Evaluation of a New Approach for Estimating the Number of U.S. Farms

Habtamu Benecha* Denise A. Abreu * Jake Abernethy * Luca Sartore † Linda J. Young *

Abstract

USDA's National Agricultural Statistics Service (NASS) employs the June Area Survey (JAS) to produce annual estimates of U.S. farm numbers. The JAS is an area-frame-based survey conducted every year during the first two weeks of June. NASS also publishes an independent estimate of the number of farms from the quinquennial Census of Agriculture. Studies conducted by NASS have shown that farm number estimates from the JAS can be biased, mainly due to misclassification of agricultural tracts during the pre-screening and data collection processes. To adjust for the bias, NASS has developed a capture-recapture model that uses NASS's list frame as the second sample, where estimation is performed based on records in the JAS with matches in the list frame. In the current paper, we describe an alternative capture-recapture approach that uses all available data from the JAS and the Census of Agriculture to correct for biases due to misclassification and to produce more stable farm number estimates.

Key Words: Misclassification Error, Area-frame, List Frame, Logistic Regression, Capture-recapture.

1 Introduction

The National Agricultural Statistics Service (NASS) uses the June Area Survey (JAS) to produce annual estimates of U.S. farm numbers. NASS also publishes an independent estimate of the number of farms from the quinquennial Census of Agriculture, which is conducted in years ending in 2 and 7. The JAS is based on an area-frame that covers the continental U.S. with every acre of land having a known probability of selection. Before the survey, all tracts of land in the sample are screened for agricultural activity and operators of tracts that are classified as agricultural are interviewed during the survey. The JAS had previously been considered to have only minimal classification error rates. However, studies conducted by NASS have shown that the misclassification rates in the JAS can be substantial and may not be ignored when farm numbers are estimated (Abreu *et al.*, 2010). Misclassification of farms occurs when agricultural (non-agricultural) tracts are classified as non-agricultural (agricultural) during the JAS pre-screening and data collection processes.

In an effort to correct for the bias due to misclassification, NASS is currently using a capture-recapture approach for estimation of farm numbers. In this approach, JAS records are matched to NASS's list-frame and the matched data are used to estimate adjustment weights for misclassification. The model-based estimates are then considered together with historical data and Census estimates when the official farm numbers are determined. The official, published numbers of U.S. farms are estimated by a group of experts called the Agricultural Statistics Board (ASB).

An alternative source of data for estimation of JAS farm numbers from the current capture-recapture model is the U.S. Census of Agriculture. Because the Census covers the vast majority of farm operations in the U.S., matching the JAS records with Census data can potentially reduce misclassification-related biases and improves the model estimates, particularly during census years.

The matched data set currently being used for estimation includes only records in the JAS with matches in the list frame. Thus, records in the list frame with no matches in the JAS are not considered for the estimation of farm numbers from the JAS. Because the list frame and the Census contain a relatively complete and up-to-date information on U.S. agricultural operations, the use of all available data from either of the

*USDA National Agricultural Statistics Service (NASS), South Building, 1400 Independence Ave., SW, Washington, DC 20250
†National Institute of Statistical Sciences, 19 T.W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC27709-4006

two data sources may provide improved farm number estimates from the JAS. Utilizing more data from the Census or the list frame can also potentially reduce the year-to-year fluctuations observed in the current model estimates.

This paper proposes an alternative capture-recapture approach (Alho, 1990) to correct for biases due to misclassification in the JAS and to produce more stable farm number estimates. The model uses data from the JAS and a second, independent source to estimate JAS farm numbers. The Census of Agriculture is used as a source of the second sample in the proposed capture-recapture model. Because the list frame and the JAS area-frame are kept independently at NASS, the Census and the JAS can generally be considered independent. Section 2 gives an overview of the JAS design and Section 3 discusses the current approach for the estimation of farm numbers. The proposed approach is presented in Section 4. Simulation studies and an application of the method to NASS data are described in Sections 5 and 6, respectively. Discussions and conclusions are provided in Section 7.

