An Alternate Approach to Assessing Misclassification in JAS

Pam Arroway[1], Denise A. Abreu[2], Andrea C. Lamas[2]
Kenneth K. Lopiano[3], Linda J. Young[3]
[1]Department of Statistics, North Carolina State University, Raleigh, NC 27695
[2]National Agricultural Statistics Service, USDA, 3251 Old Lee Hwy, Fairfax, VA 22030
[3]Department of Statistics, University of Florida, Gainesville, FL 32611

Abstract
Each year, the National Agricultural Statistics Service (NASS) obtains an estimate of the number of farms in the United States (US) based on the June Area Survey (JAS). In 2007, the JAS estimate of the number of farms was much lower than that from the quinquennial Census of Agriculture. The discrepancy was more than could be accounted for by sampling error. The JAS uses an area frame that is, by design, a complete frame of the population. Estimates incorporate sampling weights appropriate to the sample design and so are unbiased unless misclassification is present. In 2009, NASS conducted the Farm Numbers Research Project (FNRP) to study misclassification of farms and non-farms in JAS. An annual (modified) version of FNRP called the Annual Land Use Survey (ALUS) is proposed to adjust the JAS estimate of the number of farms.

Key Words: June Area Survey (JAS); Misclassification; Two-phase sampling design

1. Introduction

The National Agricultural Statistics Service (NASS) within the United States Department of Agriculture (USDA) has the responsibility for conducting surveys of the agricultural activity within the United States and publishing the results. One of the largest annual NASS survey projects is the June Area Survey (JAS). It provides information for many of the other NASS surveys. The primary purpose of the JAS is to provide direct estimates of acreage in various farming activities.

The JAS has an area sampling frame. All land in the U.S., except Alaska, is stratified by land use within a state. The specific strata types vary with state; one such stratification is given in Table 1. The primary sampling units (PSU) provide complete coverage of all agriculture activity occurring within the PSU and, consequently, all farmers in the state. Each PSU is divided into segments, which are roughly a square mile in area. Each year about 3500 segments are selected for inclusion in the sample. A selected segment stays in the sample for five years. Thus, each year about 11,000 segments are in the sample. Sampled segments are divided into tracts, each tract representing a unique land operation arrangement. During prescreening, enumerators visit each tract within the newly rotated-in segments to determine whether it has a farming operation. In June, those tracts that have been determined to have a farming operation during prescreening (about 35,000) are revisited, and crop and livestock information is collected through personal interviews.

The NISS/NASS farm numbers research team[1] proposes a yearly follow-on survey to the JAS called the Annual Land Utilization Survey (ALUS). The purpose of ALUS is to provide information about misclassification of farms and non-farms, focusing on tracts that are a) determined to be non-agricultural in June or b) are estimated in June. ALUS results could be used to directly augment JAS indications (preliminary estimates) of farm numbers. In addition, data collection will include several other variables, allowing indications of other commodities to be adjusted using ALUS.

ALUS is modeled on the 2009 Farm Numbers Research Project (FNRP). FNRP was a one-time follow-on survey to the JAS segments (Abreu, McCarthy and Colburn, 2010). The design of the JAS includes rotating in new segments each year. Segments stay in the JAS sample for five years. Each year's sample is comprised of segments from each of five rotations. Thus, the 2009 JAS contained segments that were rotated into the sample in 2009, 2008, 2007, 2006 and 2005. The sampling design of the FNRP targeted the 20% of JAS segments that were newly rotated in for 2009 ("2009 segments"). All tracts in 2009 segments that were non-agricultural or estimated in JAS were selected for FNRP.

Current NASS procedures define a tract as a unique land operating arrangement, as determined during the JAS. However, for densely populated tracts, it is possible that multiple operations (places of interest) may have been erroneously included for any particular tract during this parent survey enumeration. For the purposes of the follow-up FNRP study, for a selected tract, all places of interest were considered subtracts. Subtracts were subsampled if there were 8 or more per tract. The FNRP sample consisted of 10,204 tracts, which resulted in a total of 17,191 subtracts.

Recommendations from FNRP were to make changes to screening procedures to improve the quality of information obtained in JAS, based on analysis of the misclassification of tracts as farms/non-farms. The results of those recommendations may first be seen in the 2010 JAS, but misclassification will certainly persist. An annual follow-up will allow researchers to monitor misclassification rates for farms/non-farms and measurement error for other variables.

