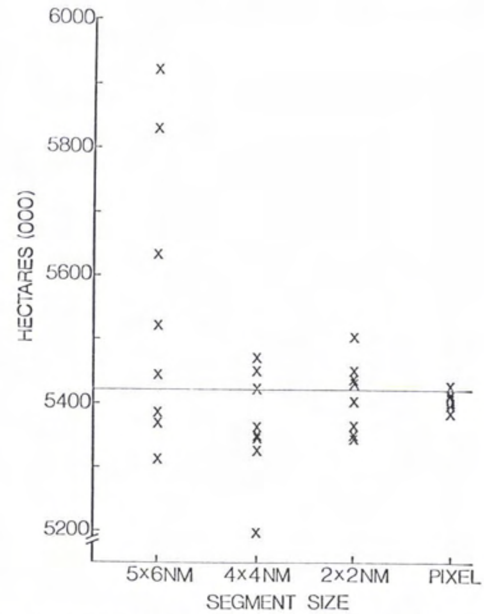

Fig. 1. Wheat area estimates for several sampling schemes compared with the population parameter (horizontal line). The size of segments was varied inversely with the number of segments in order to hold the total number of pixels sampled nearly constant.

TABLE 1. PRECISION AND BIAS OF CROP AREA ESTIMATES FOR DIFFERENT SAMPLING SCHEMES

Sampling Scheme					Difference From Population Total	
Number of Samples	Sample Unit Size	Mean (000 Ha)	Standard Deviation (000 Ha)	Coefficient of Variation (%)	Maximum (000 Ha)	Average (000 Ha)
75	5 by 6 nm	5550.9	223.7	4.0	498.2	127.5
137	4 by 4 nm	5365.0	86.3	1.6	−227.4	−58.4
560	2 by 2 nm	5409.6	55.2	1.0	80.5	−13.8
427,587	Pixel	5405.9	12.1	0.2	−39.1	−17.5

The means range from 5,365,000 ha to 5,550,000 ha (Table 1). Unlike the standard deviations, the means are not ranked in order according to the sample unit size.

The horizontal line in Figure 1 represents the total number of hectares of wheat in the classifications which were sampled. This number is the true population parameter which is to be estimated. A large systematic bias is not indicated since the population parameter falls in the center portion of the range of the estimates for all the sampling schemes, rather than most of the observations being either above or below the line. However, as indicated in Table 1, the smaller sampling units tend to produce estimates which have less bias.

Two types of nonparametric tests were performed to assess the bias of the several sampling methods. The Kruskal-Wallis rank sum test for one-way classifications was used to determine the effect of sampling plan on the area estimates (Hollander and Wolfe, 1973). No significant difference in the means was found. The sign test was performed on the estimates to determine if the mean of any of the sampling schemes was significantly different from the true area of the data sampled (Freund, 1962). Again, no statistically significant differences were found.

Although none of the sampling schemes appeared to have a systematic bias, it is important to examine the maximum deviation which was generated by each of the sampling schemes. The maximum deviation was directly related to the sampling unit size. The maximum absolute deviation for pixel samples was only about 39,000 hectares, while one of the eight 5 by 6 nm samples gave an overestimate of 498,000 hectares.

SUMMARY AND CONCLUSIONS

The estimates achieved using the 5 by 6 nm segments have the least precision of any sampling scheme tested. The estimates become more precise as the segment size decreases and more segments are taken. The precision of the 5 by 6 nm segments was significantly less than that of the pixel samples.

None of the sampling schemes was significantly biased on the average, and none of the average estimates differed significantly from the population parameter. The maximum absolute deviation, however, was directly related to sampling unit size and should be considered in selection of a sampling unit.

To assess the implications of the results of this study for operational use, other factors must be considered. In order to fully evaluate the scheme, the measurement procedure, which would include the methods of training and classification used in conjunction with a sampling plan, must also be considered. And, although the precision of estimates from choosing more but smaller segments may be higher, this gain in precision must be weighed against the costs of sample selection and classification.

A somewhat similar study was recently conducted by Perry and Hallum (1979). The objective of that study was to ascertain the effect of a change in the sampling unit size on the total number of sampling units necessary to support a wheat production estimate with a specified coefficient of variation. Their results agreed with the results of this investigation, but they concluded that no recommendation can be made until a model for the cost of selecting and analyzing segments as a function of the sampling unit size is developed.

ACKNOWLEDGMENTS

This research was sponsored by NASA Johnson Space Center (Contract NAS9-14970).

REFERENCES

Bauer, Marvin E., Marilyn M. Hixson, Barbara J. Davis, and Jeanne B. Etheridge, 1978. Area Estimation of Crops by Digital Analysis of Landsat Data, *Photogrammetric Engineering and Remote Sensing* 44:1033-1043.

Cochran, W. G., 1963. *Sampling Techniques*, Second Edition: John Wiley and Sons, Inc., New York (p. 96).

Freund, John E., 1962. *Mathematical Statistics:* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, pp. 289-290.

Hallum, C. R., R. S. Chhikara, A. H. Feiveson, and A. G. Houston, 1978. Sampling, Aggregation, and Variance Estimation for Area, Yield, and Production in LACIE, *Proc., LACIE Symposium*, Vol. 1, (JSC-16015), October 23-26, 1978, NASA, Johnson Space Center, Houston, Texas, pp. 3-18.

Hollander, Myles, and Douglas A. Wolfe, 1973. *Nonparametric Statistical Methods:* John Wiley and Sons, Inc., New York, pp. 124-129.

LACIE Staff, 1974. *Crop Assessment Subsystem Requirements, Level III Baseline,* LACIE-00200, Volume IV, NASA, Johnson Space Center, Houston, Texas, pp. B1-B5.

MacDonald, R. B., and F. G. Hall, 1980. Crop Forecasting, *Science* 208:670-679.

Perry, Charles, and Cecil Hallum, 1979. Sampling Unit Size Considerations in Large Area Crop Inventorying using Satellite-Based Data, *American Statistical Association 1979 Proc.*, Section on Survey Research Methods, August 13-16, 1979, Washington, D.C., pp. 430-433.

(Received 21 March 1980; revised and accepted 5 March 1981)

ISPRS Inter-Congress Symposia

Midway between its quadrennial Congresses, each of the technical commissions of the International Society for Photogrammetry and Remote Sensing holds a symposium. Those symposia scheduled for 1982 are listed below.

Commission I*
Advances in the Quality of Image Data
Canberra, Australia
14-16 April 1982

Commission II*
Advances in Instrumentation for Processing and Analysis of Photogrammetric and Remote Sensing Data
Ottawa, Canada
30 August-4 September 1982

Commission III*
Mathematical Models, Accuracy Aspects, and Quality Control
Helsinki, Finland
7-11 June 1982

Commission IV
Computer Assisted Photogrammetry and Cartography
Washington, D.C., U.S.A.
23-27 August 1982

Commission V*
Precision and Speed in Close-Range Photogrammetry
York, England
5-11 September 1982

Commission VI
(Still Open)
(Open)
30 August-3 September 1982

Commission VII
Operational Utilization and Interpretation of Remote Sensor Data
Toulouse, France
13-18 September 1982

* Calls for papers have been published in *Photogrammetric Engineering and Remote Sensing* for these Commission Symposia as follows:

Commission I, September 1981, page 1372
Commission II, September 1981, page 1342
Commission III, September 1981, page 1325
Commission V, March 1981, page 364