WEIGHTED SEGMENT RESEARCH FOR COVERAGE EVALUATION OF THE CENSUS OF AGRICULTURE

Paul J. Lewis and Glenn S. Wolfgang, U.S. Bureau of the Census
Paul J. Lewis, U.S. Bureau of the Census, Washington, D.C. 20233

KEY WORDS: Coverage measurement; Dual-system estimation; Open segment; Farm weight.

1. Introduction

The census of agriculture coverage evaluation program provides an assessment of the completeness and accuracy of the census of agriculture. Although the goal of each census is to enumerate all farms in the nation, incomplete mail lists contribute to errors in published census farm counts. The June Agricultural Survey (JAS), conducted by the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture, is used by the Census Bureau to provide an independent measure of farm count for estimating farms not on the census mail list. The JAS, conducted annually, uses an area sample to provide estimates of crop acreages and livestock inventories (Cotter and Nealon 1987). The area frame is a collection of segments which provide complete coverage of the land within a given geographic area.

Several methods of segment expansion are used to translate the JAS segment data to universe or population totals. The open segment estimator was used during the 1987 Coverage Evaluation Program to produce the estimates of farms not on the census mail list. Since then, NASS has adopted the weighted segment approach as their primary expansion method. The purpose of this study is to determine the potential of the weighted segment estimator for estimating farms and characteristics of farms not on the census mail list.

The study objective is met by comparing the open segment estimates to the weighted segment estimates for certain farm characteristics and by comparing their mean square errors and relative standard errors. Bootstrap methodology is used to provide a measure of the goodness of the variance estimator as well as an estimate of the bias in the estimators.

In general, the weighted segment estimates appear to have greater precision than the open segment estimates for the measured characteristics. The weighted segment estimate of farms not on the mail list was not significantly different from the open segment estimate in any of the six states where they were compared. The estimate of land in farms did differ in four states, although this is thought to be due in part to biased farm weights. For 1992, the bias in the farm weights is expected to be lower due to special efforts by NASS and the Census Bureau. Combined with the expected increase in precision, the weighted segment estimator is an attractive alternative for estimating farms and characteristics of farms not on the 1992 Census of Agriculture mail list.

2. Background

The not on the mail list study has been part of the coverage evaluation program since the census of agriculture first became enumerated by mail in 1969. The objectives and methods used for the 1987 Coverage Evaluation Program are presented by Wright *et al.* (1989). Not on the mail list estimates are generated using an independent enumeration of farms from the JAS in a coverage error model based on dual-system estimation theory (Wolter 1986).

NASS uses area sampling frames for conducting their JAS. An area frame for a state consists of a collection of land parcels defined by easily identifiable boundaries. NASS's area frame is created by dividing the land in a state into six to eight land-use strata such as intensively cultivated land, urban areas, agricultural urban areas, and rangeland. The land-use strata are identified on county highway maps and divided into primary sampling units using permanent and easily recognizable land features. Cluster analysis is used to group counties into clusters or "paper" strata with similar agricultural makeups.

NASS uses replicated sampling in the JAS. Each year a new multiple simple random sample of one segment from each paper stratum in a land-use stratum is selected and combined with the selections from the previous four years. In effect, one-fifth of the sample is redrawn every year and segments rotate out of the sample after five years. Each segment has a known probability of selection based on its size. The inverse of the probability of selection is the segment's expansion factor which inflates the segment values to population totals. The segment expansion factor is the same for all segments within a land-use stratum since each segment has the same chance of being selected. Since the land area within each segment is completely enumerated, the segment and not the farm is the basic unit of analysis for the JAS (Cotter and Nealon 1987).

Once the segments are chosen, an enumerator visits

them and establishes who operates the land within the segment, defining the ultimate sampling unit, the tract. Only one farm operation is associated with a tract; however, a farm operation may be represented by tracts outside the sampled segment. A typical segment contains portions of 2 to 4 farm operations.

