United States Department of Agriculture

National Agricultural Statistics Service

Research Division

SRB Research Report Number SRB-93-09

October 1993

Estimation of Totals Using Multiyear June Agricultural Survey Data

Raj S. Chhikara
Lih-Yuan Deng
Yilian Yuan
Charles R. Perry
William C. Iwig

ESTIMATION OF TOTALS USING MULTIYEAR JUNE AGRICULTURAL SURVEY DATA, By Raj S. Chhikara*, Lih-Yuan Deng**, Yilian Yuan***, Charles R. Perry and William C. Iwig, Sampling and Estimation Research Section, Survey Research Branch, Research Division, National Agricultural Statistics Service, United States Department of Agriculture, Washington DC 20250-2000, October 1993, Report No. SRB-93-09.

ABSTRACT

In the annual June Agricultural Surveys of the U.S. Department of Agriculture, area frame sampling is based on multiyear rotation designs with twenty percent of the sample units replaced each year. At present, the current year sample data are used for direct expansion estimation of totals for agricultural items of interest. We investigate the use of multiyear sample survey data for estimation of crop acreages and livestock. A multiyear estimation method is developed based on an analysis of variance model that takes into account the successive sampling of units in the area frame across years. The method is applied to estimate the number of total hogs and soybean acres in 1990 for four states using survey data from 1987, 1988, 1989, and 1990. These estimates are compared with those obtained using only the 1990 sample survey data. Relative efficiencies of the multiyear estimates compared to the single-year estimates show a substantial improvement in precision.

KEY WORDS

Area frame, List frame, Non-overlap domain, Multiple-frame estimator, Analysis of variance model, Relative efficiency.

ACKNOWLEDGEMENT

The authors thank Darryl Roosevelt, Chandima De Silva, M.S Kumar Peeris, and Rajeev Mishra, University of Houston-Clear Lake for their help in data handling and analysis performed at the University of Houston-Clear lake.

*Raj S. Chhikara is Professor, Division of Computing and Mathematics, University of Houston-Clear Lake.

**Lih-Yuan Deng is Associate Professor Department of Mathematical Sciences, Memphis State University.

***Yilian Yuan is Senior Analyst/Biostatistician, Department of Health Care Research and Analysis, Mid-South Foundation for Medical Care, Inc.

TABLE OF CONTENTS

LIST OF TABLES

SUMMARY

Area frame sampling for the June Agricultural Surveys is based on a multiyear rotation design which provides eighty percent overlap of sample segments from one year to the next. However, the current NASS estimators, with the exception of some ratio estimators, do not make use of sample data collected in prior years. Since a certain amount of consistency in area segment characteristics is expected from one year to another, an approach utilizing multiyear sample data is proposed to augment the survey information for estimation of crop acreages and livestock. The proposed method is developed by modeling the multiyear data using an analysis of variance model which lends itself to an increase in "effective" sample size. A new set of multiple frame estimators are constructed similar to those presently used at NASS.

Combined June Agricultural Survey area frame sample data for four years, 1987, 1988, 1989 and 1990, along with the 1990 list frame data are used to obtain the 1990 multi-year estimates for soybean acres and total hogs for Indiana or Iowa, Ohio and Oklahoma. Similarly, the 1988 multi-year estimates were obtained using sample data for two years 1987 and 1988, and the 1989 multi-year estimates using sample data for three years 1987, 1988 and 1989. All single-year estimates were obtained from only a single year of sample data using current NASS estimation methodology. The multiyear estimates compare very well with those based on the single-year data and tend to be more efficient and robust compared to the current NASS estimates.

INTRODUCTION

The National Agricultural Statistics Service (NASS) makes use of multipurpose probability surveys to estimate crop acreages, livestock inventory and other agricultural items of interest. The annual June Agricultural Survey (JAS) uses both area frame and list frame sampling units, and the estimation process involves a multiple frame methodology. The area sampling frame 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. A list sampling frame consists of known farm operators. A properly constructed list should contain names, addresses and measures of size of variates of interest. Fecso, Tortora and Vogel (1985) discussed in detail the advantages and disadvantages of the area frame and the list frame sampling as developed at NASS.

The area frame is complete and all area segments have known probabilities of selection. The main weakness is its inefficiency for items concentrated on extremely large farms and for rare commodities produced only on a few farms. This drawback can be corrected by using a supplemental sample from a list frame. However, lists are usually incomplete and difficult to compile and to maintain because of cost and constant changes of farm operators. This has led NASS to implement the multiple frame methodology originally proposed by Hartley (1962). It makes use of sample data from both the area and list frames.

With the exception of calculating a current year to previous year ratio estimator for land segments in both samples, the USDA estimation methodology is based on the use of only 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 (also known as repeated or panel surveys) 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. Gleason and Tortora (1978) proposed to estimate the crop acreages for the latest period utilizing a similar approach to improve upon the multiple frame estimator used at NASS. 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 generally is not feasible. Restricting his investigation to certain correlation structures for the covariance matrix, Wolter evaluated the gain in precision obtained using composite estimation and showed it to be substantial when there is a high correlation pattern depicted by the covariance matrix. Recently, Kott (1989) proposed to utilize this approach to improve upon current USDA estimates. Thomas, Perry, and Viroonsi (1990) applied it to estimate the total hogs for March, 1989, utilizing quarterly survey estimates over a period of two years, and obtained about a twenty percent gain in the sampling efficiency.

Thus far composite estimation has been based upon certain combinations 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 sampling design, Hartley (1980) proposed an analysis of variance approach to utilize multiyear sample data. Lythuan-Lee (1981) and Chhikara, Hallum and Lythuan-Lee (1984) implemented Hartley's idea to estimate 1978 wheat acreage for the state of North Dakota using satellite data acquired over three consecutive years, 1976, 1977 and 1978. In this application the multiyear estimate was found to be 30 percent more efficient than the single-year estimate. Sielken and Gbur (1984) further evaluated the efficiency of the multiyear estimate versus the single-year estimate, showing a substantial gain in precision under the multiyear approach.

In the present context, an evaluation study of the multiyear estimation method was conducted using simulations. The evaluations were made taking into account the number of units sampled each year, the number of years used for sample data, and the amount of overlap for the sample units. The results of this study given in Chhikara and Deng (1992) show that the multiyear estimation method can provide a substantial improvement over the single-year estimation procedure.

In the next section we briefly describe the various estimators currently used at NASS. We then discuss the multiyear approach for the area frame sampling and develop a new set of estimators similar to those presently used at NASS, but utilizing the multiyear area frame sampling data. The new estimation method was applied to survey data for four consecutive years, 1987, 1988, 1989 and 1990 to obtain separate estimates of soybean planted acres and total hogs in 1988, 1989, and 1990 for Indiana or Iowa, Ohio and Oklahoma. The numerical results are presented in this report.

CURRENT NASS ESTIMATORS

There were four distinct estimators considered by NASS in 1990: area tract, farm and weighted estimators, and a multiple-frame estimator. (The farm estimator has since been discontinued starting in June 1993.) The area tract estimator (A_1) is based on sampled inventories in the tracts, and the area farm estimator (A_2) is based on sampled inventories of the farms whose operators reside in the segment. The area weighted estimator (A_3) is derived from sampled inventories of the farms totally or partially in the sampled segments with the weight as the ratio of the tract acreage within the segment to the corresponding farm acreage. These three estimators are essentially developed from the area sampling frame supplemented by a list of large farm operators, called extreme operators. A multiple frame estimator (A_4) is obtained by combining the list frame estimator with the area frame estimator of the nonoverlap domain between the list and area sampling frames. Nealon (1984) provides a descriptive review of these estimators. Kuo (1989) and Perry, et al. (1989) studied certain composite estimators that combine the four estimators (A_1 through A_4).

The area tract estimator (A_1) of the total for a survey item is obtained by adding the data over all tracts in each sampled segment, multiplying this total by the expansion factor of the segment, adding up over all the sampled segments in the stratum, and then aggregating the stratum totals to the desired higher level. The area farm estimator (A_2) is obtained similarly. However, it uses the data only for the farm operators who reside within the sample segment boundaries. The area weighted estimator (A_3) is obtained in a manner similar to that used for the area tract estimator, except the tract data are prorated by multiplying the farm data by the ratio of a segment tract acreage to the corresponding farm acreage. All three estimators are unbiased. The area tract estimator (A_1) tends to be accurate for estimating crop acreages because it uses the accurately determined area tract data. However, it does not seem to provide a precise estimate for livestock. In fact, its use is not considered appropriate for estimating agricultural items that are associated with a farm operation. In such cases, the area farm estimator (A_2) may be preferable. However, this estimator tends to have large variability for some items and thus, leads to estimates that are not stable from year to year. An improvement upon this estimator is obtained from the use of the area weighted estimator (A_3). These area estimators can be of major concern in states with considerable amount of farmland devoted to noncrop areas. The multiple-frame estimator (A_4) is found to be the most efficient of four estimators.

