Adjusting an Area Frame Estimate for Misclassification Using a List Frame

Andrea C. Lamas[1], Denise A. Abreu[1], Hailin Sang[2], Pam Arroway[3], Kenneth K. Lopiano[4], Linda J. Young[4]

[1]National Agricultural Statistics Service, USDA, 3251 Old Lee Hwy, Fairfax VA 22030
[2]National Institute of Statistical Sciences, 19 T.W. Alexander Dr., Research Triangle, NC 27709
[3]Department of Statistics, North Carolina State University, Raleigh, NC 27695
[4]Department of Statistics, University of Florida, Gainesville, FL 32611

Abstract
In recent years, the National Agricultural Statistics Service (NASS) evaluated a variety of approaches to adjust for misclassification in its annual June Area Survey (JAS), which is based on an area frame. This misclassification is a direct cause of an undercount in the number of farms indication produced by the JAS. One approach to correct for this undercount is to use NASS's sampling list frame, which is independent of the area frame. However, recent studies showed that there are farm status inaccuracies on the list frame. These are active records that are not associated with farms. If the list frame farm status inaccuracies are not addressed, the adjusted JAS number of farms indication could become biased upwards. Using Classification and Regression Tree (CART) models, the probability that a list frame record is active can be obtained. This paper evaluates methods for classifying each active list frame record as either a farm or non-farm. One method sets a cut-off using the probabilities, a Receiver Operating Characteristic (ROC) curve, and auxiliary data. Another method uses the same probabilities along with auxiliary data in adjusting for misclassification.

Key Words: Misclassification, Area Frame, List Frame, Classification and Regression Tree, Receiver and Operating Characteristic Curve

1. Introduction

The National Agricultural Statistics Service (NASS) conducts many surveys, two of which are the June Area Survey (JAS) and the Census of Agriculture. The JAS is based on an area frame and is conducted annually. The Census of Agriculture is a dual-frame survey, using the above area frame as well as a list frame composed of all known agricultural operations, and it is conducted every 5 years. Both surveys provide an independent estimate of the number of farms in the United States. NASS defines a farm as any place from which $1,000 or more of agricultural products were produced and sold or normally would have been sold during the year, including any government agricultural payments received. Following each census, previous annual number of farms estimates are revised, if necessary, based on intercensal trends.

Figure 1 depicts the published number of farms in the United States from 2000 to 2009. Before 2007, the number of farms is shown to be decreasing. However, results from the 2007 Census indicated that the 2007 JAS estimate of the number of farms was low; resulting in an intercensal trend adjustment to the number of farms estimates that was larger than could be attributed to sampling error alone.

