ESTIMATION OF AGRICULTURAL COMMODITIES USING MULTIYEAR AREA FRAME SURVEY DATA

Michael E. Bellow and Charles R. Perry, Jr., USDA/NASS, Lih-Yuan Deng, University of Memphis
Michael E. Bellow, USDA/NASS, 3251 Old Lee Hwy., Fairfax, VA 22030

Key Words: Analysis of variance model, rotation sample design, relative efficiency

I. INTRODUCTION

The National Agricultural Statistics Service (NASS) conducts probability sample surveys to estimate many agricultural commodities in the United States. NASS's annual June Agricultural Survey (JAS) uses a multiple frame approach to sampling. The area frame is based on a land use stratification of a state's area. This frame provides full coverage of the 48 conterminous states but is inefficient for rare commodities and those represented by extremely large farms. The list frame, consisting of a list of known farm operators in a state, is much more efficient than the area frame for most commodities. However, it is usually incomplete and difficult to maintain. The multiple frame approach takes advantage of the strengths of both sampling frames.

In general, NASS's estimation methodology uses only the current year's survey data. The area frame sampling involves multiyear rotation designs with 20 percent replacement of sample units each year. Since 80 percent of units stay in the sample from one year to the next when the same frame is in place, an estimation approach that uses multiyear survey data can augment the current year's information and effectively increase the sample size. The sampling variance of estimates is thereby reduced.

Chhikara and Deng (1992) proposed an approach that applied an analysis of variance (ANOVA) model to commodity estimation using two or more years of area frame survey data. The method was evaluated using 1987-90 data for soybean and hog inventories in four central states (Chhikara et al., 1993). The main conclusion was that the multiyear model led to a more efficient estimate than the single year method. The multiyear method also produced a much more stable yearly variance estimate than the single year method.

This paper extends the earlier research on multiyear estimation. Section II gives a brief background on NASS estimators. Section III describes the multiyear model in detail. The method was compared with single year estimation at the state and national levels for various crop acreage and hog inventory items. The results of the evaluation are presented in Sections IV and V.

II. NASS ESTIMATORS

The commodity estimators currently used by NASS are described in detail by Chhikara and Deng (1992), so only a brief account is given here.

The sampling unit of NASS's area frame is known as the *segment*, a piece of land with identifiable boundaries and generally between 0.1 and 4 square miles. The reporting unit is the *tract*, an area within a segment that is under a single operation or management. An estimator of state total for a commodity can be obtained by summing the survey data over tracts within a segment, multiplying by an expansion factor, summing over segments within strata, and aggregating the stratum totals to the state level. This unbiased estimator, known as the *area tract estimator*, is considered reliable for estimating crop acreages since it uses the accurately determined tract data. However, it does not seem to work well for livestock items or commodities associated with a farm operation. In such cases, the *area weighted estimator* is preferred. This estimator is derived from sample inventories of farms totally or partially within the sample segments, with the weight being the ratio of the within-segment tract acreage to the corresponding farm acreage.

Multiple frame estimators use data from both frames, but favor the list frame. The overlap domain refers to the set of farm operators in both the area and list frames. The nonoverlap (NOL) domain is the set of farm operators in the area frame but not the list frame. A multiple frame estimator is the sum of a list frame estimator (imputed or adjusted) in the overlap domain and an area frame estimator (closed or weighted) in the NOL domain. In general, multiple frame estimators are more efficient than estimators that use only the area frame.

III. MULTIYEAR ESTIMATION MODEL

As mentioned earlier, NASS's area frame sampling has about 80 percent overlap of segments from one year to the next. There is a degree of consistency in area segment characteristics from year to year. The factors that remain fairly constant over years are the prevalent soil types in segments and the capabilities of certain operators to grow

crops. Factors that vary across years include weather and economic conditions. Multiyear estimation should achieve the largest gains in efficiency for those commodities that are most consistent across years.

