Precise Estimates of the Number of Farms in the United States

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

Abstract
Each June, the USDA's National Agricultural Statistics Service (NASS) provides an estimate of the number of farms in the United States, based primarily on the June Area Survey (JAS). Every 5 years, another indication (preliminary estimate) of the number of farms is obtained from the Census of Agriculture. In 2007, the indications from the JAS and the census were too far apart to be attributed entirely to chance. Initial reviews indicated that the differences were primarily a consequence of an undercount of small farms, especially those with minority owners. Using information from the 2007 Census and from the 2009 Farm Numbers Research Project (FNRP), misclassification in the JAS was determined to be a source of the undercount. An additional source is the treatment of incomplete responses. Possible revisions to the methodology of the JAS survey are suggested.

Key Words: survey, misclassification, undercount, non-response

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. Two of those surveys are the annual June Area Survey (JAS) and the Census of Agriculture, which is conducted every five years in years ending in 2 and 7. The JAS is one of the largest annual NASS survey projects and 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 and measures of sampling coverage. The Census of Agriculture provides detailed information on U.S. farms and ranches and the people who operate them. Both the JAS and the census provide information on a wide array of agricultural activities. In particular, each provides an indication (a preliminary estimate) of the number of farms in the U.S., and this objective is the focus of work reported here.

Before considering each survey, we need to understand the definition of a farm. In the U.S., a farm is any place from which $1000 or more of agricultural products were produced and sold or normally would have been sold during the year. This includes the corn fields, ranches, and large vegetable farms that most would

identify as a farm. It also includes the individual with a large backyard garden who sells surplus produce in the local farmers' market and the person who has five horses, perhaps for the family's recreational activities. Thus, one of the challenges of counting the number of U.S. farms is that many people who have farms according to the definition of farms, do not think of themselves as farmers.

Table 1: Land Stratification Used in the June Area Survey

Stratum Number	*Definition*
11	General cropland, greater than 75% cultivated
12	General cropland, 50 to 74% cultivated
20	General cropland, 15 to 49% cultivated
31	Ag-Urban, less than 15% cultivated, more than 100 dwellings per square mile, residential mixed with agriculture
32	Residential/Commercial, no cultivation, more than 100 dwellings per square mile
40	Open land, less than 15% cultivated
50	Non-agricultural, variable size segments

The June Area survey has an area sampling frame. All land in the U.S., except Alaska, is stratified by land use within a state. The specific stratum types vary with state; one such stratification is given in Table 1. The primary sampling units (PSU) provide complete coverage of all agricultural 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 new 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 operating 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 Census of Agriculture is a dual-frame survey conducted every five years. For the census, a list frame is developed based on a variety of sources. Sources include lists available from state and federal governments, producer associations, seed growers, pesticide applicators, veterinarians, and marketing associations. Data are collected primarily by mail. Adjustments are made for non-response. JAS is used to adjust for undercoverage.

The Census of Agriculture provides a base for the number of farms in census years, while the JAS provides an annual indication of the number of farms. At the end of each five-year period, the annual estimates are revised based on intercensal trends. As seen in Figure 1, the number of farms was steadily decreasing from 2000 to 2006, but had a dramatic increase in 2007. In that year, the indications from the census and the JAS were too far apart to be attributable to sampling error alone, leading to a major adjustment in published estimates. In reviewing the

results from the 2007 Census and the 2007 JAS, it became evident that many small farms were being missed in the JAS.

Through a cooperative agreement between NASS and the National Institute of Statistical Sciences (NISS), a research team was created to review the methodology associated with the JAS and to recommend changes that would address the undercount. The team consists of two NASS researchers, two university faculty members, a post doctoral fellow, and a graduate student. Here we report preliminary results of that team's work.