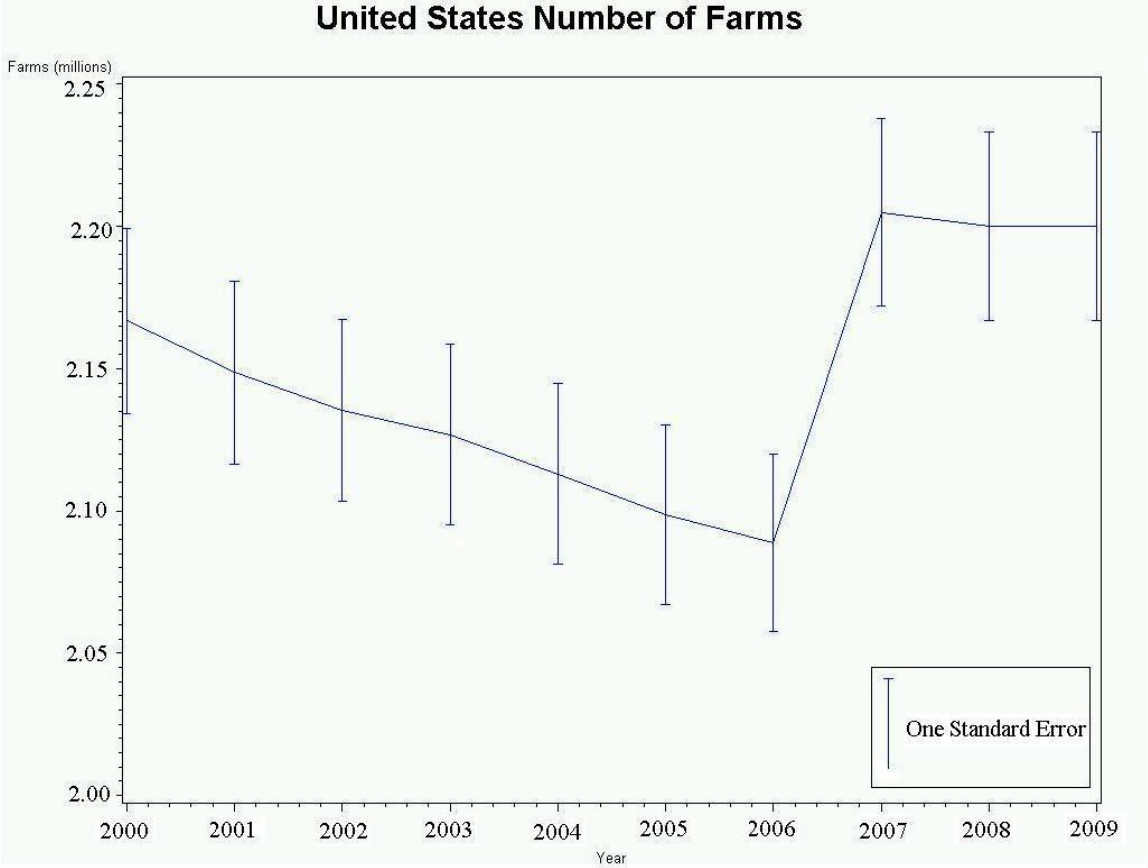

Figure 1: Published estimates of the number of U.S. farms from 2000 to 2009 with associated standard errors.

Previous studies conducted by NASS indicated that a source of this underestimate was misclassification (Abreu 2007; Johnson 2000). Misclassification occurs when an operating arrangement is identified as non-farm when there is actually agricultural activity present, or when a non-farm arrangement is incorrectly identified as a farm. A small-scale misclassification study (Abreu, McCarthy and Colburn, 2009), led to the conclusion that the JAS indication for the number of farms was biased downwards. To fully address this concern, the National Agricultural Statistics Service (NASS) and the National Institute of Statistical Sciences (NISS) formed a research team to review methodology associated with the JAS and to suggest improvements. Over the past year-and-a-half, the team has made several recommendations. One suggests the creation of a follow-up survey to the JAS that allows misclassification in the JAS to be estimated. This approach would result in the JAS having a two-phase design that would allow improved estimates for all of the JAS indications, not just the number of farms. However, due to resource constraints the Agency elected not to pursue this option at this time. As a result, a less resource-intensive method was used to leverage information contained in the NASS list frame to evaluate JAS misclassification.

Using information from the 2007 Census of Agriculture, the NASS-NISS team demonstrated that census records could be used to adjust for misclassification in the JAS indication of the number of farms (Abreu, et al. 2010; Lamas, et al. 2010). Because the Census Mailing List (CML) is derived from the list frame, the natural extension of this approach is to use the list frame for a misclassification adjustment during non-census years. The challenge with this approach is that the list frame does not have a farm / non-farm status classification. Some farm status inaccuracies

exist on the list frame. If the list frame is used to adjust for misclassification without considering the list frame farm status inaccuracies, the JAS number of farms indication could become biased upwards. In essence, the list frame must first be adjusted for these farm status inaccuracies before it can be used to measure or adjust for misclassification on the JAS.

This paper discusses two methods that adjust the JAS number of farms estimates while accounting for list frame status inaccuracies.

