Evaluating a New Approach for Estimating the Number of U.S. Farms with Adjustment for Misclassification

Denise A. Abreu, Stephen Busselberg, Andrea C. Lamas, Wendy Barboza, Linda J. Young

National Agricultural Statistics Service, USDA, 3251 Old Lee Hwy, Fairfax VA 22030

Abstract

In recent years, the National Agricultural Statistics Service (NASS) began a research effort to address an undercount in the estimate of the U.S. number of farms derived from its annual June Agricultural Survey (JAS). Misclassification of farm status was found to be a major cause of the undercount. NASS has evaluated a host of measures and methods to assess, quantify, and account for this misclassification. The approach derived from this process employs record linkage techniques, logistic regression, and NASS's annual list sampling frame. The methods developed and the subsequent results are presented here.

Key Words: Misclassification Errors, Area Frame, List Frame, Logistic Regression

1. Introduction

Each year, the National Agricultural Statistics Service (NASS) publishes an estimate of the number of farms in the United States (U.S.) based on the June Agricultural Survey (JAS). A farm is defined as a place from which $1,000 or more of agricultural products were produced and sold, or normally would have been sold, during the year, and the computation includes any government agricultural payments received. An independent estimate of the number of farms is published from the quinquennial Census of Agriculture, which is conducted in years ending in 2 and 7. At the end of each five-year period, the annual estimates based on the JAS number of farms indication are adjusted to account for intercensal trends. The annual estimate of the number of farms from the JAS has been declining steadily between censuses (especially between the 2002 and 2007 Censuses) as depicted in Figure 1. In 2007, the estimate from the JAS was significantly below that from the census; and the required intercensal trend adjustment to the JAS was unexpectedly large as shown by the circled area in Figure 1. The discrepancy between the two estimates was larger than could be attributed to sampling error alone.

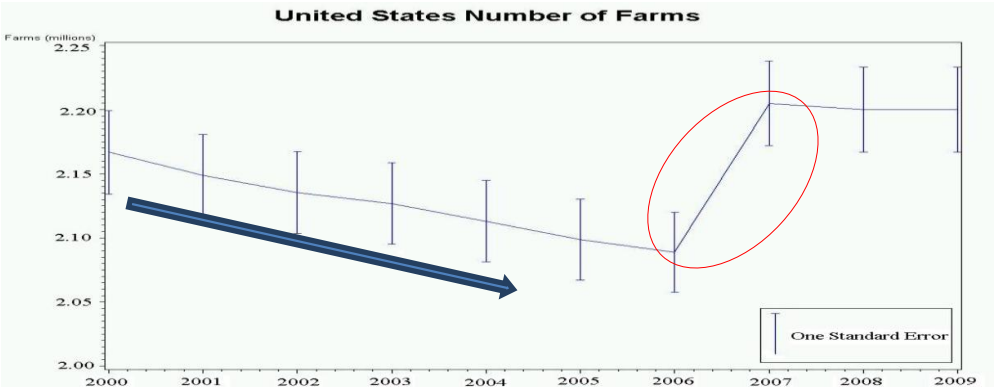

Figure 1: Published estimates of the number of U.S. farms from 2000 to 2009 with one standard error on either side of the estimate**.**

During previous studies conducted by NASS, misclassification was identified as a source of the underestimation in the JAS (Abreu 2007; Johnson 2000). Misclassification occurs (1) when an operating arrangement with qualifying agricultural activity is identified as a non-farm, or (2) when a non-farm arrangement is incorrectly identified as a farm. One study of misclassification **(**Abreu, Dickey and McCarthy, 2009**)** revealed that some agricultural operations were incorrectly classified as non-agricultural during JAS pre-screening. These results led to more intensive efforts to understand the source and extent of misclassification in the JAS so that it could be addressed. One effort was the Farm Numbers Research Project (FNRP), based on an intensive post-June survey re-screening in 2009 (Abreu, McCarthy and Colburn, 2010).