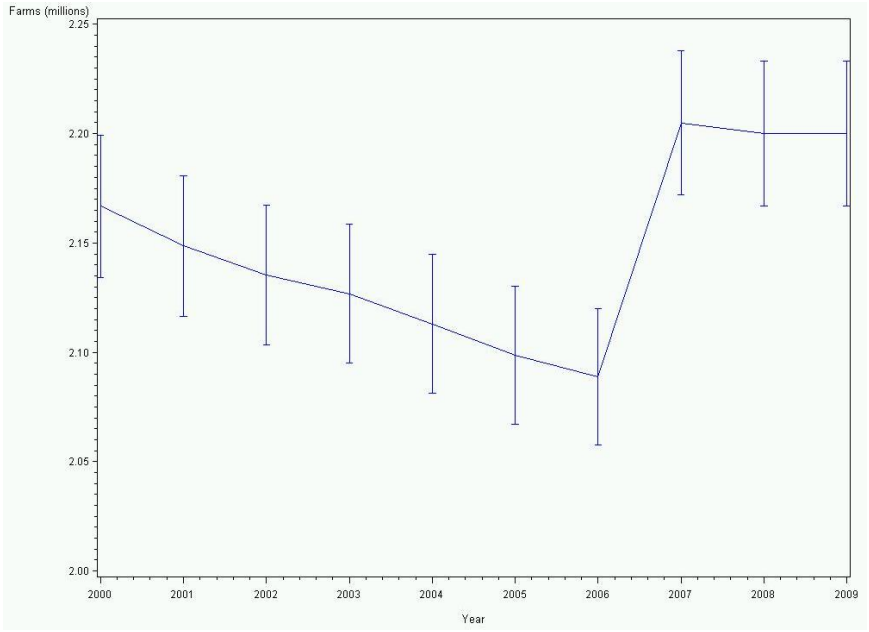

Figure 1: Published estimates of the number of U.S. farms from 2000 to 2009 and bars with a length of one standard error on either side of the estimate.

2. Misclassification

Because the JAS has an area frame, any undercount is likely the consequence of misclassification error, a consequence of declaring some farm tracts to be non-farm tracts. Each JAS agricultural tract is identified as a farm or non-farm based on whether it had $1000 in sales of agricultural products or 1,000 points based on agricultural products produced (if sales were less than $1000). All non-agricultural tracts are considered non-farms. To assess the extent to which misclassification might be a problem, an attempt was made to match each JAS tract to a census record. Most tracts were determined to be a farm or a non-farm in both surveys in which case misclassification is not present. Some tracts determined to be non-farms in the JAS were found to be farms in the census. Similarly, some tracts identified as farms in the JAS were classified as non-farms in the census. A limited amount of this disparity could be due to the fact that the JAS is conducted in June and the census in December, and some tracts could change status during this short time period. However, the vast majority of the

disagreeing classifications are likely due to misclassification on one of the two surveys. After discussion, it was decided to consider any tract determined to be a farm in either the JAS or the census to be a farm. So, the focus here is on the JAS non-farm tracts (see Figure 2).

Of the 59,223 JAS non-farm tracts, 47,928 (81%) could not be matched to a census record. This is not surprising because the list frame only contains records for known or likely farm operations. Of the 11,295 JAS non-farm tracts that matched to a census record, 8,277 matched to a census non-farm. Thus, about 73% of these tracts were found to be properly classified. The remaining 3,068 JAS non-farm tracts (27%) matched to a census farm. Of these, 1,090 were determined to be agricultural tracts in the JAS that did not have enough sales or agricultural products to qualify as a farm, but the remaining 1,978 tracts were identified as non-agricultural tracts during prescreening for the JAS.