2. Background of the June Area Survey (JAS)

The June Area Survey (JAS) has an area frame and is conducted annually. It collects information on U.S. crops, livestock, grain storage capacity and type and size of farms. Since the distribution of crops and livestock can vary widely across a state in the U.S., land is divided, in preparation for sampling, into homogeneous groups or strata. Examples include, intensively cultivated land, urban areas and range land. The general strata definitions are similar from state to state; however, minor definitional adjustments may be made depending on the specific needs of a state. Each land-use stratum is further divided into substrata by grouping areas that are agriculturally similar. This yields greater precision for state-level estimates of individual commodities. Within each substratum, the land is divided into primary sampling units (PSUs). A sample of PSUs is selected and smaller, similar-sized segments of land are delineated within these selected PSUs. Finally, one segment is randomly selected from each selected PSU to be fully enumerated. Through in-person canvassing, field interviewers divide all of the land in the selected segments into tracts, where each tract represents a unique land operating arrangement. Each tract is screened and classified as agricultural or non-agricultural. Non-agricultural tracts belong to one of three categories: (1) non-agricultural with potential, (2) non-agricultural with unknown potential, or (3) non-agricultural with no potential. A tract is considered agricultural if it has qualifying agricultural activity either inside or outside the segment. Otherwise, it's non-agricultural. An agricultural tract will subsequently be classified as a farm if its entire operation (land operated both inside and outside the segment) qualifies with at least $1,000 in sales or potential sales. All non-agricultural tracts and agricultural tracts with less than $1,000 in sales are classified as non-farms.

The JAS is a probability-based sample. Thus each tract has an inclusion probability π_i and an expansion factor $e_i = 1/\pi_i$. Within each farm tract, a proportion of a farm is observed. This proportion, the tract-to-farm ratio, is t_i = tract acres / farm acres. Both of these are used in calculating the current JAS estimate for the number of farms, which is defined as follows,

$$T = \sum_{i=1}^{l} \sum_{j=1}^{s_i} \sum_{k=1}^{n_{ij}} e_{ijk} \sum_{m=1}^{x_{ijk}} t_{ijkm}$$

where

i indexes stratum
j indexes substratum
k indexes segment
m indexes tract
l = Number of land-use strata
s_i = Number of substrata in stratum i
n_{ij} = Number of segments in substratum j within stratum i
e_{ijk} = Expansion factor or the inverse of the probability of selection for each segment in

substratum j in land-use stratum i
x_{ijk} = Number of farm tracts in the given segments
t_{ijkm} = Tract-to-farm ratio of the tract

The sampling weights are appropriate for the sample design. Therefore, this design-based estimate is unbiased unless misclassification is present.

3. NASS List Frame and Census Mail List

Annually, NASS conducts many list-based surveys. In order to conduct these surveys, NASS maintains a list of agricultural operations, on an ongoing basis. The list is built and improved by obtaining outside lists which are matched to NASS's list frame using record linkage programs. Records not on NASS's list are considered potential farms until NASS confirms they are a farm. The Census of Agriculture is NASS's largest survey. It is conducted every five years, in years ending in two and seven. The Census of Agriculture is a complete count of U.S. farms and ranches and the people who operate them. The census collects data on land use and ownership, operator characteristics, production practices, income and expenditures, and many other characteristics. The outcome, when compared to earlier censuses, helps to measure trends and new developments in the agricultural sector of our nation's economy. Census forms are sent to all known and potential agricultural operations in the U.S. The census provides the most uniform, comprehensive agricultural data for every county in the nation. It is a dual-frame survey, using the JAS as well as a list frame. On years when the census is conducted, the list frame serves as the foundation for building the Census Mail List (CML).

4. Methodology for Census Years

Because the Census Mail List is created independently from the JAS area frame, it was used to assess misclassification in the JAS. To do this, the 2007 JAS and 2007 Census reports were matched, farm/non-farm status compared, and farm status disagreement identified (Abreu et. al, 2010). Disagreement in farm status occurred when (1) tracts identified as non-farms in the JAS were identified as farms in the census or (2) tracts identified as farms in the JAS were identified as non-farms in the census. Here it was assumed that a tract that was identified as a farm in either the JAS or the census was a farm. The adjustment presented here contains the census farms identified as non-farms in the JAS.

