#22

Journal of Statistical Planning and Inference 15 (1986) 97–114
North-Holland

ESTIMATION OF STRATUM VARIANCES IN PLANNING OF CROP ACREAGE SURVEYS

Raj S. CHHIKARA

University of Houston-Clear Lake, 2700 Bay Area Boulevard, Houston, TX 77058, USA

Charles R. PERRY, Jr.

U.S. Department of Agriculture, Sample Frame Development Section, 3251 Old Lee Hwy., Fairfax, VA 22030, USA

Received September 1983; revised manuscript received February 1985
Recommended by D. Singh

Abstract: A modelling approach is used to obtain initial estimates of stratum crop acreage variances for designing crop surveys, particularly those using the remote sensing technology. The present methodology is developed based on the concept of stratum variance as a function of the sampling unit size and it uses primarily the historical crop statistics which are commonly available in most countries. Methods are proposed for the determination of stratum variances corresponding to unit sizes different from the sampling unit size and for which the historical crop statistics can be used. The methodology is applied to estimate stratum variances for wheat in the U.S. Great Plains. An evaluation of these estimates made by comparing them to those obtained from certain satellite sample data shows that the proposed method leads to reliable stratum variance estimates for a fairly large size (5×6 nautical miles area segment) sampling unit.

AMS Subject Classification: 62D05.

Key words: Variance function; Sampling unit size; Empirical modeling; LANDSAT; Crop statistics; U.S. Great Plains.

1. Introduction

The availability of remotely sensed data from satelite (LANDSAT) has made it possible to conduct crop surveys on a global basis. MacDonald and Hall (1980) discuss the experimental study of global crop forecasting conducted at the Johnson Space Center, Houston, and argue that more timely and perhaps better crop estimates can be made through the use of satellite data. Chhikara and Feiveson (1982) describe the sampling design used in this study for estimating large area wheat acreages based on LANDSAT data alone. Craig et al. (1978) utilize LANDSAT data as auxiliary information to obtain better winter wheat acreage estimates for the individual counties in Kansas. Since low to high intensely cultivated areas in a region

0378-3758/86/$3.50 © 1986, Elsevier Science Publishers B.V. (North-Holland)

can be easily delineated using LANDSAT imagery, the capability of developing an efficient stratification for crop surveys has been greatly enhanced by this new source of information. Houston and Hall (1984) describe in details the LANDSAT data and its use in agricultural surveys. A brief description of the background of LANDSAT data is given in the appendix. For a comprehensive description of the problem and the statistical methodology of crop surveys using satellite data, refer to the articles in Chhikara (1984).

Presently we consider the problem of estimating stratum variances for the proportion of acreages devoted to a certain crop when the sampling unit is of fairly large size. This problem first arose in the planning of a global wheat acreage survey with sampling unit of 5×6 nautical miles area segment (Chhikara and Feiveson, 1982). The stratum variances were estimated by assuming the binomial model where a sample unit has either all wheat or no wheat in it. Since this assumption did not hold for such large size units, stratum variances were grossly overestimated.

In this paper we investigate the variance estimation problem by assuming stratum variance as a power function of the sampling unit size. A number of empirical studies conducted in past, notably by Smith (1938) and Mahalanobis (1940), have strongly indicated 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. Mahalanobis (1968) in his 1938 survey of Jute acreage in Bengal, India, considered the following power function for the stratum crop proportion variance:

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

where p is the stratum crop proportion, x is the sampling unit size, and b and g are constants. The rationale behind the variance formula in (1) is as follows: when $x = 1/b$, the variance $\sigma_x^2 = p(1-p)$ and $1/b$ represents 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 (1) increases and σ_x^2 decreases with $p(1-p)$ as an upper bound. Thus, the bias in estimating σ_x^2 by $p(1-p)$ will depend upon how much larger the sampling unit size is from $1/b$.

The function form in (1) can simply be written as

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

where α and β are parameters to be estimated empirically. Though some rationale and certain empirical evidence exist to support this functional form for the stratum variance, no theoretical justification can be given. Cochran (1963, Ch. 9) warns agains a blind application of this model and quite correctly points out the need to test it thoroughly before using. In the present context of a LANDSAT crop survey of a large geographical area, Perry and Hallum (1979) showed that the function in (2) does provide a satisfactory model for the between-units wheat acreage variance

for sampling units ranging from 171 to 25 426 acres. For very large areal units, this type of relationship also seems to hold reasonably well as it is seen from the empirical study of Asthana (1950) for crop acreage estimation. Jessen (1942) finds this function more suitable for modeling the within sampling unit variance. Hendrick (1944) evaluates this approach for an estimation of the variance of sample mean for grouped sampling units.

In this paper we consider a sampling unit of moderate to fairly large size and discuss an approach to estimation of α and β using the available historical crop statistics. Given in Section 2 are certain methods of estimation of the stratum variance for sampling units of various sizes (assumning that the stratum contains units of equal or roughly equal size) and then empirically fitting the assumed power function to the estimated variance values. In order to achieve this fitting, a range of points is necessary to 'bracket' the desired size for the sampling unit. Point values for larger size units are obtained by considering small political subdivisions (SPD) of various sizes in the stratum; but a problem of particular interest, and a critical issue in the present paper, is the estimation of the stratum variance for a sampling unit of size smaller than the desired size, for example, the 5×6 nautical miles area segment. It is desirable to have an adequate spread of the unit sizes for fitting the power function so that one may with confidence use the fitted curve to impute stratum variance for a unit size much smaller than the 5×6 nautical miles unit used stratum variance for a unit size such smaller than the 5×6 nautical mildes unit used in LANDSAT crop acreage surveys. Two different choices for this unit size are considered and thus, two estimation methods are developed.

The present methodology is applied to obtain empirical model fits and then, to impute stratum variances for the wheat acreage proprotion in the U.S. Great Plains (USGP) region for the sampling unit of size 5×6 nautical miles. These variance estimates are compared to those obtained from a previously collected sample survey data and the results are given in Section 3.