Concurrently, this undercount issue was also addressed by a team of researchers formed to review the methodology associated with the JAS and to recommend change through a collaborative agreement with the National Institute of Statistical Sciences (NISS). This latter team consists of two NASS researchers, two university faculty members, a post doctoral fellow, and a graduate student. The team considered several measures to address the issue of misclassification on the JAS. Through matching the JAS to the Census of Agriculture list frame, the team evaluated misclassification on the JAS (Abreu et al. 2010) and then developed appropriate methodology to adjust for misclassification during non-census years (Lamas et al. 2010). In addition to misclassification, the team identified non-response as another source contributing to the JAS undercount. In (Lopiano et al. 2010), the effect of estimation of agricultural activity for some JAS sampled units is discussed, and methodology for adjusting for both non-response and misclassification is developed. Because the census is only conducted every fifth year, the team further proposed a yearly follow-on survey to the JAS called the Annual Land Utilization Survey (ALUS) (Arroway et al. 2010; Sang et al. 2011). However, due to resource constraints, the Agency elected not to pursue ALUS in 2012. As a result, a less resource-intensive method was pursued to leverage information contained in the NASS list frame to evaluate JAS misclassification.

The challenge with using the NASS list frame is that the it does not have a farm / non-farm status classification. (Abreu et al. 2011) explored the characteristics of the list frame farm status inaccuracies through matching records from the 2009 June Agricultural Survey, the 2009 list frame, and the 2009 Farm Numbers Research Project. In (Abreu et. al. 2012), logistic regression methods along with previous Census of Agriculture data were used to estimate the probability of a farm for 2011 list frame records and provided an adjusted estimate of the number of farms for the 2011 JAS. The adjustment was derived was assumed independent from the original JAS estimator of the number of farms. The estimated probabilities of farm for each list frame record were adjusted by previous census farm rates. Both of these assumptions required more research.

This report presents a more robust estimator for adjusting the JAS for misclassification using the 2012 list frame. The proposed estimator addresses concerns raised with the 2011 methodology and in addition, adjusts the JAS for non-response.

2. Estimating the Number of Farms from the June Agricultural Survey

The June Agricultural Survey (JAS) is based on an area-frame and collects information about U.S. crops, livestock, grain storage capacity, and type and size of farms. The distribution of crops and livestock can vary considerably within each state in the United States. Therefore, the precision of the survey indications can be substantially improved by dividing the land within each state into homogeneous groups (strata) and optimally allocating the total sample to the strata. The basic stratification employed by NASS involves (1) dividing the land into land-use strata such as intensively cultivated land, urban areas and range land, and (2) further dividing each land-use

stratum into substrata by grouping areas that are agriculturally similar. The JAS uses a sample comprised of designated land areas (segments) selected from this stratification. A typical segment is about one square mile (i.e., 640 acres). Each segment is outlined on an aerial photo that is provided to the appropriate field enumerator (the red outlined area in Figure 2).

Through field enumeration, a segment is divided into tracts of land, each representing a unique land operating arrangement (the blue outlined areas in Figure 2). An area screening form, which provides an inventory of all tracts within the segment and contains screening questions that determine whether or not each tract has agricultural activity, is completed for all sample segments. Using this form, all land inside the segment is screened for agricultural activity, and the screening applies to all land in the identified operating arrangement (both inside and outside the segment). Those operations (tracts) that qualify as agricultural are subsequently interviewed using the area version questionnaire, which collects detailed agricultural information about the operator's land, again both inside and outside the segment. 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 is defined as 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 agricultural sales or potential sales. All non-agricultural tracts and agricultural tracts with less than $1,000 in sales are classified as non-farms.

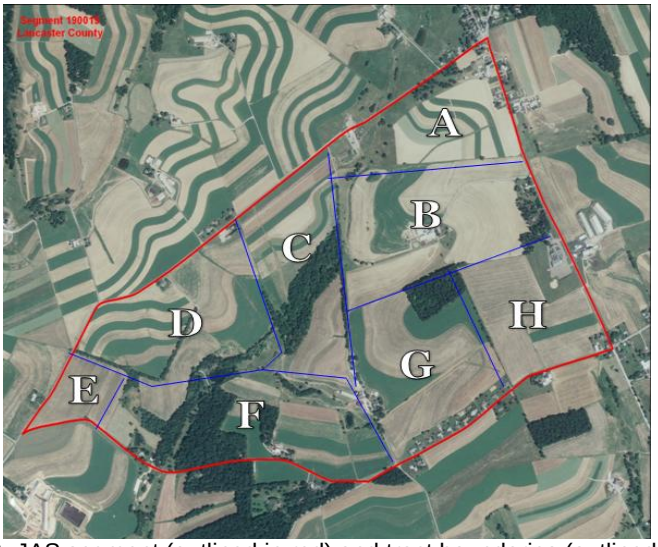

Figure 2: JAS segment (outlined in red) and tract boundaries (outlined in blue)

Because the JAS is a probability-based survey, each tract *i* has an inclusion probability π_i and an expansion factor of $1/\pi_i$. Within each farm tract, a proportion of a farm is observed (in some cases with smaller farms, the entire operation may reside entirely within the tract). This proportion, the tract-to-farm ratio for tract *i*, is t_i = tract acres / farm acres.

