MULTIYEAR ROTATION DESIGN SAMPLING IN AGRICULTURAL SURVEYS

Raj S. Chhikara, University of Houston-Clear Lake, Houston, TX. 77058
Lih-Yuan Deng, Memphis State University, Memphis, TN. 38152

KEY WORDS: Area frame, Single-year and multiyear approaches, Analysis of variance model, Relative efficiency.

ABSTRACT

We propose a multiyear estimation method based on an analysis of variance model that takes into account the successive sampling of units in the area frame across years as sampled by the U.S. Department of Agriculture. The proposed method is applied to estimate the 1989 soybean acreages using June Enumerated Survey (JES) data for three years, 1987, 1988 and 1989. These estimates are compared with those obtained using the current USDA estimation method. The proposed estimation method is also shown to be fairly robust to misspecification of the model parameter.

1. INTRODUCTION

The area sampling frame used by the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture is based on a land use stratification and provides a full coverage of the geographical area of interest. The primary sampling unit is an area segment which varies in size by land use stratum. For intensive agricultural areas, these segments are often targeted to be one square mile land areas.

The USDA estimation methodology uses the current year survey data and overlooks the fact that the area frame sampling involves multiyear rotation designs with twenty percent replacement of sample units each year. Because of a substantial overlap of sampled units from one year to another, the use of multiyear sample data would augment the sample survey information obtained in the current year and, thereby, effect an increase in the sample size. This would reduce the sampling variance of an estimate.

The estimation methodology based on successive or rotation design sampling has been considered by several investigators. The study by Patterson (1950) was the forerunner to many studies that followed. Rao and Graham (1964) and Graham (1973), among others, studied estimators derived by separating the matched and unmatched units of repeated surveys and developing a composite estimation method involving estimators for the two consecutive periods. Wolter (1979) assumed a general linear model to describe the individual panel estimators and proposed to combine these estimators into one that would have a smaller variance than the one which uses only the latest period sample data. This approach, however, would require a determination or estimation of the covariance matrix of the vector of panel estimators, which may not be feasible.

Thus far the composite estimation has been based upon certain combination(s) of periodic estimates. An alternative approach would be to pool the sample data acquired under a rotation design and construct directly a multi-period estimator. Since sample data would be cross referenced between sample units and periods, these data can be described in terms of a two-way analysis of variance model. In the context of crop surveys using satellite acquired data under a rotation design sampling, Hartley (1980) proposed the analysis of variance approach to utilize multiyear sample data for crop acreage estimation.

In the next section we describe the direct expansion estimator currently used at NASS. We then discuss in Section 3 the multiyear approach for the area frame sampling and develop a new estimator utilizing the multiyear sample data. The new estimation method was applied to survey data from three consecutive years, 1987, 1988 and 1989, to obtain estimates of planted soybean acreages in 1989 for Indiana, Ohio and Oklahoma. The numerical results are presented in Section 4.

2. AREA TRACT ESTIMATOR

In the area frame sampling, the reporting unit is a tract which is the area of land located

within a segment that is under a single operation or management. The estimator of total for a survey item is obtained by adding the tract data for each sample segment, multiplying the sum by the expansion factor for the segment, and then aggregating the expanded segment totals to the stratum and higher levels. Since its computation is based on data from the tracts confined within segments, it is sometimes referred to as the area tract estimator. This is an unbiased estimator and is considered reliable for estimating crop acreages. For background information on the land use stratification and estimation procedure, the readers may refer to Kuo (1989).

The area tract estimator is computed as follows:

$$\hat{Y} = \sum_{h \in H} \sum_{k=1}^{n_h} E_{hk} y_{hk}$$

where H is the collection of strata, E_{hk} is the expansion factor for segment k in stratum h (which simply is equal to the inverse of the probability of selection of a segment in the stratum), n_h is the number of segments sampled in stratum h, and y_{hk} is the agricultural item value for segment k of stratum h.

