SAMPLING UNIT SIZE CONSIDERATIONS IN LARGE AREA CROP INVENTORYING USING SATELLITE-BASED DATA

Charles Perry and Cecil Hallum, NASA Johnson Space Center

1.0 Introduction

The first systematic attempt to collect agricultural statistics dates back more than a century to the Census of 1840 (Benedict, 1939). From that date forward an increasing volume of agricultural statistics has been collected periodically in Census enumeration decennially to 1920 and quinquennially thereafter. A rudimentary system of annual agricultural estimation was also begun about 1840 in the Patent Office. Upon Commissioner Ellsworth's resignation in 1845, however, interest in agricultural statistics subsided in the Patent Office, and it was not until after the Department of Agriculture was organized in 1862 that annual intercensus estimates were again revived (Ebhling, 1939). Current monthly reports on crop conditions also predated the establishment of the Department of Agriculture by a few months. Orange Judd, editor of the American Agriculturalist, published summaries of crop condition reports submitted voluntarily by subscribers to his paper for the five months, May through September, 1862 (Ebhling 1939). Judd's efforts were the forerunner to the Department's program of monthly reports on crop prospects which have been issued regularly during the growing season since the first publication in July 1863.

Since 1863, the estimating work of the Department of Agriculture has expanded very greatly until today a large volume of agricultural estimates is published on a current basis. Until recent efforts of the USDA Statistical Reporting Service (now part of the Economics, Statistics, and Cooperative Services) and the Large Area Crop Inventory Experiment (LACIE) conducted at NASA/JSC, in Houston, Texas (refs. 9, 10, 11, and 14), the predominant method has been one involving the use of mailed inquiries for collection of basic data and an assortment of techniques utilized to remove bias in the transformation of basic data into published estimates. Since 1974, satellite remote sensing technology, developed in the previous decade, in conjunction with statistical survey methodology were assembled into an experimental crop inventory system (LACIE) and tested for wheat in several countries. This experiment was concluded with the LACIE Symposium conducted at NASA/JSC in October 1978 (ref. 14). For details of the sampling strategy utilized in LACIE, refer to the Proceedings of the aforementioned LACIE Symposium or to the paper by Chhikara and Feiveson in last year's Proceedings of the Annual Meeting of the ASA (ref. 3) held in San Diego.

In seeking to improve the efficiency of crop area estimation, the choice of the optimal sampling unit size has been a subject of much discussion at NASA/JSC. The purpose of this paper is to report preliminary results of the sampling unit size investigation, ongoing at NASA/JSC, that supports timely estimates on a global basis of crop acreages utilizing remotely-sensed (satellite-acquired) data. The approach taken is one of modeling the acreage variance as a function of sampling unit size based on studies by Smith (1938), Mahalanobis (1940), Jessen (1942), Cochran (1942), Hansen and Hurwitz (1942), and Asthana (1950). The size of the sampling units investigated in these earlier studies were limited in size from several square feet up to approximately forty acres. This paper reports the results of variance modeling for sampling units up to approximately 25,000 acres in size. Finally, this modeled relation is utilized in arriving at a closed-form solution to the optimal sampling unit size that minimizes cost.

2.0 The Sampling Unit Utilized in LACIE

It was decided at the outset of LACIE that sampling of areas was not only desirable but essential. It became apparent that the conversion of the satellite-acquired spectral measurements to wheat acreage estimates could not be accomplished by an automatic computerized procedure but had to be done with the participation of human intelligence (photograph interpretation by analyst-interpreters). The time-cost element of this participation had to be assessed against the efficiency of LACIE sampling techniques. It was found that the sampling error (approximately 2 percent) resulting from quite moderate sampling fractions (approximately 3 percent) was comparable if not smaller than the percentage error resulting from measurements. Cost-effectiveness and measurement considerations played a major role indicating the sampling unit size selected at the outset of LACIE.

For various reasons, it was impractical to consider using sampling units as small as one acre in size. Instead, LACIE decided to use an area unit and record the spectral measurements for all resolution elements within the area unit as the sample information. The size of the selected sampling area was 5 by 6 nautical miles. It may be argued that this unit is too large from the standpoint of sampling efficiency (it contains approximately 25,000 acres). The size of this unit may not be optimum; however, the following practical considerations dictated the use of a unit of at least a comparable size.

1. It was necessary to register the acquisition of data from segments acquired during the various passages of the satellite over the same segment. The technology of identifying the same segment in these various passages requires key points within the segment that are easily recognizable and, in turn, this requires a segment of an adequate size.

2. Again, the satellite imagery and its interpretation by the analysts, as well as the computation of signatures custom-made for the segment, requires an adequate size, as does the measurement procedure.