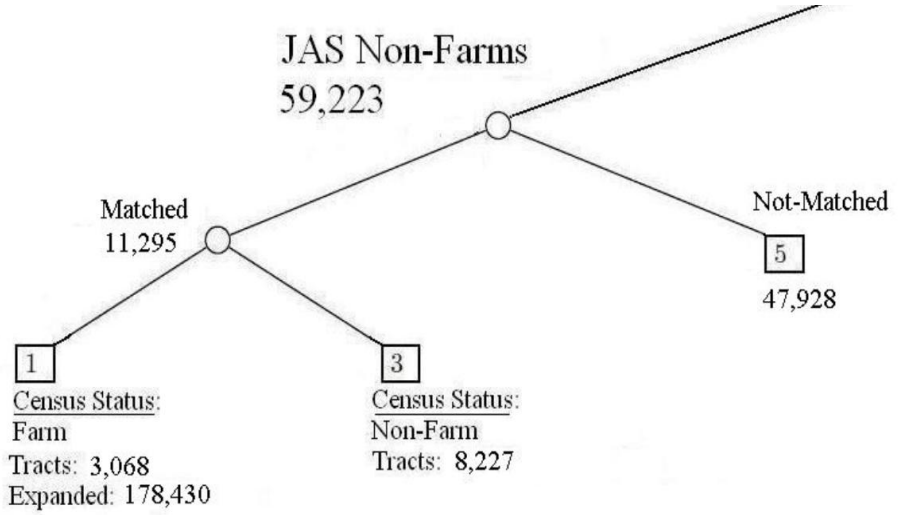

Figure 2: Results of matching JAS non-farms to census records

The traditional JAS indication of the number of U.S. farms is

$$\sum_{i \in F} \pi_i^{-1} t_i \qquad (1)$$

where F is the set of tracts identified as farms in the JAS, π_i is the probability of inclusion, and t_i is the tract-to-farm ratio. The tract-to-farm ratio is the proportion of the total farm operation in the sampled tract; that is, it is the ratio of tract size to total farm size. When assessment of misclassification, such as matching to the census, is possible, the following provides a design-based indication of the number of U.S. farms, adjusting for misclassification:

$$\sum_{i \in F} \pi_i^{-1} t_i + \sum_{i \in \bar{F}} \pi_i^{-1} y_i t_i \qquad (2)$$

where $\bar{F}$ is the set of tracts identified as non-farms in the JAS and y_i is an indicator of whether the tract is a farm. This estimate is still likely to be an undercount because not every JAS record could be matched to a census record. Some of the failures to match could be a consequence of the challenges of record linkage. A conservative approach was used in the matching process so that only records that were clearly matches were counted as having matched. With this approach, some matches were undoubtedly missed. In addition, some tracts could be misclassified as nonfarms in both surveys (see Abreu, et al. 2010 for a more detailed discussion of the matching).

The census is conducted only every fifth year so a follow-up study to assess misclassification is not possible for all years. For non-census years, the misclassification adjustment to the JAS estimate of the number of U.S. farms could possibly be based on models of the probability of misclassification based on census years for which follow-up on misclassification is possible. The model is needed to estimate the probability of misclassification and the tract-to-farm ratio in the second term in equation (2). The inclusion probability follows from the design. An initial modeling approach is as follows:

(1) Let $u \sim$ Bernoulli(π_u) be an indicator of whether ($u = 1$) or not ($u = 0$) a tract has a census follow-up (was matched with a census record)
(2) Let ($f \,|u$) ~ Bernoulli(π_f) I($u = 1$) be an indicator of whether a tract with follow-up is ($f = 1$) or is not ($f = 0$) a farm

(3) Let ($z \,|u, f$) ~ Bernoulli(π_z)I($u = 1$) I($f = 1$) be an indicator of whether a tract with follow-up and is a farm has a tract-to-farm ratio of 1 ($z = 0$) or a value less than 1 ($z = 1$).
(4) Let ($t \,|u, f, z$) ~ Beta(μ,τ)I($u = 1$)I($f = 1$)I($z = 1$) + 1 I($u = 1$)I($f = 1$)I($z = 0$) where μ is the mean and τ is the scale parameter of the beta distribution.

The parameters, π_u, π_f, π_z and μ are unknown. The probabilities and the beta mean and scale parameter were estimated using logistic and beta regression. Because only information collected for both farms and non-farms could be used to construct the models, only two covariates were available: land-use stratum (50% cultivated, 15 to 50% cultivated, agricultural urban/commercial, and less than 15% agricultural or non-agricultural) and how the tract was classified during screening (agricultural, non-agricultural with potential, non-agricultural with potential unknown, and non-agricultural with no potential). Then

$$\hat{\mathrm{E}}(t) = \hat{\pi}_f \hat{\pi}_u [(\hat{\mu} - 1)\hat{\pi}_z + 1] \tag{3}$$

