Statistical Methodology for the 2012 U.S. Census of Agriculture

Linda J. Young[1,2,3], Andrea C. Lamas[1], Denise A. Abreu[1],
Shu Wang[2], and Daniel Adrian[1]
[1]USDA National Agricultural Statistics Service, Fairfax, Virginia, USA
[2]University of Florida, Gainesville, Florida 32611-0339, USA
[3]Corresponding author: Linda J. Young, e-mail: LJYoung@ufl.edu

Abstract

The U.S. Department of Agriculture's National Agricultural Statistics Service (NASS) conducts the quinquennial U.S. Census of Agriculture, in years ending in 2 and 7. Beginning in 2009, NASS conducted a series of research projects that led to the conclusion that the assumptions underpinning the analysis of the 2007 Census were no longer valid. Consequently, NASS has adopted a unified approach to accounting for non-response, under-coverage, and misclassification using capture-recapture methodology. The two surveys used for capture-recapture are the Census and the June Area Survey (JAS). Challenges, such as resolving farm status when an operation is classified as a farm (non-farm) by the JAS and a non-farm (farm) by the Census, are discussed. Accounting for uncertainty using jackknife methods is presented.

Key words: capture-recapture, misclassification, logistic regression, jackknife

I. Introduction

The U.S. Census of Agriculture is conducted every five years. A primary objective is to estimate the number of farms in the U.S. as well as the number within each state and county. A farm is defined to be any operation with at least a $1,000 in sales of agricultural products or the potential for $1,000 in sales. In 2007, Classification and Regression Trees were used to adjust for non-response. In an independent process, NASS used its June Area Survey (JAS) to account for under-coverage. Because the JAS is based on an area frame that contains all U.S. land, it was considered complete. However, in 2009, USDA's National Agricultural Statistics Service (NASS) conducted the Farm Numbers Research Project (FNRP) that identified substantial misclassification of farm operations in the JAS (Abreu, *et al*., 2010). Although efforts have been made to reduce JAS misclassification, some remains so the assumption that the JAS area frame is able to account for all under-coverage is no longer considered valid. Consequently, for the 2012 U.S. Census of Agriculture, NASS has adopted a capture-recapture framework as a unified approach to accounting for non-response, under-coverage, and misclassification. In this paper, the challenges of resolving farm status, modeling the probabilities of capture and misclassification, and obtaining a measure of uncertainty are considered.

2. The Census Mailing List and the June Area Survey

The Census of Agriculture uses a list frame, the Census Mailing List (CML). Thus, in creating the CML, the objective is to build a complete list of agricultural operations that meet the NASS farm definition. The list frame is used for other NASS surveys so the maintenance of the list frame is a major, on-going NASS effort. The efforts intensify in Census years and, at a pre-determined point in time, the list frame is frozen. Operations unlikely to be farms are trimmed, and the remaining operations comprise the CML.

The second survey used for the capture-recapture methods is the JAS, which uses the NASS area frame. The NASS area frame covers all land in the U.S., except for Alaska. An area frame is created for each state. Within the state, the land is stratified by agricultural characteristics, *e.g.*, at least 50% cultivated, forested, etc. Segments of approximately equal size are delineated within each stratum and designated on aerial photographs. A probability sample of segments is drawn within each stratum for the NASS annual area frame survey, known as the June Area Survey (JAS) (*see* Davies, 2009, for more information on the JAS design).

Sampled segments in the JAS are personally enumerated. Each operation identified within a segment boundary is known as a tract. Each tract is identified as either agricultural or non-agricultural during JAS pre-screening. Non-agricultural tracts are further classified into one of the three following categories: with potential, with unknown potential, or with no potential. Each JAS agricultural tract is identified as a farm or non-farm in June based on whether it had $1,000 in sales of agricultural products or 1,000 points based on the potential for agricultural products produced (if sales were less than $1,000). The 2012 JAS consisted of 11,085 sampled segments, and it was supplemented with 3,292 Agricultural Coverage Evaluation Survey (ACES) segments. ACES segments were selected to reduce the coefficient of variation (CV) for Census estimates of small and minority owned farms.

The JAS estimate for the number of farms is

$$\sum_{i \in F} w_i t_i \qquad (1)$$

where w_i is the expansion factor (reciprocal of the inclusion probability) associated with tract i, t_i is the tract-to-farm ratio (tract acres divided by total farm acres), and F is the set of sampled farm tracts. The product $w_i t_i$ is referred to as the JAS weight for farm i. For the purposes of the Census of Agriculture, it is important to note that responses are obtained from all agricultural tracts in the JAS; non-response is not present.