To describe more precisely how each estimate is calculated at NASS, let D_1 be the non-overlapping domain consisting of the farm operators who are in the area frame but not in the list frame. Let D_2 be the overlapping domain with extreme operators which are on the list frame excluded, and let D_3 consist of extreme operators from the overlap domain. Since the area frame is complete, the set of domains D_1, D_2 and D_3 form a mutually exclusive and exhaustive partition of a survey population which is fully covered by the area frame. Bosecker and Ford (1976) discuss combining list and area estimates for a composite estimate for D_2. However, NASS does not currently composite estimates for the overlap domain.

The three estimators (A_1 through A_3) of the total for an agricultural item in a stratum are computed using the sample data from the area frame for domains D_1 and D_2, and the list frame data for domain D_3. The list frame data allow one to include most of the extreme operators in the sample and thus, to achieve higher efficiency for an estimator that otherwise may not be possible. All three estimators (A_1 through A_3) are of the form:

$$\hat{Y}_i = \hat{Y}_{D_1 \cup D_2, A_i} + \hat{Y}_{D_3, L}, \qquad i = 1, 2, 3. \tag{2.1}$$

The estimator $\hat{Y}_{D_1 \cup D_2, A_i}$ is defined for the joint domain which consists of the nonoverlapping domain D_1 and the overlapping domain D_2 that excludes the extreme operators. It is a direct expansion estimator computed as follows:

$$\hat{Y}_{D_1 \cup D_2, A_i} = \sum_{h \in H} \sum_{k=1}^{n_h} E_{hk} y_{i,hk} \tag{2.2}$$

where H is the collection of substrata, 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 $y_{i,hk}$ is the agricultural item value obtained in terms of estimator (A_i) as described earlier for segment k of stratum h and domains D_1 and D_2. The estimator $\hat{Y}_{D_3,L}$ is computed from the list frame samples in domain D_3 as

$$\hat{Y}_{D_3,L} = \sum_{l \in D_3} \frac{N_l}{n_l} \sum_{k=1}^{n_l} y_{lk} \tag{2.3}$$

where y_{lk} is the agriculture item value of farm k in stratum l of the list frame, and N_l is the total number of extreme operators and n_l is the corresponding sample size for the stratum.

Both the area frame and the list frame estimators are available for the overlapping domains D_2 and D_3, whereas only the area frame estimator is available for the nonoverlapping domain D_1. The individual domain estimates for an area frame can easily be computed using either area tract, or farm, or weighted data. A list frame estimator utilizes the farm data.

The area farm estimator (A_2) is based on inventory beyond the segment and thus, it is expected to be less accurate. The disadvantages of this estimator outweigh the advantages (Nealon, 1984, and Tortora, Ford and Nealon, 1984). Since the area farm estimator (A_2) is not as precise and stable as the weighted estimator (A_3), or the area tract estimator (A_1), NASS makes use of either estimator (A_1) or (A_3) in developing a multiple-frame estimator. For acreage estimation of major crops, the area tract estimator (A_1) is more precise than the area weighted estimator (A_3), and for livestock estimation, the area weighted estimator (A_3) is more precise than the area tract estimator (A_1).

Let $\hat{Y}_{D_j,L}$ denote the list estimator for domain D_j, similar to (2.3), except the summation is with respect to strata in domain $D_j, j = 2, 3$ and $\hat{Y}_{D_1,A_3}$ denote the area weighted estimator for domain D_1 as obtained by

$$\hat{Y}_{D_1,A_3} = \sum_{h \in H} \sum_{k=1}^{n_h} E_{hk} y_{3,hk}, \tag{2.4}$$

where $y_{3,hk}$ is the item value obtained in terms of the weighted estimator (A_3) for segment k of stratum h and domain D_1. The multiple-frame estimator (A_4) combines the area weighted estimator for NOL domain D_1 and the list estimators for the other two domains D_2 and D_3. This estimator is given by

$$\hat{Y}_4 = \hat{Y}_{D_1,A_3} + \hat{Y}_{D_2,L} + \hat{Y}_{D_3,L}. \tag{2.5}$$

Since the list frame is efficiently stratified using a variable related to the survey item, the multiple-frame approach generally yields more precise estimates, particularly for livestock, than any of the area estimators (A_1 through A_3). Nealon (1984) reviewed the four estimators based on practical considerations and discussed their advantages and disadvantages.

ESTIMATORS INCORPORATING MULTIYEAR ROTATION DESIGN

ANOVA Model.

Estimators presently used at NASS, with the exception of calculating a current year to previous year ratio estimate, do not incorporate any information from the rotation sampling design which has eighty percent overlap of sample segments from one year to the next. 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 tend to be invariant from year to year and the capabilities of certain operators in a segment to grow crops, etc. will tend to be persistent over years. On the other hand, factors such as weather and economic conditions will vary across years and will affect the farm outcome. To take into consideration the effect of these factors that persist over time, we propose the use of an analysis of variance (ANOVA) model approach for estimation of crop acreages or livestock that correctly avoids any unwarranted assumptions sometimes made in the analysis of other 'time series'. Intuitively, the idea behind the proposed multiyear ANOVA approach is that because of the factors that persist from year-to-year within segments, the variation of the crop acreages (or livestock numbers) from year-to-year for a particular sample segment will be less than the variance of the crop acreages between segments within a particular year.

To simplify the notation, we will suppress the subscript i used in (2.1)–(2.2) and will assume that a specified estimator is being considered. The stratum index h is also dropped. Let y_{tk} be the agricultural item value obtained in terms of the specified estimator in year t for segment k in a stratum. That is, y_{tk} represents the total over all tracts or farms in the segment. Note that y_{tk} is a segment-level value, not the farm-level value. Then the y_{tk} can be viewed as consisting of the stratum mean in year t, the kth segment effect, and an error component. One can, therefore, describe it in terms of an ANOVA model:

$$y_{tk} = \alpha_t + b_k + e_{tk} \tag{3.1}$$

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. Various 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 kth segment effect representing the deviation of the kth segment response from that of the stratum mean. e_{tk} is the model (unaccountable) error associated with segment k in year t . The error term in (3.1) 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 $\text{Var}(b_k) = \sigma_b^2$, and the e_{tk} are independently distributed with $E(e_{tk}) = 0$ and $\text{Var}(e_{tk}) = \sigma_e^2$. Furthermore, b_k and e_{tk} are independent of each other.

Estimation.

The above linear model can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{U}\mathbf{b} + \text{e}, \tag{3.2}$$

where $\mathbf{X}$ is the design matrix consisting of 0's and 1's which account for the effect due to $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_T)'$ and $\mathbf{U}$ is the design matrix of 0's and 1's which are specified according to the rotation sampling scheme and which account for the effect due to $\mathbf{b} = (b_1, b_2, ..., b_S)'$. Here we have relabelled the distinct b_k's in (3.1) as $b_1, b_2, ..., b_S$. The dimensions of $\mathbf{X}$ and $\mathbf{U}$ are $N \times T$ and $N \times S$, respectively, where $N = \sum_{t=1}^{T} n_t$ is the total number of segments and S is the total number of distinct segments sampled in T years. Note that $S \leq N$. When there is no overlap of segments between occasions, then $S = N$. Let

$$\mathbf{b}^* = \mathbf{U}\mathbf{b} + \mathbf{e}, \tag{3.3}$$

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

$$\begin{aligned}\text{Var}(\mathbf{b}^*) &= \mathbf{I}\sigma_e^2 + \mathbf{U}\mathbf{U}'\sigma_b^2 \\ &= \sigma_e^2(\mathbf{I} + \gamma\mathbf{U}\mathbf{U}') = \sigma_e^2\mathbf{W}_\gamma,\end{aligned} \tag{3.4}$$

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

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{W}_\gamma^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_\gamma^{-1}\mathbf{y}. \tag{3.5}$$

Appendix A provides a computation procedure for evaluating $\mathbf{W}_\gamma^{-1}$ and for the two terms, $\mathbf{X}'\mathbf{W}_\gamma^{-1}\mathbf{X}$ and $\mathbf{X}'\mathbf{W}_\gamma^{-1}\mathbf{y}$ that appear in Equation (3.5).

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. If it is necessary to estimate γ from the regular survey data, an estimator as described in Appendix B can be used to obtain $\hat{\gamma}$.

The variance-covariance matrix of $\hat{\boldsymbol{\alpha}}$ is given by

$$\text{Var}(\hat{\boldsymbol{\alpha}}) = (\mathbf{X}'\mathbf{W}_\gamma^{-1}\mathbf{X})^{-1}\sigma_e^2. \tag{3.6}$$

This variance formula still applies asymptotically if γ is replaced by a consistent estimator $\hat{\gamma}$. Given in Appendix B is a procedure for estimation of γ.