2 An overview of the June Area Survey

The June Area Survey is one of the largest annual surveys conducted by NASS and is used as a vehicle to collect information about U.S. crops, livestock, grain storage capacity, and farm sizes and types. The JAS uses an area-frame that covers all land in the U.S. except Alaska stratified by land use strata. The strata are further divided into substrata by grouping areas that are agriculturally similar. Within each substratum, the land is divided into primary sampling units (PSUs).

Figure 1: NASS area-frame showing tracts in a segment

For the JAS, a sample of PSUs is selected and smaller, similar-sized segments of land are sampled from the selected PSUs, to be fully enumerated. Before the survey, all tracts of land within selected segments are screened and classified as agricultural or non-agricultural. An agricultural tract will subsequently be classified as a farm if its entire operation 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 (Lamas et. al, 2010).

If π_{ijk} denotes the inclusion probability of each tract within segment k in substratum j within stratum i, farm numbers are calculated based on the JAS design as (Lamas et. al, 2010),

$$N_d = \sum_{i=1}^{L} \sum_{j=1}^{S_i} \sum_{k=1}^{n_{ij}} \frac{a_{ijk}}{\pi_{ijk}}, \tag{1}$$

where n_{ij} is the number of segments in substratum j within stratum i, S_i is the number of substrata in stratum i, L is the number of strata and $a_{ijk} = \sum_{m=1}^{n_{ijk}} t_{ijkm}$, where n_{ijk} is the number of farm tracts in segment k and t_{ijkm} is the tract to farm ratio calculated as

$$t_{ijkm} = (\text{tract acres for the } m^{th} \text{ tract})/(\text{farm acres for the } m^{th} \text{ tract}).$$

Due to misclassification, equation (1) tends to underestimate or overestimate the true number of farms (Abreu et.al, 2010). In addition, the design-based approach provides farm number estimates that tend to fluctuate from year-to-year. For example, Figures 2 and 3 show that the 2012-2016 JAS farm number estimates for the two states are highly unstable compared to the corresponding ASB estimates. Studies have shown that the biases in the JAS farm number estimates are likely caused by classification errors during the pre-screening and data collection processes. To adjust for the effects of misclassification, non-response and possible under-coverage in the estimation of farm numbers, NASS has developed a capture-recapture model as discussed in the next section.

3 NASS's current method for the estimation of the number of farms

To correct for the bias due to misclassification, NASS is currently using a capture-recapture approach for the estimation of farm numbers. In this approach, JAS records are matched to NASS's list frame by using probabilistic record linkage and a series of logistic regression models are fitted to subsets of the matched data to estimate adjustment weights for over-counting, under-counting, under-coverage and non-response. The estimated weights are then combined with inclusion probabilities and tract-to-farm ratios to estimate farm numbers for states or the U.S. as follows.

$$N_m = \sum_{i \in J,R,A,S} \frac{t_i}{\pi_i} \frac{p_i(F|S,A,R,J)}{p_i(J|S,A,R,F)p_i(R|S,A,F)p_i(A|S,F)}. \tag{2}$$

In equation (2), t_i denotes a tract-to-farm ratio or the proportion of a farm represented by the i^{th} tract, π_i is the inclusion probability for tract i; S denotes the event that the tract is in the sample, A is for the event that the tract passed agricultural screening and data collection processes , R denotes the event that the tract responded, J denotes the event that a tract is recorded as a farm in the JAS and F denotes the event that the tract is truly a farm.

The probability component $p_i(F|S,A,R,J)$ is an adjustment for misclassification causing over-count 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. The quantity $p_i(A|S,F)$ is a 'coverage' adjustment to account for coverage errors due to initial misclassifications of farm-containing agricultural tracts as a non-agricultural tracts.