3. Capture-Recapture Methods

Capture-recapture methodology is the foundation of the methods used to adjust the 2012 Census of Agriculture for under-coverage, non-response, and misclassification. The ideas presented here draw heavily on the work conducted by the U.S. Census Bureau in preparation for the Accuracy and Coverage Evaluation of Census 2000 (U.S. Census Bureau, 2004) and for coverage measurement for Census 2010 (U.S. Census Bureau, 2008; National Research Council, 2008) as well as traditional capture-recapture methods developed for estimation of animal populations (*see* Chao, 2001; Seber, 2002). To implement capture-recapture methods, two independent surveys are required. The Census of Agriculture (based on the CML) and the JAS are taken to be those two surveys.

The JAS provides a wealth of information that can be used to model the probability that a JAS farm is captured by the Census. In addition to assuming that the two surveys (JAS and Census) are independent, the second basic assumption is that the proportion of JAS farms with a given set of characteristics captured by the Census is equal to the proportion of U.S. farms with those same characteristics captured by the Census. Thus, the CML records that overlap with JAS segments are matched with the JAS tracts (the JAS sample). The CML records that match with JAS tracts represent the Census sample. Note: The Census sample is a subset of the CML records and includes only those records matching a JAS tract (*see* Figure 1). Both agricultural and non-agricultural tracts are included in the matched dataset.

To illustrate the basic concept, suppose for the moment that each U.S. farm has the same probability, say $\pi = 0.5$, of being on the CML and responding as a farm on the Census. Then, through the Census, about half of all U.S. farms will be "captured." To obtain an estimate of the number of U.S. farms, the number of farms on the CML and

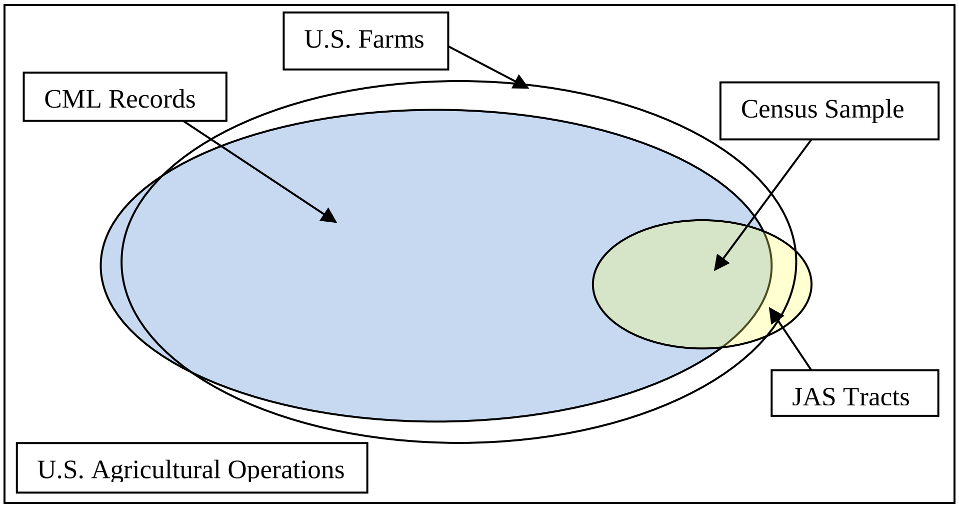

Figure 1. The Census sample is comprised of the Census records that match JAS tracts

responding to the Census is doubled, which is equivalent to dividing the number of responding farms by $\pi = 0.5$, the probability that a farm is on the CML and responds to the Census. Note: It does not matter why a farm is not recorded as a farm on the Census. It could be that it did not respond when mailed a Census form. It could be it did not receive a form. What matters is whether or not it was identified as a CML farm, a farm that is on the CML and responds as a farm on the Census.

The probability π that a U.S. farm is captured by the Census is not known and must be estimated. Based on the second assumption, if half of all U.S. farms are captured by the Census, half of all JAS farms are captured by the Census. By matching, all JAS tracts to CML records, it can be determined which JAS farms were captured by the Census and which were not. Then the probability a farm is captured by the Census, which for now is assumed to be the same for JAS farms and all U.S. farms is

$$\pi_1 = \frac{F_C}{N} \approx \frac{F_{JC}}{F_J}$$

where N is the number of U.S. farms (the parameter of interest) and F_C, F_J, and F_{JC} are the number of U.S. farms captured by the Census, the JAS, and both the JAS and Census, respectively.

Challenges quickly arise in applying the capture-recapture methods as described thus far. First, in the Census sample, a substantial portion of the records are identified as a farm (non-farm) on the Census and as a non-farm (farm) on the JAS. Such records are said to have conflicting or unresolved farm status. To resolve the farm status for these records, a logistic model of the probability an operation is a farm based on the records with resolved farm status is developed. The resulting missing data model is then used to estimate the probability p_i that each of the agricultural operations with unresolved farm status is a farm. The JAS weight is multiplied by p_i to obtain the weight for that record as a farm, and it is multiplied by $(1 - p_i)$ to obtain the weight for that record as a non-farm.