To adjust for misclassification on census years, consider the following design-based estimate:

$$T_0 = T + \sum_{i=1}^{l} \sum_{j=1}^{s_i} \sum_{k=1}^{n'_{ij}} e_{ij} \sum_{m=1}^{z_{ijk}} t_{ijkm}$$

where

z_{ijk} =Number of non-farm tracts in the given segment
t_{ijkm} = tract-to-farm ratio of the tract
T = JAS estimator without considering the misclassified non-farms.

This estimate provides a design-based estimate for misclassification in 2007, which can be compared to methods using the list frame to adjust for misclassification. The design-based estimate for misclassification using the census is 178,431 farms.

5. Methodology for Non-Census Years

On years when the census of agriculture is not conducted, results from the census cannot be used to adjust for misclassification because you cannot assume farm status remains the same between years. Therefore, NASS's list frame, which is maintained every year (including non-census years) was tested for use to adjust for misclassification. But, unlike the census, the list frame does not maintain a farm/non-farm status. To address this issue, Classification and Regression Tree (CART) models were developed on the 2007 Census Mail List that created a probability that a list frame record is a farm. This probability can be used to address the limitation in the list frame (Garber, 2009).

The JAS and the CML (derived from the list frame) were previously matched for 2007 and the probabilities of being a farm were calculated for 2007. From this work a design-based estimate of misclassification of 178,431 farms was calculated for 2007 (as provided in section 4 of this report). Two methods were developed using Garber (2009) probability of a list frame record being a farm, to adjust for misclassification on the JAS. The results of these two methods for non-census years are tested on 2007 data to be able to compare to the design-based estimate for 2007.

Method 1 used a receiver operating characteristic (ROC) curve to establish a cut-off that determines whether a list frame record is a farm or non-farm. The ROC curve is obtained by plotting sensitivity against specificity (see Figure 2). Sensitivity is the probability the list frame record is associated with a farm given that it was classified as a farm. Specificity is the probability that the list frame record represents a non-farm given that it was classified as a non-farm. The lower left point of the ROC curve (0,0) represents misclassifying all records by declaring all farm list frame records to be non-farms and all non-farm list frame records to be farms. In contrast, the upper right point of an ROC curve (1,1) is ideal because it represents always correctly classifying a list frame record as either a farm or non-farm. In practice, there is a trade-off in these two measures: As the sensitivity goes up, specificity goes down and vice versa. (Fawcett 2005). For this method, if the probability of being a farm falls above the cut-off, the record is assigned a farm status. If the probability of being a farm falls below the cut-off, the record is assigned a non-farm status. Any record that is not matched to a list frame record is assumed to be a non-farm. The adjustment for misclassification is:

$$T_1 = T + \sum_{i=1}^{l} \sum_{j=1}^{s_i} \sum_{k=1}^{n'_{ij}} e_{ij} \sum_{m=1}^{z_{ijk}} t_{ijkm} I(p_{ijkm} > c)$$

where

n'_{ij} = Number of segments where a list frame record was matched to a non-farm
z_{ijk} = Number of matched non-farm tracts in the given selected segment
p_{ijkm} = Probability that the list frame record represents a farming operation
c = cut-off probability in the interval (0,1)
$I(p_{ijkm} > c)$ = indicator function which has value 1 if the probability $p_{ijkm} > c$, otherwise, it has value 0
T = JAS estimator without considering the misclassified non-farms.

Using the 2007 information, the ROC curve for this method is given in figure 2 below.