2. Empirical modeling of stratum variance

Suppose a stratum consists of several SPD's for which historical crop data are available and that these are of various sizes. Assume that the stratum contains units of size equal to the size of a SPD. Then each SPD may be taken as a sample of one from a hypothetical stratum made up exclusively of units of that size. Thus, the squared deviation corresponding to the i-th SPD of size x_i and crop acreage proportion p_i,

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

where p is the stratum crop acreage proportion, is a reasonable, if somewhat *adhoc*, estimate of the stratum variance, $\sigma_{x_i}^2$. The crop acreage proportion may be based on the most recent historical data. Since the between-years variance for the stratum

crop acreage proportion is often much smaller than its between-units variance in a year, one can ignore the former and use historical crop acreages to obtain the estimate in (3). However, one may improve upon this estimate by using averages from more than one year historical data for p_i and p. Another way to improve is to first group the SPD's where x_i varies minimallly within each group of SPD's, and then estimate the stratum variance by combining the within-group and the between-group variances as following: suppose a stratum consists of L SPD's which form k groups. Let $\bar{x}_i$ be the average size, $S_{\bar{x}_i}^2$ be the within-group variance and M_i be the number of SPD's in the i-th group. If S_B^2 is the between-group variance for the stratum, then $S_{\bar{x}_i}^2 + M_i S_B^2$ can be used for an estimate of $\sigma_{\bar{x}_i}^2$, $i = 1, 2, \ldots, k$. But this approach requires at least three groups of SPD's in a stratum to estimate the parameters and thus, cannot be adopted when $k < 3$. For example, in the application discussed in the next section, there are several strata consisting of only a few SPD's, each resulting in a single group. Thus we have used the expression in (3) to estimate $\sigma_{x_i}^2$ for our model-fits.

Suppose x_0 is the unit size that is smaller than the desired size for the sampling unit. A natural choice for x_0 is either the field size or the measurement unit size. For the LANDSAT observations, the measurement unit is a rectangular area of 1.1 acre in size and is called *pixel*. For each of these two choices for x_0, we now obtain an approximation to the stratum variance, $\sigma_{x_0}^2$.

2.1. Stratum variance for field size unit

If all fields are of the same size and shape and the sampling unit is randomly placed such that it intersects only one field, then the stratum variance corresponding to the field size unit x_0 is given by the binomial variance as discussed in Section 1. However, in a LANDSAT type area sampling, the sample unit is randomly located and is expected to intersect more than one field. Thus, a closer approximation to $\sigma_{x_0}^2$ than the binomial variance is desirable. Below we develop one such approximation. We assume that (1) a stratum is divided into square units, each equal to the average field size, say x_0, (2) the contents of adjacent units are uncorrelated with respect to the crop of interest, and (3) the sample unit is randomly placed with its boundaries aligned with the square unit grid coordinates. Under these assumptions, a randomly placed sample unit consists of areas from at most four adjacent units as shown in Figure 1 and its acreage devoted to the specific crop of interest can be obtained as follows:

Let X and Y be the intercept lengths of a randomly placed unit on the grid with respect to the top right square unit. Then X and Y are two independent uniformly distributed random variables over [0, 1]. So the crop acreage of the sample unit can be expressed as

$$A = \sum_{i=1}^{4} a_i A_i$$

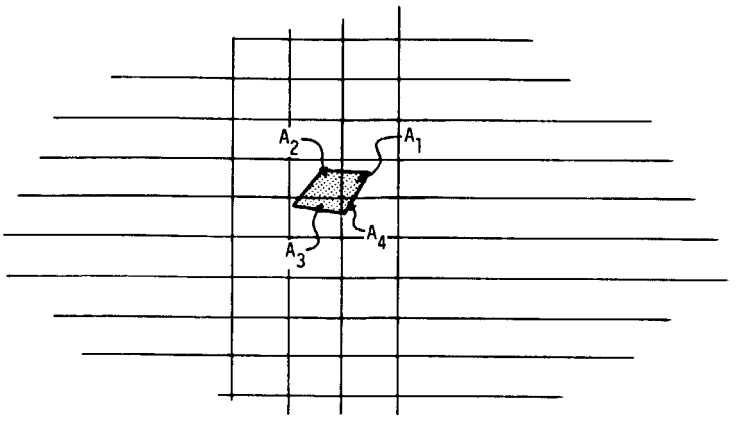

Fig. 1. An illustration of a randomly placed sample unit.

where $A_1 = XY$, $A_2 = (1-X)Y$, $A_3 = (1-X)(1-Y)$, $A_4 = X(1-Y)$ and the a_i are the Bernoulli variables, each taking value of 1 with probability of a square unit belonging to the crop of interest, say P, and value of 0 with probability $1-P$. Clearly, $E[A] = P$ and

$$\begin{aligned}\mathrm{Var}(A) &= E\left[\mathrm{Var}\left(\sum_i a_i A_i \,|\, A_i, i=1,2,3,4\right)\right] \\ &\quad + \mathrm{Var}\left[E\left(\sum_i a_i A_i \,|\, A_i, i=1,2,3,4\right)\right] \\ &= E\left[\sum_i A_i^2 \,\mathrm{Var}(a_i)\right] + \mathrm{Var}\left[\sum_i A_i E(a_i)\right] \\ &= P(1-P)\sum_i E(A_i^2) + P^2\,\mathrm{Var}\left(\sum_i A_i\right).\end{aligned}$$