Both the open and the weighted segment approaches require that data be obtained on the entire farm. They differ in that the open segment estimator includes farm data only when the residence of the farm operator is located within the boundaries of the sampled area segment (called resident farm operators or RFOs). The weighted segment estimator includes a portion of data from all farms within the sampled segment regardless of where the operator's residence is. The portion of a farm's acres within a sampled segment is the tract acres. The weighted segment farm weight is the tract acres divided by the total farm acres.

The primary advantage of the open segment approach is that it requires fewer interviews to be conducted. However, it is less precise and can be affected by errors in association between operators and their residences. The weighted segment approach provides more precise estimates, but it is more costly to conduct and has been shown to have a slight upward bias due to underreporting of total farm acres. Acreages outside the sample tract are harder to identify or verify, and respondents tend not to include in their estimate of total farm acres woodland, wasteland, or idleland. The bias is expected to be less of a problem in 1992 because of increased training of NASS enumerators to emphasize the importance of collecting accurate total farm acres. Nealon (1984) provides a thorough comparison of the advantages and disadvantages of each estimator.

The open segment estimate of the value of some characteristic for all farms in a segment is:

$$y_{ijk} = \sum_{m=1}^{f_{ijk}} b_{ijkm} y_{ijkm} \qquad (2.1)$$

where f_{ijk} is the number of tracts in the k^{th} segment, j^{th} paper stratum, and i^{th} land-use stratum; b_{ijkm} equals 1 if the farm operator's residence is located within the segment, 0 otherwise; and y_{ijkm} is the value of the characteristic of the entire farm for the m^{th} tract in the k^{th} segment, j^{th} paper stratum, and the i^{th} land-use stratum.

The weighted segment estimate of the value of some characteristic for all farms in a segment is:

$$y_{ijk} = \sum_{m=1}^{f_{ijk}} a_{ijkm} y_{ijkm} \qquad (2.2)$$

where a_{ijkm} is the farm weight (tract acres divided by total farm acres) for the m^{th} tract in the k^{th} segment, j^{th} paper stratum, i^{th} land-use stratum.

The coverage error model assumes that the census mail list and the JAS area frame are independent of one another and that every farm has the same chance of being included in either survey independently of any other farm. Other assumptions are also made and are described in detail by Wolter (1986).

By matching the JAS respondents (and their farms) to the census mail list, each case can be classified into one of the cells in the coverage error model, shown in Figure 2.1. The cell values represent the expanded or weighted number of JAS farms. The number of farms in the population, T, is the sum of the four cells. N_{11} is the number of farms on the census mail list and in the JAS sample, N_{12} is the number of farms on the census mail list but not in the JAS, N_{21} is the number of farms in the JAS but not on the census mail list, and N_{22} is the unobservable number of farms not on the census mail list or in the JAS sample. The row marginal N_{1+} is the total number of farms on the census mail list and the column marginal N_{+1} is the total number of farms in the JAS.

Figure 2.1 Coverage Error Model.

	JAS		
Census	Farm	Nonfarm	All
On mail list	N_{11}	N_{12}	N_{1+}
Not on mail list	N_{21}	N_{22}	NML
All	N_{+1}	N_{+2}	T

The assumptions of independence permit the calculation of unknown terms in the model. The number of farms not on the mail list, NML, is estimated by :

$$\hat{NML} = \hat{N}_{21} \frac{N_{1+}}{\hat{N}_{11}} \qquad (2.3)$$

Some characteristic x of farms not on the mail list, $\hat{NML}_x$, is estimated in a similar manner:

$$\hat{NML}_x = \hat{S}_x \frac{N_{1+}}{\hat{N}_{11}} \qquad (2.4)$$

where $\hat{S}_x$ is the estimate of the total of the characteristic x for farms not on the mail list but in the JAS.

3. Methods

The study population for the weighted segment research was created from a match of 1987 JAS sample cases to the 1987 Census of Agriculture mail list. The JAS sample cases were included in the enumeration for the 1987 census and the census respondent data used to calculate estimates of farms not on the census mail list.