A second challenge occurs because the probability of capture is not the same for all U.S. farms. As examples, large farms have a higher probability of capture than small farms, and commodity farms, such as those growing corn or wheat, have a higher probability of capture than specialty farms, such as those growing Christmas trees or nuts. Several approaches have been used to adjust for this differential catchability. One approach is to partition the farms into groups so that the probability of being captured by the Census is about the same within each group (Alho1990, 1994; Alho, *et al.* 1993; U.S.

Census Bureau, 2004). Although the members within each of the constructed groups have similar capture probabilities, some variation remains.

In 2007, NASS used classification trees to form groups with similar probabilities of non-response (Cecere, 2009). This same approach could be used to form groups of similar capture probabilities. However, concerns about the bias associated with the estimates and the challenge of obtaining proper measures of uncertainty for this approach led the team to seek another alternative. Logistic regression was chosen to model the probability of capture. This has been used extensively to model the capture probabilities in wildlife studies (Chao, 2001; Armstrup, *et al*., 2005) and, in 2010, it was used in the Accuracy and Coverage Evaluation by the U.S. Census Bureau (U.S. Census Bureau, 2008). If the variables used to model the probability of capture are all categorical, then groups of members with similar capture probabilities are formed, as with demographic analysis and classification trees. However, continuous variables can also be used so that each member could have its own capture probability.

Misclassification of farms provides another challenge in the conduct of the Census of Agriculture. That is, some farms are said to be non-farms, and some non-farms are said to be farms. Recall that the JAS has responses for all agricultural tracts. Thus, the only farms that are captured by the Census but not the JAS are those that have been misclassified. The potential for misclassification is evident when agricultural operations surveyed in both the JAS and Census do not have the same classification. The 2007 Classification Error, which was based on 67 records, indicated that classification errors were made during both the JAS and the Census, but the Census tended to provide the correct classification more often than the JAS (Abreu, *et al.* 2009).

Misclassification can be confounded with the transition of non-farms to farms (births) and of farms to non-farms (deaths) during the six months between the conduct of the JAS and the Census. Seber (2002) and Otis, *et al*. (1978) considered the effect of births and deaths on the population estimate. Suppose that some farms become non-farms between June, when the JAS is conducted, and the end of the year, when the Census is conducted. As long as these are occurring at random so that the average probability of a farm surviving until the time of the Census is the same for farms that are part of the JAS and those that are not included in the JAS, the estimator is still valid. If some operations become farms between the JAS and Census, the population estimator is a valid estimator for population size at the time of the Census. If some operations become farms and others transition from farms to non-farms between the JAS and the Census, the population size at the time of the Census will tend to be over-estimated.

To illustrate the effect of misclassification, for the moment, suppose again that the probability of capture is the same for all farms. The probability π_1 of a U.S. farm being captured is then estimated by

$$p_1 = \frac{F_{JC}}{F_J}$$

Misclassification during the JAS affects both the numerator and denominator of the estimator, and misclassification during the Census affects the numerator, leading to bias in the estimator. Thus, the logistic model of the probability of capture must account for under-coverage, non-response, and misclassification.

Census misclassification occurs when $F_C \neq N_C$, where F_C denotes the agricultural operations identified as farms by their responses to the census questionnaire and N_C is the number of operations responding to the Census that are truly farms. In this case, the estimate of the number of farms would need to be adjusted by a factor of

$$\pi_2 = \frac{N_C}{F_C}$$

where π_2 is the proportion of agricultural operations correctly responding as farms on the Census. Because N_C is unknown, π_2 is unknown. As with capture, the probability of misclassification differs with farm operator and operation characteristics. Logistic regression is used to model the probability of misclassification given these characteristics. Note: The logistic model for capture accounts for the misclassification of a farm due to failure to identify a CML operation as a farm whereas the logistic model for misclassification adjusts for non-farm CML operations incorrectly identified as a farm.

The final estimate of the number of U.S. farms is the number of Census respondents, adjusted for non-response, under-coverage, and both misclassification of farms as non-farms and of non-farms as farms. To make this adjustment, the predicted probabilities from the logistic models for capture and misclassification are combined to form a weight for each of the CML farms. Then the dual system estimator (*DSE*) is the sum of these weights; that is,

$$DSE = \sum_{j \in \text{CML Farm}} \frac{p_{2j}}{p_{1j}}$$

where p_{1j} and p_{2j} are, respectively, the predicted probabilities of π_1 and π_2 for the *j*th CML farm. Estimates for the *i*th unit, such as state *i* or county *i*, denoted by DSE_i, are obtained by summing CML farms within that unit.