Due to symmetry, $\sum E[A_i^2] = 4E[A_1^2]$. Since $\sum A_i = 1$, $\mathrm{Var}(\sum A_i) = 0$. Hence

$$\mathrm{Var}(A) = 4P(1-P)E[A_1^2] = \tfrac{4}{9}P(1-P).$$

The last expression is obtained because

$$E[A_1^2] = E[X^2Y^2] = E[X^2]E[Y^2] = (\tfrac{1}{3})^2.$$

Since P is the probability that a unit belongs to the crop of interest and all units are assumed equal, P is equal to the stratum crop proportion, p. Thus, an approximation of $\sigma_{x_0}^2$ is given by

$$\sigma_{x_0}^2 = \tfrac{4}{9}p(1-p). \tag{4}$$

2.2. Stratum variance for pixel size unit

When x_0 is equal to the size of measurement unit, referred to as a pixel, the assumptions made to derive the approximation in (4) become unnecessary and so the variance $\sigma^2_{x_0}$ can be derived in a somewhat exact form as discussed in the appendix. In this case, the stratum variance is approximated by the following expression:

$$\sigma^2_{x_0} = \alpha_1(1-p)^2 + \alpha_2 p^2 + \alpha_3(0.3682 - p + p^2) \tag{5}$$

where α_1, α_2, and α_3 are defined and expressed in terms of the stratum crop proportion p and the stratum field size distribution. See the appendix for details.

2.3. The model fit

The least squares estimates of parameters α and β of the model in (2) are easily obtained for a stratum by using the variance estimates obtained from (3) and one of the equations (4) or (5). Since the model fit is highly influenced by the choice of x_0 and a close approximation of $\sigma^2_{x_0}$ is obtained, particularly using equation (5), we assume that the curve $\sigma^2_x = \alpha x^\beta$ passes through the point $(x_0, \sigma^2_{x_0})$. Thus

$$\alpha = \sigma^2_{x_0}/x_0^\beta \tag{6}$$

and the model in (2) reduces to

$$\sigma^2_x = \left(\frac{x}{x_0}\right)^\beta \sigma^2_{x_0} \tag{7}$$

which involves a single parameter, β.

One may obviously consider transforming (7) in the logarithmic form,

$$\ln \sigma^2_x = \ln \sigma^2_{x_0} + \beta(\ln x - \ln x_0), \tag{8}$$

so that the parameter β is estimated by a linear least squares fit rather than a non-linear least squares fit-method. In our application, the model fit in logarithmic form resulted in a substantially higher residual mean square error as compared to that in the original form (7) [Chhikara and Perry (1980)]. Hendrick (1944) also obtained similar results. A possible reason may be that the error component of the model in (7) is not necessarily multiplicative. Some other possible explanations are given in Smith (1938) and Hendrick (1944). Hence, the equation in (8) is not considered for the model fit.

Suppose B denotes the non-linear least squares estimate of parameter β in a model fit of (7). Then the power curve

$$\hat{\sigma}^2_x = Ax^B, \tag{9}$$

where

$$A = \sigma^2_{x_0}/x_0^B,$$

provides an empirical model for the stratum variance.

3. Stratum variance estimates for wheat in U.S. Great Plains

3.1. Stratification and historical crop data

During the crop year 1978, a sample survey of wheat acreage in the U.S. Great Plains (USGP) was carried out using LANDSAT data at the Johnson Space Center, Houston. The USGP region initially was stratified into 27 agrophysical units (APU) according to agronomical and meteorological considerations. Nonagricultural land was excluded from the area frame using LANDSAT imagery. This stratification was further refined by intersecting the APU with the state boundaries to account for the state difference. For each refiend stratum, the counties, their sizes (measured in terms of 5×6 nautical mile area segments over the agricultural land), and the wheat proportions were determined. The wheat acreages given in the 1974 Agricultural Census Report (U.S. Bureau of Census (1977)) were used in computing the wheat proportions since these acreages were more accurate and consistent at the county level than the 1977 estimates by the Statistical Reporting Service of the U.S. Department of Agriculture. The crop field size information was not available and the average field size for a stratum was computed by the ratio of its total crop land divided by the total number of farm operators in the stratum. In specific, the field size was estimated by the ratio

$$f = \sum_{1}^{k} A_i \Big/ \sum_{1}^{k} N_i \tag{10}$$

where N_i and A_i were the number of farm operators and the crop acreage respectively, for crop type i and k was the number of major crop types in the stratum. Next, the proportion of wheat acreage and the between-county variance for this proportion were computed for each stratum. Table 1 gives these data for all refined strata in USGP.

It is seen from Table 1 that the number of counties per stratum varies considerably across the strata ranging from 2 to 44. So are the other stratum characteristics, the size (given by the number of agricultural segments), average field size, wheat acreage proportion and between-county standard deviation. Because of considerable variability in strata characteristics there is no consistent trend of an increasing standard deviation with an increase in the wheat acreage proportion, contrary to what one may expect. In two cases (strata 16 and 103 in Nebraska) the listed proportion of 0 is due to its rounding off to two decimal places. Though similar data for the counties in the region are not shown here, we observed that the county size varied between 1 and 87 agricultural segments and the county wheat proportion between 0 and 0.6.

It may be mentioned that the ratio in (10) provided a crude estimate of the average field size for a stratum. These field size estimates were on the average much larger than the estimates computed from a limited set of ground data reported by Pitts and Badhwar (1980). One reason may be that a farm operator accounted for by crop

type, may have more than one field of a crop type and thus, the ratio in (10) would overestimate the crop field size.

