Modeling Misclassification in the June Area Survey

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

Abstract

Each year, the National Agricultural Statistics Service (NASS) conducts the June Area Survey (JAS), which is based on an area frame. The JAS provides information on U.S. agriculture, including an estimate of the number of farms in the U.S. NASS also conducts the Census of Agriculture every five years in years ending in 2 and 7. The census uses a list frame, and also produces an estimate of the number of farms. In 2007, the two estimates were further apart than could be attributed to sampling error alone. Using data from the 2007 JAS and the 2007 Census, misclassification of tracts as agricultural or non-agricultural can be identified. A model estimating the JAS undercount of the number of farms was developed and used to provide a revised estimate. The development of the model and its potential use for adjusting the JAS estimate for misclassification in non-census years are discussed.

Key Words: June Area Survey; Misclassification; Generalized Linear Models

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. A farm is defined 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. 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.

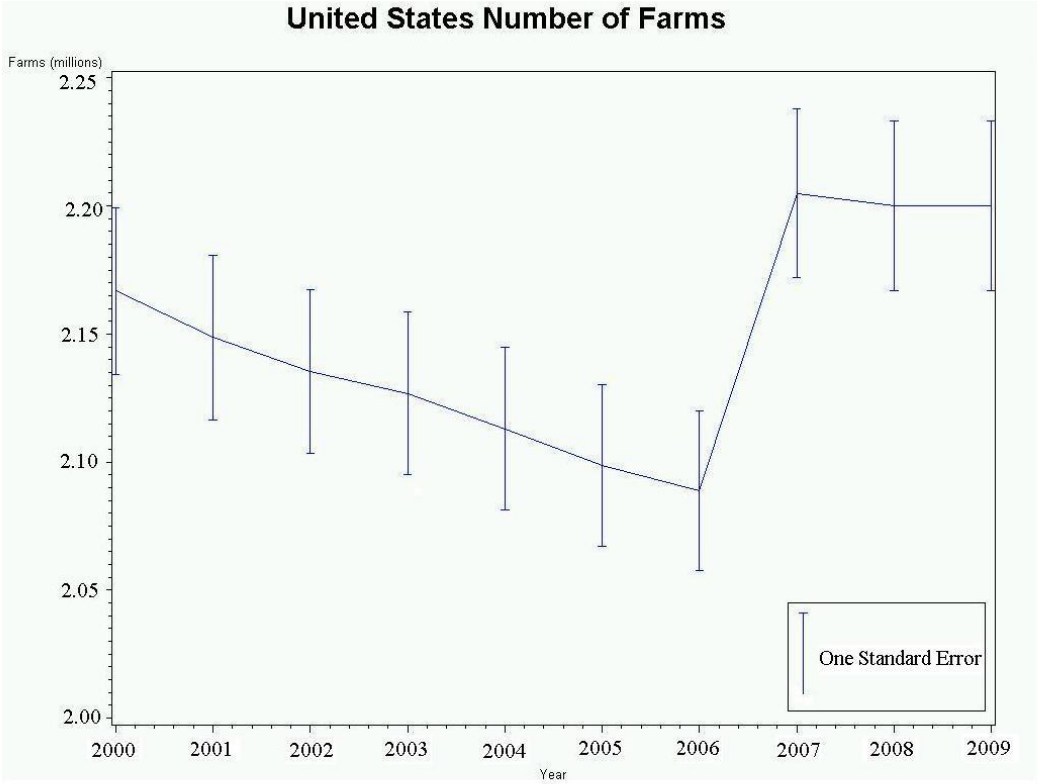

Figure 1: Published estimates of the number of U.S. farms from 2000 to 2009 and bars with length of one standard error.

Previous studies conducted by NASS indicated that a possible source of this underestimate is misclassification. One such study is the Classification Error Survey (CES) conducted in 2007, which was based on a final set of only 67 respondents. The CES results suggested that during the screening procedures of the JAS, some agricultural operations were incorrectly classified as non-agricultural. 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.

The purpose of this work is to gain a better understanding of the misclassification present in the JAS and to use this information to propose an adjusted estimate of the number of farms for the JAS.

