AUG 14 1980

AgRISTARS

FC-J0-00468
JSC-16371

81-10035
CR-163552

A Joint Program for Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing

Foreign Commodity Production Forecasting

July 1980

STRATUM VARIANCE ESTIMATION FOR SAMPLE ALLOCATION IN CROP SURVEYS

PREPRINT FOR THE ANNUAL MEETING OF THE AMERICAN STATISTICAL ASSOCIATION
– SURVEY RESEARCH METHODS SECTION –
HOUSTON, TEXAS, 11-14 AUGUST 1980

(E81-10035) STRATUM VARIANCE ESTIMATION FOR SAMPLE ALLOCATION IN CROP SURVEYS (Lockheed Engineering and Management) 20 p HC A02/MF A01 CSCL 02C

N81-12517
Unclas
G3/43 00035

Lyndon B. Johnson Space Center
Houston, Texas 77058

STRATUM VARIANCE ESTIMATION FOR SAMPLE ALLOCATION IN CROP SURVEYS

Charles R. Perry*
NASA - Lyndon B. Johnson Space Center
Houston, Texas 77058

and

Raj S. Chhikara
Lockheed Engineering and Management Services Company, Inc.**
Houston, Texas 77058

ABSTRACT

The problem of determining stratum variances needed in achieving an optimum sample allocation for crop surveys by remote sensing is investigated by considering an approach based on the concept of stratum variance as a function of the sampling unit size. A methodology using the existing and easily available information of historical crop statistics is developed for obtaining initial estimates of stratum variances. The procedure is applied to estimate stratum variances for wheat in the U.S. Great Plains and is evaluated based on the numerical results thus obtained. It is shown that the proposed technique is viable and performs satisfactorily, with the use of a conservative value for the field size and the crop statistics from the small political subdivision level, when the estimated stratum variances were compared to those obtained using the Landsat (land satellite) data.

Keywords: Sample allocation
Initial stratum variance estimation
Sampling unit size
Remote sensing
Survey design
Variance function

---

*National Research Council Senior Resident Research Associate on leave from Texas Lutheran College, Seguin, Texas.

**Under Contract NAS 9-15800 to the National Aeronautics and Space Administration, Lyndon B. Johnson Space Center, Houston, Texas.

1. INTRODUCTION

In any cost-effective stratified sampling design, the optimal sample size and its allocation between the different strata depend on the within-stratum variances, the stratum size, and the precision required for the estimate. With the development of an area sampling frame, strata sizes are known in terms of the total number of sampling units per stratum. The precision goal is fixed in advance and hence known. However, prior to the survey, no direct knowledge of within-stratum variances is available; therefore, it is necessary to estimate them. Usually, a pilot survey is conducted and, subsequently, the information resulting from the pilot study is utilized in planning a full-scale sample survey. In this report, a methodology for indirectly estimating stratum variances using existing agricultural statistics and other ancillary information is proposed and evaluated for wheat in the U.S. Great Plains (USGP).

In most countries, crop statistics are computed annually either through complete enumeration or by employing sample survey methodology. However, the geographical level and the type of crop statistics reported vary considerably from one country to another. For example, reliable crop statistics for area, yield, and production are available in the United States at the county level. In contrast, crop statistics are not available for China at a political subdivision level lower than the country level. Canada, India, and several other countries provide fairly reliable annual crop statistics at a geographic level similar to the U.S. county. Yet, even among these countries, the type of crop statistics produced is varied; for example, in Australia, annual crop statistics contain no information on harvested acreage. Consequently, no fixed procedure can be applied to each and every country for determining the within-stratum variances.

During the first year, little to no previously analyzed Landsat data are available on a crop region for making within-stratum variance estimates; thus, a technique is needed for making initial within-stratum variance estimates without the use of previously analyzed Landsat data. The description and the evaluation of such a technique are presented in this paper. Details of the proposed technique are given in section 2. The technique is motivated by the empirical models employed by Perry and Hallum (ref. 1) in their study on sampling unit size. The technique is designed to make optimal use of the available data (even if limited by its reliability) for estimating within-stratum variances on crop regions that otherwise would not be estimated because previously analyzed Landsat data are not available.