3.2. Curve fitting

For each stratum, estimated values of stratum variance for different county size units were computed from Equation (3). Two separate curve fits were made for the model given in (7). In one case the average field size given in Table 1 was used for the smaller unit x_0 and (4) was used for determining $\sigma^2_{x_0}$. This will be called

Table 1
Wheat proportions and other data for refined strata in U.S. Great Plains

State	Refined stratum[a]	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	0.13	0.088
	101	21	227	126	0.03	0.031
Kansas	7	10	226	276	0.39	0.121
	8	8	179	288	0.30	0.061
	9	13	258	460	0.25	0.049
	11	18	409	239	0.21	0.040
	12	17	311	152	0.22	0.107
	13	18	271	57	0.07	0.032
	14	11	161	52	0.07	0.033
	15	2	37	173	0.29	0.120
	60	3	75	390	0.20	0.033
	102	4	74	73	0.04	0.007
Minnesota	15	15	238	34	0.02	0.019
	19	16	317	60	0.06	0.053
	20	13	308	189	0.23	0.090
Montana	21	3	141	502	0.23	0.045
	22	6	212	363	0.11	0.035
	23	13	662	490	0.15	0.067
	104	32	503	213	0.04	0.030
Nebraska	10	9	203	340	0.18	0.118
	11	15	297	131	0.09	0.042
	14	9	137	47	0.08	0.029
	15	44	651	56	0.04	0.051
	16	4	114	64	0.00	0.003
	17	3	89	189	0.09	0.067
	103	7	115	83	0.00	0.001
North	19	20	582	292	0.28	0.055
Dakota	20	7	214	268	0.34	0.041
	21	24	831	259	0.19	0.069
	22	2	30	263	0.14	0.097

Table 1 (continued)

State	Refined stratum[a]	Number of counties	Number of agricultural segments	Average field size in acres	Proportion of wheat acreage	Between-county standard deviation
Oklahoma	3	5	42	93	0.06	0.041
	7	22	401	232	0.37	0.151
	9	2	84	380	0.19	0.063
	13	3	23	69	0.07	0.058
	60	11	219	250	0.22	0.058
	102	26	131	75	0.02	0.021
South	15	7	99	44	0.01	0.007
Dakota	16	22	441	186	0.06	0.068
	17	10	358	352	0.06	0.037
	18	5	204	249	0.05	0.014
	19	12	283	139	0.14	0.060
	21	6	197	208	0.09	0.030
	104	5	89	179	0.03	0.012
Texas	2	13	230	84	0.03	0.032
	3	28	458	105	0.04	0.035
	4	23	525	170	0.06	0.066
	5	12	153	201	0.12	0.088
	9	7	161	476	0.18	0.087
	60	5	55	385	0.15	0.074
	61	13	219	216	0.07	0.079
	101	28	228	89	0.01	0.009
	102	26	290	76	0.01	0.013

[a] The numbers assigned to the APU are being used for refined strata in a state.

Method 1. In the second case x_0 was taken to be a pixel and Equation (5) for determining $\sigma^2_{x_0}$. This will be called *Method 2.* The field size and other units were converted in terms of pixels. The stratum proportion of wheat acreage given in Table 1 was used for computing the two values of $\sigma^2_{x_0}$ in Equations (4) and (5). A nonlinear least squares estimation of parameter β was considered to obtain A and B of the model fit in (9).

Figure 2 shows the actual model fits and the estimated values of stratum variances obtained from Equation (3) for a representative refined stratum. The two variance curves cut across at two points, first at a point very close to the field size and then at a point which is to the right of a single segment size.[1] The crossing of the two curves corresponding to the field size reflects the consistency between the two approximations of $\sigma^2_{x_0}$ discussed in the paper. However, the two curves initially are quite different; the curve obtained under Method 2 declines much faster than does the curve obtained under Method 1. Both curves decline very slowly after about half the size of a segment for the sampling unit, and they almost are parallel or identical

[1] The horizontal scale is the number of segments, each of size 5×6 nautical miles.

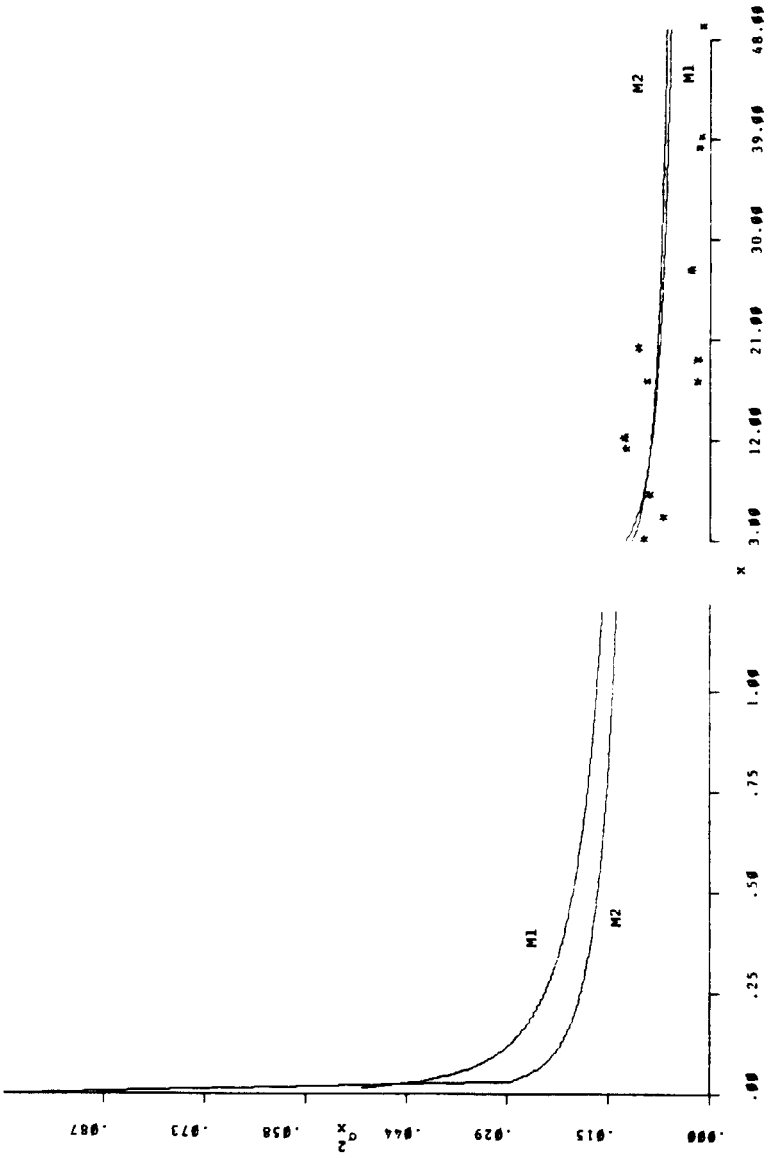

Fig. 2. The two model fits for refined stratum 10 in Colorado. The horizontal scale of sampling unit size is shown in terms of the number of segments and is broken in two parts in order to show the behavior of a variance curve in the lower range of unit sizes as well as to depict the full range of unit sizes used in the curve fitting. Legend: M1 = Method 1, and M2 = Method 2.

toward the end. Since the county size units are substantially large, the two curves show almost the same goodness of fit for the stratum variance estimates computed for these units.