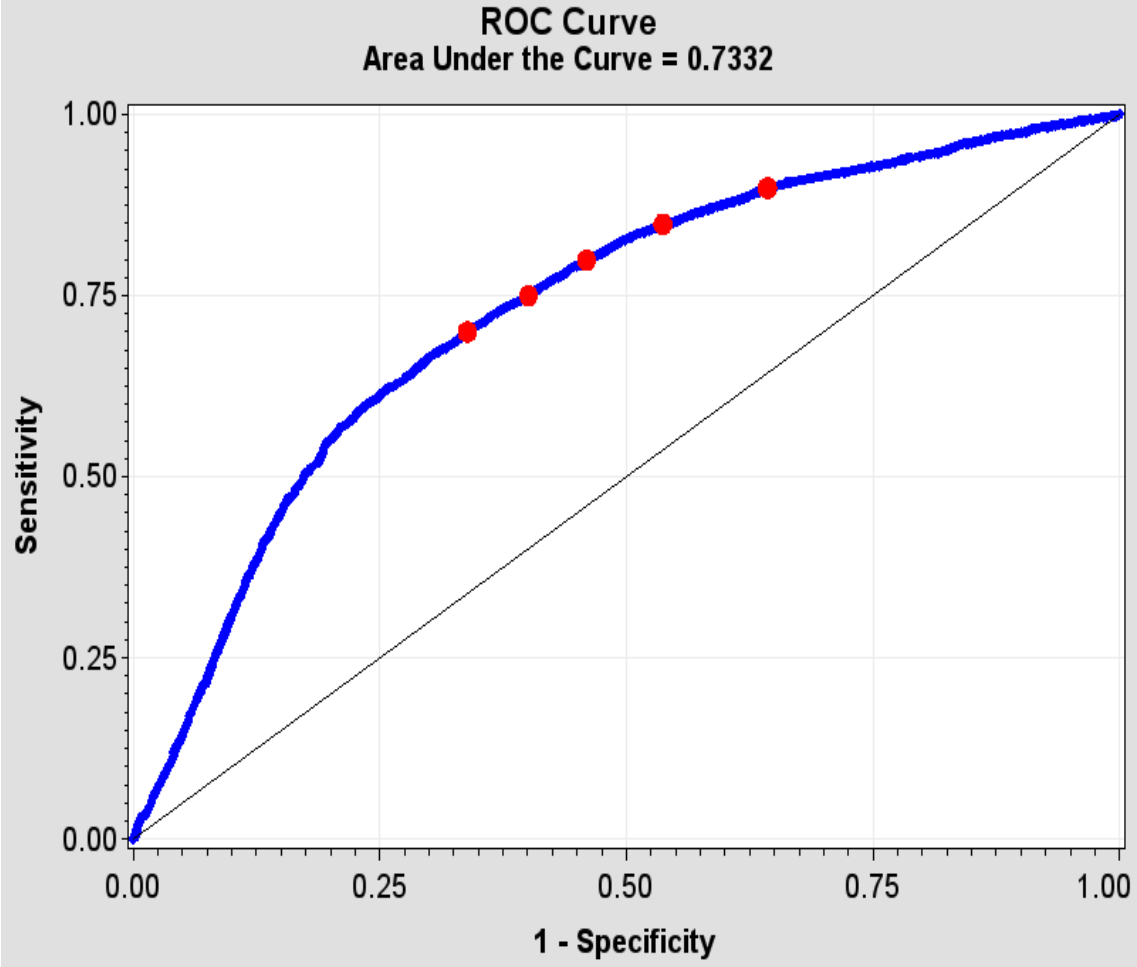

Figure 2: ROC curve using 2007 data

With the given cut-offs, specificities, and sensitivities, the results are in table 1 below.

Table 1: Number of Farms Adjustment by Cut-off Probability from ROC curve

Cut-off Probability	**Specificity (Prop. of non-farms correctly classified as a non-farm)**	**Sensitivity (Prop. of farms correctly classified as a farm)**	**Farms Adjustment**
0.868	0.66	0.70	121,020
0.843	0.60	0.75	133,657
0.780	0.54	0.80	147,183
0.687	0.46	0.85	162,760
0.495	0.36	0.90	179,183