Both of these are used in calculating the current JAS estimate for the number of farms (denoted as T), defined below,

$$T = \sum_i \frac{t_i}{\pi_i}$$

where i indexes tract on the JAS,
t_i = Proportion of a farm represented by tract i,
π_i = Sample inclusion probability for tract i,

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

However, when the magnitude of potential misclassification on the area frame became evident, NASS instituted a series of measures to reduce, if not eliminate, this misclassification. To decrease the number of operations misclassified as non-farms during the screening process, field enumerators received enhanced training, and the time allocated for screening was increased from one to two weeks.

For the JAS, questionnaire data are manually imputed when an operator cannot be reached or refuses to respond. Because of this, the quality of the response may depend on the method used for data estimation. A question was added to the JAS questionnaire that identifies the source of the imputed information (i.e., tax assessor's information, previously reported data, etc). The purpose is to allow later evaluation of the quality of the imputation from various sources.

In addition, a question was added to the JAS pre-screening form to further categorize non-agricultural tracts. Enumerators are to choose from the following categories: residential, woods, idle open land, pasture, water (lakes, rivers, etc.), reported non-ag by respondent, vacant houses, obvious non-agricultural (schools, cemeteries, prisons, airports, road/highways, interstate, etc.), grassland, hunting preserve, government land, and other (explain land use). After the 2012 Census of Agriculture, the relationship between misclassification and the category of non-agricultural tract will be studied in an effort to further identify ways to reduce misclassification. The following estimator is proposed for the 2012 JAS:

$$T = \frac{t_i}{\pi_i} \frac{p_i(F|SARJ)}{p_i(J|SARF)p_i(R|SAF)p_i(A|SF)}$$

where i indexes tract on the JAS,
t_i = Proportion of a farm represented by tract i,
π_i = Sample inclusion probability for tract i,
S = Tract is in the sample,
A = Tract passes Ag screening process,
R = Tract responds to the survey,
F = Tract contains a portion of a farm.
J = Tract is identified as a farm on the JAS

The estimator adjusts for the two types of misclassification and for non-response. The probability component $p_i(F|SARJ)$ is the adjustment for misclassification causing an overcount of farms.

This type of measurement error occurs during the data collection phase when a tract is identified as a farm, and in fact there is no farming operation in existence. Misclassification that causes an undercount can occur in two different phases; during the pre-screening process or during the data collection process. If an enumerator classifies a tract as non-agricultural during the pre-screening process, these tracts are not followed up during the data collection phase of the survey in June. Misclassification of tracts containing farms as non-agricultural during the pre-screening phase creates undercoverage, as no data is collected from non-farms during the data collection phase. This paper therefore refers to the probability adjustment $p_i(A|SF)$ – due to an initial misclassification of a farm-containing agricultural tract as a non-agricultural tract – as a "coverage" adjustment. Once a tract containing a farm has been classified as containing agricultural activity during the pre-screening, it has another opportunity to be misclassified as a non-farm during the data collection phase. The probability component $p_i(J|SARF)$ is the adjustment for misclassification causing undercount during the data collection phase. Finally, farm tracts on the JAS that are inaccessible or refuse to answer the survey are manually imputed. Instead of manually imputing these records, let us consider them non-respondents to the survey and apply a non-response adjustment to each respondent. Thus, $p_i(R|SAF)$ is the adjustment for non-response. All four probability adjustments are either conditioned on a record containing a portion of a farm (F) or, in the case of the probability adjustment for overcount, the response in question is whether the tract record actually contains a portion of a farm. Due to misclassification of farms, this condition is uncertain. For this purpose, it is important to have another source to serve as a validation of farm status, and NASS's annual list frame is used here in that context.