Thus,

$$\sum_{i \in F} \pi_i^{-1} t_i + \sum_{i \in \bar{F}} \pi_i^{-1} \hat{\mathrm{E}}(t_i) \tag{4}$$

is the model-based estimate of the number of U.S. farms. Notice that modeling only affects the second part, which is the adjustment for the undercount. Because tracts that were not matched to a census record are probably less likely to be a farm than those that were matched, the tract-to-farm ratio for tracts that are not matched to a census record is taken to be zero. However, this is likely to result in a continuing undercount when farm tracts were not matched to a census record (see Lamas, et al 2010, for more details)

3. Non-Response

Non-response in the JAS occurs when an agricultural tract operator is either inaccessible for or refuses an interview. Currently, agricultural activity in these tracts is estimated. Tract-level data are estimated by enumerators. To obtain farm-level data, field offices search auxiliary sources. When available, these sources generally provide good information, but they are not available for all tracts. In those cases, median imputation is used to complete farm-level information.

In 2009, the Farm Numbers Research Project (FNRP) was conducted. As part of that effort, enumerators were able to obtain farm-level data on 595 tracts that had been estimated in June. Actual and estimated farm-level data were compared. The estimated and actual tract-to-farm ratios showed a high level of discordance for these tracts. In addition, the actual tract-to-farm ratio was highly correlated with other farm-level covariates. Currently, it is not possible to distinguish JAS records completed based on auxiliary sources from those records completed using imputation; that is, we are unable to determine the quality of farm-level information at the analysis stage. A proposal has been put forward that would allow the approach used to complete farm-level information to determined in the future. Until this is available, all estimated tracts will be treated as non-respondents.

In Figure 3, non-response is outlined. To account for this non-response, first consider that the usual JAS estimate *without* the estimated tracts may be written as

$$\sum_{i \in R} \pi_i^{-1} y_i t_i \qquad (5)$$

where R is the set of respondents, π_i is the inclusion probability of respondent i, y_i is an indicator of whether the tract is a farm, and t_i is the tract-to-farm ratio. Let ϕ_i be the probability of response from respondent. Then the usual JAS estimate, adjusted for non-response, is