2. Major Findings of FNRP

An analysis of the impact of JAS screening procedures used in FNRP was completed by Abreu, McCarthy and Colburn (2010). A major finding of this work is that, assuming misclassification rates are the same for all rotations, the JAS indication of number of farms would increase by approximately 580,000 farms using FNRP data. We will refer to this as the "FNRP adjustment" to the JAS indication; see Table 1. The bulk of these farms were "found" in tracts that had been identified as non-agricultural with no potential in JAS. On the order of 45% of tracts are pre-screened into this category in a typical JAS. In FNRP, 6% of the sampled subtracts selected from this category were determined to be farms, resulting in 500,000 of the FNRP adjustment. Another 75,000 of the FNRP adjustment came from tracts pre-screened as non-agricultural with either potential for agriculture or unknown potential. The remaining increase in the JAS indication came

[1] NASS has a two year collaborative research program with the National Institute of Statistical Sciences (NISS) called the Cross-Sector Research in Residence Program. This program is composed of three academic-government teams focusing on important NASS research issues. One of the teams was entrusted to work on potential improvements to the methodology and design of the June Area Survey.

from tracts that had to be estimated in JAS. Although most tracts (92%) that had been estimated as farms in JAS were confirmed as such in FNRP, approximately 30% of those that had been estimated as non-farms were identified as farms in FNRP. The net increase in the JAS farm numbers indication from estimated tracts was about 5,000 farms.

Table 1: FNRP Results by type of tract

Type of Tract	FNRP Sample Size (subtracts)	Number of FNRP farms	Net Expanded Number of Farms
Estimated as farm	1,591	1,466	(7,822)
Estimated as non-farm	121	37	13,032
Non-agricultural: potential	487	95	38,346
Non-agricultural: unknown potential	364	56	37,479
Non-agricultural: no potential	14,628	905	500,338
FNRP Total	17,191	2,559	581,373

Rates of "conversion" (subtracts that were identified as non-farms in JAS and as farms in FNRP) varied by state and stratum. Nationally, tracts that had agricultural potential or unknown potential in JAS had conversion rates of about 20% and 15%, respectively. However, within a state, sample sizes were typically less than 10 tracts per stratum, making estimates of conversion rates unreliable at the state level. The conversion rate for tracts that were identified as having no agricultural potential in JAS was 6% overall. About 95% of the strata had conversion rates of less than 17% for no-potential subtracts. However, tracts without potential from strata in the 40's (low rates of cultivation) contributed over 237,000 of the increase to the JAS indication (over half of the FNRP adjustment).

FNRP results are used as guidelines for the ALUS design, but ALUS will be able to detect different types of trends as well. Due to the experience the enumerators gained in conducting FNRP, the changes in JAS protocols made following FNRP and the fact that FNRP included only 2009 segments, results from ALUS may be quite different.

3. Overall Sampling Design

As in JAS, ALUS will be a stratified sample of segments, using JAS strata and sampling across rotations. Segments that are eligible for inclusion in ALUS must have at least one tract that was pre-screened as non-agricultural (regardless of potential) or that was estimated in JAS (as either farm or non-farm). For a selected segment, all tracts satisfying one of these criteria will be re-evaluated using the FNRP questionnaire. In the 2009 JAS, over 90% of all segments would have been considered eligible for ALUS. This collection of eligible segments will be called the ALUS population.

The sample allocation of segments to each state-stratum combination considers two factors: the proportion of the ALUS population in the stratum and the proportion of the FNRP adjustment from non-agricultural tracts in the stratum. The latter simultaneously accounts for the number of converted non-agricultural tracts and the expansion factors associated with them, allowing states and strata that contributed most to the FNRP adjustment to be targeted. Tracts that were estimated as farms or non-farms in JAS contributed little to the FNRP adjustment, so this information is not included in choosing allocations to the strata in the ALUS sample.