3. The NASS List Frame

NASS conducts hundreds of list-based surveys each year. The agency maintains a list of farmers and ranchers from which the samples for these list-based surveys are selected. This list frame also serves as the foundation for the development of the Census Mail List (CML). NASS builds and improves the list on an ongoing basis by obtaining outside source lists. Sources include lists from state and federal government agencies, producer associations, marketing associations, and a variety of other agricultural sources. NASS also obtains special commodity lists to address specific list deficiencies. These outside source lists are matched to the NASS list using record linkage programs. Most names on newly acquired lists are already on the NASS list. Records not on the NASS list are treated as potential farms until NASS can confirm their existence as a qualifying farm. Each operation on the list frame is categorized as active, potential farm (criteria), or inactive. Active list records are assumed to have a high probability of representing active farming operations. Potential farm or criteria records are records whose involvement in agriculture is unknown. Inactive list records may be associated with landlords, deceased operators, farms no longer in business, etc. Many of the active records represent agricultural establishments that operate land but do not have sufficient production to be classified as a farm in a specific year. However, they are maintained on the list frame as active records to help ensure high coverage of farms for the Census of Agriculture every five years. Potential farm or criteria records are periodically screened to determine whether or not they are involved in agriculture. Pure active status inaccuracies also exist on the list frame; that is, some records identified as "active" are out-of-business or no longer operate any agricultural land or facilities.

4. Matching 2012 JAS to the 2012 List Frame

To help validate the farm status, probabilistic record linkage was used to match the 93,409 agricultural and non-agricultural tracts on the 2012 JAS to over 4.5M records on the 2012 list frame in the 48 conterminous states. The JAS is only conducted in Hawaii during census years, and Alaska does not have an area frame. Records were brought together into link groups, each of

which possibly represented a single operation. Subsequently, link groups were classified into one of three distinct types: definite match, possible match or non-match (Broadbent et. al. 1999). Possible matches were sent to our Frames Management Group (FMG) staff for review and were further classified as matches or non-matches. All non-matches were excluded from further analysis.

When matching, the ideal scenario is to have one area record match one list record. However, after the initial matching, some link groups had more than one tract and others had more than one list frame record. Although the area file was set up to have only one tract per link group, in some cases, more than one tract occurred in a link group, indicating that different tracts matched to the same list records. To address this issue, tracts were split into separate groups and all list records that matched were assigned to both split groups. When multiple list records matched one tract, the list frame records were ranked and based on their active/inactive status, the "best" one was selected using the following rules:

Ranks Used to Assign the Best of Several List Frame Records to a JAS Tract

Rank	List Record Type	Description
1	Active target	Assumed to be farming operations
2	Potential CML	Non-respondents to any of the agricultural surveys conducted routinely to update active status of the list frame
3	Active partner	Partners associated with active target
4	Inactive	Deceased operators, farms no longer in business, idle facilities, landlords, etc.
5	Other	Hired managers, etc.

5. Matching Results

The results of this matching procedure yielded 43,108 matches between the JAS and the list frame. Note that there were 44,721 non-agricultural tracts that did not match any list frame record due to lack of name and address information. These records were considered non-matches. In addition, 5,580 JAS agricultural tracts did not match to the list frame. From the matches, there were 7,721 estimated (i.e., JAS non-respondent) tracts and 35,387 non-estimated tracts.

Table 1 shows the breakdown of the matched tracts by type of agricultural tract as identified in the JAS. Recall that during JAS screening procedures, non-agricultural tracts are classified into the following three types: potential for agriculture unknown, having potential for agriculture, and not having potential for agriculture. Non-agricultural tracts without potential comprised 13.4 percent of all the matches, while agricultural tracts identified as non-farms comprised 2.8 percent.

Table 1. Matched JAS Tracts and List Frame Records by Type of Agriculture as Identified by the JAS

Type of Agricultural Tract	**Number of Tracts Matched**	**Percent**
Agricultural tracts identified as farms	35,510	82.4
Agricultural tracts identified as non-farms	1,193	2.8
Non-agricultural tracts w/ potential	442	1.0
Non-agricultural tracts w/ unknown potential	186	0.4
Non-agricultural tracts w/out potential	5,777	13.4
Totals	43,108	100.0

Table 2 shows the breakdown of the matched tracts by the type of list frame record. Results show that over 85% of the matches were to active records; while matches to inactive and criteria records were much smaller approximately, 7 and 8 percent, respectively.

Table 2. Matched JAS Tracts and List Frame Records by Type of List Frame Record

Type of List Frame Record	Number Tracts Matched	Percent
Active	36,911	85.6
Criteria	2,855	6.6
Inactive	3,342	7.8
Totals	43,108	100.0

Table 3 presents the JAS farms and non-farms and the type of list frame record they matched. Recall records on the list frame are classified as either active, inactive, or criteria. Active list records are assumed to have a high probability of representing active farming operations. Potential farm or criteria records are records whose involvement in agriculture is unknown. Inactive list records may be associated with landlords, deceased operators, farms no longer in business, etc. Farm status is not on the list frame. Thus, an inactive record most likely represents a non-farm and an active record is most likely a farm. The cells highlighted show "farm" status discrepancies between the list frame and the JAS.