Due to cost and processing constraints, only six states were used in the study. The six can be classified as either domestic crop and land coverage (DCLC) states or nondomestic crop and land coverage (non-DCLC) states. The distinction between them is that non-DCLC states tend to have more area in woodland and idleland and less land in commercial cropland. Iowa (IA), Missouri (MO), and Illinois (IL) are the DCLC states and Minnesota (MN), North Carolina (NC), and Ohio (OH) are the non-DCLC states.

A total of 7,606 JAS weighted segment cases were obtained from NASS and matched to the census mail list. Table 3.1 lists the number of total sample cases, the number of matched sample cases, and the number of nonmatched sample cases for each state. Not all of the cases were used in the estimation process. Census nonfarms, nonrespondents to the census, and cases which were undeliverable as addressed were not used because they could violate the assumptions of the model.

Table 3.1 JAS sample cases by state.

	Total	Matched	Nonmatched
IA	1520	1485	35
IL	1483	1435	38
MN	1295	1257	38
MO	1282	1262	20
NC	961	917	44
OH	1065	1007	58

A comparison is made between the open and the weighted segment estimates for the number of farms, land-in-farms, total value of products sold, cattle and calves inventory, hogs and pigs inventory, acres of corn, and acres of soybeans not on the mail list. Several statistics are used to make the comparisons. The relative standard error (RSE) is computed (as shown in equation 3.1 below) for all of the listed variables.

$$RSE = \frac{\sqrt{\hat{Var}(N\hat{M}L_x)}}{N\hat{M}L_x} \qquad (3.1)$$

The mean square error (MSE) is computed (as shown in equation 3.2 below) for the number of farms and land-in-farms.

$$MSE = \hat{Var}(N\hat{M}L_x) + [\hat{Bias}(N\hat{M}L_x)]^2 \qquad (3.2)$$

For both the RSE and the MSE, the variance estimator is based on a Taylor series expansion of the not on the mail list estimator.

Bootstrap techniques are used to provide an independent estimate of the variance to assess the worthiness of the Taylor series estimator as it is applied to the open segment and weighted segment estimates. It is also used to provide an estimate of the bias in equation 3.2. The bootstrap technique does not rely on the usual parametric assumptions (e.g. normally distributed observations, etc.) and it does not require knowing the true variance formula. All that is necessary is knowing how the sample was drawn. This has an obvious advantage when the variance formula is complex or hard to derive.

The bootstrap technique used here is due to Rao and Wu (1988) and follows from Thomas *et al.* (1990). The procedure consists of three steps:

1) Draw a random sample of segments of size n_{ij} - 1 from each paper stratum. Adjust each segment's weight by $n_{ij}/(n_{ij} - 1)$ to account for the change in sample size used for the bootstrap. Calculate:

$$N\hat{M}L_x^* = \hat{S}_x^* \frac{N_{1+}}{\hat{N}_{11}^*} \qquad (3.3)$$

2) Independently replicate (1) a large number of times, say B, to produce many (500 - 1000) bootstrap estimates.

3) Calculate the variance of the bootstrap estimates to give the bootstrap estimate of variance:

$$\hat{Var}(N\hat{M}L_x^*) = \frac{1}{B-1}\sum_{b=1}^{B}\left[N\hat{M}L_x^*(b) - N\bar{M}L_x^*\right]^2 \qquad (3.4)$$

where $N\hat{M}L_x^*(b)$ is the sample estimate for the b^{th} bootstrap replication, and $N\bar{M}L_x^*$ is the mean of the bootstrap replications.

4) Compute the bootstrap bias estimate:

$$\hat{Bias}(N\hat{M}L_x) = N\hat{M}L_x - N\bar{M}L_x^* \qquad (3.5)$$

The bootstrap bias estimate is computed for both

the open segment and the weighted segment estimates. The bias estimate is a functional statistic of the quantity being estimated, in this case $N\hat{M}L_x$. It is the true bias of $N\hat{M}L_x$ if the function used to generate $N\hat{M}L_x$ is equal to the actual unknown function (Efron 1982). It is used here to provide a general idea as to which estimator is less biased given the selected sample and the specific coverage error model used. Following from Rao and Wu (1988), the bootstrap bias estimate is the difference between the model generated estimate and the bootstrap generated estimate. A positive bias indicates that the estimator is overestimating the true value of the variable while a negative bias indicates an underestimation.