Single-year estimates currently used at NASS can be obtained by setting $\gamma = 0$ (i.e., no segment effect) in Equation (3.5) for $\hat{\boldsymbol{\alpha}}$, and are given by

$$\hat{\boldsymbol{\alpha}}_s = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, ...\bar{y}_{T\cdot})', \tag{3.7}$$

where $\bar{y}_{t\cdot}$ is the sample mean for year t. In order to evaluate the performance of $\hat{\boldsymbol{\alpha}}_s$ as an alternative to $\hat{\boldsymbol{\alpha}}$, one computes the variance-covariance matrix of $\hat{\boldsymbol{\alpha}}_s$ under model (3.2), which would be

$$\mathrm{Var}(\hat{\boldsymbol{\alpha}}_s) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}_\gamma\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\sigma_e^2. \tag{3.8}$$

From the generalized Gauss-Markov theorem, we have

$$\mathrm{Var}(\mathbf{c}'\hat{\boldsymbol{\alpha}}) \leq \mathrm{Var}(\mathbf{c}'\hat{\boldsymbol{\alpha}}_s), \tag{3.9}$$

for any vector $\mathbf{c}$. In particular, the multiyear estimator has a variance smaller than or equal to the variance of the single-year estimator for the current year mean obtained by taking $\mathbf{c}'$ to be the $1 \times T$ row vector,

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

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 $\hat{\alpha}_{T_s} = \mathbf{v}_T'\hat{\boldsymbol{\alpha}}_s$, respectively. Thus, the multiyear estimator $\hat{\alpha}_T$ is more efficient than the single-year estimator $\hat{\alpha}_{T_s}$, as one would expect because the proposed estimation procedure utilizes all sample data obtained under the multiyear rotation design. If there is no consistency in the individual segment totals over years, which would lead to small year-to-year segment effects b_k in the ANOVA model, then $\gamma \approx 0$ and $\mathbf{W}_\gamma \approx I$. This implies that the variances of the single-year and multiyear estimates are nearly the same and there is no gain from the multiyear approach. If individual segment totals are fairly consistent over years, but there is considerable variation between segments within a year, then $\gamma > 0$.

Remark: The misspecification of model parameter γ is an important issue. Analysis showed that there was minimal effect on the multiyear estimator due to misspecification of γ.

A Numerical Illustration.

Listed below are soybean acres for three years, 1987, 1988 and 1989, for the area segments sampled in a substratum using the NASS rotation sampling design. The data format is a two-way classification showing various sample segments versus the year(s) in which each segment is sampled. The sample means, $\bar{y}_{t.}$ for year t = 1987, 1988 and 1989 and $\bar{y}_{.k}$ for segment $1, 2, \ldots,$ and 14, are given on the two margins. The overall sample mean, $\bar{y}_{..} = 131.72$, is also shown.

Table 1: Total soybean acres for sample segments in a stratum during 1987-89

	YEAR			
Segment	1987	1988	1989	Mean $\bar{y}_{.k}$
1	92.00			92.00
2	121.10			121.10
3	180.00	151.60		165.80
4	122.00	119.90		120.95
5	140.90	144.80	122.50	136.07
6	139.90	116.00	134.00	129.97
7	119.00	86.50	134.00	113.17
8	156.70	58.30	220.50	145.17
9	125.60	144.40	230.00	166.67
10	95.00	134.00	126.70	118.57
11		121.50	98.50	110.00
12		147.50	137.30	142.40
13			46.00	46.00
14			185.70	185.50
Mean $\bar{y}_{t.}$	129.22	122.45	143.52	$\bar{y}_{..} = 131.72$

In terms of the ANOVA model given in Equation (3.1), a sample observation, say y_{tk}, in Table 1 can be described as a sum of three components: (1) the sample average for the year t, $\alpha_t = \bar{y}_{t.}$, (2) the segment effect, $b_k = \bar{y}_{.k} - \bar{y}_{..}$ due to segment k, and (3) the residual or unaccountable error, $e_{tk} = y_{tk} - \bar{y}_{t.} - \bar{y}_{.k} + \bar{y}_{..}$. Thus, in terms of these components, the sample observation can be expressed as follows:

$$y_{tk} = \bar{y}_{t.} + (\bar{y}_{.k} - \bar{y}_{..}) + (y_{tk} - \bar{y}_{t.} - \bar{y}_{.k} + \bar{y}_{..}),$$

$$k = 1, 2, ..., 14, \text{ and } t = 1987, 1988, 1989.$$

To illustrate numerically the ANOVA model, each observation is decomposed into the model components α_t, b_k, and e_{tk} as stated above. For example, considering the first observation 92.00 and its corresponding marginal totals, and the overall sample mean given in Table 1, the model component values are as follows:

$$\begin{aligned} 92 &= 129.22 + (92.00 - 131.72) + (92.00 - 129.22 - 92.00 + 131.72) \\ &= 129.22 - 39.72 + 2.50 \end{aligned}$$

Here the first component (α_t) value 129.22 is the sample mean for year 1987 in which the sample observation 92.00 was made, the second component (b_k) value -39.72 is the deviation of the sample mean for segment 1 from the overall sample mean as given in the last column in Table 1, and the error component (e_{tk}) value 2.50 is the residual that is left and remains unaccountable in terms of the year and segment effects. Other sample observations can similarly be described in terms of these three components of the ANOVA model. Accordingly, we have the following set of decompositions for the observations given in Table 1:

$$\begin{pmatrix} 92.00 \\ \vdots \\ 95.00 \\ 151.60 \\ \vdots \\ 147.50 \\ 122.50 \\ \vdots \\ 185.70 \end{pmatrix} = \begin{pmatrix} 129.22 \\ \vdots \\ 129.22 \\ 122.45 \\ \vdots \\ 122.45 \\ 143.52 \\ \vdots \\ 143.52 \end{pmatrix} + \begin{pmatrix} -39.73 \\ \vdots \\ -13.16 \\ 34.07 \\ \vdots \\ 10.67 \\ 4.34 \\ \vdots \\ 53.97 \end{pmatrix} + \begin{pmatrix} 2.51 \\ \vdots \\ -21.06 \\ -4.92 \\ \vdots \\ 14.38 \\ -25.36 \\ \vdots \\ -11.79 \end{pmatrix}$$

Altogether there are 30 observations listed in the left hand column. In terms of the ANOVA model (3.1), all thirty observations are to be jointly utilized to estimate the mean value α_t for the latest year, 1989. On the other hand, the sample average, $\bar{y}_{t.} = 143.52$

provides an estimate of the 1989 mean value for the stratum when sample data for that year are utilized.

The deviations b_k for the fourteen sample segments in Table 1 provide an estimate of σ_b^2 as $\hat{\sigma}_b^2 = 1196.28$ and the thirty residuals e_{tk} provide an estimate of σ_e^2 as $\hat{\sigma}_e^2 = 728.69$. Making use of the estimation procedure described in Appendix B but exercised at the substratum level, an initial estimate of γ is $\hat{\gamma} = 0.30$. This suggests that certain segment characteristics persist across three years and therefore the use of data for segments sampled in the previous two years 1987 and 1988 will augment the information contained in the sample data obtained in 1989. This, in turn, effects an increase in the sample size and thereby would reduce the variance for the estimate when based on multiyear data.

Table 2 lists the two estimates, one using the current single-year approach and the other using the proposed multiyear approach. In the latter case, the model parameter γ was estimated to be 0.012 using the procedure given in Appendix B and $\hat{\alpha}_t$ was computed using the formula in (3.5). The estimation of γ as described in Appendix B is considered at the landuse stratum level and not at the substratum level since this stabilizes the estimates of γ across substrata.

Table 2: Estimated soybean acres and their standard errors for the stratum

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

* Omitted because these are not relevant estimates since input of 1989 data should influence only the estimate for 1989 and not 1988 and 1987.

The results in Table 2 show that the estimates for this particular stratum under the two approaches are almost the same, yet there is a significant change in their standard errors. First, the standard error varies substantially from one year to another under the single-year approach. Secondly, the 1989 estimate under the multiyear approach has a significantly smaller standard error than under single-year approach. Moreover, the standard error under the multiyear approach is expected to be much less variable across years.

NUMERICAL RESULTS

Survey Data and Estimates.

We computed the 1988,1989 and 1990 estimates for total hogs and soybean acres separately for four states, Indiana, Iowa, Ohio and Oklahoma, using list and area frame JAS data for 1987, 1988, 1989 and 1990. The 1988 multiyear estimates utilize data for two years, 1987 and 1988; the 1989 estimates utilize 1987, 1988 and 1989 data, and the 1990 estimates are made using all four years, except in Iowa where the segments were not sampled using a rotation design in all four years because a new area frame was implemented for Iowa in 1989. Hence, the multiyear estimates for Iowa are computed for 1990 using the sample data from 1989 and 1990. Moreover, 1990 estimates were not made for Indiana due to the change in the sample design between 1989 and 1990 so that no segments were overlapped.

We restrict our computations to those estimators that are of major utility to NASS. Some of these estimators differ slightly in form from those described in Section 2 as discussed later. We also do not make any comparison with the current year to previous ratio estimates since the present NASS ratio estimator does not utilize the historic data in the same manner that we have suggested here.

The multiyear 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 the two items of planted soybean acres and total hogs where all the codes and variables necessary for their 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 as depicted for example, in Table 1.

For the four states, the soybean acres and total hogs 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 defined by

$$\text{R. E.} = \text{Var}(\hat{\alpha}_{T_s})/\text{Var}(\hat{\alpha}_T).$$

where $\text{Var}(\hat{\alpha}_{T_s})$ is the variance of the single-year estimator $\hat{\alpha}_{T_s}$ and $\text{Var}(\hat{\alpha}_T)$ is the variance of the multiyear estimator $\hat{\alpha}_T$. Here subscript T stands for the year for which an estimate is desired.