3. Estimators Incorporating Multiyear Rotation Design

3.1. ANOVA Model.

As we mentioned earlier, the area frame sampling has substantial overlap of sampled units from one year to another. As Hartley (1980) noted, there is a certain amount of consistency in area segment characteristics from one year to another. For example, the suitability of the segment prevalent soil types will be invariant from year to year or the capabilities of certain operators in a segment to grow crops, etc. will be largely persistent. On the other hand, factors such as weather and economic conditions will vary across years and will affect the farm outcome. Taking these aspects into consideration we propose the following multiyear model approach to estimation of crop acreages that avoids any unwarranted assumptions sometimes made, and correctly so, in the analysis of other 'time series'.

To simplify the notation, we consider estimation for a stratum. Let y_{tk} be the agricultural item value in year t for segment k in a stratum. Then the y_{tk} can be viewed consisting of the stratum mean in year t, the segment effect, and an error component. One can, therefore, describe it in terms of an analysis of variance model:

$$y_{tk} = \alpha_t + b_k + e_{tk} \tag{1}$$

where $k = 1, 2, ..., n_t$ and $t = 1, 2, ..., T$. Here n_t denotes the number of segments sampled for the stratum in year t. Other terms in the ANOVA model are as follows: α_t is the mean value for the characteristic of interest over all the segments in the stratum for year t. b_k is the segment effect representing the deviation of the kth segment response from that of the stratum mean. e_{tk} is the model error associated with segment k in year t. The error term in (1) comprises the sampling error and any interaction that may exist between years and segments.

We assume that the b_k are independently distributed with $E(b_k) = 0$ and $\mathrm{Var}(b_k) = \sigma_b^2$, and the e_{tk} are independently distributed with $E(e_{tk}) = 0$ and $\mathrm{Var}(e_{tk}) = \sigma_e^2$. Furthermore, b_k and e_{tk} are independent of each other.

3.2. Estimation.

The above linear model can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{U}\mathbf{b} + \mathbf{e}, \tag{2}$$

where $\mathbf{X}$ is the design matrix consisting of 0's and 1's which account for the effect due to α_t's and $\mathbf{U}$ is the design matrix of 0's and 1's which are specified according to the rotation sampling scheme. The dimensions of $\mathbf{X}$ and $\mathbf{U}$ are $N \times T$ and $N \times S$, respectively, where $N = \sum_{t=1}^{T} n_t$ and S is the total number of distinct segments sampled in T years. Note that $S \leq N$. Let

$$\mathbf{b}^* = \mathbf{U}\mathbf{b} + \mathbf{e}, \tag{3}$$

then $E(\mathbf{b}^*) = 0$ and

$$\begin{aligned} \mathrm{Var}(\mathbf{b}^*) &= \mathbf{I}\sigma_e^2 + \mathbf{U}\mathbf{U}'\sigma_b^2 \\ &= \sigma_e^2(\mathbf{I} + \gamma\mathbf{U}\mathbf{U}') = \sigma_e^2\mathbf{W}_\gamma, \end{aligned} \tag{4}$$

where $\gamma = \sigma_b^2/\sigma_e^2$. The weighted least-squares estimator of $\boldsymbol{\alpha}$ is given by

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{W}_\gamma^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_\gamma^{-1}\mathbf{y}. \tag{5}$$

The parameter γ in $\mathbf{W}_\gamma$ can be estimated by $\hat{\sigma}_b^2/\hat{\sigma}_e^2$ preferably using some previously obtained survey or pilot study data. The variance-covariance matrix of $\hat{\boldsymbol{\alpha}}$ is given by

$$\text{Var}(\hat{\boldsymbol{\alpha}}) = (\mathbf{X}'\mathbf{W}_\gamma^{-1}\mathbf{X})^{-1}\sigma_e^2. \qquad (6)$$

This variance formula still applies asymptotically if γ is replaced by a consistent estimator $\hat{\gamma}$.