Comparisons between the RSEs, MSEs, and bootstrap estimates are presented in the next section. The difference between the open segment estimate and the weighted segment estimate is tested for significance using Student's t. For all comparisons, a reduced level of significance (0.017) is used for individual state tests to achieve the overall significance level of 0.10. This accounts for simultaneous comparisons within the six states and controls the Type I testing error.

4. Results

The open segment and weighted segment estimates of the number of farms missed is fairly consistent within each of the six states, differing by no more than 10.5% (Minnesota) in any of them. However, the weighted segment estimates of land-in-farms missed are substantially higher than the open segment estimates in Illinois, Minnesota, Missouri, and Ohio. The comparisons are shown in Table 4.1.

Table 4.1. Not on the mail list estimates and RSEs for number of farms and land-in-farms (LIF).

		Open Segment		Weighted Segment			
		$N\hat{M}L_x$	RSE	$N\hat{M}L_x$	RSE	Diff.	
IL	Farms	5226	23.3	5155	19.3	71	
	LIF	298955	42.7	543265	26.5	-244310	*
IA	Farms	5573	22.2	5220	23.4	353	
	LIF	371530	38.6	346552	32.3	24978	
MN	Farms	5711	23.2	6304	20.8	-593	
	LIF	305598	31.3	669609	27.3	-364011	*
MO	Farms	3043	32.2	2874	28.0	169	
	LIF	324659	44.1	415272	33.5	-90613	*
NC	Farms	7502	23.5	7721	19.7	-219	
	LIF	449557	36.0	446991	26.1	2566	
OH	Farms	10258	18.5	11306	15.5	-1048	
	LIF	498155	28.0	742645	21.9	-244490	*

* difference significant at overall $\alpha = 0.10$

Table 4.2 provides comparisons of the open segment verses the weighted segment estimates for the commodities in the six states. Included in the tables are the estimates for the total value of products sold (TVP), number of cattle and calves, number of hogs and pigs, acres of corn, and acres of soybeans. Their RSEs are given along with the difference between the open segment and weighted segment estimates.

Table 4.2. Not on the mail list estimates and RSEs for commodity data.

		Open Segment		Weighted Segment		
		$N\hat{M}L_x$	RSE	$N\hat{M}L_x$	RSE	Diff.
IL	TVP	24866	52.3	49645	33.2	-24779*
	Cattle	14820	47.2	24340	64.2	-9520
	Hogs	6876	76.1	7809	59.8	-933
	Corn	32653	60.8	90392	39.1	-57739*
	Soyb.	39836	76.8	71085	36.4	-31249
IA	TVP	61063	42.9	73915	44.3	-12852
	Cattle	28768	39.8	35643	34.6	-6875
	Hogs	196231	62.6	140383	54.3	55848*
	Corn	104386	59.0	70531	38.5	33855*
	Soyb.	55224	54.5	61289	46.5	-6065
MN	TVP	10868	36.2	31068	33.9	-20200*
	Cattle	5749	58.0	14656	57.7	-8907*
	Hogs	76880	97.1	4036	60.7	72844*
	Corn	12775	57.9	47497	37.4	-34772*
	Soyb.	6321	100.4	70202	48.9	-63881*
MO	TVP	21798	45.8	19497	40.6	2301
	Cattle	49367	48.2	26667	43.5	22700*
	Hogs	1064	72.7	992	72.6	72
	Corn	0	.	1684	75.0	-1684
	Soyb.	21297	100.3	33895	64.9	-12598*
NC	TVP	14760	64.0	16312	42.1	-1552
	Cattle	16640	38.4	17596	27.6	-956
	Hogs	3313	84.0	2378	81.4	935
	Corn	8987	52.6	6791	40.4	2196
	Soyb.	27211	100.3	22753	82.8	4458
OH	TVP	56192	33.5	85776	30.6	-29584*
	Cattle	18635	36.9	23609	29.5	-4794
	Hogs	90192	64.7	179246	73.3	-89054*
	Corn	23857	65.7	80825	39.2	-56968*
	Soyb.	59890	54.7	111020	33.0	-51130*