2

2. PRESENT METHODOLOGY

A procedure for indirectly estimating the stratum variances used in an initial allocation is presented. There are three basic underlying ideas. First, obtain estimates of the stratum variance for a set of sampling unit sizes, including both large and small size sampling units; second, establish empirically a relationship between the sampling unit size and the stratum variance; and third, use the empirical model to obtain an estimate of the stratum variance for the desired sampling unit size, which is a segment.

In the context of crop estimation, Smith (ref. 2) and Mahalonobis (ref. 3), independently of each other, thought that the stratum between-units variance could be modeled as a power function of the sampling unit size. A number of empirical studies [Smith, Mahalonobis, Jessen, Hansen et al., and Asthana (refs. 2, 3, 4, 5, and 6, respectively)] strongly indicate that the power function provides a simple, yet satisfactory, mathematical model for the functional dependence of the stratum between-units variance on the sampling unit size. The first application of this functional form specifically to the between-units crop proportion variance was made by P. C. Mahalonobis (ref. 3) in his 1938 study of jute production for Bengal (India). He considered the following function for the stratum between-units crop proportion variance.

$$\sigma_x^2 = \frac{\tilde{p}(1 - \tilde{p})}{(bx)^g} \tag{1}$$

where

$\tilde{p}$ = the stratum crop proportion

x = the sampling unit size

The sample sizes considered in this study were 1, 2.25, 4, 6.25, and 9 acres.

The rationale behind the variance formulation in equation (1) follows. When $x = 1/b$, the variance $\sigma_x^2 = \tilde{p}(1 - \tilde{p})$ and 1/b represent the largest area (e.g., crop field) for which the crop proportion is either 0 or 1. As x increases in size away from 1/b, the denominator in equation (1) increases and σ_x^2

decreases with $\tilde{p}(1 - \tilde{p})$ as an upperbound. If it is assumed that fields in a stratum are not mixed and all its fields are approximately of equal size, the difference between the average field size and the sampling unit size being considered should be indicative of the decrease in σ_x^2 from $\tilde{p}(1 - \tilde{p})$; a smaller decrease in σ_x^2 is expected with a small difference between the sampling unit size and 1/b. Consequently, the bias in estimating σ_x^2 by $\tilde{p}(1 - \tilde{p})$ will be smaller for the sampling unit size closer (on high side) to 1/b, and it is zero when the sampling unit size is less than or equal to 1/b.

This same model was employed by Perry and Hallum (ref. 1) in their sampling unit size study. Their study concluded that indeed the power function does provide a satisfactory model for the between-units wheat acreage (or proportion) variance for sampling unit sizes ranging from 171 to 25 426 acres. Several other studies, particularly those by Jessen (ref. 4) and Asthana (ref. 6), show this general relationship to hold reasonably well even for very large areal units, a county for example.

The relationship in equation (1) can be rewritten as

$$\sigma_x^2 = \alpha x^{\beta} \tag{2}$$

where

x = the sampling unit size
σ_x^2 = the stratum crop proportion variance corresponding to x
and α and β are parameters to be empirically determined for each stratum.

In developing this model for the different strata, it would be ideal to have knowledge of σ_x^2 over a wide range of sampling unit sizes, x. For most countries, this is not feasible because it would require expensive sampling or complete enumeration to be performed, thus defeating the purpose of employing the model in the first place. Therefore, one is led in least-squares estimation of the stratum parameters α and β to choose sampling unit sizes for which σ_x^2 can be estimated directly from existing agricultural statistics or

4

can be mathematically modeled and then estimated from existing agricultural statistics.