$$\sum_{i \in R} \pi_i^{-1} \phi_i^{-1} y_i t_i \qquad (6)$$

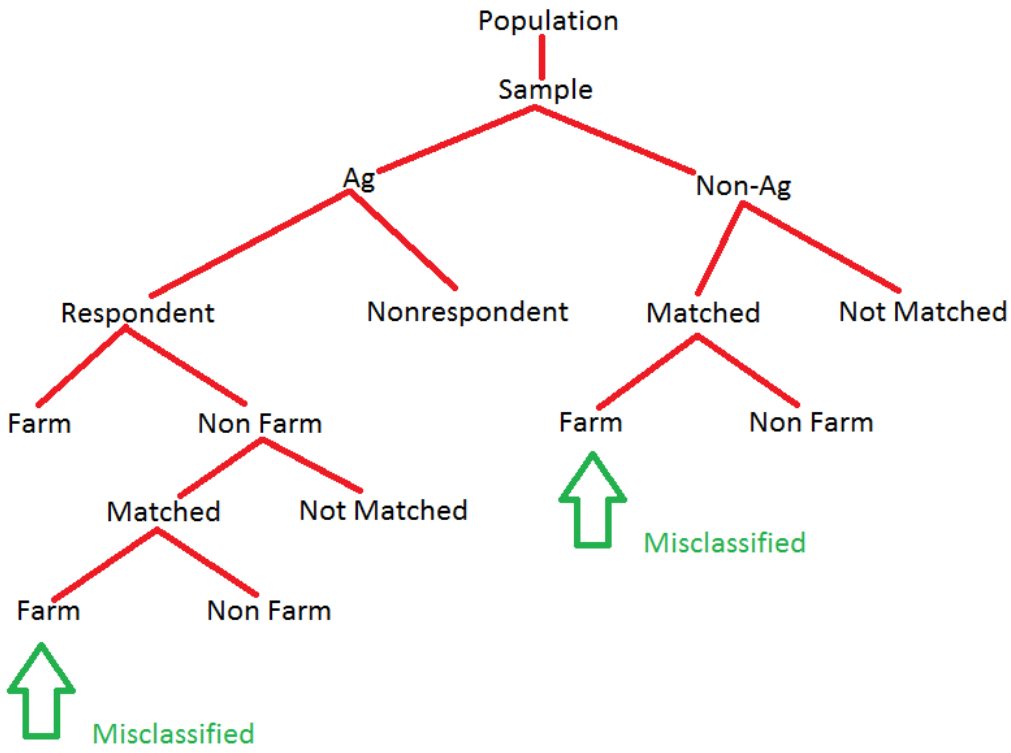

Figure 3: Conceptualization of sample with non-response and misclassification.

The probability of response must be estimated, and a logistic regression model was developed for this purpose. Only information common to both respondents and non-respondents could be used in this modeling process. Thus, land-use stratum, state, and tract-level items were the only explanatory variables that could be used in the model. As with misclassification, land-use stratum was collapsed into five categories. Because farm-level information was not available for non-respondents, these variables could not be included in the model. Based on the 2009 FNRP, the probability of response is correlated with farm-level information so the inability to include these variables in the analysis is a limitation.

After modeling the response probability, the JAS indication of the number of U.S. farms can be adjusted for non-response as follows:

$$\sum_{i \in R} \pi_i^{-1} \hat{\varphi}_i^{-1} y_i t_i \qquad (7)$$

where $\hat{\phi}_i$ is the estimated probability of a response for tract *i*.

By putting together the methods for misclassification and non-response, the JAS indication of farm numbers can be adjusted for both. For the design-based approach, equation (2) shows the adjustment for misclassification. Using that with equation (7), we have

$$\sum_{i \in M \cup R} \pi_i^{-1} \hat{\phi}_i^{-1} y_i t_i \qquad (8)$$

where M is the set of all JAS tracts that have had their responses updated as a result of the matching procedure and R continues to be the set of all JAS tracts associated with respondents to the JAS. For the model-based approach, adding an adjustment for non-response to the adjustment for misclassification as shown in equation (4), we have

$$\sum_{i \in F} \pi_i^{-1} \hat{\varphi}_i^{-1} t_i + \sum_{i \in \bar{F}} \pi_i^{-1} \hat{\varphi}_i^{-1} \hat{\mathrm{E}}(t_i)$$

An estimate of variance has been developed for the design-based estimate, and work is underway to derive the variance for the model-based adjusted estimate (see Lopiano, et al. 2010 for more details on modeling for non-response).

Both the design-based and model-based approaches for adjusting for misclassification and non-response were applied to the 2007 JAS. The results from the two were similar. For each, the original JAS indication represented 76% of the final adjusted number and the non-response and misclassification adjustments accounted for 8% and 16% of the adjusted indication of farm numbers. These were within error of the 2007 Census of Agriculture published number of 2,204,792 U.S. farms.

The 2007 Census list frame was used to develop the models for adjusting for misclassification and non-response, and these were applied to obtain adjustments for the 2009 JAS indication of the number of U.S. farms. During that process, a critical assumption was that misclassification and non-response behaved the same in both 2007 and 2009. Because the system had been in place for a number of years and no large changes were made, this assumption seemed reasonable. However, in 2009, the same enumerators who participated in the JAS were also involved in the FNRP. During that process, they revisited all tracts that were either estimated or determined during the screening process to be non-agricultural. Most witnessed that some of their tracts changed classification. This is likely to have resulted in a change in behaviour beginning with the 2010 JAS. Further, NASS is instituting a host of measures designed to reduce misclassification. During this time of change, using a model developed in one year in subsequent years is unlikely to yield satisfactory estimates.

NASS has a continuing process of updating the list frame for its yearly list-based survey and to provide a base for the Census of Agriculture. The natural question is whether this list frame could be used in non-census years to adjust for misclassification. The challenge with this is that the list frame in non-census years contains numerous records that are not farms. Based on previous studies, it is thought that about 30% of the records on the census list frame are not farms.

Whether adjustments can be made for these records is a current research topic. Thus, the viability of using the list frame in non-census years must be assessed.