The above model fits are typical across the refined strata in the region. Though the measure of goodness of fit of estimated variance values is not very satisfactory (for example, $R^2 \leq 0.50$), the performance of a model fit in imputing stratum variance for a sampling unit of size 5×6 nautical miles or smaller is presently of main concern. Despite of low values of R^2, the fitted models, particularly for

Method 2, provide quite reasonable stratum variance estimates for the sampling unit of size 5×6 nautical miles as discussed next.

3.3. Comparison between model-imputed and LANDSAT-based stratum variance estimates

Listed in Table 2 are the values of A and B for the stratum model fits and the estimated standard deviation for the sampling unit of size 5×6 nautical miles for each method. The standard deviation estimates correspond to the imputed stratum variances obtained from curve fits (Equation (9)) with $x = 22\,932$[2]. One finds that

Table 2
Empirical models and stratum standard deviation 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	0.242	−0.269	0.127	0.108	−0.204	0.118
	101	0.058	−0.355	0.041	0.023	−0.273	0.039
Kansas	7	0.289	−0182	0.216	0.221	−0.215	0.160
	8	1.124	−0.447	0.113	0.197	−0.313	0.092
	9	1.825	−0.512	0.103	0.182	−0.337	0.078
	11	0.888	−0.456	0.095	0.157	−0.353	0.068
	12	0.222	−0.211	0.164	0.162	−0.210	0.141
	13	0.109	−0.343	0.059	0.058	−0.320	0.048
	14	0.124	−0.381	0.052	0.061	−0.328	0.048
	15	0.684	−0.403	0.109	0.189	−0.253	0.122
	60	1.881	−0.563	0.081	0.155	−0.408	0.051
	102	0.204	−0.620	0.020	0.034	−0.527	0.013
Minnesota	15	0.035	−0.371	0.029	0.022	−0.332	0.028
	19	0.082	−0.293	0.066	0.054	−0.233	0.073
	20	0.375	−0.306	0.132	0.166	−0.239	0.122
Montana	21	2.485	−0.565	0.093	0.172	−0.351	0.071
	22	0.994	−0.533	0.069	0.098	−0.335	0.058
	23	0.532	−0.365	0.117	0.125	−0.248	0.102
	104	0.125	−0.397	0.048	0.034	−0.287	0.044
Nebraska	10	0.230	−0.221	0.158	0.144	−0.187	0.148
	11	0.133	−0.344	0.076	0.076	−0.297	0.062
	14	0.179	−0.454	0.043	0.068	−0.362	0.042
	15	0.043	−0.225	0.067	0.038	−0.213	0.067
	16	0.016	−0.623	0.005	0.003	−0.473	0.005
	17	0.220	−0.344	0.084	0.079	−0.242	0.083
	103	0.018	−0.865	0.002	0.001	−0.614	0.001

[2] A segment of size 5×6 nautical miles consists of 22 932 pixels.

Table 2 (continued)

State	Refined stratum	Method 1			Method 2		
		A	B	Standard deviation estimate	A	B	Standard deviation estimate
North	19	0.777	−0.389	0.125	0.190	−0.313	0.090
Dakota	20	1.238	−0.459	0.111	0.210	−0.373	0.070
	21	0.402	−0.328	0.122	0.147	−0.258	0.105
	22	0.285	−0.306	0.115	0.112	−0.248	0.096
Oklahoma	3	0.166	−0.427	0.048	0.057	−0.321	0.047
	7	0.325	−0.216	0.193	0.216	−0.178	0.191
	9	0.702	−0.392	0.117	0.150	−0.312	0.081
	13	0.084	−0.291	0.067	0.057	−0.270	0.062
	60	0.647	−0.389	0.114	0.162	−0.307	0.086
	102	0.073	−0.478	0.024	0.022	−0.343	0.026
South	15	0.024	−0.481	0.014	0.009	−0.436	0.011
Dakota	16	0.097	−0.254	0.087	0.058	−0.199	0.089
	17	0.370	−0.453	0.063	0.060	−0.296	0.056
	18	0.441	−0.578	0.036	0.042	−0.420	0.025
	19	0.258	−0.324	0.100	0.115	−0.270	0.087
	21	0.380	−0.426	0.073	0.080	−0.340	0.051
	104	0.430	−0.679	0.022	0.031	−0.468	0.017
Texas	2	0.054	−0.327	0.045	0.028	−0.261	0.045
	3	0.058	−0.291	0.056	0.033	−0.264	0.048
	4	0.071	−0.203	0.096	0.055	−0.196	0.088
	5	0.191	−0.275	0.110	0.101	−0.219	0.106
	9	0.321	−0.269	0.147	0.140	−0.237	0.113
	60	0.558	−0.396	0.102	0.121	−0.272	0.089
	61	0.068	−0.143	0.127	0.060	−0.183	0.098
	101	0.030	−0.484	0.015	0.007	−0.380	0.013
	102	0.029	−0.414	0.021	0.011	−0.345	0.019

the variance estimates for Method 1 are generally higher than for Method 2 across the refined strata.

For the 1978 wheat survey of USGP, LANDSAT data for more than 400 area segments each of size 5×6 nautical miles, were collected and their proportions of wheat acreage were estimated. Sample segments were randomly selected in each stratum. LANDSAT stratum variance estimates were made except for eight strata where either none or only one sample segment data were available in a stratum.