* difference significant at overall $\alpha = 0.10$

The open segment and weighted segment estimates of hogs and pigs not on the mail list differ significantly in Iowa, Minnesota, and Ohio. The weighted segment estimate of acres of corn not on the mail list is also higher in Iowa, Illinois, Minnesota, and Ohio. There are no large differences between any of the estimates in North Carolina. Minnesota differs on all commodities and Ohio differs on all but cattle and calves inventory.

These differences could be due to a number of things including incorrect data values, large differences between the RFOs and the non-RFOs within a segment, or biased farm weights. Both Minnesota and Ohio are non-DCLC states, hence farm weights in these two states might be more susceptible to the underreporting bias described earlier.

In general, the RSEs for the weighted segment estimates appear to be lower. This is probably due to the decreased variability in the entire farm values and because more observations are being used to calculate the weighted segment estimates. The RSE for a weighted segment estimate is smaller than the RSE for the corresponding open segment estimate in 37 of the 42 comparisons.

The ratio of the Taylor series derived variance estimate to the bootstrap derived variance estimate is given in Table 4.3 for the estimates of farms and land-in-farms not on the mail list. A number less than 1 indicates that the Taylor series variance estimate is smaller than the bootstrap variance estimate. None of the Taylor series variance estimates fall outside of a 90% confidence interval placed around the bootstrap variance estimate, evidence that Taylor series methods provide a fairly accurate variance estimator. Table 4.3 also provides the value of the bootstrap estimate of bias as a percentage of the estimator. None of the bias estimates are significantly large, and although an appropriate test was not identified, there does not appear to be a clear pattern of overestimation or underestimation by either estimator.

The relative efficiency (RE) of the weighted segment estimator as compared to the open segment estimator using the ratio of their MSEs is also shown in Table 4.3. A number less than 100 indicates that the weighted segment is more efficient. While no tests were done, the MSE for the number of farms missed suggests that the weighted segment estimator was more efficient than the open segment estimator for all six states, although it was only marginally so in Iowa and Minnesota. For those states which had a substantially higher weighted segment estimate of land-in-farms missed (Illinois, Minnesota, and Ohio), the open segment estimator was more efficient. The apparent trend is that the greater the percent difference between the open segment and weighted segment estimates, the less efficient the weighted segment estimator.

It is important to note that the bootstrap bias estimate does not evaluate bias due to violations of the coverage error model assumptions or the bias in the farm weight of the weighted segment estimator. The bias due to underestimating (overestimating) the JAS farms was discussed briefly in the methods section. This nonsampling bias is much more difficult to quantify than the bootstrap bias estimates. The usual method in survey sampling is to conduct a reinterview of the respondents, asking the original set of questions and reconciling any differences between the two sets of answers. The bias in the original response is then the difference between it and the reconciled response. A reinterview was not done for this study although it is being investigated for the 1992 Coverage Evaluation program. Another way to reduce bias is to carefully word each question and/or add clarifying questions or text to the questionnaire so as to elicit the correct response during the initial survey.

Table 4.3. Bootstrap comparisons for the open segment and weighted segment estimates.