Table 2: Guidelines for ALUS Allocation Scheme

Strata	Proportion of ALUS-eligible segments in 2009 JAS	Proportion of FNRP adjustment from non-agricultural tracts	Suggested Proportion of ALUS sample
10s	53%	16%	27%
20s	26%	34%	30%
30s	3%	<1%	3%
40s	17%	50%	40%
50s	<1%	<1%	0%
Total	10,168 segments	576,000 farms	

In JAS, the sampling scheme favors cultivated areas. For ALUS, the sampling will lean more heavily on moderately and less cultivated strata where the largest portion of the FNRP adjustment originates. For example, strata 10s (10, 11, …) are highly cultivated areas. The exact stratum definition varies from state to state, but this may be more than 50% cultivated land. In JAS, over half of the selected segments are from these strata. However, 10s made up only 16% of the FNRP adjustment arising from non-agricultural tracts, so only about 27% of the ALUS sample will come from these strata. The sample will be evenly distributed over the five rotations, with approximately 20% of the ALUS sample selected from each. This will allow modeling of the effect of the number of years a segment has been in the survey on misclassification rates.

Within each stratum of the ALUS population, segments will be selected with probability proportional to size (pps) sampling where the size measure of a segment is defined as

$$\text{size}= \begin{pmatrix}\text{number of tracts pre-screened}\\ \text{as non-agricultural}\end{pmatrix}+\begin{pmatrix}\text{number of tracts}\\ \text{estimated as non-farm}\end{pmatrix}+0.1*\begin{pmatrix}\text{number of tracts}\\ \text{estimated as farm}\end{pmatrix}.$$

The rationale for the size measure comes from FNRP, but does not depend heavily on the specific results of that study. As noted previously, non-agricultural tracts/subtracts made up the vast majority of the FNRP adjustment. Estimated tracts had less impact on the FNRP adjustment. These tracts do not affect allocation (sample size) in each stratum, but are used in helping to select segments once allocations are determined. In a typical JAS, few tracts are estimated as non-farms (around 400 in 2009), but one-third of these estimated tracts converted to farms in FNRP. Thus, estimated tracts will be over-sampled relative to their contribution to the FNRP adjustment. Because most tracts (92%) estimated as farms in JAS were confirmed as farms in FNRP, ALUS will not target segments that have *only* ALUS tracts from estimated farms. This is reflected in the multiplier of 0.1 on the number of tracts estimated as farms in the size measure. If a segment is selected, all ALUS eligible tracts will be sampled, including those estimated as farms.

Within selected tracts, sampling rates of subtracts will be the same as FNRP. That is, if the tract contains 7 dwellings or less, then all are sampled. If the tract contains 8-20 dwellings, half are sampled. If there are more than 20 dwellings, one-sixth are sampled.

4. Example Allocations

Both the standard error and the cost of the proposed ALUS were investigated using FNRP data. This required development of specific example allocations for each state and stratum. In practice, ALUS allocations will need to be determined each year after JAS data are collected because a segment's weight for the pps sampling will depend on its JAS classification.

For different national sample sizes, a proposed stratified allocation of segments was developed using strata that are combinations of state and JAS strata. National sample sizes ranging from 500 to 5000 segments were considered. Note that the approximate size of the FNRP sample was about 2200 segments. The numbers of segments allocated to each stratum (across the nation) are summarized in Table 3.

Table 3: Example (Approximate) Allocation of Segments to Strata

Sample size (segments)	JAS stratum				
	10s	20s	30s	40s	50s
5000	1350	1500	150	2000	0
4500	1215	1350	135	1800	0
4000	1080	1200	120	1600	0
3500	845	1050	105	1400	0
3000	810	900	90	1200	0
2500	675	750	75	1000	0
2000	540	600	60	800	0
1000	270	300	30	400	0
500	135	150	15	200	0

The design attempts to maintain a minimum number of segments (between 1 and 4 depending on the total sample size) in each state and stratum combination. JAS 2009 data were used to approximate the number of segments in the ALUS population for a particular stratum in a typical JAS. If the allocation was larger than the ALUS population in the 2009 JAS, then the allocation was reduced. In practice, these allocations would need to be adjusted at least slightly each year based on the JAS data and resulting ALUS population sizes. In addition, any stratum that had a sample size of zero in FNRP was not included in the example allocations. This was done for estimation purposes and is not recommended for actual ALUS allocations.