3. LACIE addressed the problem of how the variance of the statistical sample could be reduced by using areas of smaller size; the gains did not justify changing from the above segment size to a much smaller area in view of the aforementioned and other practical limitations.

With future plans for system capabilities that permit a relaxation of many of the constraints that existed in LACIE, additional consideration can be given to alternative sampling unit sizes which is the subject of the remainder of this paper.

3.0 Model Form Selected for Investigation

The guiding theory for selecting the proper size of cluster has been investigated by a number of statisticians. Several attempts have been made to work out the relationship between the variance of the mean of a single cluster and its size. The first one was due to Fairfield Smith (1938). He found the relationship to be satisfactory on yield data for different size plots. Jessen (1942) showed that most economic characters relating to farm data follow a slightly different law from that of Fairfield Smith. He postulated that the mean square among elements within a cluster is a monotonic increasing function of the size of the cluster. The same relationship developed by Jessen was independently suggested by Mahalanobis (1940). This was also the finding of Asthana (1950) who has fitted Jessen's law to describe the mean square within clusters for acreage under wheat for a large number of villages. The algebraic solution of the problem of choosing the optimum number and size of clusters was given by Cochran (1942), confirming the conclusions based on Jessen's empirical calculations. The fact that Jessen's approach was not universally applicable was soon evidenced when Hansen and Hurwitz (1942) presented examples which showed that for certain items in urban sampling the variance function was quite different from that used by Jessen. In any case, the success of these studies dictated our choice of model and the subsequent investigation in this paper.

The above studies indicated that the use of the power function is a strong candidate for providing a simple yet satisfactory mathematical model for the functional dependence of the population unit-to-unit variance on the sampling unit size. The size of the sampling units in these earlier studies were limited to sizes ranging from several square feet to approximately 40 acres. This paper investigates the utility of the power function in modeling the variance as a function of sampling units ranging all the way up to more than 25,000 acres.

The remaining sections of this report cover the approach used to determine the model fit, an evaluation of the model using ground truth data collected from the 1977-78 wheat crop year of the Large Area Crop Inventory Experiment in the U.S. Great Plains, and, finally, derivation of the optimal sampling unit size under certain cost considerations.

4.0 Approach for Estimation of Model Parameters

This section gives a brief description of the Analysis of Variance Techniques (see Cochran [1977]) used to obtain estimates of the cluster-to-cluster wheat area variance for different size clusters and the approach used to fit the power function. In the following discussion, let N denote the total number of 5 by 6 nautical mile segments constituting the sampling frame (i.e., the agricultural area of a stratum) and consider each to be further subdivided into M subunits of equal size (discounting left over areas). Finally, letting n denote a random sample of n segments from the stratum and A_{ij} denote the crop area in segment i (i=1,...,n) for subunit j (j=1,...,M), then S_b^2, S_w^2, and S^2 provide unbiased estimates for σ_b^2, σ_w^2, and σ^2, respectively, (see Cochran [1977]) where:

$$S_b^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{M} (A_{i.} - A_{..})^2}{n-1} \tag{4.1}$$

$$S_w^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{M} (A_{ij} - A_{i.})^2}{n(M-1)} \tag{4.2}$$

$$S^2 = \frac{N-1}{NM-1} S_b^2 + \frac{N(M-1)}{NM-1} S_w^2 \tag{4.3}$$

Historically (refs. 1, 4, 7, 8, 12, and 13), the model

$$S^2 (x) = Ax^B \tag{4.4}$$

has been found to work quite well in relating the areal subunit size, x, to the subunit-to-subunit crop area variance, $S^2(x)$ (A and B are estimated parameters). Using the 5 by 6 nautical mile data collected from the 1977-78 wheat crop in the U.S. Great Plains for input to equations (4.1)-(4.3), A and B in (4.4) were estimated by the method of least squares.

5.0 Evaluation of Fitted Model

Digitized ground truth for a random sample of 124, 5 by 6 nautical mile segments from nine states (see Table 5.1) was utilized in equations (4.1)-(4.3) to estimate A and B in (4.4) for subunits ranging in size from 171 to 25,463 acres.

STATE	NUMBER OF GROUND TRUTH SEGMENTS
COLORADO	9
KANSAS	13
MINNESOTA	13
MONTANA	18
NEBRASKA	15
NORTH DAKOTA	19
OKLAHOMA	13
SOUTH DAKOTA	15
TEXAS	9
TOTAL	124

Table 5.1: Summary of Data by State

Estimates of the variance using the fitted equation were in close agreement with the estimate obtained from the analysis of variance technique with coefficients of determination being very close to one for all states. The relative errors, sum of relative errors, and the mean of the absolute relative errors were all negligibly small for each state. The subunit-to-subunit variance was estimated directly from the data set for other subunit sizes not used in the approximation of A and B. These estimates also proved to be in very close agreement with the projected values estimated from the fitted models. Table 5.2 summarizes the estimates for A and B for each of the nine states. Table 5.3 details the results for Texas (similar results were obtained for the remaining eight states investigated). Assuming equal costs (per sampling unit), Table 5.4 summarizes the 9-state allocation (under a Neyman allocation) and sampling rate results as a function of the sampling cluster size.