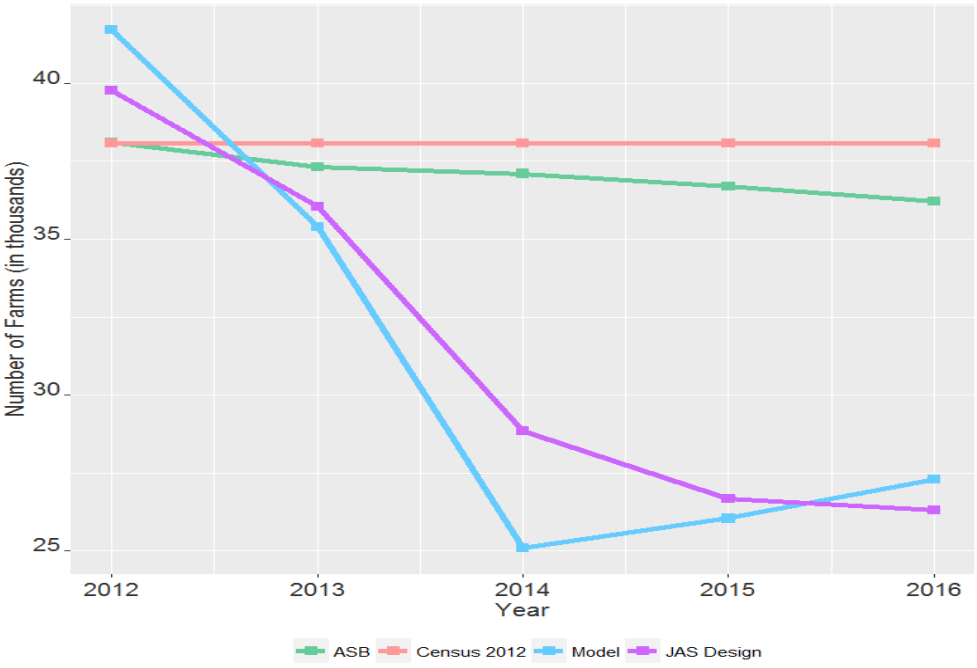

Figure 2: Annual farm number estimates for State 1 from four sources

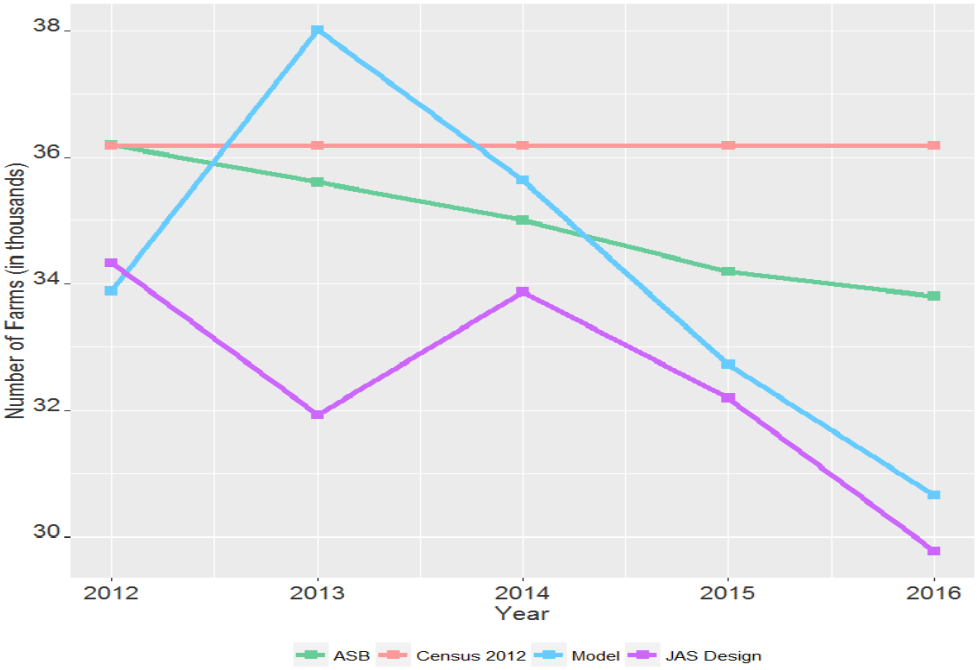

Figure 3: Annual farm number estimates for State 2 from four sources

The probability component $p_i(J|S, A, R, F)$ is the adjustment for misclassification causing under-count 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, the model adjusts for non-response by the quantity $p_i(R|S, A, F)$.