		Open Segment		Weighted Segment		
		Ratio	%Bias	Ratio	%Bias	RE
IL	Farms	1.012	-2.4	1.151	-1.5	67
	LIF	1.179	-2.3	0.967	-0.6	127
IA	Farms	1.338	-0.7	1.310	0.2	98
	LIF	1.095	-1.0	1.098	-0.3	61
MN	Farms	0.964	0.9	1.021	-0.3	98
	LIF	0.907	2.3	0.962	-0.8	364
MO	Farms	1.095	3.0	0.890	4.7	69
	LIF	1.120	4.4	0.941	1.1	94
NC	Farms	0.948	1.6	0.867	-0.2	74
	LIF	1.000	2.2	0.918	-1.1	52
OH	Farms	1.162	-1.8	1.015	-0.2	85
	LIF	1.052	-2.8	1.094	0.4	134

5. Conclusions

Both the open segment and weighted segment approaches can be used to generate estimates of farms and characteristics of farms not on the census mail list. Each type of estimator has advantages and disadvantages, some of which are discussed in this paper. The open segment estimator is characterized by a lack of precision while the weighted segment tends to have an upward bias in the farm weight due to underreporting of total farm acres. Comparisons of the variances, relative standard errors (RSEs), mean square errors (MSEs), and bootstrap bias estimates demonstrate these tendencies. Another finding of the study is that Taylor series methods provide what appear to be good estimators for the variance estimates.

Based on the evidence presented here, there is no reason to believe that the weighted segment approach will give poorer estimates of farms and characteristics of farms not on the mail list. The increased precision

of the weighted segment estimator might outweigh all of the other problems created, especially if extra effort is put into ensuring that total farm acres are accurately reported by JAS respondents reducing the amount of nonsampling bias in the data.

6. Further Research

More work is needed on ways to reduce bias in the weighted segment estimator. Some proposed methods include using NASS's follow-on surveys to confirm the total farm acres reported in the JAS, adding questions to the JAS instrument which clarify the definition of total farm acres, or designing a reinterview study to verify total farm acres. Another possibility is to replace the JAS estimate of total farm acres with the census estimate of total farm acres. Also, differences between the JAS and census estimates of total farm acres could be reconciled.

The data from the 1992 JAS will contain information on the resident farm operator status. This same study could be duplicated for the 1992 Coverage Evaluation Program, although this time, the weighted segment expansion will be used in all states. The open segment estimator could be contrasted with the weighted segment estimator for the same six states to provide additional information about each estimator.

7. Acknowledgements

Ann Vacca made significant contributions to this paper. Dr. Cynthia Z.F. Clark and Karen E. Wright provided the basis for the study as well as an outline for the paper. Dr. Charles R. Perry was instrumental in developing the Taylor series and bootstrap estimates of the variance. The staff of the Program Research and Development Branch of the Census Bureau's Agriculture Division also assisted. Everyone's help is greatly appreciated.

8. List of References

Cotter, J. and J. Nealon. 1987. *Area Frame Design for Agricultural Surveys.* National Agricultural Statistics Service, USDA, Washington, D.C. 67 pp.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans.* CBMS-NSF Monograph #38, Society for Industrial and Applied Mathematics. 92 pp.

Nealon, J. 1984. *Review of the Multiple and Area Frame Estimators.* SF & SRB Staff Report No. 80, Statistical Reporting Service, USDA, Washington, D.C. 47 pp.

Rao, J.N.K. and C.F.K Wu. 1988. *Resampling Inference With Complex Survey Data.* JASA, **83**(401):231-241.

Steele, R.G.D. and J.H. Torrie. 1980. *Principals and Procedures of Statistics, 2nd edition.* McGraw-Hill, New York. 633 pp.

Thomas, D.R., C.R. Perry, and B. Viroonsri. 1990. *Estimation of Totals for Skewed Populations in Repeated Agricultural Surveys: Hogs and Pigs.* National Agricultural Statistics Service, USDA, Washington, D.C. 101 pp.

Wolter, K.M. 1986. *Some Coverage Error Models for Census Data.* JASA, **81**(394):338-346.

Wright, K.E., W.C. Davie, J.D. Sandusky, and E.A. Vacca. 1989. *1987 Census of Agriculture Coverage Evaluation Estimation.* American Statistical Association 1989 Proceedings of the Section on Survey Methods, 599-604.