Table 3. JAS farm status by List Frame "farm" status.

	List Frame Inactive	List Frame Active	List Frame Criteria	Total
JAS Non-Farm	2,438	3,973	1,187	7,598
JAS Farm	904	32,938	1,668	35,510
Total	3,342	36,911	2,855	43,108

*Number of records with disagreeing (or unknown) farm status is highlighted.

The proposed 2012 JAS farm numbers estimator contains adjustment components that are conditional probabilities. The conditions of these probabilities are met with absolute certainty because it is known which records are in the sample, responded to the survey and passed the agricultural screening. However, the only condition that is an exception is – the event that the tract contains a portion of a farm (F). It is for this reason that JAS tract records must be compared with another data source to obtain confirmation on the farm status in order to meet this condition. The secondary data source in non-census years is the list frame; however, there is no variable which equates to farm status in common in both the JAS records and the matching list frame records. This is needed in order to develop a dataset of records that have agreement on farm status. As noted earlier, instead of farm status (i.e., is a farm vs. is not a farm), the list frame contains an active status. Thus, it is necessary to develop methodology to account for the uncertainty in the farm status condition. To account for this uncertainty, the probability that a record is a farm is modeled.

6. Modeling the Probability that a Record Contains a Farm: $\hat{p}_i(F)$

In order to develop a binary response variable to model the probability that a record is a farm, given different constructs of semi-comparable farm status variables ("farm status" from the JAS and "active status" from the list frame); it is necessary to make the following two assumptions:

- Active list frame records matching JAS farms are defined as true farms (response IsFarm=1).

- Inactive list frame records matching JAS non-farms are defined as true non-farms (response IsFarm=0).

Records that do not show consistent farm status and active status are not used in modeling the probability a tract contains a portion of a farm. Using the data for the records that agree in status according to the two assumptions listed above (2,438 + 32,938 records), the probability that a record is a farm is modeled using logistic regression.

$$\hat{p}_i(F) = \left(1 + e^{-x_i'\beta}\right)^{-1}$$

For the record weighting, the sampling weights are normalized to the number of observations in the modeling dataset. JAS data, list frame data, population census data and data from the Cropland Data Layer (CDL) were used to identify a set of explanatory variables (x_i) to be used in the model.

The Cropland Data Layer (CDL) is a raster-formatted, geo-referenced, crop specific land cover classification derived from satellite data acquired from April through September each year (Boryan et. al. 2011). The CDL is produced annually and made available to the public on the CropScape web portal for all 48 states in the conterminous US (Han et. al. 2012). Using CDL crop-specific covariate data created from multi-year CDL data, percentages of land cultivated, corn, wheat, developed open space, urban, and water were calculated for each segment (Boryan et. al. 2013).

Using a stepwise selection method in the logistic regression, the following covariates were selected for modeling the probability that a record is a farm ($p_i(F)$).

Step	EffectEntered	ItemDescription
1	LOGTA	Log of Tract Acres
2	PERC_FORE	CDL Percent Forest
3	HSD310212	Persons per household, 2008-2012
4	PERC_CORN	CDL Percent Corn
5	SBO015207	Women-owned firms, percent, 2007
6	PERC_LO_UR	CDL Percent Low Intensity Urban
7	PERC_CULT	CDL Percent Cultivated
8	LND110210	Land area in square miles, 2010
9	RHI225212	Black or African American alone, percent, 2012
10	RHI525212	Native Hawaiian and Other Pacific Islander alone, percent, 2012
11	HSG445212	Homeownership rate, 2008-2012
12	AGE775212	Persons 65 years and over, percent, 2012
13	PVY020212	Persons below poverty level, percent, 2008-2012
14	LFE305212	Mean travel time to work (minutes), workers age 16+, 2008-2012
15	HSG495212	Median value of owner-occupied housing units, 2008-2012
16	SBO115207	American Indian- and Alaska Native-owned firms, percent, 2007
17	VET605212	Veterans, 2008-2012
18	RHI425212	Asian alone, percent, 2012

Given a model for the estimated probability $\hat{p}_i(F)$ that a tract record contains a portion of a farm, the four probability components in the proposed estimator are modeled using the normalized estimated probability of containing a portion of a farm as the record weight to account for the uncertainty of the "contains a farm (F)" condition. The normalization is to the number of matched records in the modeling dataset that meet the criteria of the respective conditional probabilities. All tract records are then scored for $\hat{p}_i(F)$ including the records used to obtain the model. This reflects the farm status uncertainty contained within those records used in the model due to different farm status constructs.