Hartley (1980) proposed an ANOVA approach to crop estimation using multiyear data acquired from earth orbiting satellites. Lycthuan-Lee (1981) implemented Hartley's idea, estimating North Dakota wheat acreage using 1976-78 satellite data. Chhikara and Deng (1992) adapted this methodology to estimation of commodities using multiyear survey data collected from the area frame.

The multiyear ANOVA model of Chhikara and Deng is given by:

$$y_{tk} = \alpha_t + b_k + e_{tk}$$
$$(k = 1, 2, ..., n_t;\ t=1,2, ...,T)$$

where:

T = number of years
n_t = sample size in year t

In matrix form, the model can be written as:

$$\mathbf{y} = \mathbf{X}\alpha + \mathbf{U}\mathbf{b} + \mathbf{e} \qquad (3.1)$$

where:

$\mathbf{y} = (y_{11}, y_{12}, .., y_{Tn_t})$
$\alpha = (\alpha_1, ..., \alpha_T)$
S = no. distinct segments sampled over T years
$\mathbf{b} = (b_1, ..., b_S)$
$\mathbf{e} = (e_{11}, e_{12}, ..., e_{Tn_t})$

X is an NxT design matrix consisting of 0's and 1's accounting for the fixed year effect α. **U** is an NxS design matrix of 0's and 1's specified according to the rotation sampling scheme and accounting for the random segment effect **b**. The assumptions are that **b** has mean **0** and covariance matrix $\sigma_b^2 I$, and the random error **e** has mean **0** and covariance matrix $\sigma_e^2 I$. The total error **q** = **Ub** + **e** has mean **0** and covariance matrix $\sigma_e^2 \mathbf{W}$, where:

$$\mathbf{W} = \mathbf{I} + \gamma\mathbf{U}\mathbf{U}', \quad \gamma = \sigma_b^2 / \sigma_e^2$$

The parameter γ is usually not known beforehand and must be estimated from survey data. The estimator used here is:

$$\hat{\gamma} = (S/N)[(MS_b / MS_e) - 1]$$

where MS_b is the mean square due to segment and MS_e is the mean square due to error. Although the model is applied separately within each substratum (subdivision of a stratum), γ is estimated at the stratum level to stabilize estimation across substrata.

The weighted least squares estimator of α is:

$$\hat{\alpha} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}$$

The covariance matrix of $\hat{\alpha}$ is:

$$\mathcal{C}(\hat{\alpha}) = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\sigma_e^2$$

The single year estimator used by NASS is obtained by assuming no segment effect, i.e., setting $\gamma = 0$:

$$\tilde{\alpha} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

The covariance matrix of $\tilde{\alpha}$ under the multiyear model is given by:

$$\mathcal{C}(\tilde{\alpha}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\sigma_e^2 \qquad (3.2)$$

(Chhikara et al., 1993). The diagonal elements of this matrix are the single year variances for years 1,...,T under the multiyear model. Alternatively, the single year variance for a given year can be estimated by the standard formula using only the current year's survey data. That estimator will be referred to as the survey-based single year variance estimator, and the one computed from equation (3.2) as the model-based single year variance estimator.

Of prime interest is $\hat{\alpha}_T$, the multiyear estimator for the final (current) year. The variance of this estimator is always less than or equal to the model-based variance of the single year estimate. Assuming that γ is known, $\hat{\alpha}_T$ is the best linear unbiased estimator (BLUE) of α_T under the multiyear model. A simulation study performed by Chhikara and Deng (1992) showed the multiyear estimation method to be fairly robust to misspecification of γ.

The optimal number of years of survey data to use is related to the percentage of sample segments replaced each year. By comparing results for T=2,3,4 and 5 in a simulation study, Chhikara and Deng concluded that under NASS's current sample design, the best efficiency would be achieved for T=5.

When multiple frame estimation is used, the multiyear method applies only to the area frame (NOL) component of estimates. The list frame (overlap) component is the same as for single year estimation. Consequently, the gains due to the multiyear method should be lower for multiple frame estimators than for area frame estimators.