The single-year estimate currently used at NASS and described in Section 2 can be obtained by setting $\gamma = 0$ (i.e., no segment effect) in Equation (5) for $\hat{\boldsymbol{\alpha}}$, and are given by

$$\hat{\hat{\boldsymbol{\alpha}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\bar{y}_1, \bar{y}_2, ...\bar{y}_T)', \qquad (7)$$

where $\bar{y}_t$ is the sample mean for year t. In order to evaluate the performance of $\hat{\hat{\boldsymbol{\alpha}}}$ as an alternative to $\hat{\boldsymbol{\alpha}}$, one computes the variance-covariance matrix of $\hat{\hat{\boldsymbol{\alpha}}}$ under model (2), which would be

$$\text{Var}(\hat{\hat{\boldsymbol{\alpha}}}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}_\gamma\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\sigma_e^2.$$

From the generalized Gauss-Markov theorem, we have

$$\text{Var}(\mathbf{c}'\hat{\boldsymbol{\alpha}}) \leq \text{Var}(\mathbf{c}'\hat{\hat{\boldsymbol{\alpha}}}),$$

for any vector $\mathbf{c}$. In particular, the multiyear estimator has a smaller or equal variance than the single-year estimator for the current year mean obtained by taking $\mathbf{c}'$ to be

$$\mathbf{v}_T' = (0, 0, ..., 0, 1)_{1\times T}.$$

In other words, for the latest year the multiyear estimator and the single-year estimator are $\hat{\alpha}_T = \mathbf{v}_T'\hat{\boldsymbol{\alpha}}$ and $\hat{\hat{\alpha}}_T = \mathbf{v}_T'\hat{\hat{\boldsymbol{\alpha}}}$, respectively. Thus, the multiyear estimator $\hat{\alpha}_T$ is more efficient than the single-year estimator $\hat{\hat{\alpha}}_T$, as one would expect because the proposed estimation procedure utilizes all sample data obtained under the multiyear rotation design.

4. Numerical Results

4.1. Survey Data and Estimates.

Area frame JES data for 1987, 1988 and 1989 were utilized to compute the 1989 estimates for planted soybeans acreages for three states, Indiana, Ohio and Oklahoma. The three-year data sets were merged and cross referenced from one year to another with respect to rotation design sample units. A new data set was developed for planted soybean acreages where all the codes and variables necessary for estimation were retained. This new data set allowed us to obtain a two-way classification table showing the various sample segments versus the years in which each one was observed. Table 1 shows the rotation design configuration for an area frame stratum for the three years 1987-1989.

[Refer to Table 1]

The mean soybeans acreage was estimated from the three years data given in Table 1 using both the single-year and multiyear approaches. The estimates and their standard errors in each case are presented in Table 2.

[Refer to Table 2]

It is noted that the 1989 estimates under both methods are about the same. However, there is a substantial difference in their standard errors. The relative efficiency given by $(17.53)^2/(12.46)^2 = 1.98$ suggests that the multiyear approach is almost two times as efficient as the single-year approach. Moreover, the multiyear estimate has a much more stable yearly variance estimate than does the single-year estimate. Thus, it provides an additional advantage over the currently used single-year estimate.

For the three states, the soybeans acreage estimates were computed using the single-year and the multiyear estimation methodologies as previously discussed. Also computed were their standard errors (S.E.) and the relative efficiency of a multiyear estimate compared to the corresponding single-year estimate. The relative efficiency is computed as

$$\text{R. E.} = \left[SE(\hat{\alpha}_T)/SE(\hat{\alpha}_T)\right]^2.$$

The numerical results are listed in Table 3.

[Refer to Table 3]

The two estimators at the state level differ insignificantly, particularly for the two major producing states of Indiana and Ohio.

The computed relative efficiencies of the area tract estimates are 1.13, 1.74 and 1.25 for Ohio, Oklahoma and Indiana, respectively.

Thus the multiyear approach provides a much more efficient estimate than does the current approach.

4.2. Effect of error in estimating γ.

The value of γ used in obtaining a multiyear estimate was determined using the analysis of variance method applied to the three years survey data and then further iterated so that the computed variance of $\hat{a}_T$ is minimum. Because of sampling variability, the estimation of γ will be subject to error. To evaluate the effect on the multiyear estimates due to error in estimating γ, these estimates of the three states and the corresponding standard errors were computed by varying $\hat{\gamma}$. Table 4 shows the computed estimates and standard errors corresponding to various values of $\hat{\gamma}/\gamma$, where $\hat{\gamma}$ is the value used in the estimation procedure and γ is the best value determined using sample data.