In the United States, crop statistics are available at the county level, and a strataum normally consists of many counties. Thus, the between-counties variance can be easily computed and used as an estimate of stratum variance corresponding to a sampling unit approximately equal to the average county size. However, since the counties often vary considerably in size, the stratum variance should vary statistically as the sampling unit size varies from the smallest to the largest county. This statistical variability may be preserved by using a one-point estimate of σ_x^2 for each county in the stratum. The one-point estimates are obtained as follows. Consider the county as a sampling unit

where

x_i = the size of the i^{th} county in a stratum
p_i = the proportion of crop acreage for the i^{th} county in the stratum
$\tilde{p}$ = the proportion of crop acreage in the stratum

Then the squared deviation

$$s_{x_i}^2 = (p_i - \tilde{p})^2 \tag{3}$$

provides an estimate of $\sigma_{x_i}^2$ for the sampling unit size x_i. Although these county-level estimates can be expected to provide guidance in estimating the stratum variance for a sampling unit approximately the size of a county, they alone can not be expected to be sufficient to predict the stratum variance for a sampling unit of the size of a smaller area segment because it will be outside the sampling unit size range for the counties.

The next three estimates are developed for use with small sampling unit sizes. Any one of these estimates along with the one-point variance estimates from equation (3) are used for the least-squares estimation of the parameters α and β. The resulting regression curve is evaluated for the sampling unit size of interest (segment) to obtain the corresponding stratum variance estimate.

Later, it will be observed empirically that the last two relationships provide fairly reliable stratum variance estimates.

First, suppose that all fields are of the same size and shape and the sampling unit is randomly placed with the exception that it intersects only one field. Then the stratum variance corresponding to the field size, x_0, is given by the binomial variance

$$\sigma^2_{x_0} = \pi(1 - \pi) \tag{4}$$

where π is the proportion of the fields belonging to the crop type of interest. For a fixed crop proportion $\tilde{p}$ and a fixed sampling unit size, the between-units variance is maximized when the sampling unit proportions are all either 0 or 1. Thus, equation (4) provides an upperbound of $\tilde{p}(1 - \tilde{p})$ for the stratum variance regardless of the sampling unit size. This feature and the method in general are illustrated in figure 1.

Second, in a Landsat type sampling process, the sampling unit is randomly located and is expected to intersect more than one field. Thus, a closer approximation to $\sigma^2_{x_0}$ than that given in equation (4) is desirable. An exact determination of the variance $\sigma^2_{x_0}$ is not feasible. However, a realistic approximation can be developed under the following assumptions: (1) all fields are square and equal in size to the sampling unit size, x_0, (2) the contents of any four adjacent fields are uncorrelated with respect to the crop of interest, and (3) the sampling unit is randomly placed with the exception that its sides are parallel to the field boundaries. It follows easily as proved by Chhikara and Perry (ref. 7) that

$$\sigma^2_{x_0} = \frac{4}{9}\,\tilde{p}(1 - \tilde{p}) \tag{5}$$

where $\tilde{p}$ is the stratum crop proportion.