IV. STATE LEVEL RESULTS

Multiyear estimates of eight crop acreage items and four hog inventory items were computed for the 48 conterminous states, using JAS data from the four-year period 1992-95. Since two separate estimators were computed for the crop items, there were 20 commodity/estimator combinations. The results for one crop and one hog item in several states are shown here for illustrative purposes.

Table 1 compares the single year and multiyear area tract estimates of planted corn acreage, in the top five corn states according to the single year estimates for 1995. The ratios between the multiyear and single year estimates are shown, along with the standard errors. Table 2 gives the results for total hogs in the top five hog producing states for 1995, per the single year estimates. All standard errors shown are survey-based. The multiple frame estimator used was the sum of the area weighted estimator in the NOL domain and a nonresponse adjusted list frame estimator in the overlap domain. Table 2 also shows the NOL component of the single year and multiyear hog estimates and their standard errors.

Three separate measures of relative efficiency (RE) are shown in the tables. The survey-based RE is the ratio of the survey-based variance of the single year estimator to the multiyear variance. The model-based RE is the ratio of the model-based variance of the single year estimator to the multiyear variance. The survey-based RE is a questionable measure of the improvement achieved by using multiyear estimation, since it is highly sensitive to outliers and underestimation of the single year standard error. This observation is illustrated by Table 2, where the survey-based RE in the NOL domain is less than one for four of the five states. The multiyear variance estimate is much more stable over years, and hence more reliable. The model-based estimate of single year variance, obtained from equation (3.2), is also much more stable over years than the corresponding survey-based estimate. Hence, assuming the multiyear model is valid, the model-based RE should be a more reliable measure of effectiveness than the survey-based RE. However, little is known about robustness of the model-based RE against departure from model assumptions. One way of addressing this issue is to use a resampling method. After consideration of several options, a balanced bootstrap on model residuals was chosen as the most feasible method to apply here. Bootstrapping regression residuals is described by Efron and Tibshirani (1993) and Shao and Tu (1995). The balanced bootstrap, where each observed residual is constrained to appear the same number of times in the set of all bootstrap samples, improves the efficiency of the results.

Direct application of the bootstrap is difficult due to the correlated errors of the multiyear model. Therefore a transformation was applied to the data so that an equivalent model with diagonal error covariance matrix could be used (Seber, 1977).

The nonsingular matrix V satisfying $W = VV'$ was first computed using the eigenvalues and eigenvectors of W. Premultiplication of model (3.1) by V^{-1} yields the transformed model:

$$\mathbf{z} = \mathbf{B}\alpha + \mathbf{f}$$

where $\mathbf{z} = \mathbf{V}^{-1}\mathbf{y}$, $\mathbf{B} = \mathbf{V}^{-1}\mathbf{X}$, and $\mathbf{f} = \mathbf{V}^{-1}(\mathbf{U}\mathbf{b} + \mathbf{e})$. The error term $\mathbf{f}$ has mean $\mathbf{0}$ and covariance matrix $\sigma_e^2 \mathbf{I}$. The adjusted residuals of the transformed model are given by:

$$\hat{\mathbf{f}}_a = [N/(N\text{-}tr(\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'))]^{1/2} (\mathbf{z} - \mathbf{B}\hat{\alpha})$$

One can show that $(\hat{\mathbf{f}}_a' \hat{\mathbf{f}}_a)/ N$ is an unbiased estimator of the random error variance σ_e^2.

A balanced set of 500 bootstrap samples was selected from the empirical distribution assigning probability 1/N to each of the N adjusted transformed residuals. For each replication, the selected bootstrap residuals were substituted into the transformed model to construct bootstrap values of $\mathbf{z}$, which were premultiplied by V to obtain bootstrap values of $\mathbf{y}$. The model was then fitted to obtain both multiyear and single year bootstrap estimates of α.

The above procedure was applied separately within each substratum. The bootstrap values of α were summed over substrata and strata to obtain the bootstrap state level totals. The means and variances of those state totals over all replications were then computed. The bootstrap relative efficiency was computed as the ratio between the bootstrap single year and multiyear variances.