For soybean acres, we give estimates in Tables 3, 4 and 5 corresponding to the area tract estimator (A_1). There are two estimators developed using the tract data; one utilizes strictly the sampled tract data, and is denoted as "Area Tract", which is slightly different from the previously defined estimator (A_1), and the other one is the multiple-frame estimator denoted as "NOL+List", where the first component is the area tract estimator for the nonoverlapping domain D_1 and the second component is developed from the list frame data for domains D_2 and D_3. Note that the list component of the multiple-frame estimator will be the same for both the multiyear and single-year estimates since the list frame sampling does not involve any over years rotation design. Thus, relative efficiency is not computed and shown separately for the area and list components of the multiple-frame estimator.

Table 3: 1988 Soybean Acres Estimates (in thousands).

Frame	Estimator	Single-year approach		Multiyear approach		R. E.
		Estimate	S. E.	Estimate	S. E.	
Ohio						
Area	Area Tract	3938.8	175.5	3924.9	171.1	1.05
Area	NOL	865.7	106.1	821.1	105.9	
List	List	3003.1	40.3	3003.1	40.3	
Multiple	NOL + List	3868.8	113.5	3824.3	113.3	1.01
Oklahoma						
Area	Area Tract	351.3	75.2	340.3	69.6	1.17
Area	NOL	131.5	42.3	130.6	34.3	
List	List	207.5	39.6	207.5	39.6	
Multiple	NOL + List	339.0	57.9	338.0	52.4	1.22
Indiana						
Area	Area Tract	4147.8	176.9	4137.1	164.0	1.16
Area	NOL	799.1	93.2	797.9	99.8	
List	List	4399.0	157.0	4399.0	157.0	
Multiple	NOL + List	5198.1	182.6	5196.9	186.0	0.96

Table 4: 1989 Soybean Acres Estimates (in thousands).

Frame	Estimator	Single-year approach		Multiyear approach		R. E.
		Estimate	S. E.	Estimate	S. E.	
Ohio						
Area	Area Tract	4075.6	174.2	4055.6	167.3	1.08
Area	NOL	902.3	102.5	861.0	98.2	
List	List	3266.5	125.0	3266.5	125.0	
Multiple	NOL + List	4168.8	161.6	4127.5	159.0	1.03
Oklahoma						
Area	Area Tract	351.05	83.6	318.0	67.3	1.55
Area	NOL	140.3	47.9	128.2	34.8	
List	List	155.8	26.8	155.8	26.8	
Multiple	NOL + List	296.1	54.9	284.0	43.9	1.56
Indiana						
Area	Area Tract	4509.9	178.2	4517.3	163.9	1.18
Area	NOL	755.8	90.2	753.5	96.5	
List	List	4495.5	133.5	4495.5	133.5	
Multiple	NOL + List	5251.3	161.1	5249.0	164.7	0.96

Table 5: 1990 Soybean Acres Estimates (in thousands).

Frame	Estimator	Single-year approach		Multiyear approach		R. E.
		Estimate	S. E.	Estimate	S. E.	
Ohio						
Area	Area Tract	3344.1	162.2	3356.9	157.6	1.06
Area	NOL	638.4	72.4	652.4	94.5	
List	List	2919.7	102.2	2919.7	102.2	
Multiple	NOL + List	3558.1	125.2	3572.0	139.2	0.81
Oklahoma						
Area	Area Tract	233.1	72.4	221.6	63.8	1.29
Area	NOL	58.5	28.8	44.8	32.7	
List	List	199.0	29.0	199.0	29.0	
Multiple	NOL + List	257.5	40.9	243.8	43.7	0.88
Iowa						
Area	Area Tract	7814.2	228.2	7814.2	244.9	0.87
Area	NOL	1536.2	127.9	1536.2	142.5	
List	List	7268.6	181.8	7268.6	181.8	
Multiple	NOL + List	8804.8	222.3	8804.8	231.0	0.93

Tables 6, 7 and 8 provide the total hogs estimates. The estimates that are of primary interest are area tract ($A1$), area weighted ($A3$), and NOL weighted multiple frame ($A4$). These were computed using the three estimators, tract, weighted and multiple-frame estimators denoted as "Tract", "Weighted", and "NOL + List", respectively. The area tract estimate for $D_1 \cup D_2$ is computed using the tract hogs whereas the other two estimates were computed for the two domains D_1 and D_2 using the weighted sample data for total hogs on the entire farm. The EO part of the list estimate (D_3) was computed using data for the total hogs and was combined with other domain estimates for all three estimates. Because the tract hogs data are not considered as precise as the data for total hogs, USDA relies more heavily on the weighted or multiple-frame estimator to finalize its estimate of the total hogs in a state.

Table 6: 1988 Hog Estimates (in thousands).

Frame	Estimator	Single-year approach		Multiyear approach		R. E.
		Estimate	S. E.	Estimate	S. E.	
Ohio						
Area	Area Tract	1473.2	401.2	1485.6	353.9	
	NOL	526.4	141.8	469.6	140.0	
	NOL+NEO	1575.9	227.9	1595.3	204.8	
List	List	1858.8	116.5	1858.8	116.5	
	EO	390.3	22.1	390.3	22.1	
Multiple	Tract	1863.5	401.8	1875.8	354.6	1.28
	Weighted	1966.1	229.0	1985.6	206.0	1.24
	NOL + List	2385.2	183.5	2328.4	182.2	1.01
Oklahoma						
Area	Area Tract	203.0	78.4	203.4	87.9	
	NOL	67.1	23.7	64.4	20.1	
	NOL+NEO	119.9	35.8	129.2	33.5	
List	List	125.2	8.0	125.2	8.0	
	EO	59.8	0.7	59.8	0.7	
Multiple	Tract	262.8	78.4	263.2	87.9	0.80
	Weighted	179.7	35.8	189.0	33.5	1.14
	NOL + List	192.3	25.0	189.6	21.6	1.33
Indiana						
Area	Area Tract	2388.1	439.2	2405.5	428.4	
	NOL	528.2	107.8	526.7	101.4	
	NOL+NEO	2695.3	272.8	2637.2	441.6	
List	List	1858.8	116.5	1858.8	116.5	
	EO	1521.4	49.9	1521.4	49.9	
Multiple	Tract	3909.6	442.0	3926.9	431.3	1.05
	Weighted	2387.0	158.7	2385.5	154.4	1.06
	NOL + List	4216.7	277.3	4158.7	444.4	0.39

Table 7: 1989 Hog Estimates (in thousands).

Frame	Estimator	Single-year approach		Multiyear approach		R. E.
		Estimate	S. E.	Estimate	S. E.	
Ohio						
Area	Area Tract	2088.1	604.4	2024.8	414.4	
	NOL	605.8	274.9	563.4	176.8	
	NOL+NEO	1680.8	324.4	1682.9	227.6	
List	List	1453.3	78.2	1453.3	78.2	
	EO	439.5	32.8	439.5	32.8	
Multiple	Tract	2527.7	605.3	2464.3	415.7	2.12
	Weighted	2120.3	326.0	2122.5	229.9	2.01
	NOL + List	2059.1	285.8	2016.6	193.3	2.19
Oklahoma						
Area	Area Tract	273.5	121.4	284.9	90.3	
	NOL	49.7	19.1	43.6	19.9	
	NOL+NEO	81.7	28.6	99.2	29.8	
List	List	165.2	21.8	165.2	21.8	
	EO	70.4	0.7	70.4	0.7	
Multiple	Tract	343.9	121.4	355.3	90.3	1.81
	Weighted	152.1	28.6	169.7	29.9	0.92
	NOL + List	215.0	28.9	208.9	29.5	0.97
Indiana						
Area	Area Tract	2147.9	535.7	2068.8	448.6	
	NOL	1015.5	710.1	941.9	391.4	
	NOL+NEO	2945.3	733.7	2832.2	534.4	
List	List	3970.7	143.0	3970.7	143.0	
	EO	1543.5	54.6	1543.5	54.6	
Multiple	Tract	3691.3	538.5	3612.3	451.9	1.42
	Weighted	4488.8	735.8	4375.7	537.2	1.88
	NOL + List	4986.2	724.3	4912.6	416.7	3.02

Table 8: 1990 Hog Estimates (in thousands).