7. Modeling the Probability Adjustment for Undercount, Overcount, Coverage, and Non-response

Once each record's probability of farm, $\hat{p}_i(F)$, is obtained, the four probability adjustment components of the following proposed estimator were modeled:

$$T = \frac{t_i}{\pi_i} \frac{p_i(F|SARJ)}{p_i(J|SARF)p_i(R|SAF)p_i(A|SF)}$$

Stepwise logistic regression is conducted to obtain the expected probabilities for each record for an undercount, overcount, coverage, and non-response adjustment. The set of explanatory variables selected for each model is displayed in Table 4.

Table 4. Set of Covariates Selected for Each Probability Adjustment

Overcount	Undercount	Coverage	Non-response
Log of Cattle	Log of Cattle	Log Total Tract Acres	Log of Cattle
Log of Cropland	Log Equine	CDL Percent Forest	Log Equine
Log of Government Payments	Log Other Equine	CDL Percent Open Space	Log Other Equine
Log of Land in a Conservation Reserve Program (CRP)	Log of Cropland	Persons per household, 2008-2012	Log of Cropland
Log Land Owned	Log Sheep	Rural Urban Code	Log of Government Payments
Log Total Tract Acres	Log Total Tract Acres	Population, percent change - April 1, 2010 to July 1, 2012	Log Sheep
JAS Stratum	Log CRP	Building permits, 2012	Log Hog
	JAS Stratum	Native Hawaiian- and Other Pacific Islander-owned firms, percent, 2007	Log Land Rented From Others
		Female persons, percent, 2012	Log Land Rented to Others
		White alone, not Hispanic or Latino, percent, 2012	Log CRP
		Persons below poverty level, percent, 2008-2012	Log Land Owned
		JAS Stratum	Log Total Tract Acres
		CDL Percent Cotton	Farmtype*Sales
		Retail sales per capita, 2007	JAS Stratum
			Rural Urban Code

For the undercount, non-response, and coverage adjustments; records for building the models are weighted by the estimated probability of containing a portion of a farm (covered in section 6) normalized to the number of records in the covariate dataset.

For the overcount probability adjustment, $p_i(F|SARJ)$, the response variable is whether a tract is farm or not (F). However, this is one of the conditions for the other three adjustment probabilities. The response variable is assigned and weighted according to the following two cases:

Case 1:
JAS farm status = "Farm" and active list status = "Active"
Response = 1 with weight $\hat{p}_i(F)$

Case 2:
For all other combinations of JAS farm status and list frame active status levels, records are duplicated and assigned
Response = 1 with weight $\hat{p}_i(F)$
Response = 0 with weight $1 - \hat{p}_i(F)$

The weights are normalized to the number of records (which includes the duplicates) in the overcount covariate dataset.

For the final probability adjustment models, Figure 3 shows the total contribution of each scored farm record to the total estimate is graphed by the predicted probabilities for each adjustment below.