6

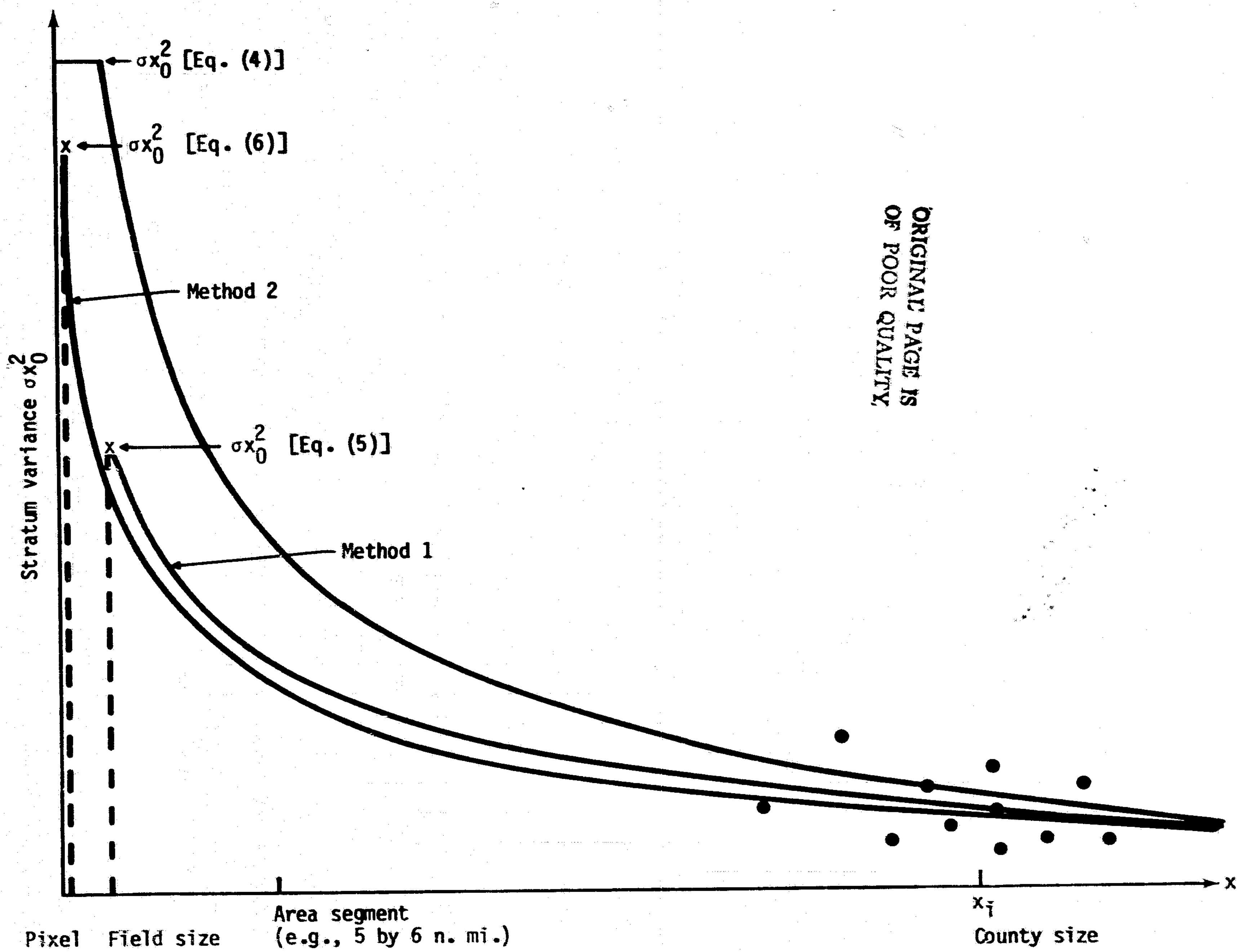

Figure 1.- An illustration of the fitted model.

Third, when the sampling unit size x_0 is small relative to the size of the fields, then it is possible to derive the variance in a somewhat exact form as described in the appendix. In this case, the estimate corresponding to the

small sampling unit x_0, referred to as a pixel, is approximated by the equation

$$\sigma_{x_0}^2 = \alpha_1(1 - \tilde{p})^2 + \alpha_2\tilde{p}^2 + \alpha_3(0.3682 - \tilde{p} + \tilde{p}^2) \qquad (6)$$

where α_1, α_2, and α_3 are defined and evaluated in terms of the crop proportion and the field size distribution.

As outlined earlier, equation (3) combined with any one of the equations (4), (5), or (6) provide stratum-variance estimates over widely separated sampling unit sizes from which the parameters α and β can be determined using a least-squares fit. An estimate of the stratum variance corresponding to a specified sample unit size, x, is then obtained by evaluating along the fitted curve

$$\hat{\sigma}_x^2 = AX^B \qquad (7)$$

where A and B are the least-squares estimates of the parameters α and β.

It will be seen from the numerical results that use of both equations (5) and (6) lead to fairly reliable variance estimates. Yet, equation (5) is probably preferable if accurate determination of the field sizes can be made or if the field sizes are large. Otherwise, it is probably better to use equation (6) since it is less sensitive to error in the field size measurements.

8

3. VARIANCE ESTIMATION FOR WHEAT IN THE U.S. GREAT PLAINS

The methodology of the previous section was applied to estimate stratum variances for wheat in the USGP. Two estimation methods were created by considering the county size units with the field size unit in one case (method 1), and the county size units with the smaller size unit in the other case (method 2). The variance inputs for the least-square fit in equation (7) were obtained from equation (3) and that given by equation (5) or (6) as applicable.