2. Background of the June Area Survey and Census of Agriculture

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, such as 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,

$$\sum_{i=1}^{l} \sum_{j=1}^{s_i} \sum_{k=1}^{n_{ij}} e_{ijk} a_{ijk}$$

where

i indexes stratum
j indexes substratum
k indexes segment
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 the selection for each segment in substratum j in land-use stratum i

$$a_{ijk} = \sum_{m=1}^{x_{ijk}} t_{ijkm}$$

m indexes tract
$x_{ijk} =$ Number of *farm* tracts in the given segment

$$t_{ijkm} = \text{ Tract-to-farm ratio of the tract} = \frac{\text{tract acres for the } m^{\text{th}} \text{ tract}}{\text{farm acres for the } m^{\text{th}} \text{ tract}}$$

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

In addition to the JAS, NASS conducts a Census of Agriculture every five years (for years ending in 2 and 7). 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 employs a dual frame: an independent list frame of all known agricultural operators and the area frame from the JAS. The area frame is used as a measure of incompleteness of the census list frame. In this work, it is shown that the census list frame can also be used as a follow-up to the JAS and to assess potential misclassification of the JAS non-farms.

3. Methodology

Because the census list frame is created independently from the JAS area frame, it can be 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 final census farm status was considered the follow-up to the JAS. However, the adjustment presented here refers only to census farms identified as non-farms in the JAS.

To adjust for misclassification, consider the following estimate:

$$\sum_{i=1}^{l} \sum_{j=1}^{s_i} \sum_{k=1}^{n_{ij}} e_{ijk} a_{ijk} + \sum_{i=1}^{l} \sum_{j=1}^{s_i} \sum_{k=1}^{n_{ij}} e_{ijk} y_{ijk}$$

where

$y_{ijk} = \sum_{m=1}^{z_{ijk}} t_{ijkm}$

$z_{ijk} =$ Number of *non-farm* tracts in the given segment

$t_{ijkm} =$ Tract-to-farm ratio of the tract

In the current JAS estimate, all non-farm tracts have t_i = 0. Therefore, for misclassified non-farm tracts, t_i is incorrectly identified as 0 when t_i is actually greater than 0. Thus the second term is reported as 0, when it is actually greater than 0, leading to an undercount when misclassification is present.

When a follow-up is conducted, we are able to adjust for misclassification directly by obtaining the true t_i for all non-farms. Using the true t_i's, the second term is calculated for all non-farms and the number of farms is adjusted for misclassification. However, this is still potentially an undercount because not every record can be matched. In the adjustment, all unmatched tracts are given t_i = 0, since it cannot be assumed that the rate of conversion for matched tracts and unmatched tracts are the same. If some unmatched tracts are truly farms, their true t_i is greater than 0, leading to an undercount. Otherwise, if all unmatched tracts are true non-farms then the true $t_i = 0$, and the estimate is accurate.

In years when a follow-up cannot be conducted, the tract-to-farm ratios cannot be directly estimated. Therefore, the second term must be estimated for non-farm tracts by other means. Using the data from the matching procedure conducted in 2007, a model was developed that captures the misclassification behavior and yields an expected tract-to-farm ratio for 2009.

Because the tract-to-farm ratio (t_{ijkm}) is unobserved in non-farm tracts, y_{ijk} in the second sum is unknown. Here y_{ijk} is estimated using

$$\hat{y}_{ijk} = \sum_{m=1}^{z_{ijk}} \hat{E}(t)_{ijkm}$$

where $\hat{E}(t)_{ijkm}$ is the estimated expected tract-to-farm ratio of farm tracts in stratum i, substratum j, segment k, tract m that were classified as non-farms in the JAS.

The challenge is to obtain a good estimate of $E(t)_{ijkm}$ for all i, j, k, m. To do this, a hierarchical model was developed.

Consider a tract that was identified as non-farm in the JAS. Let **X** be a set of covariates.

Let u be an indicator of whether or not a tract had census follow-up ($u = 1$ if the tract had census follow-up and $u = 0$ if the tract did not have census follow-up). Furthermore, suppose

$$u \sim \text{Bernoulli}\ (\pi_u)\,,$$

where π_u depends on **X**.

Let f be an indicator of whether the tract qualifies as a farm ($f = 1$ if the tract is a farm and $f = 0$ if the tract is a non-farm). Conditional on u being 1 (the tract had a census follow-up), let

$$(f\,|u = 1) \sim \text{Bernoulli}\ (\pi_f),$$

where π_f also depends on **X**. Thus, $f\,|u$ has the following density,

$$f_1(f|u) = \pi_f^f (1-\pi_f)^{1-f} I(u=1).$$

Let z be an indicator that the tract-to-farm ratio is *not* equal to 1 ($z = 1$ if the tract-to-farm ratio is less than 1 and $z = 0$ if the tract-to-farm ratio is 1). Thus, conditional on u being 1 and f being 1,

$$(z|f = 1, u = 1) \sim \text{Bernoulli}(\pi_z),$$

where π_z again depends on **X**. Thus, $(z|f, u)$ has the following density,

$$f_2(z|f,u) = \pi_z^z (1-\pi_z)^{1-z} I(u=1) I(f=1).$$

Finally, let *t* denote the tract-to-farm ratio. Conditional on *z*, *f* and *u* all being 1, let

$$(t|z = 1, f = 1, u = 1) \sim \text{Beta}(\mu, \phi),$$

where μ and ϕ depend on **X**. It is important to note that Beta(μ, ϕ) has the following density,

$$\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu))} t^{\mu\phi-1}(1-t)^{(1-\mu)\phi-1}$$