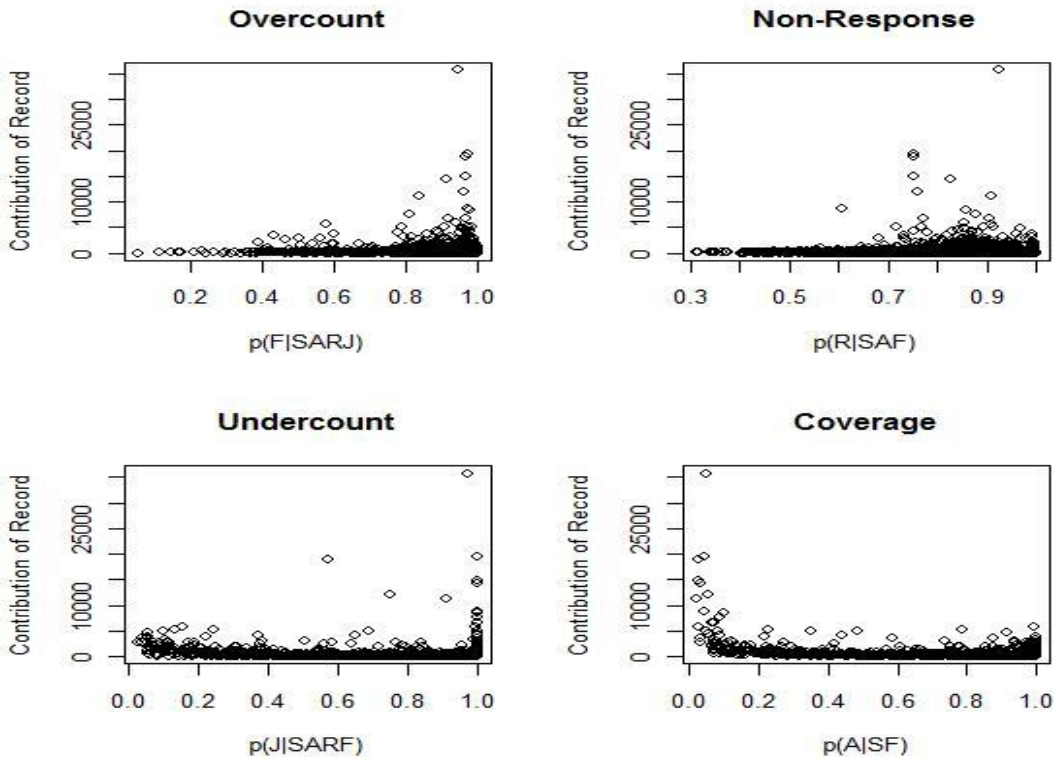

Figure 3: Total contribution of each scored farm to the total estimate by the predicted probabilities for each type of adjustment.

Examination of the graphs yields the conclusion that the records with extreme contributions relative to the others are associated with small coverage probabilities, or that the estimated probability of passing the agricultural screening is very low. The primary cause of this is most likely the lack of modeling covariates. The records that fail the screening are labeled non-agricultural and do not have recorded information other and tract size, stratum, segment level information from the CDL, and population census information. For this reason, the only unit-level covariate available for modeling this adjustment is the tract acres.

Table 5 gives the quantiles of the 30,448 records that responded as farms for the 2012 JAS sample.

Table 5. Quantiles of the probability adjustments

Quantile	Overcount	Undercount	Non-Response	Coverage
100% Max	1.0000000	1.0000000	0.9991530	1.0000000
99%	0.9999912	1.0000000	0.9942270	1.0000000
95%	0.9999611	1.0000000	0.9599520	1.0000000
90%	0.9999235	1.0000000	0.9313590	1.0000000
75% Q3	0.9997573	1.0000000	0.8839990	1.0000000
50% Median	0.9989365	1.0000000	0.8289640	1.0000000
25% Q1	0.9934707	0.9998824	0.7656100	1.0000000
10%	0.9677797	0.9959088	0.6953560	1.0000000
5%	0.9315792	0.9473821	0.6431900	1.0000000
1%	0.7710320	0.5065736	0.5428750	0.4899400
0% Min	0.0537534	0.0244979	0.3129230	0.0201670

The lowest quantile shaded in gray shows some records may be receiving overinflated adjustments due to low expected probabilities. For both coverage and undercount the maximum adjustment was set to 2, which corresponds to an expected probability of 0.5. Only 313 records out of 30,448 have an estimated probability for the coverage adjustment less than 0.5, and 298 records out of 30,448 have an estimated probability for the undercount adjustment less than 0.5. There are 17 records that have both coverage and undercount adjustments less than 0.5. Applying a minimum expected probability of 0.5 affects 594 records (313+298-17) out of 30,448 and prohibits an overinflated contribution to the final estimate by any one extreme record.

8. Farm Numbers Results

From Section 2, the current JAS estimate for the number of farms (denoted as T_1), is defined as follows,

$$T_1 = \sum_i \frac{t_i}{\pi_i}$$

This estimate is unbiased unless misclassification, non-response, and undercoverage are present. These have been found to be present on the JAS. Using the modeled probabilities presented in Section 7, an estimator for the number of farms from the JAS with an adjustment for misclassification, non-response and coverage can be constructed as follows:

$$T = \frac{t_i}{\pi_i} \frac{p_i(F|SARJ)}{p_i(J|SARF)p_i(R|SAF)p_i(A|SF)}$$

where

i indexes tract on the JAS,

t_i = Proportion of a farm represented by tract i,

π_i = Sample inclusion probability for tract i,

S = Tract is in the sample

A = Tract passes Ag screening process

R = Tract responds to the survey

F = Tract is truly a farm

J = Tract is identified as a farm on the JAS

Using the estimated probabilities for each of the adjustments at the record level for each tract containing a farm, the misclassification adjustment is 9.4 percent at the U.S. level, i.e., $(T_2 - T1100\%T1=9.4\%$. Due to the confidential nature of how NASS derives the JAS annual estimate of the number of farms, no further results can be presented here.