STATE	A	B
COLORADO	0.040	1.67
KANSAS	0.040	1.70
MINNESOTA	0.044	1.82
MONTANA	0.030	1.72
NEBRASKA	0.029	1.81
NORTH DAKOTA	0.027	1.58
OKLAHOMA	0.089	1.80
SOUTH DAKOTA	0.017	1.72
TEXAS	0.066	1.74
Median Value of B = 1.72		
Minimum Value of B = 1.58		
Mean Value of B = 1.73		
Maximum Value of B = 1.82		

Table 5.2: State-Level Parameter Estimates of A and B in $S^2(x)=Ax^B$

STATE MODEL	$S^2(x) = 0.0658\ x^{1.7351}$		
SUB UNIT AREA	ESTIMATED VARIANCE	PROJECTED VARIANCE	PERCENT RELATIVE ERROR
39.67	36.8112	39.0906	6.2
9.92	3.7381	3.5271	-5.6
4.40	0.8955	0.8603	-3.9
2.47	0.3195	0.3151	1.4
1.58	0.1442	0.1454	0.8
1.09	0.0752	0.0765	1.7
0.81	0.0453	0.0456	0.8
0.61	0.0278	0.0279	0.2
0.48	0.0187	0.0188	0.2
0.39	0.0130	0.0131	1.0
0.31	0.0089	0.0088	-0.7
0.27	0.0066	0.0067	1.2

Table 5.3: Summary of Results for Texas

Under stratified random sampling, the acreage estimator, $\hat{A}$, has the form

$$\hat{A} = \left[\sum_{j=1}^{L} \frac{1}{n_j} \sum_{i=1}^{n_j} \hat{A}_{ij} \right] N_j \qquad (5.1)$$

where

L = the total number of strata

n_j= the number of sampling units selected from stratum j

$\hat{A}_{ij}$=the crop acreage estimate for the ith sampling unit in stratum j

and

N_j= the total number of sampling units in the sampling frame of stratum j.

Similarly, from (5.1), the variance, $\sigma_{\hat{A}}^2$, of A is given by

$$\sigma_{\hat{A}}^2 = \sum_{j=1}^{L} N_j^2 \left(1 - \frac{n_j}{N_j}\right) \frac{\sigma_{\hat{A}_j}^2}{n_j}$$
$$\doteq \sum_{j=1}^{L} N_j^2 \frac{\sigma_{\hat{A}_j}^2}{n_j} \qquad (5.2)$$

Replacing N_j and $\sigma_{\hat{A}_j}^2$ in (5.2) with

$$N_j = \frac{A_j}{x_j} \qquad (5.3)$$

and

$$\sigma_{\hat{A}_j}^2 = a_j x_j^{b_j} \qquad (5.4)$$

where

A_j = the total area of the sampling frame in the jth stratum

x_j = the total area of each sampling unit in stratum j

and a_j and b_j are parameters estimated using the approach discussed earlier, $\sigma_{\hat{A}}^2$ takes the form

$$\sigma_{\hat{A}}^2 = \sum_{j=1}^{L} \frac{A_j^2}{n_j} a_j x_j^{b_j - 2} \qquad (5.5)$$

A cost function that appears more realistic in the case of acquiring and processing (i.e., estimating sampling unit level crop acreages) satellite-based data is the following:

$$C = \sum_{j=1}^{L} n_j \left(C_{Bj} + x_j C_{wj}\right) \qquad (5.6)$$

where n_j and x_j are as described earlier and

C_{Bj} = the cost per sampling unit in stratum j regardless of its size (i.e., overhead costs, etc.)

C_{wj} = the cost per elemental unit (one acre in this study) making up the sampling units in stratum j.

Using the Lagrangian multiplier method to minimize C subject to equation (5.5) holding results in the following values for x_j, n_j; and C_{min}:

$$x_j = \frac{C_{Bj}}{C_{wj}} \left(\frac{1}{b_j - 1} - 1\right) \qquad (5.7)$$

$$n_j = \sqrt{\frac{C_{min}}{\sigma_{\hat{A}}^2} \frac{A_j^2 a_j}{C_{wj}} (2-b_j) \left[\frac{C_{Bj}}{C_{wj}} \left(\frac{1}{b_j-1} - 1\right)\right]^{b_j-3}} \qquad (5.8)$$

$$C_{min} = \frac{1}{\alpha_{\hat{A}}^2} \left[\sum_{j=1}^{L} \frac{C_{Bj} A_j}{(b_j - 1)} \left[\frac{C_{Bj}}{C_{wj}} \left(\frac{1}{b_j-1} - 1\right)\right]^{b_j-3} \sqrt{\frac{a_j(2-b_j)}{C_{wj}}} \right]^2 \qquad (5.9)$$