Although a third method of estimation is possible by using results from equation (3) with that from equation (4), it was not considered because of the unrealistic basis of equation (4). The fitted curve was forced through the point $(x_0, \sigma^2_{x_0})$ since it acts as an intercept and is the single most influential point. Thus, the A in equation (7) was replaced by $\sigma^2_{x_0}/x_0^B$, and the least square estimate of B was obtained by minimizing the sum of squared deviations of variances given by the model from those resulting from the use of equation (3) for all counties in a stratum.

The USGP region initially was stratified into 27 agrophysical units (APU). This stratification was further refined by intersecting the APU with the state boundaries to account for the state difference. For each refined stratum, the counties, their sizes (measured in terms of 5- by 6-nautical-mile area segments over the agricultural land), and the wheat proportions were determined for obtaining input to equation (3). The wheat acreages given in the 1974 Agricultural Census Report were used in computing the wheat proportions. The average field size, the proportion of wheat acreage, and the between-county variances were computed for each stratum. The stratum-level data are given in table 1.

The average field size (more precisely, the distribution of field size) varies from strata to strata and was difficult to determine. The following technique, employing the 1974 Agriculture Census Report data, was used to estimate the average field size for a given stratum. Suppose N_i and A_i, respectively,

9

TABLE 1.- REFINED STRATA DATA INPUT FOR VARIANCE ESTIMATION FOR WHEAT IN THE USGP

State	Refined stratum	Number of counties	Number of agricultural segments	Average field size in acres	Proportion of wheat acreage	Between-county standard deviation
Colorado	9	3	150	450	0.16	0.020
	10	20	558	345	.13	.088
	101	21	227	126	.03	.031
Kansas	7	10	226	276	.39	.121
	8	8	179	288	.30	.061
	9	13	258	460	.25	.049
	11	18	409	239	.21	.040
	12	17	311	152	.22	.107
	13	18	271	57	.07	.032
	14	11	161	52	.07	.033
	15	2	37	173	.29	.120
	60	3	75	390	.20	.033
	102	4	74	73	.04	.007
Minnesota	15	15	238	34	.02	.019
	19	16	317	60	.06	.053
	20	13	308	189	.23	.090
Montana	21	3	141	502	.23	.045
	22	6	212	363	.11	.035
	23	13	662	490	.15	.067
	104	32	503	213	.04	.030
Nebraska	10	9	203	340	.18	.118
	11	15	297	131	.09	.042
	14	9	137	47	.08	.029
	15	44	651	56	.04	.051
	16	4	114	64	.00	.002
	17	3	89	189	.09	.067
	103	7	0	83	.00	.001
North Dakota	19	20	582	292	.28	.055
	20	7	214	268	.34	.041
	21	24	831	259	.19	.069
	22	2	30	263	.14	.097
Oklahoma	3	5	42	93	.06	.041
	7	22	401	232	.37	.151
	9	2	84	380	.19	.063
	13	3	23	69	.07	.058
	60	11	219	250	.22	.058
	102	26	131	75	.02	.021
South Dakota	15	7	99	44	.01	.007
	16	22	441	186	.06	.068
	17	10	358	352	.07	.037
	18	5	204	249	.05	.014
	19	12	283	139	.14	.060
	21	6	197	208	.09	.030
	104	5	89	179	.03	.012
Texas	2	13	230	84	.03	.032
	3	28	458	105	.04	.035
	4	23	525	170	.06	.066
	5	12	153	201	.12	.088
	9	7	161	476	.18	.087
	60	5	55	385	.15	.074
	61	13	219	216	.07	.079
	101	28	228	89	.01	.009
	102	26	290	76	.01	.013

10

are the number of operators and the 1974 crop acreage for the i^{th} crop in a stratum. Then, average field size, f_0, for the stratum is estimated by

$$\hat{f}_0 = \left[\sum_{i=1}^{k} A_i \bigg/ \sum_{i=1}^{k} N_i\right] \tag{8}$$

where k is the number of major crops in the stratum. The field size estimates resulting from this computation are listed in table 1.