Letting $c = \hat{\gamma}/\gamma$, we consider $c = 0$, 0.5, 1.0, 1.5 and 2.0, where $c = 0$ corresponds to the single-year estimation method.

[Refer to Table 4]

For $c = 0.5, 1.5$ and 2.0 versus $c = 1.0$, which corresponds to the case of no misspecification, results in Table 4 shows that these estimates hardly vary. Similar is the case with respect to their standard errors. Hence, we conclude that the multiyear estimates display a high level of robustness to misspecification of model parameter γ.

5. Summary

For the rotation design sampling as implemented at NASS, a multiyear approach to estimation of crop acreages is proposed. A multiyear estimation method is developed based on a two-way analysis of variance model. We show the multiyear estimator to be more efficient and robust. The proposed model-based approach is easy to implement as demonstrated here by the application made using the real data obtained from NASS.

Acknowledgement

The authors' work was partially supported with the funds provided by the National Agricultural Statistics Service of the U.S. Department of Agriculture under a cooperative research program at the University of Houston–Clear Lake. The authors would like to thank Mr. Bill Iwig of USDA for his insightful explanation of the USDA data collection and estimation procedures.

References

Graham, J. E. (1973), "Composite estimation in two cycle rotation sampling designs," *Communications in Statistics, A* **1**, 419-431.

Hartley, H. O. (1980), "A survey of multiyear estimation procedures," Technical Report DS1, Department of Mathematics, Duke University, North Carolina.

Kuo, L. (1989), "Composite estimation of totals for livestock surveys," *Journal of the American Statistical Association* **84**, 421-429.

Patterson, H. D. (1950), "Sampling on successive occasions with partial replacement of units," *Journal of the Royal Statistical Society B* **12**, 241-255.

Rao, J. N. K. and J. E. Graham (1964), "Rotation designs for sampling on repeated occasions," *Journal of the American Statistical Association* **59**, 492-509.

Wolter, K. M. (1979), "Composite estimation in finite population," *Journal of the American Statistical Association* **74**, 604-613.

Table 1: Total soybeans acreages for sample segments in a stratum during 1987-89

	YEAR		
Segment	1987	1988	1989
1	92.00		
2	121.10		
3	180.00	151.60	
4	122.00	119.90	
5	140.90	144.80	122.50
6	139.90	116.00	134.00
7	119.00	86.50	134.00
8	156.70	58.30	220.50
9	125.60	144.40	230.00
10	95.00	134.00	126.70
11		121.50	98.50
12		147.50	137.30
13			46.00
14			185.70

Table 2: Estimated soybeans acreages and their standard errors for the stratum

	Single-year approach		Multiyear approach	
Year	$\hat{\alpha}_T$	S. E.	$\hat{\alpha}_T$	S. E.
1987	129.22	8.42	129.13	—
1988	122.45	9.46	122.33	—
1989	143.52	17.53	143.55	12.46

Table 3: 1989 soybeans acreage estimates (in thousands).

	Single-year approach		Multiyear approach		R. E.
State	Estimate	S. E.	Estimate	S. E.	
Ohio	4075.6	174.2	4069.6	163.9	1.13
Oklahoma	351.1	83.6	315.0	63.4	1.74
Indiana	4509.9	178.2	4492.5	159.5	1.25

Table 4: Multiyear estimates of soybeans acreages with γ misspecified as $\hat{\gamma}(= c\gamma)$

	c				
State	0.0	0.5	1.0	1.5	2.0
Ohio	4076	4071	4070	4069	4069
	(178)	(165)	(164)	(165)	(167)
Oklahoma	351	322	315	311	309
	(76)	(64)	(63)	(64)	(65)
Indiana	4510	4496	4493	4490	4489
	(172)	(161)	(159)	(160)	(162)

* values in parentheses are for standard errors.