The matched data set currently being used for estimation includes only the JAS records with matches in

the list frame, and records in the list with no matches in the JAS are not considered in the capture-recapture model. Thus, much of the information in the list frame is not used for the estimation of farm numbers from the JAS. Because the list frame and the Census contain relatively complete and up-to-date information on U.S. agricultural operations, the use of all available data from either of the two data sources can potentially provide improved estimates of farm numbers. Utilizing more data from the Census or the list frame for the estimation of farm numbers can also potentially reduce the year-to-year fluctuations observed (as illustrated in Figures 2 & 3) in the estimates from the current model.

4 An alternative capture-recapture approach for estimation of farm numbers

In this section we discuss an alternative capture-recapture approach that utilizes all available records from the JAS and the Census to correct for biases due to misclassification and to produce more stable farm number estimates. The method plausibly assumes that the JAS and the Census are independent and takes into account heterogeneities among farms by estimating farm-level inclusion probabilities using logistic regression models (Huggins, 1989; Alho, 1990). The Census data can be matched with JAS records for census years and the list frame may replace the Census for non-census years. However, data from the list frame need further processing to be used in the model. Let N be the total number of farms in the population. Define a farm as an operation that was classified as a farm in the Census or the JAS. For farm f_i, $i = 1, 2, 3, \ldots, N$, in the population, define the following indicator functions (Alho 1990).

$$c_{10i} = \begin{cases} 1, & \text{if farm } f_i \text{ was captured only by the JAS} \\ 0, & \text{otherwise} \end{cases}$$

$$c_{01i} = \begin{cases} 1, & \text{if farm } f_i \text{ was captured only by the Census} \\ 0, & \text{otherwise} \end{cases}$$

$$c_{11i} = \begin{cases} 1, & \text{if farm } f_i \text{ was captured by the JAS and the Census} \\ 0, & \text{otherwise} \end{cases}$$

$$c_{00i} = \begin{cases} 1, & \text{if farm } f_i \text{ was not captured by either of the two lists} \\ 0, & \text{otherwise} \end{cases}$$

Let n_J, n_C and n_B respectively denote the numbers of farms captured by the JAS, by the Census and by both the JAS and the Census. Traditional capture-recapture models would estimate the total number of farms in the population by $\hat{N} = \frac{n_J n_C}{n_B}$, assuming that all farms have equal probabilities of capture. However, farms have heterogeneous capture probabilities. For example, larger farms have higher probabilities of capture than smaller ones.

Let θ_{Ji} and θ_{Ci} respectively denote the probabilities of capture by the JAS and the Census for farm f_i. Under the assumption of independence, the probability of capture by both the JAS and the Census, θ_{Bi}, is given by $\theta_{Bi} = \theta_{Ji}\theta_{Ci}$ and the probability of capture by at least one of the lists is $\phi_i = \theta_{Ji} + \theta_{Ci} - \theta_{Bi}$. Thus, the probability of non-capture for farm f_i is $1 - \phi_i$. If $\mathbf{C}_i = (c_{10i}, c_{01i}, c_{11i}, c_{00i})$ is a vector of capture statuses for farm f_i, then $\mathbf{C}_i$ has a multinomial distribution (Alho, 1990) with parameter vector $\boldsymbol{\theta}_i = (\theta_{Ji}(1-\theta_{Ci}), \theta_{Ci}(1-\theta_{Ji}), \theta_{Bi}, 1-\phi_i)$. That is,

$$\mathbf{C}_i | \boldsymbol{\theta}_i \sim \text{Multinomial}(1, \theta_{Ji}(1-\theta_{Ci}), \theta_{Ci}(1-\theta_{Ji}), \theta_{Bi}, 1-\phi_i). \tag{3}$$

Alho (1990) employs conditional likelihoods to estimate these probabilities based on data from the captured farms. When $c_{00i} = 0$ (i.e., when the farm was captured at least once), the likelihood contribution from the i^{th} farm is

$$L_i(\boldsymbol{\theta}_i|.) = \frac{\left\{\theta_{Ji}(1-\theta_{Ci})\right\}^{c_{10i}}\left\{\theta_{Ci}(1-\theta_{Ji})\right\}^{c_{01i}}\theta_{Bi}^{c_{11i}}}{\phi_i}.$$