Method 2 used all of the estimated probabilities of being a farm, regardless of size. JAS non-farm tracts that match a list frame record have their probabilities applied to its appropriate expansion factor. Tracts not matched to a list frame records are assumed to be non-farms and are not adjusted for. Using this method, the adjustment for misclassification is:

$$T_2 = T + \sum_{i=1}^{l} \sum_{j=1}^{s_i} \sum_{k=1}^{n'_{ij}} e_{ij} \sum_{m=1}^{z_{ijk}} t_{ijkm} p_{ijkm}$$

where

n'_{ij} = Number of segments where a list frame record was matched to a non-farm
z_{ijk} = Number of matched non-farm tracts in the given selected segment
p_{ijkm} = Probability that the list frame record represents a farming operation
T = JAS estimator without considering the misclassified non-farms

Using method 2, the adjustment using this method is 179,810 farms.

Both method 1 and method 2 assume that JAS non-farm tracts that do not match to list frame records are non-farms. Also the quality of these two estimators depends on the estimation of the probability, p_{ijkm}. These methods also have caveats. Since these methods use the list frame, if a tract is not on the list frame, it will not be adjusted. Also, as the tract-to-farm ratio is necessary for expansion and unavailable from the JAS, the total farm size from the list frame is used to calculate the tract-to-farm ratios in the adjustment.

6. Results and Conclusions

The adjustments for misclassification from both methods are close to the design-based adjustment for misclassification. The results of all the methods are shown in table 2.

Table 2: Number of Farms Adjustment Using Each Method

Method	**Farms Adjustment**
Design-based	178,431
ROC Curve (Method 1)	(121,020 – 179,183)
All Probabilities (Method 2)	179,810

The results of the 2007 analysis showed to be promising in the attempt to use the list frame to adjust for misclassification in the JAS. However, due to the uncertainty of setting the cut-offs with the ROC curve, the conclusion is that the method using all the probabilities (method 2) should be used. However, given that the list frame is dynamic and these methods should be tested on other year's data before reaching any strong conclusions. Therefore, these methods are being tested on the 2011 JAS records and the 2011 list frame records in order to fully evaluate the methods.

7. References

Abreu, Denise A., Andrea C. Lamas, Hailin Sang, Kenneth K. Lopiano, Pam Arroway, and Linda J. Young (2011). On the Feasibility of Using NASS's Sampling List Frame to Evaluate Misclassification Errors of the June Area Survey. Research and Development Division. RDD Research Report #RDD-11-01. Washington, DC: USDA, National Agricultural Statistics Service.

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 2010 Joint Statistical Meetings.

Abreu, D. A., J. S. McCarthy, and 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.

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

Abreu, D. A. (2007). Results from the 2002 Classification Error Study. Research and Development Division. RDD Research Report #RDD-07-03. Washington, DC:USDA, National Agricultural Statistics Service.

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 2010 Joint Statistical Meetings.

Davies, Carrie (2009). Area Frame Design for Agricultural Surveys. Research and Development Division. RDD Research Report #RDD-09-06. Washington, DC: USDA, National Agricultural Statistics Service.

Fawcett, Tom (2005). An introduction to ROC analysis.

Garber, Samuel Chad (2009). Census Mail List Trimming using SAS Data Mining. Research and Development Division. RDD Report Number RDD-09-02.

Johnson, J.V. (2000). Agricultural Census Classification Error Estimation Using an Area Frame Approach. Data Quality Research Section. Unpublished Manuscript. Washington, DC: National Agricultural Statistics Service, USDA.

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

Lopiano, Kenneth K., Andrea C. Lamas, Denise A. Abreu, Hailin Sang, Pam Arroway, and Linda J. Young (2011). Proposal for Using the List Frame to Adjust for Misclassification in the June Area Survey. Research and Development Division. RDD Research Plan. Washington, DC: USDA, National Agricultural Statistics Service.

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

Young, Linda J., Denise A. Abreu, Pam Arroway, Andrea C. Lamas, and Kenneth K. Lopiano (2010). Precise Estimates of the Number of Farms in the United States. Proceedings of the 2010 Joint Statistical Meetings.