Following methods suggested by the U.S. Census Bureau (2004), jackknife methods are used to assess the uncertainty associated with the Census estimates at the national, state, and county levels. To conduct the jackknifing, *k* mutually exclusive and exhaustive groups of JAS segments are formed. The groups are selected using a stratified random design so that each group reflects the survey design and includes segments from across the U.S. In turn, each group, $j = 1, 2, \ldots, k$, is deleted and the $DSE_i^{(j)}$ is computed for each unit *i* at the specified geographical level, such as nation, state, or county, using the remaining $(k – 1)$ groups. An estimate of the variability associated with the estimated *DSE* is then

$$\sigma_i^2 = \frac{k-1}{k} \sum_{j=1}^{k} (DSE_i^{(j)} - DSE_i)^2$$

.

4. Conclusions

In 2007, 30% of the published number of 2.2 million U.S. farms represented a correction for non-response and under-coverage. The capture-recapture methods used for the 2012 U.S. Census of Agriculture account for farms missed by both the JAS and the Census. Although this alone would lead to an anticipation of an even greater adjustment in 2012, improvements in the Census processes may reduce the size of the adjustment. However, small farms continue to be challenging to identify, and a sizeable adjustment for them is needed.

Demographic characteristics of farm operators, such as age, sex, race and ethnicity, as well as farm characteristics, such as land in farms and type and size of farm, are of primary interest in the U.S. Census of Agriculture. The capture-recapture methods should provide more precise estimates of the numbers of operations with the characteristics of interest.

The measure of uncertainty is biased downwards. It does not account for the uncertainty associated with resolving farm status or for model uncertainty. Further, the bias often contributes more than the standard error when considering the mean squared error associated with the synthetic estimators proposed here. The work by Seiss and Mule

(2012) is being reviewed as a possible approach to determining the mean squared error associated with the *DSE*.

References

Abreu, D.A., N. Dickey and J. McCarthy (2009). 2007 Classification Error Survey for the United States Census of Agriculture. RDD Research Report Number RDD-09-03. Washington, DC: USDA, National Agricultural Statistics Service.

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 Number RDD-10-03. Washington, DC: USDA, National Agricultural Statistics Service.

Alho, J.M. 1990. Logistic regression in capture-recapture models. *Biometrics* 46:623-635.

Alho, J.M. 1994. Analysis of sample based capture-recapture experiments. *Journal of Official Statistics* 10: 245-256.

Alho, J.M., M.H. Mulry, K. Wurdeman, and J. Kim. 1993. Estimating heterogeneity in the probabilities of enumeration for dual-systems estimation. *Journal of the American Statistical Association* 88: 1130-1136.

Armstrup, Steven C., Trent L. McDonald, and Bryan F.J. Manly (eds). 2005. *Handbook of Capture-Recapture Analysis.* Princeton University Press: Princeton, NJ.

Cecere, Will. 2009. 2007 Census of Agriculture non-response methodology. *Proceedings of the Government Statistics Section, JSM 2009*. Pp. 2762-2769.

Chao, Anne. 2001. An overview of closed capture-recapture methods. *Journal of Agricultural, Biological, and Environmental Statistics* 6: 158-175.

Davies, Carrie (2009). Area Frame Design for Agricultural Surveys. Research and Development Division Internal Document.

National Research Council. 2007. *Research and Plans for Coverage Measurement in the 2010 Census: Interim Assessment: Panel on Correlation Bias and Coverage Measurement in the 2010 Decennial Census.* Ed. Robert M. Bell and Michael L. Cohen. The National Academies Press: Washington, DC.

National Research Council. 2008. *Coverage Measurement in the 2010 Census: Panel on Correlation Bias and Coverage Measurement in the 2010 Decennial Census.* Ed. Robert M. Bell and Michael L. Cohen. The National Academies Press: Washington, DC.

Otis, D.L., K.P. Burnham, G.C. White, and D.R. Anderson. 1978. Statistical inference from capture data on closed animal populations. *Wildlife Monographs*: 62:1-138.

Seber, G.A.F. 2002. *The Estimation of Animal Abundance and Related Parameters*, 2nd edition. The Blackburn Press: Caldwell, New Jersey.

Seiss, Mark and Thomas Mule. 2012. Census coverage measurement: Measurement of error in synthetic DSE estimates of states, counties and places. DSSD 2010 Census Coverage Measurement Memorandum Series. U.S. Census Bureau.

U.S. Census Bureau. 2004. *Accuracy and Coverage Evaluation of Census 2000: Design and Methodology*. September, 2004. Online: http://www.Census.gov/prod/2004pubs/dssd03-dm.pdf

U.S. Census Bureau. 2008. *2010 Census Coverage Measurement Estimation Methodology*. October, 2008. Online: http://www.Census.gov/coverage_measurement/pdfs/2010-E-18.pdf.