The stratum standard deviation estimates given in Table 2 were compared with those estimated from the LANDSAT survey data. Estimates for Method 2 compared quite well and not so well for Method 1, except when a LANDSAT variance estimate was obtained using less than five segments and hence, making the comparison somewhat unreliable. The relative difference of a model imputed estimate to the corresponding LANDSAT estimate was at most 25 percent for Method 2 and

65 percent for Method 1 for those strata which had at least 10 sample segments available.

Suppose $\hat{\sigma}_{jk}$ denotes the estimated standard deviation for stratum j using method k and S_j is the LANDSAT-based sample standard deviation for the stratum. Considering the difference $(\hat{\sigma}_{jk} - S_j)$ for each method, the average and variance of the differences obtained for all refined strata were computed. Only refined strata with two or more sample segments were considered for this comparison. The results as given in Table 3 show that the estimates using Method 2 are in agreement with those computed from the sampled LANDSAT data, but the estimates given by Method 1 do not seem to be in agreement; this is because the relative size of the average difference to its standard deviation is small for Method 2 and substantially larger for Method 1.

Table 3
Average difference and its variance for standard deviation estimates across strata

Method	Average difference	Variance
1	0.0110	0.00109
2	0.0013	0.00123

If the average difference is an indication of the likely bias introduced by a method, then Method 2 seems to be unbiased and Method 1 seems to overestimate. The poorer performance by Method 1 is partly due to its sensitivity to the field size which was highly overestimated using Equation (10); Chhikara and Perry (1980) show that Method 1 leads to a very small average difference as in Method 2 if the estimate of field size is reduced by a factor of four.

4. Concluding remarks

Presently, the values obtained for B have the range of $-0.679 \leq B \leq -0.143$ for Method 1 and $-0.473 \leq B \leq -0.178$ for Method 2. Interestingly enough, these values compare quite well with the range of values, $0.123 \leq g \leq 0.548$, obtained by Mahalanobis (1968, Ch. 3) for his model as stated in Equation (1). (The sign difference is because g is an exponent in the denominator of his function and will have an opposite sign to B.) He estimated the variance parameter g from the Jute acreage data collected from several districts of Bengal and used the sample unit sizes of 1, 2.25, 4, 6.25, and 9 acres, which are very small (with the exception of pixel in our Method 2) compared to the sample unit sizes presently used. This further suggests that values of B are very stable and the proposed methodology is viable.

In summary, two methods are proposed 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. Each method uses existing and easily available information of historical crop statistics in developing this relationship. Because the anchor point x_0 is a critical point in our model fit and would affect the stratum variance estimation significantly, Method 1 should not be used unless accurate field size determinations can be made. Overall, the use of Method 2 is preferable.

Appendix

A.1. *Background of LANDSAT Data*

Since 1972 NASA has launched a series of land observatory satellites, called LANDSAT, to develop a remote sensing technology for monitoring various types of earth resources on local as well as global basis. Each satellite can provide 18-day repetitive coverage of the earth surface. The LANDSAT coverage of an area is in the form of a scene consisting of its scanlines with a certain number of resolution elements per scanline. The size of a resolution element is approximately 1.1 acres and is known as *pixel*. The sensor system on board the LANDSAT is a multispectral scanner (MSS) which measures the reflectance of a scene in four different wavelength bands. The spectral measurements are converted to digital counts and transmitted to receiving stations on earth. The measurements are influenced by the vegetation, soil type and atmospheric conditions, and when these are statistically modeled and correlated with the features on ground, the assessment of earth resources from acquiring and analyzing MSS data for an area becomes feasible.

The image analysis techniques are used to label spectral classes by crop types on ground. A segment of several square miles in area is required to delineate discernible patterns and identify possible crop types. A crop can be distinguished from others in a scene by monitoring the temporal development of its fields from planting through harvest since LANDSAT data for the scene are collected every 18 days. The entire method of crop acreage estimation involves techniques of scene registeration, image analysis and data classification, and is documented in the LACIE Symposium (1979).

A.2. *Derivation of variance, equation* (5)

Let the sampling unit size be x_0 (preferably much smaller than the average field considered in the other case) and Y be the proportion of acreages devoted to a specific crop in a randomly selected sample unit. Three outcomes are possible: (1) the unit contains only the crop of interest, $Y=1$, (2) it does not contain the crop at all, $Y=0$, or (3) it partially contains the crop, $0<Y<1$. Let α_1, α_2, and α_3 are the respective probabilities of outcomes in (1), (2) and (3). Suppose p is the crop acreage proportion in the stratum. Then $E[Y]=p$ and the variance of Y,

$$\sigma_{x_0}^2=\alpha_1(1-p)^2+\alpha_2 p^2+\alpha_3 E[(Y-p)^2\,|\,0<Y<1]. \tag{A.1}$$

We now express α_1, α_2, and α_3, and $E[(Y-p)^2 \mid 0<Y<1]$ in terms of the crop proportion p and the field size distribution of the stratum.

Assume that the stratum has area A and the crop fields of length l_i and width w_i have relative frequencies f_i, $i=1,2,\ldots,N$. As shown by the illustration of a field in Figure A1, suppose b is the expected 'width of a sample unit falling on the field boundary. Since the sample units are small relative to crop field sizes, it is assumed that $b<l_i$ and $b<w_i$ for all i and the distance between any two fields of the crop is greater than or equal to b.