4. An Alternative: An Annual Follow-up Survey

An alternative to matching JAS records to the list frame is to conduct an annual follow-up survey that will allow corrections for misclassification and non-response. Such a survey has been proposed and is called the Annual Land Utilization Survey (ALUS). It builds on the 2009 FNRP. Although the FNRP focused on those non-agricultural tracts entering the sample in 2009, ALUS considers all non-agricultural tracts. However, unlike FNRP which included all non-agricultural and estimated tracts in that rotation, a sample of the non-agricultural and estimated tracts would be used in ALUS. In Table 2, the relative contribution of major strata groups to the FNRP adjustment as well as the pool of non-agricultural tracts available are displayed.

The proposed ALUS design is similar to that of the FNRP. Segments would be allocated proportionally across states (according to their contribution to the FNRP adjustment) and across rotations. ALUS strata would be the same as the JAS strata, that is, by state, land-use, and rotation. Within ALUS strata, segments would be selected with probability proportional to size where size is defined to be the sum of (1) the number of non-agricultural tracts, (2) the number of tracts estimated as farms, and (3) a tenth of the number of tracts estimated as non-farms. Within a selected segment, all tracts that are ALUS-eligible would be revisited. ALUS would be considered a second-phase sample from the JAS. Using two-phase estimators, the JAS estimate of farm numbers would be adjusted. By building on FNRP, ALUS should have improved estimates (see Arroway, et al., for a more complete discussion of ALUS.

Table 2: Percentages of FNRP Adjustment from Non-agricultural Tracts and of ALUS-eligible Segments in 2009 JAS

Strata	*Percentage of FNRP Adjustment from Non-Agricultural Tracts*	*Percentage of ALUS-eligible Segments in 2009 JAS*
10s: Highly cultivated	16%	53%
20s: Moderately cultivated	34%	26%
30s: Agricultural/urban	<1%	3%
40s: Low cultivation	50%	17%
50s: Non-ag (“known”)	<1%	<1%
TOTAL	FNRP adjustment from non-agricultural = 576,000 Farms	10,168 >90% of all JAS segments

5. Conclusions

Misclassification in the JAS has led to underestimates of the number of U.S. farms. Adjustments have been proposed using census records for evaluation of misclassification in census years. Although modeling from census years has been

used for these adjustments in non-census years, NASS is making efforts to reduce misclassification. Consequently, models constructed during census years for non-census years will no longer be appropriate. NASS keeps a list frame, but it is not evident that it can be used to obtain accurate adjustments for misclassification.

To more appropriately account for non-response, it is critical that tracts with good estimates for farm-level information be separated from other estimated tracts. Once these can be separated, the estimated tracts without good farm-level information should be treated as non-respondents and the JAS indication adjusted. However, the unavailability of farm-level information makes estimating the probability of response challenging.

Many of the challenges of misclassification and non-response could be addressed with a follow-up (second stage), survey, and the ALUS has been proposed for this purpose. However, the ALUS would be expensive in terms of money, time, and personnel. If some other method of adjustment could be used in at least some of the years, then that approach would be preferred. At present, this would seem to require the use of the list frame in non-census years. The ability to use the list frame to produce accurate corrections remains an open issue needing further study.

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, Denise A., Pam Arroway, Andrea C. Lamas, Kenneth K. Lopiano, and Linda J. Young. 2010. Using the Census of Agriculture list frame to assess misclassification in the June Area Survey. *Proceedings of the Joint Statistical Meetings*.

Arroway, Pam, Denise A. Abreu, Andrea C. Lamas, Kenneth K. Lopiano, and Linda J. Young. 2010. An alternate approach to assessing misclassification in JAS. *Proceedings of the Joint Statistical Meetings*.

Lamas, Andrea C., Denise Abreu, Pam Arroway, Andrea C. Lamas, Kenneth K. Lopiano, and Linda J. Young. 2010. Modeling misclassification in the June Area Survey. *Proceedings of the Joint Statistical Meetings*.

Lopiano, Kenneth K., Denise Abreu, Pam Arroway, Andrea C. Lamas, and Linda J. Young. 2010. Adjusting the June Area Survey for non-response and misclassification. *Proceedings of the Joint Statistical Meetings*.