9. Jackknife Standard Errors

In order to estimate the standard errors of the farm numbers estimates, a delete-a-group jackknife approach was employed. Each segment from the JAS sample that was successfully matched to the list frame was assigned to a random group, with a total of ten groups. Groups were assigned so that each state by stratum combination is represented in all ten groups. This group assignment is relevant because the records were stratified in the sample design by percent of agricultural cultivation. This ensures that each of the ten mutually exclusive random groups was representative of the sample. The jackknife standard error estimate was then calculated as

$$STDERR_{JK}(\theta) = \sqrt{\frac{M-1}{M}\sum_{i=1}^{M}\left(\theta_{(i)} - \theta\right)^2}$$

where M is the number of jackknife groups, θ is the estimate, and $\theta_{(i)}$ is the estimate calculated by omitting group i (Lohr 1999).

The standard error associated with this proposed estimator is 11,102 farms.

10. Conclusions and Future Work

The estimator presented here is more robust than previous estimators researched at NASS. This estimator adjusts for misclassification and coverage as well as for non-response. The use of logistic regression modeling provides a solid, reproducible technique to modeling the farm probability for records with disagreeing farm status. Using this framework, estimates of the number of farms can be produced for subsequent years of this survey. Future work will include a more detailed research and analysis of adjustment or capping methods for extreme weights.

11. References

Abreu, Denise A., Andrea C. Lamas, Shu Wang, Linda J. Young. (2012). Estimating the Number of U.S. Farms with Adjustment for Misclassification. Proceedings of the 2012 Joint Statistical Meetings.

Abreu, Denise A., Andrea C. Lamas, Hailin Sang, Kenneth K. Lopiano, Pam Arroway, 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.

Abreu, Denise A., Andrea C. Lamas, Hailin Sang, Pam Arroway, Kenneth K. Lopiano, Linda J. Young. (2011). Is It Feasible to Use a Sampling List Frame to Evaluate Misclassification Errors on an Area-Frame-Survey? Proceedings of the 2011 Joint Statistical Meetings.

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.

Boryan, Claire. Abreu, Denise A., Andrea C. Lamas, Hailin Sang, Pam Arroway, Kenneth K. Lopiano, Linda J. Young. (2011). Is It Feasible to Use a Sampling List Frame to Evaluate Misclassification Errors on an Area-Frame-Survey? Proceedings of the 2011 Joint Statistical Meetings.

Boryan, Claire and Z. Yang (2013). Deriving crop specific covariate data sets from multi-year NASS geospatial cropland data layers. Pproceedings of the IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2013, Melbourne, Australia.

Boryan, Claire. Z. Yang, R. Mueller, and M. Craig (2011). Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service Cropland Data Layer Program. Geocarto International, 26 (5): 341-358.

Broadbent, K and Iwig, W. (1999), "Record Linkage at NASS Using Automatch". *1999 FCSM Research Conference, http://*www.fcsm.gov/99papers/broadbent.pdf

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.

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

Han , W., Z. Yang, L. Di, and R. Mueller (2012). CropScape: A web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. Computer and Electronics in Agriculture, 84: 111–123.

Kovacevic, Milorad S., Lenka mach, and Georgia Roberts. 2008. Bootstrap variance estimation for predicted individual and population-average risks. Proceedings of the Section on Survey Research Methods, JSM 2008. Pp. 2289-2296.

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, Hailin Sang, Pam Arroway, Kenneth K. Lopiano, and Linda J. Young (2011). Adjusting an Area Frame Estimate for Misclassification Using a List Frame. Proceedings of the 2011 Joint Statistical Meetings.

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, Pam Arroway, and Linda J. Young (2011). Adjusting the June Area Survey Estimate for the Number of U.S. Farms for Misclassification and Non-response. Research and Development Division. RDD Report Number RDD-11-04.

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.

Lohr, Sharon L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Sang, Hailin, Pam Arroway, Kenneth K. Lopiano, Denise A. Abreu, Andrea C. Lamas, and Linda J. Young (2011). Annual Land Utilization Survey (ALUS): Design and Methodology. Research and Development Division. RDD Report Number RDD-11-02.

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.