Frame	Estimator	Single-year approach		Multiyear approach		R. E.
		Estimate	S. E.	Estimate	S. E.	
Ohio						
Area	Area Tract	2118.8	508.4	2103.2	461.7	
	NOL	663.2	262.2	665.5	247.8	
	NOL+NEO	1785.2	315.7	1795.5	300.8	
List	List	1349.2	60.0	1349.2	60.0	
	EO	411.7	23.7	411.7	23.7	
Multiple	Tract	2530.4	509.0	2514.9	462.3	1.21
	Weighted	2196.9	316.6	2207.2	301.8	1.10
	NOL + List	2012.3	269.0	2014.6	255.0	1.11
Oklahoma						
Area	Area Tract	207.6	91.0	215.1	93.3	
	NOL	45.4	18.5	42.0	18.3	
	NOL+NEO	45.5	18.5	55.2	27.2	
List	List	126.4	8.2	126.4	8.2	
	EO	76.9	0.4	76.9	0.4	
Multiple	Tract	284.4	91.0	292.0	93.3	0.95
	Weighted	122.4	18.5	132.1	27.2	0.46
	NOL + List	171.4	20.2	168.3	20.0	1.02
Iowa						
Area	Area Tract	11573.7	1252.0	11573.7	1348.0	
	NOL	2629.8	405.3	2629.8	425.6	
	NOL+NEO	12434.7	838.1	12434.6	886.2	
List	List	1226.5	61.8	1226.5	61.8	
	EO	10772.4	280.5	10772.4	280.5	
Multiple	Tract	12800.2	1253.5	12800.2	1391.4	0.81
	Weighted	13661.1	840.4	13661.1	888.4	0.89
	NOL + List	13402.2	492.9	13402.2	509.7	0.93

Discussion.

The multiyear estimates are in fair agreement with the single-year estimates for both soybean acres and total hogs for all three years, 1988, 1989, 1990. The variances, and hence the relative efficiencies do vary considerably. This is primarily due to the instability of the variance estimate under the single-year estimation method. The computed relative efficiencies of the area tract estimates are higher than the corresponding relative efficiencies obtained for the multiple-frame estimates. This is expected because the multiple frame estimator utilizes the multiyear sample data only for the nonoverlap domain D_1.

In Table 6, the 1988 multiple frame estimate of total hogs in Indiana has a very low R. E. value of 0.39. This is primarily due to one sample observation in the NOL+NEO domain that had an extremely large expanded value influencing the multiyear variance significantly.

It may be pointed out that the computed relative efficiency for the multiple-frame estimate is less than 1.0 in several cases, implying a higher variance under the multiyear estimation method. A low value of computed R. E. is because of a large yearly variation in sample standard errors for the single-year estimates for NOL domain D_1, where the single-year estimate has a considerably under-estimated standard error corresponding to one year than other years. For example, this can be seen from the earlier illustration as shown in Table 2. So the R. E. value of 0.39 in Table 6 as pointed above does not necessarily imply that multiyear estimate is inefficient. The multiyear estimate has a much more stable, and hence reliable yearly variance estimate than does the NASS single-year estimate and thus, providing an additional advantage in the use of the multiyear estimator.

A comparison between the use of two years of survey data for the 1988 estimates, three years data for the 1989 estimates, and four years data for the 1990 estimates shows that the gain in relative efficiencies does not necessarily improve with more years of survey data. For example, the computed relative efficiencies for 1990 estimates are generally lower than for 1989 estimates for both Ohio and Oklahoma. But overall the multiyear approach provides a more efficient estimate than does the current approach with three as the optimum number of years to use.

Notes: (1) Since data were collected in 1990 for all but one of 1989-90 segments, the 1990 Iowa estimates are essentially the same under both approaches. The Iowa area frame sample allocation was increased in 1990 from 308 to 395 by rotating in 88 segments and only rotating out one segment. (2) The total hogs R. E. is computed only for the multiple-frame estimates since these are the main estimates used by NASS.

EFFECT OF ERROR IN ESTIMATING γ

As remarked earlier, the misspecification of model parameter γ is an important issue. Presently, we have estimated γ at the landuse stratum level in obtaining a multiyear estimate. This estimation of γ was carried out for each landuse stratum using the analysis of variance method as described in Appendix B. This estimate of γ was used for each substratum in the landuse stratum. 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 and the corresponding standard errors were computed by varying $\hat{\gamma}$. Tables 9 and 10 show the computed estimates and standard errors of different multiyear estimators corresponding to various values of $\hat{\gamma}/\gamma$, where $\hat{\gamma}$ is the value used in the estimation procedure.

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. A comparison of both the soybean acres and hogs estimates in all three states for $c = 0.5, 1.5$ and 2.0 versus $c = 1.0$, which corresponds to the case of no misspecification, shows that these estimates hardly vary. Similar is the case with respect to their standard errors. Hence, the multiyear estimates display a high level of robustness to misspecification of model parameter γ.

Table 9: 1989 Multiyear Estimates Of Soybean Acres With γ Misspecified As $\hat{\gamma}(= c\gamma)$

	c				
Estimator	0.0	0.5	1.0	1.5	2.0
			Ohio		
Area Tract	4076	4056	4056	4058	4060
	(179)	(168)	(167)	(169)	(172)
NOL	902	870	861	857	856
	(110)	(98)	(98)	(100)	(103)
			Oklahoma		
Area Tract	351	325	318	314	312
	(76)	(67)	(67)	(69)	(72)
NOL	140	130	128	127	127
	(41)	(35)	(35)	(36)	(37)
			Indiana		
Area Tract	4510	4513	4517	4520	4521
	(172)	(165)	(164)	(165)	(167)
NOL	756	755	754	753	752
	(98)	(97)	(97)	(97)	(98)

Table 10: 1989 Multiyear Estimates Of Hogs With γ Misspecified As $\hat{\gamma}(= c\gamma)$

	c				
Estimator	0.0	0.5	1.0	1.5	2.0
			Ohio		
Area Tract	2088	2041	2025	2017	2012
	(463)	(416)	(414)	(423)	(436)
NOL	606	571	563	560	559
	(204)	(177)	(177)	(181)	(187)
NOL+NEO	1681	1680	1683	1685	1687
	(261)	(229)	(228)	(233)	(240)
			Oklahoma		
Area Tract	273	281	285	287	288
	(104)	(88)	(90)	(94)	(98)
NOL	50	46	44	42	42
	(21)	(20)	(20)	(20)	(20)
NOL+NEO	82	95	99	101	102
	(34)	(30)	(30)	(31)	(32)
			Indiana		
Area Tract	2148	2095	2069	2053	2043
	(478)	(447)	(449)	(456)	(465)
NOL	1015	971	942	927	919
	(418)	(397)	(391)	(391)	(392)
NOL+NEO	2945	2874	2832	2808	2792
	(560)	(535)	(534)	(540)	(547)

* values in parentheses are for standard errors.

CONCLUSIONS

For the rotation design sampling as implemented at NASS, a multiyear approach to estimation of crop acreages and livestock inventory is evaluated. A multiyear estimation method is developed based on a two-way analysis of variance model. We have given a set of estimators similar to those currently used at NASS and show these multiyear estimators to be more efficient and robust. The proposed model-based approach is easy to implement as demonstrated here by the application made using the actual NASS survey data.

RECOMMENDATION

The multiyear estimation methodology shows potential for improving upon the current NASS estimates based on area frame sampling. The study clearly suggests that consideration should be given to implementation of the proposed multiyear estimation method and evaluate it further for five to ten major agricultural items. It will be desirable to perform evaluations on a workstation under the UNIX operating system available at NASS. Appendix C provides documentation of the SAS program prepared to compute multiyear area frame estimates. The list component of an estimate is to be computed as presently done using only the current survey year data.

REFERENCE

Bosecker, R. R. and B. L. Ford (1976), "Multiple frame estimation with stratified overlap domain, " in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 219-224.

Chhikara, R. S., C. R. Hallum and T. G. Lythuan-Lee (1984), "Sampling designs for Landsat crop surveys, " *Communications in Statistics, A* **13**, 2905-2935.

Fecso, R., R. D. Tortora and F. A. Vogel (1985), "Recent and future activities to improve sampling frames for agriculture, " in *Proceedings of the Survey Research Method Section, American Statistical Association*, pp. 132-141.

Gleason, C. P. and R. D. Tortora (1978), "Successive sampling of two overlapping frames, " in *Proceedings of the Survey Research Method Section, American Statistical Association*, pp. 320-324.

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

Hartley, H. O. (1962), "Multiple frame surveys, " in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 203-206.

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

Kott, P. S. (1989), "Unbiased estimation using past survey data, " Unpublished manuscript.

Kuo, L. (1986), "Composite estimation of totals for livestock surveys, " SRS Staff Report No. SF&SRB-92, Statistical Reporting Service, U. S. Department of Agriculture, Washington, D. C.

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

Lythuan-Lee, T. G. (1981), "Development of rotation sampling design for the estimation of crop acreages, " AgRISIARS Report FC-LI-04155, JSC-17421, JCS, Houston,Texas, September, 1981.

Nealon, J. P. (1984), "Review of the multiple and area frame estimators, " SRS Staff Report No. 80, Statistical Reporting Service, U. S. Department of Agriculture, Washington, D. C.

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

Perry, C. R., P. W. Cook, M. L. Holko and S. D. Wiyatt (1989), "Composite estimators in the June hog series, " NASS Staff Report No. SRB-89-06, U. S. Department of Agriculture, Washington, D. C.

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.

Searle, S. R. (1982), *Matrix Algebra Useful for Statistics*, John Wiley & Sons, New York.

Sielken, Jr., R. L. and E. E. Gbur (1984), "Multiyear, through the season crop acreage estimation using estimated acreage in sample segments, " *Communications in Statistics, A* **13**, 2961-2974.