Listed in table 2 are individual stratum standard deviation estimates obtained for the sampling unit size of 5- by 6-nautical-mile area using each method. The coefficient values of A and B are also given. The comparison between the two sets of estimates shows that with only four exceptions the method 1 stratum-variance estimates are larger. This result is expected of the methodology, as depicted in figure 1. In addition, an examination of A and B values across the strata suggests that A is significantly influenced by the stratum crop proportion and B is highly dependent upon the between-county variance. (See table 1 for information on the stratum crop proportion and the between-county variance.) This indicates that there is a positive correlation between the crop proportion and the value of A, as well as between the value of B and the between-county variance. The correlation is exhibited more in the case of method 2 than in the other method.

It should be noted that the parameter B takes on values between -1 and 0. When the largest area with crop proportion near 0 or 1 is considered for the sampling unit, the intraclass correlation is near 1, and the stratum variance is close to the binomial form and almost equal to A; therefore, $B \doteq 0$. On the other hand, if the sampling unit is chosen to be a large cluster made of randomly selected elements, the interclass correlation is zero and the stratum variance is equal to A/x, where x is the sampling unit size; therefore, $B \doteq -1$. An intuitive understanding of the observed dependence of B on the between-county variance component follows. Because a larger between-county variance component is indicative of a possible smaller within-county variance component and, thus, a lower intraclass correlation, it follows that a smaller value for B may be expected when the between-county variance is small.

TABLE 2.- WITHIN-STRATUM VARIANCE ESTIMATES FOR METHODS 1 AND 2

State	Refined stratum	Method 1			Method 2		
		A	B	Standard deviation estimate	A	B	Standard deviation estimate
Colorado	9	1.716	-0.572	0.074	0.127	-0.447	0.038
	10	.242	-.269	.127	.108	-.204	.118
	101	.058	-.355	.041	.023	-2.73	.039
Kansas	7	.289	-.182	.216	.221	-.215	.160
	8	1.124	-.447	.113	.197	-.313	.092
	9	1.825	-.512	.103	.182	-.337	.078
	11	.888	-.456	.095	.157	-.353	.068
	12	.222	-.211	.164	.162	-.210	.141
	13	.109	-.343	.059	.058	-.320	.048
	14	.124	-.381	.052	.061	-.328	.048
	15	.684	-.403	.109	.189	-.253	.122
	60	1.881	-.563	.081	.155	-.408	.051
	102	.204	-.620	.020	.034	-.527	.013
Minnesota	15	.035	-.371	.029	.022	-.332	.028
	19	.082	-.293	.066	.054	-.233	.073
	20	.375	-.306	.132	.166	-.239	.122
Montana	21	2.485	-.565	.093	.172	-.351	.071
	22	.994	-.533	.069	.098	-.335	.058
	23	.532	-.365	.117	.125	-.248	.102
	104	.125	-.397	.048	.034	-.287	.044
Nebraska	10	.230	-.221	.158	.144	-.187	.148
	11	.133	-.344	.076	.076	-.297	.062
	14	.179	-.454	.043	.068	-.362	.042
	15	.043	-.225	.067	.038	-.213	.067
	16	.016	-.623	.005	.003	-.473	.005
	17	.220	-.344	.084	.079	-.242	.083
	103	.018	-.865	.002	.001	-.614	.001
North Dakota	19	.777	-.389	.125	.190	-.313	.090
	20	1.238	-.459	.111	.210	-.373	.070
	21	.402	-.328	.122	.147	-.258	.105
	22	.285	-.306	.115	.112	-.248	.096
Oklahoma	3	.166	-.427	.048	.057	-.321	.047
	7	.325	-.216	.193	.216	-.178	.191
	9	.702	-.392	.117	.150	-.312	.081
	13	.084	-.291	.067	.057	-.270	.062
	60	.647	-.389	.114	.162	-.307	.086
	102	.073	-.478	.024	.022	-.343	.026
South Dakota	15	.024	-.481	.014	.009	-.436	.011
	16	.097	-.254	.087	.058	-.199	.089
	17	.370	-.453	.063	.060	-.296	.056
	18	.441	-.578	.036	.042	-.420	.025
	19	.258	-.324	.100	.115	-.270	.087
	21	.380	-.426	.073	.080	-.340	.051
	104	.430	-.679	.022	.031	-.468	.017
Texas	2	.054	-.327	.045	.028	-.261	.045
	3	.058	-.291	.056	.033	-.264	.048
	4	.071	-.203	.096	.055	-.196	.088
	5	.191	-.275	.110	.101	-.219	.106
	9	.321	-.269	.147	.140	-.237	.113
	60	.558	-.396	.102	.121	-.272	.089
	61	.068	-.143	.127	.060	-.183	.098
	101	.030	-.484	.015	.007	.380	.013
	102	.029	-.414	.021	.011	-.345	.019