From Figure A1, we note that the pure crop area and the mixed area associated with a field of length l_i and width w_i are given by $(l_i-b)(w_i-b)$ and $(l_i+b)(w_i+b)-(l_i-b)(w_i-b)$, respectively. Next, the total number of fields of length l_i and width w_i is equal to pAf_i/l_iw_i. Thus, it follows that

$$\begin{aligned}
\alpha_1 &= \frac{1}{A}\left[\sum_{i=1}^{N}\left(\frac{f_i pA}{l_i w_i}\right)(l_i-b)(w_i-b)\right] \\
&= p\sum_{i=1}^{N} f_i \frac{(l_i-b)(w_i-b)}{l_i w_i}, \\
\alpha_3 &= \frac{1}{A}\left\{\sum_{i=1}^{N}\left(\frac{f_i pA}{l_i w_i}\right)[(l_i+b)(w_i+b)-(l_i-b)(w_i-b]\right\} \\
&= p\sum_{i=1}^{N} \frac{2bf_i(w_i+l_i)}{w_i l_i}, \\
\alpha_2 &= 1-\alpha_1-\alpha_3.
\end{aligned} \tag{A.2}$$

To facilitate the evaluation of $E[(Y-p)^2 \mid 0<Y<1]$, assume that a sample unit falling on a field boundary is configured as in Figure A2. The directed distance from

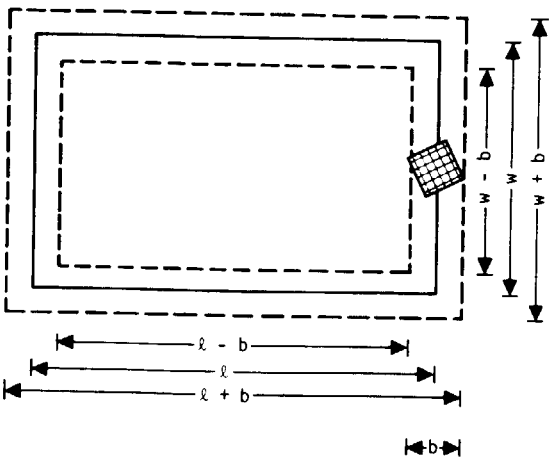

Fig. A1. An illustration of a crop field for the crop of interest and a randomly placed unit on the field boundary. Legend: l=length, w=width, and b=expected width of a sample unit over the field boundary.

the center of the unit to the field boundary is denoted by x, where x is taken to be positive if the center of the unit is not in the field, and x is taken to be negative if the center of the unit is in the field. The smallest angle that a diagonal makes with the horizontal is denoted by θ. Now it is easy to see that $|x| \le d\cos\theta$ and $0 \le \theta \le \frac{1}{4}\pi$, where d is the half of the diagonal length for the square sample unit.

The area of the sample unit contained within the crop field can be expressed as a function of x and θ for $0 \le \theta \le \frac{1}{4}\pi$ and $0 \le x \le d\cos\theta$ using simple geometric observations as follows:

$$A(\theta, x) = \begin{cases} (d\cos\theta - d\sin\theta)[\tan(\frac{1}{4}\pi - \theta) + \tan(\frac{1}{4}\pi + \theta)]\left(\dfrac{d\cos\theta + d\sin\theta}{2} - x\right) \\ \quad \text{if } 0 \le x \le d\sin\theta, \\ \frac{1}{2}(d\cos\theta - x)^2[\tan(\frac{1}{4}\pi - \theta) + \tan(\frac{1}{4}\pi + \theta)] \\ \quad \text{if } d\sin\theta \le x \le d\cos\theta. \end{cases} \tag{A.3}$$

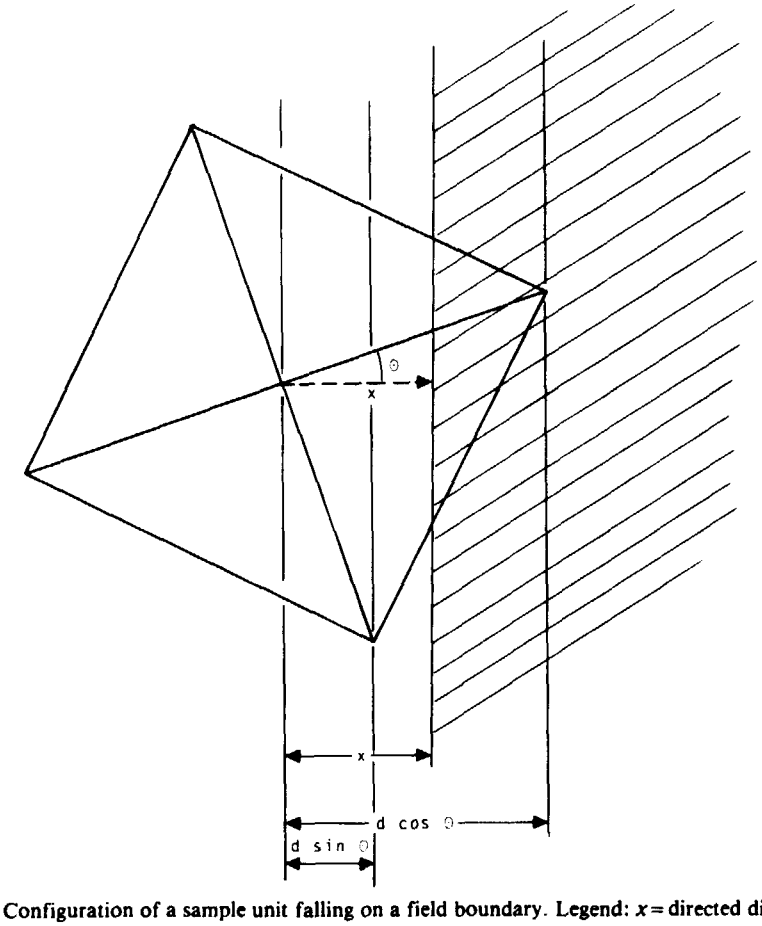

Fig. A2. Configuration of a sample unit falling on a field boundary. Legend: x = directed distance from the center of the sample unit to the field boundary, θ = smallest angle that a diagonal makes with the horizontal axis, and d = half the length of diagonal of the sample unit.