Thomas, D. R., C. R. Perry and B. Viroonsri (1990), "Estimation of totals for skewed populations in repeated agricultural surveys: hogs and pigs, " NASS Staff Report No. SRB-90-02, U. S. Department of Agriculture, Washington, D. C.

Tortora, R. D., B. L. Ford and J. P. Nealon (1984), "Comparison of area frame estimators in agricultural surveys: sampling versus nonsampling errors, " Unpublished manuscript.

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

APPENDIX A. COMPUTATION OF $\mathbf{W}_\gamma$

Note that the dimension of $\mathbf{W}_\gamma$ defined in Equation (3.4) is $N \times N$, which, in general, is a large matrix, especially when T is large. The direct computation of the inverse of $\mathbf{W}_\gamma$ may present a numerical problem. Using the special structure of $\mathbf{W}_\gamma$ given in Equation (3.4), we can find the inverse as following: Using the matrix inversion formula (see Searle, 1982, p. 151)

$$(\mathbf{I} + \mathbf{AB})^{-1} = \mathbf{I} - \mathbf{A}(\mathbf{I} + \mathbf{BA})^{-1}\mathbf{B},$$

we have

$$\begin{aligned}
\mathbf{W}_\gamma^{-1} &= \mathbf{I} - \gamma\mathbf{U}(\mathbf{I} + \gamma\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}' \\
&= \mathbf{I} - \gamma\mathbf{U}(\mathbf{I} + \gamma\mathrm{diag}(m_1, ..., m_S))^{-1}\mathbf{U}' \\
&= \mathbf{I} - \mathbf{U}\mathbf{D}_1\mathbf{U}', \qquad \text{(A.1)}
\end{aligned}$$

where m_k is the total number of yearly survey periods for which segment k falls in the sample and the diagonal matrix,

$$\mathbf{D}_1 = \mathrm{diag}(\frac{\gamma}{1 + \gamma m_1}, \frac{\gamma}{1 + \gamma m_2}, \cdots \frac{\gamma}{1 + \gamma m_S}).$$

Making use of (A.1), one can simplify the matrix expressions in Equations (3.5) and (3.6) as follows.

$$\begin{aligned}
\mathbf{X}'\mathbf{W}_\gamma^{-1}\mathbf{X} &= \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{U}\mathbf{D}_1\mathbf{U}'\mathbf{X} \\
&= \mathbf{D}_2 - \mathbf{X}'\mathbf{U}\mathbf{D}_1\mathbf{U}'\mathbf{X},
\end{aligned}$$

where

$$\mathbf{D}_2 = \mathrm{diag}(n_1, ..., n_T).$$

Clearly, this would avoid a direct tedious computation of $\mathbf{W}_\gamma^{-1}$. Similarly,

$$\mathbf{X}'\mathbf{W}_\gamma^{-1}\mathbf{y} = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{U}\mathbf{D}_1\mathbf{U}'\mathbf{y}$$

$$= \begin{pmatrix} y_{1.} \\ . \\ . \\ . \\ y_{T.} \end{pmatrix} - \mathbf{X}'\mathbf{U} \begin{pmatrix} d_1 y_{.1} \\ . \\ . \\ . \\ d_S y_{.S} \end{pmatrix}$$

where $y_{t.}$ is the total of y's for year t, $y_{.k}$ is the total of y's for segment k, and $d_i = \frac{\gamma}{1+\gamma m_i}, i = 1, 2, ..., S$.

APPENDIX B. ESTIMATION OF γ

Note that the multiyear estimator $\hat{\alpha}_T$ is the best linear unbiased estimator(BLUE) of α_T, provided γ is specified correctly in matrix $\mathbf{W}_\gamma$. In general, γ is unknown and needs to be estimated from some survey data. We will use $\hat{\boldsymbol{\alpha}}(\gamma_0)$, and $\hat{\alpha}_T(\gamma_0)$ for the latest period, to denote the multiyear estimators when γ is specified as γ_0.

If it is necessary to estimate γ using the regular survey data, one can first get an initial estimate $\hat{\gamma}$ from the multiyear rotation sampling design data using the two-way analysis of variance method. This is done first by computing the mean square due to segment and the mean square due to error as follows:

$$\mathrm{MS}_b = \frac{\sum_{k=1}^{S} m_k(\bar{y}_{.k} - \bar{y}_{..})^2}{S-1}$$

and

$$\mathrm{MS}_e = \frac{\sum_{t=1}^{T} \sum_{k\in S_t} (y_{tk} - \bar{y}_{.k} - \bar{y}_{t.} + \bar{y}_{..})^2}{N - S - T + 1},$$

where S is the number of distinct segments sampled in T years, S_t is the set of sampled segments in the tth year, m_k is the number of times the kth segment is observed in T years, $\bar{y}_{.k}$ is the sample average for segment k, and $\bar{y}_{..}$ is the grand average of all segments in T years. Although this computation of MS_e is most suitable for a full randomized block design, we use it here to compute simply an initial estimate of γ.

When there is a full randomized block design, it is well-known from the random-effect ANOVA model that

$$E(\mathrm{MS}_b) = \sigma_e^2 + T\sigma_b^2.$$

Accordingly, a reasonable estimate of σ_b^2 may be obtained by

$$\hat{\sigma}_b^2 = \frac{1}{T'}[\mathrm{MS}_b - \mathrm{MS}_e]_+,$$

where $[x]_+ = \max(x, 0)$ and $T' = N/S$. Since $N = \sum_{t=1}^{T} n_t$, T' represents on average the number of years each segment is observed. Hence, an initial estimate of γ is obtained by

$$\hat{\gamma} = \frac{1}{T'}\left[\frac{\mathrm{MS}_b}{\mathrm{MS}_e} - 1\right]_+ .$$

APPENDIX C: SAS PROGRAM TO COMPUTE THE MULTIYEAR AREA FRAME ESTIMATES

This SAS program is set up to compute the multiyear area frame estimates for soybeans using Ohio area tract data from four years. It can easily be modified for other states and/or agricultural items by appropriately changing the state and item names in the program.

The software program consists of several parts organized in the order suggested by the multiyear estimation methodology. First, it computes the expanded value of an item for each sampled segment in different years and then merges the expanded sample data into a single data set in the form of a two-way classification. This is followed by computation of the model parameter γ at the landuse stratum level. The estimated value $\hat{\gamma}$ is applied to each substratum to compute the estimate of the item total and its variance for substrata. The stratum total estimates are obtained and then aggregated to compute the state level estimates.

The program also computes the estimates and their variance estimates under the current NASS procedure for the area sampling frame.