Tables 1 and 2 also show the bootstrap RE values. For the hog estimates, the bootstrap RE is given only for the NOL domain. Comparison of the model-based and

bootstrap RE's shows close agreement, with the largest discrepancy being 0.04 for the North Carolina hog estimate and no other difference exceeding 0.02. From these results and others not shown here, the model-based RE can be considered a reliable measure.

Of the top five corn states listed in Table 1, only Nebraska's model-based RE of 1.29 was appreciably greater than one. Nebraska was also the only state where the difference between the two estimators exceeded one percent. Of the top five hog states in Table 2, only North Carolina had a model-based RE exceeding 1.01 for the MF estimate. However, North Carolina's result appears to be an outlier-induced anomaly as the multiyear estimator was more than six times as large as the single year estimator in the NOL domain.

Overall, the multiyear method showed noteworthy gains in efficiency over single year estimation only in a small fraction of cases. Of the 20 commodity/estimator combinations evaluated, the best state level results were obtained for the area tract estimators of planted corn and harvested hay. Ten states had model-based RE greater than 1.25 for planted corn, and twelve states for harvested hay. The highest state level RE of any item evaluated was 1.48, for the multiple frame estimator of sows farrowed (Dec. 1994 to Feb. 1995) in Alabama.

V. NATIONAL LEVEL RESULTS

The 1995 state level multiyear and single year estimates of twelve commodities were aggregated to the national level. Four states received new area frames during the 1992-95 period: Oklahoma (1993), California (1994), New York (1995) and South Carolina (1995). The multiyear estimates for those states were computed using only the years when the new frames were in effect, e.g., 1993-95 for Oklahoma.

Tables 3 and 4 compare the single year and multiyear estimation methods for the eight crop acreage items and four hog inventory items at the national level in 1995. Two estimators were computed for each crop item; the area tract estimator and a multiple frame estimator. The latter was the sum of the area tract estimator in the NOL domain and an imputed list frame estimator in the overlap domain. The SE's and RE's shown are the model-based values, in light of the bootstrap results discussed earlier.

Table 3 shows that the highest RE's occurred for the area tract estimates of harvested acreage of durum wheat (1.18), all hay (1.17) and alfalfa hay (1.15). Those same three items showed the largest percent differences between the single year and multiyear estimates. As expected, the RE for each crop was higher with the area tract estimator than the multiple frame estimator. From Table 4, the RE's of the four hog items were all below 1.05, and the percent differences were very small.

VI. SUMMARY AND CONCLUSIONS

The multiyear estimation method was evaluated for a number of crop acreage and hog inventory items at the state and national levels using 1992-95 area frame survey data. Relative efficiencies and estimator ratios were used to compare multiyear estimation with single year estimation. The RE values may in fact be slightly optimistic since they depend to some degree on model accuracy, which has never been verified. State level results showed that the multiyear method caused appreciable gains in efficiency only in a minority of cases. At the national level, the estimated model-based RE's applied to the area tract estimator of crop acreages ranged from 1.07 to 1.18. Gains for the multiple frame estimator of crop acreages and hog inventories were less than 1.05. Relative efficiencies of this magnitude do not warrant operational use by NASS.

REFERENCES

Chhikara, R.S. and Deng, L.Y. (1992). "Estimation Using Multiyear Rotation Design Sampling in Agricultural Surveys". *Journal of the American Statistical Association*, 87, 924-932.

Chhikara, R.S., Deng, L.Y. , Yuan, Y., Perry, C.R. and Iwig, W.C. (1993). "Estimation of Totals Using Multiyear June Agricultural Statistics Data". U.S. Dept. of Agriculture, NASS Report No. SRB-93-09.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York, N.Y., Chapman & Hall.

Hartley, H.O. (1980). *A Survey of Multiyear Estimation Procedures*. Technical Report DS1, Duke University, Dept. of Mathematics.