The stratum-variance estimates given in table 2 were compared with the within-stratum variances computed from Landsat estimates of wheat proportions of randomly selected 5- by 6-nautical-mile area segments in each stratum. Only refined strata with two or more sample segments were considered.

Suppose S_{jk} is the estimated standard deviation for the j^{th} stratum using the k^{th} method, and σ_j is the sample-based standard deviation estimate for the j^{th} stratum. Consider the set of differences, $\{(S_{jk} - \sigma_j)\}$, for each method. The mean and variance of each set of differences were computed. Assuming the difference to be an estimate of the error in estimating the within-stratum variance by a method, then they (i.e., mean and variance for the difference) provide an estimate of the possible bias and the variance expected in estimating a stratum variance using this method. Listed in table 3 are the estimated bias and variance for each method.

The results in table 3 show that more accurate stratum-variance estimates were obtained using method 2. This result is somewhat surprising because the use of field size unit is more appropriate than the smaller size unit unless the spatial distribution of a crop is not influenced by the average field size. Moreover, the poorer performance by method 1 may have been due to its sensitivity to the field size which was crudely estimated for each stratum using equation (8). In fact, the field size estimates computed from the ratio of crop acreages to farm operators were on the average four times larger than field size estimates computed from a limited set of ground truth given by Pitts and Badhwar (ref. 8). Note that a farm operator (accounted for by crop type) may have more than one field of a given crop type, hence, the average field size can be expected to be smaller than the value estimated using equation (8). The numerical results tend to confirm this.

Regardless of the method used, the stratum field sizes must be determined and the best possible information should be used for the evaluation. If data on crop statistics and cropping practices from which the field size, f_0, can be estimated are unavailable, then Landsat imagery can be employed to obtain an estimate of average field size for a stratum.

TABLE 3.- THE ESTIMATED BIAS AND VARIANCES IN ESTIMATING STRATA VARIANCES

Method	Bias estimate, average difference	Estimated variance of the difference
1	[a]0.0110	0.00109
2	.0013	.00123

[a]Significant against the 5-percent level t-test.

14

4. CONCLUSION AND SUMMARY

The present study proposes a new method to obtain initial variance estimates for sample allocations in designing crop surveys. The approach is to develop empirically a relationship between the stratum variance and the sampling unit size.

A procedure is devised that uses existing and easily available information of historical crop statistics in developing this relationship. Consideration is given to the field size in order to effect a modification in stratum variance that is necessary for small sampling unit sizes.

The numerical results tend to show that methods 1 and 2 perform about equally well and that either method produces realistic stratum variance estimates, given reliable input data. However, method 1 is more sensitive to the field size variable and should be used if accurate field size determinations can be made. Otherwise method 2 is preferable.

In summary, the study suggests that (1) the technique is viable, (2) care should be exercised to ensure the reliability of the input data, and (3) the field sizes must be realistically estimated either from historical statistics or Landsat imagery.

6. REFERENCES

1. Perry, C. R. and Hallum, C. R.: LACIE Sampling Unit Size Considerations in Large Area Crop Inventorying, Using Satellite-Based Data. Proceedings of the Annual Meeting of the American Statistical Association, Washington, D.C., August, 13-16, 1979.