Next, the program for the list frame estimates is appended so that the multiple frame estimates are computed analogous to the current NASS method.

```
options mprint symbolgen nodate
linesize=131 /*80 132 168*/ pagesize=60;

LIBNAME SAVE '.';

/* ohi SOYABEAN AREA TRACT 4 YEARS
   MEXPFCTR OF YEAR 90           */


/*
     The following parts of program compute the value of YR87
     sampled in 1987 of every segment for every stratum.
*/

DATA CROSS87;
RENAME SEGMENT=SEG;
SET SAVE.DSJ7ohi;              /* input data file */
IF STRATAY >= 1101;
IF 1 <= LHHOLNOL <= 2 then WT=CSOYXXPL*MEXPFCTR;
ELSE WT=0;
PROC MEANS NOPRINT;
     CLASS  SEG;
     BY STRATAY;
     VAR WT;
     OUTPUT OUT=WTYR87 SUM=YR87;

DATA WT87;                     /* output data set */
SET WTYR87;
IF ( SEG=. ) THEN DELETE;

/*
     The following parts of program compute the value of YR88
     sampled in 1988 of every segment for every stratum.
*/

DATA CROSS88;
RENAME SEGMENT=SEG;
SET SAVE.DSJ8ohi;
IF STRATAY >= 1101;
IF 1 <= MOLNOLAC <= 2 then WT=CSOYXXPL*MEXPFCTR;
ELSE WT=0;
PROC MEANS NOPRINT;
     CLASS  SEG;
     BY STRATAY;
     VAR WT;
```

```
     OUTPUT OUT=WTYR88 SUM=YR88;

DATA WT88;
SET WTYR88;
IF ( SEG=. ) THEN DELETE;

/*
     The following parts of program compute the value of YR89
     sampled in 1989 of every segment for every stratum.
*/

DATA CROSS89;
RENAME SEGMENT=SEG;
SET SAVE.DSJ9ohi;          /* input data file */
IF STRATAY >= 1101;
IF 1 <= MOLNOLAC <= 2 then WT=CSOYXXPL*MEXPFCTR;
ELSE WT=0;
PROC MEANS NOPRINT;
     CLASS  SEG;
     BY STRATAY;
     VAR WT;
     OUTPUT OUT=WTYR89 SUM=YR89;

DATA WT89;                /* output data set */
SET WTYR89;
IF ( SEG=. ) THEN DELETE;




/*
     The following part of program compute the value of YR90
     sampled in 1990 of every segment for every stratum.
*/

DATA CROSS90;
RENAME SEGMENT=SEG;
SET SAVE.DSJ0ohi;        /* input data file */
IF STRATAY >= 1101;
IF 1 <= MOLNOLAC <= 2 THEN WT=CSOYXXPL*MEXPFCTR;
   ELSE WT=0;
PROC MEANS NOPRINT;
     CLASS  SEG;
     BY STRATAY;
     VAR WT;
     OUTPUT OUT=WTYR90 SUM=YR90;

DATA WT90;                /* output data set */
SET WTYR90;
IF ( SEG=. ) THEN DELETE;

/*
     The following parts of the program find the expansion factor
     for every stratum.  The minimal value of variable MEXPFCTR
```

```
     of every stratum is chosen if there are more than one values
     of this variable. Data set F2 contains the minimum value of
     MEXPFCTR. Data set F3 contains the sorted result.
*/

DATA F1 (KEEP=MEXPFCTR STRATAY);
SET CROSS90;
PROC MEANS NOPRINT;
BY STRATAY;
VAR MEXPFCTR;
OUTPUT OUT=F2 MIN= ;

DATA F3 (KEEP=STRATAY MEXPFCTR);
SET F2;
PROC SORT DATA=F3 NODUPLICATES;
BY STRATAY;
    PROC PRINT DATA=F3;

/*
     This part of program merges the data of different years into
     one data set called W1.
*/

DATA YEAR87(KEEP=SEG STRATAY );
SET CROSS87;

DATA YEAR88(KEEP=SEG STRATAY );
SET CROSS88;

DATA YEAR89(KEEP=SEG STRATAY );
SET CROSS89;

DATA YEAR90(KEEP=SEG STRATAY );
SET CROSS90;

DATA FINAL;
SET YEAR87 YEAR88 YEAR89 YEAR90;
PROC SORT DATA=FINAL NODUPLICATES;
BY STRATAY  SEG;

PROC MEANS NOPRINT;
BY STRATAY;
VAR SEG;
OUTPUT OUT=A1 N=NN;  /* N means number of segments in every
stratum */

DATA A2 (KEEP=STRATAY NN);
SET A1;
IF (STRATAY=.) THEN DELETE;

DATA A2(KEEP=STRATAY NN);
MERGE F3 A2;
BY STRATAY;
```

```
IF MEXPFCTR=. THEN DELETE;


DATA A3;
SET A2;
STRAT=INT(STRATAY/100)*100;
RUN;

PROC SORT DATA=A3;
BY STRAT STRATAY;
PROC MEANS NOPRINT;
BY STRAT;
VAR STRATAY;
OUTPUT OUT=A4 N=NSTUM; * NSTUM is number of STRATAY in stratum;
RUN;

DATA A5(KEEP=STRAT NSTUM);
SET A4;
IF STRAT=. THEN DELETE;
RUN;

DATA W1 (KEEP=STRATAY SEG YR87 YR88 YR89 YR90);
MERGE WT87 WT88 WT89 WT90;  /* merges the data in dataset W1 */
BY STRATAY SEG;

DATA W1;
MERGE F3 W1;
BY STRATAY;
IF MEXPFCTR=. THEN DELETE;
DROP MEXPFCTR;

PROC IML;

/*
     This subroutine computes the initial estimate of gamma.  The
     input is a matrix Y which includes the information of one
     stratum.  The output is the estimate of initial gamma.
*/

START GAMI(Y);
W2=Y;
C=NCOL(W2);        /*gets number of columns of W2 */
WW=J(NROW(W2),C-2,0);   /* creates matrix WW */
WW=W2[ ,3:C];      /* cuts out first two columns */
T=C-2;             /* T represents number of years */
S=NROW(WW);        /* S represents number of different segments */
MI=J(S,T,-1)#WW;
N1=(MI>0);
N2=N1[+, ];     /* N2 means number of missing value in each year
*/
N3=J(1,T,S);
NT=N3-N2;       /* NT means number segments having values in year */
N=NT[ ,+];
N5=N1[ ,+];
```

```
MK=J(S,1,T)-N5;
M1=J(S,T,1)-N1; /* M1=X ,the design matrix */
WT=M1#WW;                    /* WT=MATRIX CORRESPONDING TO Y */
SEGAVR=WT[ ,+]/MK;           /* SEGMENT AVERAGE OF Y */
WT1=WT[+, ];
WT2=WT1[ ,+];
YAVR=WT2/N;                   /* GRAND AVERAGE OF Y*/
WT3=WT[+, ];
YT=J(1,T,0);
DO I=1 TO T;
 IF NT[1,I]>0 THEN YT[1,I]=WT3[1,I]/NT[1,I];
END;
MS1=SHAPE(YT,S,T);
MS2=SHAPE(SEGAVR`,T,S);
MS2=MS2`;
MS3=J(S,T,YAVR);
MS=WT-MS1-MS2+MS3;
MS=M1#MS;
MS=MS#MS;
MS=MS[ ,+];
MS=MS[+, ];
GG=J(1,6,0);
GG[1,1]=W2[1,1];
GG[1,2]=MS;         * SSE;
GG[1,3]=N-S-T+1;    * N-S-T+1= D.F. OF SSE;
S1=SEGAVR-J(S,1,YAVR);
S2=S1#S1;
S3=MK#S2;
SSB=S3[+, ];
GG[1,4]=SSB;
GG[1,5]=S-1;    * S=# OF SEG ;
GG[1,6]=T;
RETURN(GG); * GG=[STRATAY SSE DFE SSB DFB YEARS];
FINISH;



/*   This subroutine calculates the estimate of alpha and
     variance of it.  The input are a matrix Y which includes the
     information of the stratum and the estimate of gamma. The
     output is a matrix Z which includes the data of the last
     year.
     Z[1,1]=variance of alpha corresponding to the last year
     Z[1,2]=alpha corresponding to the last year
     Z[1,3]=the last diagonal value of inverse of X'WX
     Z[1,4]=estimate of varinace of error (Se2)
*/

START VARIAN(Y,GAMMA);
M1=Y;
C=NCOL(M1);
MM=M1[ ,3:C];
SS=NROW(MM);
TT=C-2;
```

```
MI=J(SS,TT,-1)#MM;
N1=(MI>0);
N2=N1[+, ];
N3=J(1,TT,SS);
NT=N3-N2;
N=NT[ ,+];
N5=N1[ ,+];
MK=J(SS,1,TT)-N5;
M1=J(SS,TT,1)-N1;
WT=M1#MM;
UX=M1;
XU=M1`;
D1=I(SS);
DO I=1 TO SS;
  D1[I,I]=GAMMA/(1+GAMMA*MK[I,1]);
END;
D2=I(TT);
DO I=1 TO TT;
  D2[I,I]=NT[1,I];
END;
XWX=D2-XU*D1*UX;
DO I=1 TO TT;
  IF NT[1,I]=0 THEN XWX[I,I]=0.01;
END;
A=DET(XWX);
IF A^=0 THEN IXWX=INV(XWX);
  ELSE IXWX=I(TT);
XY=WT[+, ];
XY=XY`;