Table 4 summarizes the anticipated results and costs of each sample size. The anticipated national level CV and standard error on the number of farms adjustment are calculated following the method used for FNRP. That is, the appropriate formulae for follow-on surveys were used. (Kott, 1990)

Table 4: Summary of Proposed ALUS Allocations

Approximate Sample Size (segments)	Anticipated CV of ALUS adjustment	Anticipated Standard Error of ALUS adjustment	Anticipated Cost to States
5000	6.5	37,000	$909,000
4500	6.8	39,000	$818,000
4000	7.1	41,000	$726,000
3500	7.7	44,000	$639,000
3000	8.2	47,000	$550,000
2500	9.0	52,000	$458,000
FNRP: 2200	*10.9*	*63,000*	*$412,000*
2000	10.0	58,000	$366,000
1000	13.8	80,000	$194,000
500	18.8	108,000	$107,000

JAS 2008 cost data are available for each state on a per segment basis. These costs are based on an enumerator visiting every tract within a selected segment. For ALUS, we assume that 56% of tracts in a selected segment will be ALUS-eligible. The value of 56% is derived from calibrating the cost of a FNRP size sample (2200 segments) to match the actual cost of FNRP ($412,000). Anticipated costs are summarized in Table 4. Cost data are only available at the state level, not at the stratum level. The anticipated cost assumes that segments in all strata have the same cost. Although this calculation is quite rough, more sophisticated methods would probably not result in marked improvement of the cost approximations. Note that these estimates only include approximate cost to the states. In FNRP, real estate parcel data, which cost $92,000 for a one-year license, were used to improve the quality of the names and addresses for non-agricultural tracts. This cost is not considered here.

5. Using ALUS results to adjust JAS estimates

The FNRP questionnaire was designed to target misclassification of farm status, to capture data on the type of farms that were believed to have been misclassified and for multiple modes of data collection (face-to-face, phone, mail, etc). It is essentially a shortened version of the JAS questionnaire. NASS should consider redesigning this questionnaire for ALUS to collect data on as many JAS variables as possible for farms newly identified in ALUS. This will allow ALUS results to be used to adjust more than just farm numbers indications. In particular, use of the full JAS questionnaire, a shortened version of the JAS questionnaire[2] or an extended version of the FNRP questionnaire should be considered. If an operation is still inaccessible or refuses for ALUS, data should be collected about the source of data used to estimate for that operation. Regardless of the questionnaire, it will be necessary to identify data coming from the original JAS separately from ALUS data.

[2] The Agricultural Coverage Evaluation Survey (ACES) was a supplemental survey to the 2007 JAS. The additional ACES segments targeted farming operations that typically had lower coverage rates on the Census of Agriculture list frame. ACES segments were sent a shortened version of the JAS questionnaire.

The combination of JAS and ALUS can be considered a two-phase sample. JAS is the first phase of the sample; then a sub-sample of JAS segments are selected for ALUS. Provided that each phase makes use of a probability sampling design for which the inclusion probabilities are known, standard results can be used to construct a design-based estimator. (Sarndal and Swensson, 1987) This methodology can be applied not only to estimates of number of farms but to all variables collected in the ALUS. Thus, although the primary impetus for this work is to improve estimates of the number of farms, it can improve estimates of other important variables. In particular, farms that are "missed" in JAS will not have values for many JAS variables. Those that are newly identified in ALUS will have accompanying data that can be used to adjust any variables common to both ALUS and JAS. For this reason, the FNRP questionnaire should be reviewed within NASS to determine whether other information should be gathered during ALUS. It is expected that misclassification and non-response will still occur in ALUS. However, this follow-up survey will provide valuable information for adjusting estimates and should reduce the amount of non-response.

Researchers have proposed alternate methods of adjusting JAS for misclassification error. Notably, the approach of matching JAS records to the annual list frame is a competing approach. The results of ALUS would likely provide lower CVs for indications of the number of farms, would provide annual monitoring of classification error that may inform the data collection process and would provide improved indications for other variables. However, the cost of conducting ALUS is non-trivial. Researchers intend to pursue further comparisons of these two approaches.

Acknowledgements

The authors of this paper are all members of a team brought together under a cooperative agreement between the National Institute of Statistical Sciences and USDA's National Agricultural Statistics Service.

References

Abreu, D. A., J. S. McCarthy, L. A. Colburn (2010). Impact of the Screening Procedures of the June Area Survey on the Number of Farms Estimates. Research and Development Division. RDD Research Report #RDD-10-03. Washington, DC: USDA, National Agricultural Statistics Service.

Kott, P. S. (1990). Mathematical Formulae for the 1989 Survey Processing System (SPS) Summary. NASS Staff Report, SRB-90-08, National Agricultural Statistics Service, USDA.

Sarndal, C-E. and B. Swensson (1987). A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. *International Statistical Review*. v 55(3):279-29