The likelihood from all the captured farms is then

$$L(\boldsymbol{\theta}|.) = \prod_{f_i \in \mathcal{C}} \frac{\left\{\theta_{Ji}(1-\theta_{Ci})\right\}^{c_{10i}}\left\{\theta_{Ci}(1-\theta_{Ji})\right\}^{c_{01i}}\theta_{Bi}^{c_{11i}}}{\phi_i},$$

where $\mathcal{C}$ is the set of all captured farms. The capture probabilities θ_{Ji} and θ_{Ci} can be modeled as functions of covariates $\mathbf{Z}_{1i}$ and $\mathbf{Z}_{2i}$ as:

$$\begin{aligned}\text{logit}(\theta_{Ji}) &= \mathbf{Z}'_{1i}\boldsymbol{\beta}_1 \\ \text{logit}(\theta_{Ci}) &= \mathbf{Z}'_{2i}\boldsymbol{\beta}_2\end{aligned}$$

After the model is fitted, ϕ_i can be estimated by $\hat{\phi}_i = \hat{\theta}_{Ji} + \hat{\theta}_{Ci} - \hat{\theta}_{Ji}\hat{\theta}_{Ci}$. The total number of farms can then be estimated by using the Horvitz-Thompson estimator

$$\hat{N} = \sum_{f_i \in \mathcal{C}} \frac{1}{\hat{\phi}_i}.$$

5 Simulation studies

Simulations were performed to study the performance of the proposed method in the estimation of farm numbers. The probabilities of capture by the JAS and the Census (i.e., θ_{Ji} and θ_{Ci}) were assumed to depend on a total of 20 covariates. These probabilities were estimated from the model

$$\begin{aligned}\text{logit}(\theta_{Ji}) &= \mathbf{Z}'_{1i}\boldsymbol{\beta}_1 \\ \text{logit}(\theta_{Ci}) &= \mathbf{Z}'_{2i}\boldsymbol{\beta}_2,\end{aligned} \tag{4}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of parameter values. The outcome for the i^{th} (i.e., capture or non-capture of the farm by the JAS, the Census or both) was generated from a multinomial distribution with size 1 and probability vector ($\theta_{Ji}(1-\theta_{Ci})$, $\theta_{Ci}(1-\theta_{Ji})$, $\theta_{Ji}\theta_{Ci}$, $1-\theta_{Ji}-\theta_{Ci}+\theta_{Ji}\theta_{Ci}$).

Table 1: Percent relative mean biases, simulation standard deviations and mean squared errors for farm number estimates from the capture-recapture model

Population Size (N)	Percent Rel. Bias	Sim. Std	MSE
500	11.278	235.390	55287.0
1000	3.473	82.059	6720.2
1500	2.085	91.438	8344.26
2000	1.734	102.539	10493.2
2500	0.920	113.2019	12789.03
5000	0.612	144.286	20776.8
8000	0.083	169.0899	28305.5
10000	-0.050	190.003	35739.0

The model was then applied to the simulated data to estimate the number of farms. The sample sizes considered ranged from 500 to 10,000. Table 1 shows the percent relative mean biases, simulation standard deviations and the mean squared errors of the farm number estimates. As can be seen from the table, the model based estimates of farm numbers are very close to the true values for all sample sizes considered.

6 Application to the 2012 JAS

The proposed capture-recapture model is illustrated by using data from the 2012 JAS for the estimation of farm numbers. Because 2012 was a census year, the JAS records were matched to the Census records to perform estimation. For this analysis, data from 12 states were considered and the farm numbers were estimated separately for these states. The covariates considered in the models were total value of production (TVP) categorized into five groups, farm type, and the sex, age and race of the operator. Table 2 shows ratios of farm number estimates from the current model (the proposed model) and the ASB estimates. Due to the confidential nature of parts of the data, actual state labels are not provided. Bar charts of the differences between the model based estimates and the ASB estimates are also shown in Figure 4 and Figures 5, 6 & 7 show bar charts of the estimated farm numbers from the three sources. While the estimates from the current and the proposed models are comparable for some of the states, the two methods provide substantially different estimates for the majority of the states considered. For example, the proposed model and the ASB estimates for State 6 are very close to each other, but the estimate from the current model is much smaller.