UY=WT[ ,+];
XWY=XY-XU*D1*UY;
WT2=WT#WT;
WT2=WT2[ ,+];
WT2=WT2[+, ];
YWY=WT2-UY`*D1*UY;
ALPHA=IXWX*XWY;
Z=J(1,4,0);
Z[1,2]=ALPHA[TT,1];
IF ((N-TT) > 0) THEN SE2=(YWY-ALPHA`*XWY)/(N-TT);
   ELSE SE2=0.0001;
Z[1,4]=SE2;
VV=IXWX[TT,TT]*SE2;
Z[1,1]=VV;
Z[1,3]=IXWX[TT,TT];
RETURN (Z);
FINISH;




/*
     This subroutine figures out the estimate of the total of the
```

```
      stratum for the last year and the variance for the last year
      using both methods proposed by USDA and UHCL. The input are
      a matrix Y which includes the the information of the
      stratum, FPCL which equals to the expansion factor of the
      last year, ALPHAL which denotes the estimate of alpha for
      the last year, IXWXL for the last diagonal value of inverse
      of X'WX, and SE2L which is estimate of varinace of error .
      The output is a matrix RE.
 RE[1,1]=total of the stratum of the last year using method of
USDA   RE[1,2]=variance of the estimate using method of USDA
 RE[1,3]=total of the stratum of the last year using method of
USCL   RE[1,4]=variance of the estimate using method of USCL
*/

START TOT(Y,FPCL,ALPHAL,IXWXL,SE2L);
W2=Y;
FPC=1-1/FPCL;
IF (ALPHAL < 0) THEN ALPHA=0;
               ELSE ALPHA=ALPHAL;
IXWX=IXWXL;
SE2=SE2L;
C=NCOL(W2);
WW=J(NROW(W2),C-2,0);
WW=W2[ ,3:C];
T=C-2;
S=NROW(WW);
MI=J(S,T,-1)#WW;
N1=(MI>0);
N2=N1[+, ];
N3=J(1,T,S);
NT=N3-N2;
NTT=NT[1,T];
M1=J(S,T,1)-N1;
WT=M1#WW;
WT3=WT[+, ];
YT=J(1,T,0);
DO I=1 TO T;
 IF NT[1,I]>0 THEN YT[1,I]=WT3[1,I]/NT[1,I];
END;
YTT=YT[1,T];
A2=YTT;
Y1=J(S,1,YTT);
Y2=WT[ ,T]-Y1;
Y2=M1[ ,T]#Y2;
Y2=Y2#Y2;
Y3=Y2[+, ];
IF (NTT-1)>0 THEN S2=Y3/(NTT-1);
   ELSE S2=0;
TOT2=A2*NTT;
VAR2=S2*NTT*FPC;
TOT1=ALPHA*NTT;
VAR1=SE2*(NTT**2)*FPC*IXWX;
RE=J(1,4,0);
RE[1,1]=TOT2;  /*USDA TOTAL    */
```

```
RE[1,2]=VAR2;  /*USDA VARIANCE */
RE[1,3]=TOT1;  /*UHCL TOTAL    */
RE[1,4]=VAR1;  /*UHCL VARIANCE */
RETURN(RE);
FINISH;

/* Main Program begins */
USE W1;
/* W1 contains STRATAY, SEG, and sampling data of each year*/
READ ALL INTO W;  /* Read dataset W1 into matrix W */
DO I=3 TO NCOL(W);
  DO J=1 TO NROW(W);
    IF (W[J,I]=.) THEN W[J,I]=-1; /* Change missing values to -1
*/
  END;
END;


USE F3;
/* F3 contains the stratum and corresponding value of MEXPFCTR */
READ ALL INTO F; /* Read dataset F3 into matrix F */
USE A2;
/* A2 contains stratum and the corresponding number of segments
*/
READ ALL INTO S; /* Read A2 into matrix S */
USE A5;
READ ALL INTO STUM;
print stum;
*STUM contains the stratum and the corresponding number of
STRATRY;
NST=NROW(STUM);
S2=J(NST+1,1,0);
S2[1:NST,1]=STUM[ ,2];
L1=0;
L2=S2[1,1];
NS=NROW(S);
S1=J(NS+1,1,0);
S1[1:NS,1]=S[ ,2];
SGGV=J(NS,3,0);
SGGV[ ,1]=S[ ,1];
SAISF=J(NS,4,0);
SAISF[ ,1]=S[ ,1];
TOTV=J(NS,5,0);
TOTV[ ,1]=S[ ,1];
K1=0;
K2=S1[1,1];
GGG=J(NS,6,0);
DO I=1 TO NS;
   Y=W[(K1+1):K2, ];
/* Y contains segments, sampling data for one stratum*/
   PRINT Y;
   GGG[I, ]=GAMI(Y);
   T=GGG[I,6];
   GGG[I,1]=INT(GGG[I,1]/100)*100;
```

```
   K1=K2;
   K2=K2+S1[I+1,1];
END;
DO J=1 TO NST;
   Y1=GGG[(L1+1):L2, ];
   Y2=Y1[+, ];
   MSE=Y2[1,2]/Y2[1,3];
   MSB=Y2[1,4]/Y2[1,5];
   IF MSE=0 then
    GNEW=0.1;
   else
    GNEW=(MSB/MSE-1)/T;
   IF GNEW<=0 THEN GNEW=0.1;
   SGGV[(L1+1):L2,2]=GNEW;
   L1=L2;
   L2=L2+S2[J+1,1];
END;
K1=0;
K2=S1[1,1];
DO I=1 TO NS;
   Y=W[(K1+1):K2, ]; * Y contains segments, sampling data for one
strataу;
   GAMMA=SGGV[I,2];
   AL=VARIAN(Y,GAMMA); * Get the informatoion related to the
gamma;
   SGGV[I,3]=AL[1,1];
   SAISF[I,2]=AL[1,2];
   SAISF[I,3]=AL[1,3];
   SAISF[I,4]=AL[1,4];
   TOTV[I,2:5]=TOT(Y,F[I,2],AL[1,2],AL[1,3],AL[1,4]);
       /* Call subroutine TOTV */
   K1=K2;
   K2=K2+S1[I+1,1];
END;
NAMESG={STRATAY GAMMA VARIANCE};
NAMESA={STRATAY ALPHA IXWXTT SE2 FPC};
NAMETV={STRATAY USDAT USDAV UHCLT UHCLV};
PRINT SGGV[COLNAME=NAMESG];
PRINT SAISF[COLNAME=NAMESA];
PRINT TOTV[COLNAME=NAMETV];
UHCL=TOTV[+, ];
UHCL[1,1:4]=UHCL[1,2:5];
UHCL[1,5]=UHCL[1,2]/UHCL[1,4];/* ratio=Var(USDA)/Var(UHCL) */
NAMEU={USDAESTIMATE USDAVAR UHCLESTIMATE UHCLVAR RATIO};
PRINT UHCL[COLNAME=NAMEU];
UH=UHCL;
UH[1,2]=SQRT(UH[1,2]);
UH[1,4]=SQRT(UH[1,4]);
NAMEH={USDAESTIMATE USDASTD UHCLESTIMATE UHCLSTD RATIO};
PRINT UH[COLNAME=NAMEH];
QUIT;   /* Main program ends. */

/*
     The following program is written by Bill Iwig and updated by
```

```
    Charles Perry of USDA. It computes the product of list part.
*/


DATA TEMP1;
SET SAVE.DSJOohi(keep= state
stratay segment  MPOPCNTS MOLNOLAC mresphog
 CSOYXXPL);
IF STRATAY <= 98;
IF 1 <= MRESPHOG <= 5; USABLE = 1;
RUN;

PROC SORT DATA=TEMP1;
BY STATE STRATAY SEGMENT;
RUN;

DATA TEMP11;
SET TEMP1;
BY STATE STRATAY SEGMENT;
IF FIRST.SEGMENT^=LAST.SEGMENT THEN OUTPUT TEMP11;
IF (FIRST.SEGMENT=0 AND LAST.SEGMENT=0) THEN OUTPUT TEMP11;
RUN;

DATA TEMP1;
SET TEMP1;
CSOYXXPL=CSOYXXPL*MOLNOLAC;
RUN;

DATA TEMP11;
SET TEMP1;
BY STATE STRATAY SEGMENT;
IF FIRST.SEGMENT^=LAST.SEGMENT THEN OUTPUT TEMP11;
IF (FIRST.SEGMENT=0 AND LAST.SEGMENT=0) THEN OUTPUT TEMP11;
RUN;

PROC MEANS NOPRINT DATA=TEMP1;
BY STATE stratay segment;
ID MPOPCNTS;
VAR CSOYXXPL USABLE;
OUTPUT OUT=TEMP1 SUM(CSOYXXPL)=  MEAN(USABLE)= ;
RUN;

DATA TEMP1;
SET TEMP1;
BY STATE STRATAY SEGMENT;
IF USABLE^=1 THEN USABLE=0;
PROC MEANS NOPRINT;
  BY STRATAY;
  VAR USABLE;
  OUTPUT OUT=B2 SUM=NUSABLE;
RUN;

DATA B3;
  MERGE TEMP1 B2;
```

```
  BY STRATAY;

EXPHOGS=CSOYXXPL*MPOPCNTS/NUSABLE;
IF STRATAY >= 80 THEN EOHOGS=EXPHOGS;
  ELSE EOHOGS=0;
LISTSQ=EXPHOGS**2;
EOSQ=EOHOGS**2;
RUN;

PROC MEANS NOPRINT;
  BY STATE;
  VAR EXPHOGS EOHOGS;
OUTPUT OUT=TOT SUM=LIST_EXP EO_EXP;
RUN;

PROC MEANS DATA=B3 NOPRINT;
  BY  STATE STRATAY;
  VAR LISTSQ EOSQ EXPHOGS EOHOGS MPOPCNTS NUSABLE;
  OUTPUT OUT=VARSQ SUM=LISTSQ EOSQ LISTDE EODE X1 X2
          MEAN=X3 X4 X5 X6 MPOPCNTS NUSABLE;
RUN;

DATA B6; SET VARSQ;
KEEP STATE STRATAY LISTVAR EOVAR;
LISTVAR=((MPOPCNTS-NUSABLE)/MPOPCNTS)*(NUSABLE/(NUSABLE-1))
         *(LISTSQ-LISTDE**2/NUSABLE);
EOVAR=((MPOPCNTS-NUSABLE)/MPOPCNTS)*(NUSABLE/(NUSABLE-1))
         *(EOSQ-EODE**2/NUSABLE);
RUN;

PROC MEANS NOPRINT;
  BY STATE;
  VAR LISTVAR EOVAR;
  OUTPUT OUT=VARI SUM=LISTVAR EOVAR;
RUN;

PROC PRINT DATA=VARI;
VAR LISTVAR EOVAR;

DATA B7;
SET VARI;
LISTSTD=SQRT(LISTVAR);
EOSTD=SQRT(EOVAR);
RUN;

DATA B8;
MERGE TOT B7;
PROC PRINT DATA=B8;
VAR LIST_EXP LISTSTD EO_EXP EOSTD;
RUN;
```