CLUSTER SIZE IN ACRES	CLUSTER SIZE AS PERCENT OF 5x6 N.MI. SEGMENT	TOTAL ALLOCATION	SAMPLING RATE
25,463	100%	487	3.54%
22,918	90%	501	3.28%
20,371	80%	517	3.01%
17,825	70%	536	2.73%
15,278	60%	559	2.44%
12,732	50%	587	2.14%
10,185	40%	624	1.82%
7,639	30%	674	1.47%
5,092	20%	753	1.10%
2,546	10%	908	.66%
1,019	4%	1,163	.34%
113	.0045%	2,108	.07%
1.13	.000045%	7,325	.002%

Table 5.4: The Estimated Total U.S. Allocation and Sampling Rate as a Function of Sampling Cluster Size

6.0 Summary and Conclusions

Empirical results from remotely-sensed (satellite-acquired) data indicate that the power function (various forms of which were initially, and successfully, utilized by Smith [1938], Jessen [1942], and others [ref. 1, 4, 7, and 12]) is satisfactory in modeling the within-stratum between cluster variance for a surprisingly large range of sampling cluster sizes. This modeled form was then utilized to gain insight into the relationship between the sampling rate and the sampling unit size under two separate cost structures.

Although concern in this paper is devoted entirely to modeling the sampling variance, it is not to be misconceived that measurement error variance is insignificant and, hence, ignored. Further effort is justified (and currently underway) to attempt to model variations due to measurement error. Sufficient information exist from the measurement results obtained using the sampling unit crop area measurement procedure utilized at NASA/JSC (ref. 14) to warrant further investigation into attempting to characterize this variance as a function of sampling unit size also. Until further insight is gained into this relationship, determinations of the optimal sampling unit sizes will continue to be determined primarily from ranges dictated by various engineering and/or other system constraints.

7.0 References

1. Asthana, R. S. (1950). The size of sub-sampling unit in area estimation. Unpublished thesis, Indian Council of Agricultural Research, New Delhi.

2. Benedict, M. R. (1939). Development of Agricultural Statistics in the Bureau of the Census. Journ. of Farm Economics, Vol. 21, pp. 735-60.

3. Chhikara, R. S. and Feiveson, A. H. (1978). Landsat-Based Large Area Crop Acreage Estimation - An Experimental Study. Proceedings of the Survey Research Methods Section of the Annual ASA Meeting, San Diego, Calif.

4. Cochran, W. G. (1942). Sampling theory when the sampling units are of unequal sizes. J. Amer. Statist. Ass., 37, 199-212.

5. Cochran, W. G. (1977). Sampling Techniques, J. Wiley & Sons, Inc., New York (3rd Edition).

6. Ebhling, Walter H. (1939). Why the Government Entered the Field of Crop Reporting and Forecasting. Journ. of Farm Economics, Vol. 21, pp. 718-34.

7. Hansen, M. H. and Hurwitz, W. N. (1942). Relative Efficiencies of Various Sampling Units in Population Inquiries. J. Amer. Statist. Ass., 37, 89-94.

8. Jessen, R. J. (1942). Statistical Investigation of a Sample Survey for Obtaining Farm Facts. Iowa Agric. Exper. Stn. Res. Bull., 304.

9. MacDonald, R. B.; Hall, F. G.; and Erb, R. B. (1975). The Use of Landsat Data in a Large Area Crop Inventory Experiment (LACIE). Proceedings of Second Symposium on Machine Processing of Remotely Sensed Data, C. D. McGillem, ed., IEEE, New York, pp. 1B-1 to 1B-23.

10. MacDonald, R. B.; Hall, F. G.; and Erb, R. B. (1976). The Large Area Crop Inventory Experiment (LACIE), an assessment after one year of operation. Proceedings of 10th International Symposium on Remote Sensing of Environment, Vol. 1, ERIM, Ann Arbor, Mich., pp. 17-37.

11. MacDonald, R. B.; and Hall, F. G. (1977). LACIE: A proof of concept experiment in global crop monitoring. Paper presented at Midcon/77 Electronic Show and Convention, Chicago.

12. Mahalanobis, P. C. (1940). A sample survey of the acreage under jute in Bengal. Sankhya, 4, pp. 511-30.

13. Smith, H. F. (1938), An empirical law describing heterageneity in the yields of agricultural crops. J. Agric. Sci., 28, pp. 1-23.

14. Proceedings of the LACIE Symposium, October 1978, NASA/JSC, Houston, Texas.