Table 2: Ratios of farm number estimates from the current model (the proposed model) and from the ASB

	Source	
State	Current Model	Proposed Model
State 1	1.001	1.053
State 2	0.936	1.006
State 3	1.139	0.950
State 4	0.697	1.069
State 5	0.938	0.956
State 6	0.749	0.965
State 7	1.095	1.013
State 8	1.333	1.079
State 9	1.001	1.019
State 10	0.985	0.974
State 11	1.072	1.189
State 12	1.063	0.999

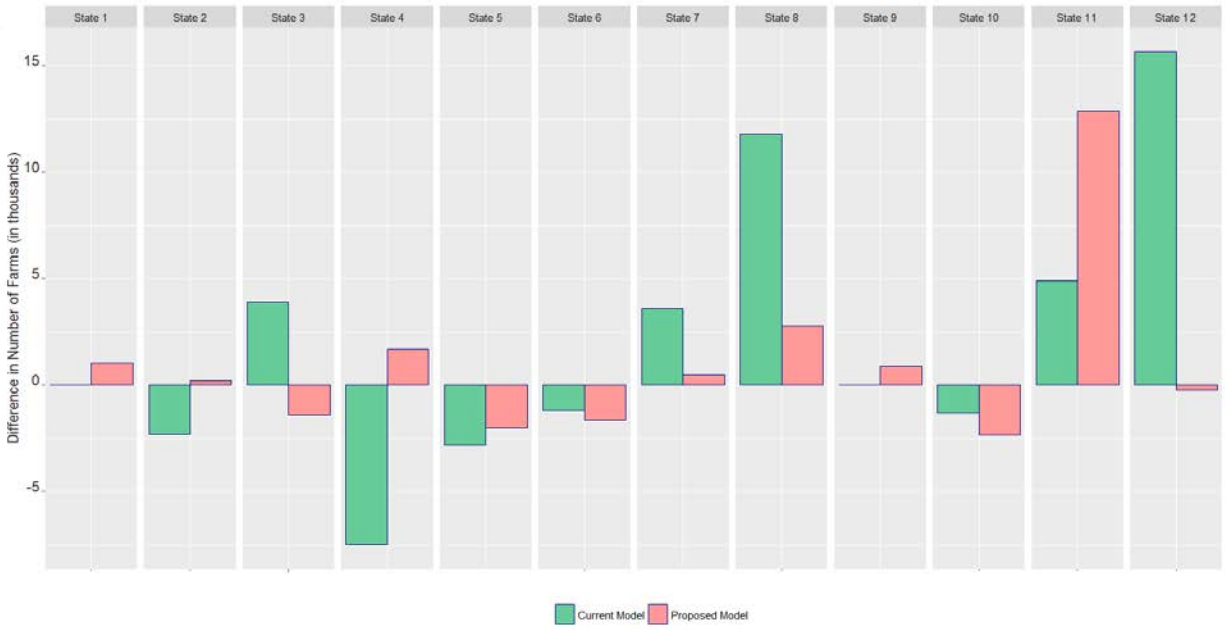

Figure 4: Differences of farm number estimates between the current/proposed model and the ASB estimates