This formula is readily extended to negative values of x and then adjusted for the total area of the sample unit, x_0, to obtain the following expression for the proportion of the unit contained within the crop field:

$$p(\theta, x) = \begin{cases} \dfrac{A(\theta, x)}{x_0}, & 0 \le x \le d\cos\theta, \\ 1 - \dfrac{A(\theta, -x)}{x_0}, & -d\cos\theta \le x \le 0. \end{cases} \tag{A.4}$$

Observe that any angle $0 \le \theta \le \frac{1}{4}\pi$ corresponds to two positions of the square: one where the angle is measured below the horizontal and the other where the angle is measured above the horizontal. Thus, it follows that

$$E[Y \mid 0 < Y < 1] = \frac{4}{\pi}\int_0^{\pi/4}\left[\frac{1}{2d\cos\theta}\int_{-d\cos\theta}^{d\cos\theta} p(\theta, x)\,dx\right]d\theta,$$

$$E[Y^2 \mid 0 < Y < 1] = \frac{4}{\pi}\int_0^{\pi/4}\left[\frac{1}{2d\cos\theta}\int_{-d\cos\theta}^{d\cos\theta} [p(\theta, x)]^2\,dx\right]d\theta$$

The first integral when evaluated gives $E[Y \mid 0 < Y < 1] = \frac{1}{2}$. Evaluation of the second integral is considerably more involved, requiring several steps, as discussed in Chhikara and Perry (1980). Here we omit the details and give the resulting expression:

$$\begin{aligned} E[Y^2 \mid 0 < Y < 1] = \frac{4}{\pi}\Bigg\{ & \tfrac{1}{8}\pi - \frac{d^2}{3x_0}\left[(\ln 2) + \left(\tfrac{1}{8}\pi - \frac{\ln 2}{4}\right)\right] \\ & + \frac{d^4}{x_0^2}\left[\left(\tfrac{1}{8}\pi - \tfrac{1}{4}\right) + \tfrac{1}{3}\left(\ln 2 - \tfrac{1}{4} - \tfrac{1}{8}\pi\right)\right] \\ & + \frac{d^4}{5x_0^2}\left[2 - \tfrac{1}{2}\pi - \frac{3\ln 2}{2}\right]\Bigg\}. \end{aligned} \tag{A.5}$$

Taking the sampling unit to be one unit square, one has $x_0 = 1$ and $d = 1/\sqrt{2}$. Then, the right side expression in (A.5) is approximately equal to 0.3682. Using this approximation for $E[Y^2 \mid 0 < Y < 1]$ and the value of $\frac{1}{2}$ for $E[Y \mid 0 < Y < 1]$ obtained earlier, the expression for $\sigma_{x_0}^2$ in Equation (5) of Section 2.3 follows. Next, as it can be seen from Figure 2a that the expected 'width' of a sample unit falling on field boundaries is given by

$$b = \frac{4}{\pi}\int_0^{\pi/2} 2d\cos\theta\,d\theta = 1.2732.$$

This completes the formulas for the probabilities α_1, α_2, and α_3, and hence, the derivation of $\sigma_{x_0}^2$.

In the derivation of $\sigma_{x_0}^2$, it was assumed that the sample unit did not fall on a field corner. This, of course, introducs an error. The magnitude of this error was

estimated to be less than 5 percent relative to the variance in (5). Full details are given in Chhikara and Perry (1980).

Acknowledgement

The authors appreciate very much the suggestions and guidance received from their former colleagues A.H. Feiveson and C.R. Hallum at the Johnson Space Center and the late Professor H.O. Hartley during the course of this study. They thank the two referees for their helpful comments made on an earlier version.

The research work of Raj S. Chhikara was supported under NASA contract, NASA-15800. Charles R. Perry did his work while he was NRC Senior Resident Research Associate with NASA/JSC, Houston.

References

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

Chhikara, R.S. (1984). (Ed.). Crop Surveys Using Satellite Data (A Special Issue). *Comm. Statist. Theory Meth.* 13 (23), 2857-2996.

Chhikara, R.S. and A.H. Feiveson (1982). A sample survey of global wheat acreage using satellite (LANDSAT) data. *Sankhyā Ser. B* 44, 304-329.

Chhikara, R.S. and C.R. Perry (1980). Estimation of within-stratum variance for sample allocation. AGRISTARS Report, JSC-16343, July 1980, Johnson Space Center, Houston, TX.

Cochran, W.G. (1963). *Sampling Techniques*, Second Edition. John Wiley & Sons, New York.

Craig, M.E., R.S. Sigman and M. Cardenas (1978): Area Estimates by LANDSAT: Kansas 1976 winter wheat. Research Report (August 1978), Economics, Statistics, and Cooperative Service, U.S. Department of Agriculture, Washington, DC.

Hendricks, W.A. (1944). The relative efficiency of groups of farms as sampling units. *J. Amer. Statist. Assoc.* 39, 367-376.

Houston, A.G. and F.G. Hall (1984). Use of satellite data in agriculture surveys, *Comm. Statist. Theory Meth.* 13 (23), 2857-2880.

Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. Iowa Agricultural Experiment Station, Research Bulletin 304.

LACIE Symposium (1978). *Proceedings of Technical Sessions*. NASA, JSC 14551, NTIS, Springfield, VA.

MacDonald, R.B. and F.G. Hall (1980). Global crop forecasting. *Science* 208, 670-679.

Mahalanobis, P.C. (1940). A sample survey of the acreage under jute in Bengal. *Sankhyā* 4, 511-530.

Mahalanobis, P.C. (1968). *Sample Census of Area Under Jute in Bengal, 1940*. Statistical Publishing Society, Calcutta.

Perry, C.R. and C.R. Hallum (1979). LACIE sampling unit size considerations in large area crop inventorying using satellite-based data. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 430-433.

Pitts, D.E. and G. Badhwar (1980). Field size, length, and width distributions based on LACIE ground-truth data. *Remote Sensing of Environment* 10, 201-213.

Smith, H.F. (1938). An empirical law describing heterogeneity in the yields of agriculture crops. *J. Agricultural Sci.* 28, 1-23.

U.S. Bureau of Census (1977). *1974 Census of Agriculture, Parts 1-51*. U.S. Department of Commerce, Washington, D.C.