2. Smith, H. F.: An Empirical Law Describing Heterogeneity in the Yields of Agriculture Crops. Journal of Agricultural Science, vol. 28, 1938, pp. 1-23.

3. Mahalonobis, P. C.: A Sample Survey of the Acreage Under Jute in Bengal. Sankhya (New Delhi, India), vol. 4, 1940, pp. 511-530.

4. Jessen, R. J.: Statistical Investigation of a Sample Survey for Obtaining Farm Facts. Iowa Agricultural Experimental Station, Research Bulletin 304, 1942.

5. Hansen, M. H. and Hurwitz, W. N.: Relative Efficiencies of Various Sampling Units in Population Inquiries. Journal of American Statistics, no. 37, 1942, pp. 89-94.

6. Asthana, R. S. The Size of Sub-Sampling Unit in Area Estimation. Indian Council of Agricultural Research (New Delhi, India), 1950 (unpublished thesis).

7. Chhikara, R. S.; and Perry, C. R.: Estimation of Within-Stratum Variance For Sample Allocation. NASA Technical Report, JSC-16343 (to be published).

8. Pitts, D. E. and Badhwar, Gautam: Field Size, Length, and Width Distributions Based on LACIE Ground-Truth Data. Submitted to Remote Sensing of Environment, August, 1979.

APPENDIX

Developed in this appendix is a statistical model for the within-stratum variance for sampling units which are very small relative to the field size of the crop of interest. Crop X will refer to the crop of interest. The model is developed using the definitions and assumptions of the following conceptual experiment.

A square area unit with diagonal 2d is randomly selected from the area of a stratum having a proportion p for crop X. A random variable P is defined over the sample space of the experiment as follows. P has value p if the randomly selected square has proportion p for crop X. Probabilities α_1, α_2, and α_3 are associated, respectively, with the following events: the square selected is pure and contains only crop X; the square selected is pure and does not contain crop X; and the square selected is mixed. With this notation, it is observed that

$$\alpha_1 = \mathrm{Prob}(P = 1)$$

$$\alpha_2 = \mathrm{Prob}(P = 0)$$

$$\alpha_3 = \mathrm{Prob}(0 < P < 1)$$

$$\alpha_1 + \alpha_2 + \alpha_3 = 1$$

$$E(P) = \tilde{p}$$

$$\mathrm{Var}(P) = \alpha_1(1 - \tilde{p})^2 + \alpha_2\tilde{p}^2 + \alpha_3 E_{P|0<P<1}(P - \tilde{p})^2$$

where the expectation in the last equation is understood to be taken over the collection corresponding to the mixed squares. Tractable analytic expressions for the probabilities α_1, α_2, and α_3 and the expected value $E_{P|0<P<1}(P - \tilde{p})^2$ in terms of the stratum field size distribution and the crop proportion, $\tilde{p}$, for crop X were derived in Chhikara and Perry (ref. 7).

It was shown that the following expression provides a good approximation of Var(P).

$$\mathrm{Var}(P) \doteq \alpha_1(1 - \tilde{p})^2 + \alpha_2\tilde{p}^2 + \alpha_3(0.3682 - \tilde{p} + \tilde{p}^2)$$

where

$$\alpha_1 = \frac{1}{A}\left[\sum_{i=1}^{N}\left(\frac{f_i\tilde{p}A}{\ell_i w_i}\right)(\ell_i - b)(w_i - b)\right]$$

$$= \tilde{p}\sum_{i=1}^{N} f_i \frac{(\ell_i - b)(w_i - b)}{\ell_i w_i}$$

$$\alpha_3 = \frac{1}{A}\left\{\sum_{i=1}^{N}\left(\frac{f_i\tilde{p}A}{\ell_i w_i}\right)\left[(\ell_i + b)(w_i + b) - (\ell_i - b)(w_i - b)\right]\right\}$$

$$= \tilde{p}\sum_{i=1}^{N} \frac{2bf_i(w_i + \ell_i)}{w_i\ell_i}$$

$$\alpha_2 = 1 - \alpha_1 - \alpha_3$$

and

f_i = frequency of fields with length ℓ_i and width w_i

A = stratum size

NASA-JSC

18