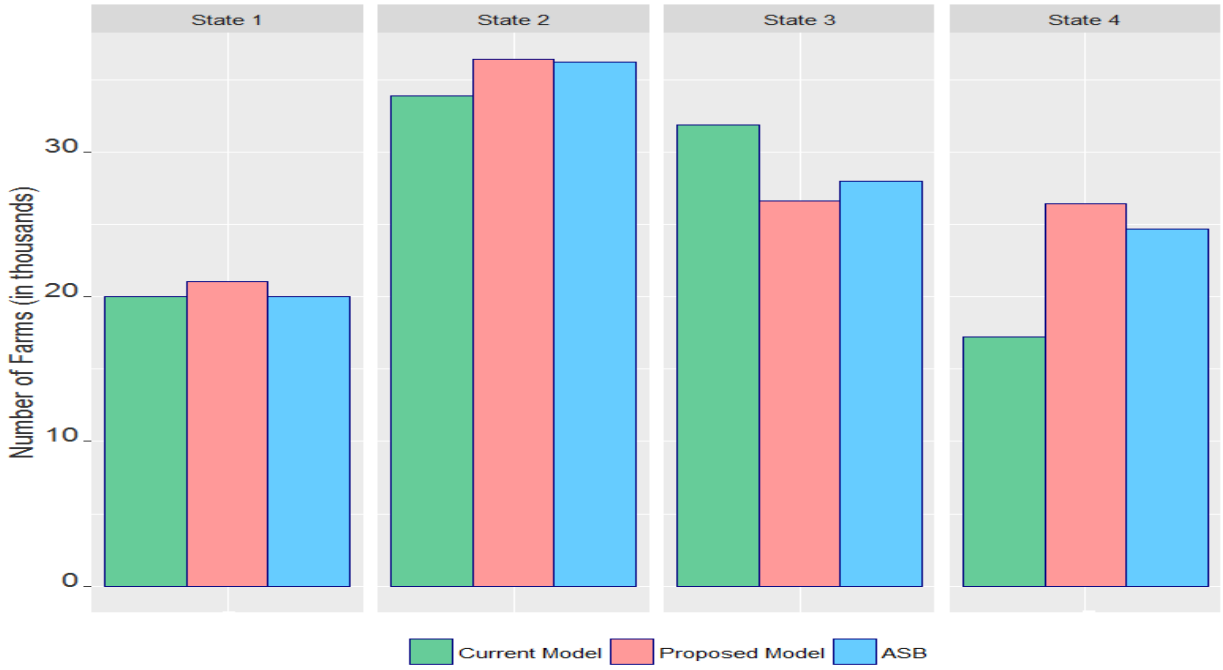

Figure 5: Farm number estimates from the three sources

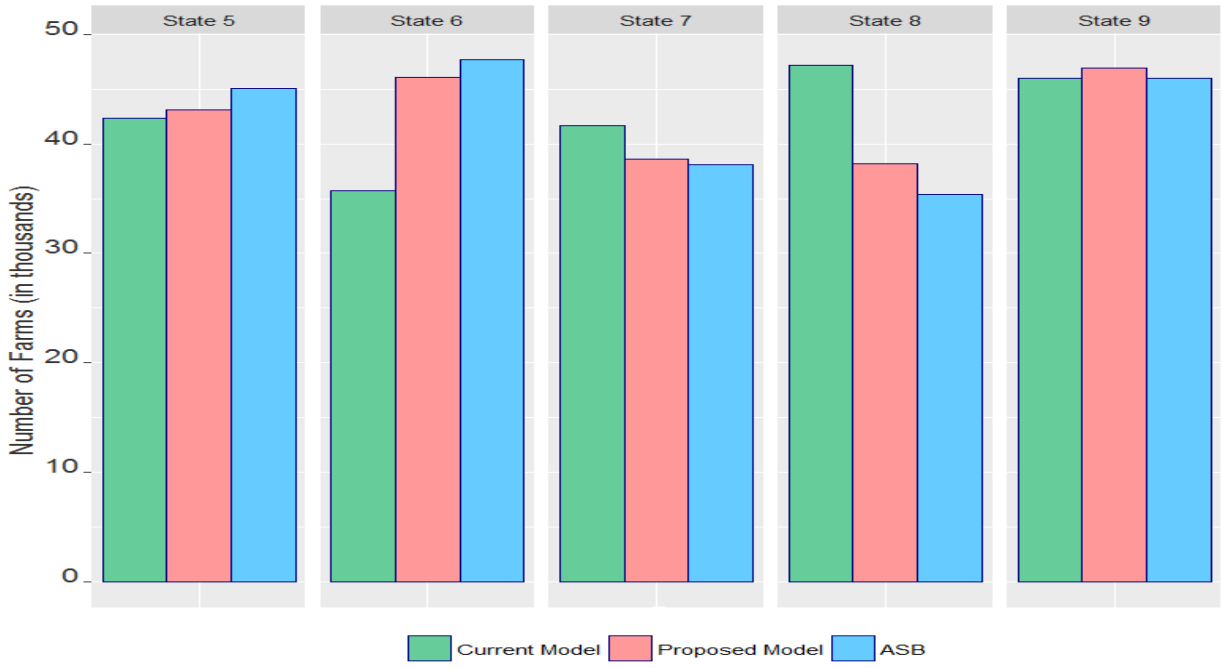

Figure 6: Farm number estimates from the three sources

Figure 7: Farm number estimates from the three sources

7 Discussion and conclusion

The proposed method uses the JAS and Census records for the estimation of farm numbers from the JAS. To account for the varying JAS inclusion probabilities among farms, the method estimates heterogeneous

probabilities of capture based on logistic regression models. Because all the Census records are used in the estimation process, the proposed method can potentially reduce estimation biases due to misclassification and provide estimates with less year-to-year fluctuations.

Simulations showed that the proposed method provides farm number estimates with low biases when the model is correctly specified. The method was also applied for the estimation of farm numbers from the 2012 JAS by using the 2012 Census as the second source of data. A limited number of covariates were used as predictors in the model for the 12 states considered, but considering more predictors may potentially improve the performance of the model. Because the JAS or Census inclusion probabilities are likely to be related to some of the thousands of variables in the respective data sets, the inclusion or exclusion of important covariates can affect farm number estimates. Variable selection methods, such as stepwise regression and LASSO, may be employed to determine the covariates to be included in the final model.

The list frame can be considered as a potential source of the second sample for the estimation of JAS farm numbers, particularly during estimation years that are far from census years. However, each operation in the list frame is categorized as 'active', 'criteria' (potential farm), or 'inactive' and the status of the operation as a farm or a non-farm is not available in the list. As a result, one needs to estimate the farm statuses of operations before employing the list frame data for the estimation of farm numbers from the JAS. The proposed model assumes that the number of farms in a state or the U.S. remains the same during the time between the JAS and the completion of the Census. Because farms, especially small ones, are continually coming into and going out of business, this assumption may not be plausible. Thus, the model should be modified to account for the time lapse between the JAS and the Census.