Lycthuan-Lee, T.G. (1981). *Development of Rotation Sample Designs for the Estimation of Crop Acreages*. Technical Report No. 15409, Lockheed Engineering and Management Services Company, Inc.

Seber, G.A.F. (1977). *Linear Regression Analysis*. New York, N.Y., John Wiley & Sons.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. New York, N.Y., Springer.

Table 1: 1995 Planted Corn Acreage Estimator Comparison for Top Five Corn States (values in thousands).

State	Domain	Ratio - M/S (%)	Single Year SE (SB)	Multiyear SE (SB)	RE (SB)	RE (MB)	RE (Boot)
Iowa	AF	100.7	260.3	232.4	1.25	1.07	1.08
Illinois	AF	100.6	247.9	244.8	1.03	1.04	1.04
Nebraska	AF	102.5	314.9	239.5	1.73	1.29	1.30
Indiana	AF	99.3	183.9	173.6	1.12	1.05	1.03
Minnesota	AF	100.5	199.1	196.0	1.03	1.02	1.01

(SB = survey-based, MB = model-based, AF = area frame, M/S = {multiyear}/{single year})

Table 2: 1995 Total Hog Estimator Comparison for Top Five Hog States (values in thousands).

State	Domain	Ratio - M/S (%)	Single Year SE (SB)	Multiyear SE (SB)	RE (SB)	RE (MB)	RE (Boot)
Iowa	NOL	88.2	234.4	476.0	0.24	1.02	1.01
	MF	98.9	395.8	572.9	0.48	1.01	
North Carolina	NOL	622.2	14.7	145.0	0.01	1.28	1.24
	MF	101.9	115.0	184.5	0.39	1.17	
Illinois	NOL	96.2	196.5	165.1	1.42	1.01	1.02
	MF	99.6	301.5	282.1	1.14	1.00	
Minnesota	NOL	94.9	74.0	92.7	0.64	1.03	1.03
	MF	99.8	160.0	169.5	0.89	1.01	
Nebraska	NOL	105.2	93.0	125.3	0.55	1.01	1.02
	MF	100.2	155.3	176.5	0.77	1.01	

(SB = survey-based, MB = model-based, AF = area frame, MF = multiple frame, M/S = {multiyear}/{single year})

Table 3: 1995 U.S. Level Crop Acreage Estimator Comparison (values in thousands). SE, RE values model-based.

Item	Type	Ratio - M/S (%)	Single Year SE	Multiyear SE	RE
Alfalfa Hay (H)	AF	101.1	535.4	499.9	1.15
	MF	99.8	411.7	404.5	1.04
Hay (H)	AF	101.6	859.7	796.4	1.17
	MF	100.1	688.4	671.9	1.05
Corn (H)	AF	100.3	716.9	682.2	1.10
	MF	100.0	549.7	544.9	1.02
Corn (P)	AF	100.4	725.1	688.0	1.11
	MF	100.0	564.5	559.7	1.02
Soybeans (P)	AF	99.9	662.9	641.5	1.07
	MF	99.6	606.8	603.0	1.01
Winter Wheat (H)	AF	100.6	701.6	677.9	1.07
	MF	99.7	481.1	476.7	1.02
Durum Wheat (H)	AF	102.3	224.1	205.9	1.18
	MF	100.1	153.7	153.5	1.00
All Wheat (H)	AF	100.5	839.2	803.3	1.09
	MF	99.9	593.5	587.9	1.02

Table 4: 1995 U.S. Level Hog Estimator Comparison (values in thousands). SE, RE values model-based.

Item	Type	Ratio - M/S (%)	Single Year SE	Multiyear SE	RE
Pig Crop (Dec.-Feb.)	MF	100.0	462.9	456.6	1.03
Sows Farrowed (Dec.-Feb.)	MF	100.0	56.1	55.4	1.03
Total Breeding Stock	MF	100.0	149.4	147.0	1.03
Total Hogs and Pigs	MF	100.1	818.5	804.1	1.04

(H = harvested, P = planted, AF = area tract estimator, MF = multiple frame estimator)