References

Abreu, D. (2007). Results from the 2002 classification error study. *Research and Development Division. RDD Research Report: RDD-07-03*.

Abreu, D., Arroway, P., Lamas, A., Lopiano, K., and Young, L. (2010a). Using the census of agriculture list frame to assess misclassification in the june area survey. In *Proceedings of the Joint Statistical Meetings*.

Abreu, D., Busselberg, S., Lamas, A., Barboza, W., and Young, L. (2014). Evaluating a new approach for estimating the number of u.s. farms with adjustment for misclassification. In *Proceedings of the Joint Statistical Meetings*.

Abreu, D., Dickey, N., and McCarthy, J. (2009). 2007 classification error survey for the united states census of agriculture. Technical report, United States Department of Agriculture, National Agricultural Statistics Service.

Abreu, D., McCarthy, J., and Colburn, L. (2010b). 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*.

Alho, J. (1990). Logistic regression in capture-recapture models. *Biometrics*, pages 623–635.

Arroway, P., Abreu, D., Lamas, A., Lopiano, K., and Young, L. (2010). An alternate approach to assessing misclassification in the JAS. In *Proceedings of the Joint Statistical Meetings*.

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76(1):133–140.

Lamas, A., Abreu, D., Arroway, P., Lopiano, K., and Young, L. (2010). Modeling misclassification in the June Area Survey. In *Proceedings of the Joint Statistical Meetings*.

Lopiano, K., Lamas, A., Abreu, D., Arroway, P., and Young, L. (2011). Adjusting the june area survey estimate of the number of us farms for misclassification and non-response. Technical report, United States Department of Agriculture, National Agricultural Statistics Service.

Young, L., Abreu, D., Arroway, P., Lamas, A., and Lopiano, K. (2010). Precise estimates of the number of farms in the united states. In *Proceedings of the Joint Statistical Meetings*.