Under this parameterization the mean is μ. Thus, $(t|f, z, u)$ has the following density,

$$\begin{aligned} f_3(t|f,z,u) = {} & \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu))} t^{\mu\phi-1}(1-t)^{(1-\mu)\phi-1} I(z=1)I(f=1)I(u=1) \\ & + 1I(z=1)I(f=1)I(u=1) \end{aligned}$$

The first term in the above sum corresponds to tracts with a tract-to-farm ratio less than 1, i.e. $z = 1$, where the tract-to-farm ratio follows a beta distribution. The second part of the sum corresponds to where the tract-to-farm ratio is 1, i.e. $z = 0$.

The unobserved tract-to-farm ratio of a non-farm tract, *t*, is of primary interest because *t* is unobserved. Therefore *E(t)* is used to estimate it. Based on the hierarchy described above, the expected value of *t* is calculated as follows:

$$\begin{aligned} E(t) &= E_u(E_f(E_z(E_t(t|f,z,u)))) \\ &= E_u(E_f(E_z(\mu I(z=1)I(f=1)I(u=1) + 1I(z=0)I(f=1)I(u=1)))) \\ &= E_u(E_f(\mu\pi_z I(f=1)I(u=1) + (1-\pi_z)I(f=1)I(u=1))) \\ &= E_u(\mu\pi_z\pi_f I(u=1) + (1-\pi_z)\pi_f I(u=1)) \\ &= \mu\pi_z\pi_f\pi_u + (1-\pi_z)\pi_f\pi_u \\ &= \pi_f\pi_u((\mu-1)\pi_z + 1) \end{aligned}$$

One major assumption of this model is that the tract-to-farm ratio is 0 when no follow-up was done. This assumption is partially justified because follow-up was an attempt to match a JAS tract to a census record. Failure of a JAS tract to match a census record is assumed to be a result of that tract truly being a non-farm. Thus, the unobserved tract-to-farm ratio would be 0. If all JAS tracts had census follow-up ($\pi_u = 1$), this assumption would not be necessary. However, because π_u is less than 1, it is likely this adjustment will still be an underestimate.

Given the model, the next step is to develop an estimator for *E(t)*. Suppose $\hat{\mu}$, $\hat{\pi}_z$, $\hat{\pi}_f$, and $\hat{\pi}_u$ are independent estimates of μ, π_z, π_f, and π_u. An estimate for *E(t)* would therefore be

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

Thus, the challenge is to develop estimates for μ, π_z, π_f, and π_u. Based on the distributional assumptions, generalized linear models were employed to estimate each of the unknown parameters. Because the information available for non-farm tracts is limited, only covariates that were collected on each non-farm tract can be used. The two covariates included were the land-use stratum and the tract's agricultural classification. The stratum took on one of four values indicating whether or not the tract falls into a stratum between 10 and 19 (>50% cultivated), 20 and 29 (15-50% cultivated), 30 and 39 (agricultural urban/commercial), or 40 and 49 (<15% cultivated or non-agricultural). In terms of agricultural classification, the JAS tract falls into one of four categories:

- Agricultural
- Non-Agricultural with Potential
- Non-Agricultural with Potential Unknown
- Non-Agricultural with No Potential

Recall that *i* indexes the tract's stratum (10-19, 20-29, 30-39, 40-49) and *j* indexes the tract's JAS status (agricultural, non-agricultural with potential, non-agricultural with potential unknown, non-agricultural with no potential).

The relationship between μ and the covariates is modeled with a beta regression model with a logit link. That is,

$$\log\left(\frac{\mu(i,j)}{1-\mu(i,j)}\right) = \alpha_i + \beta_j.$$

Similarly, π_z, π_u and π_f are modeled using the following logistic regression models.

$$\log\left(\frac{\pi_z(i,j)}{1-\pi_z(i,j)}\right) = \alpha_i^z + \beta_j^z,$$

$$\log\left(\frac{\pi_u(i,j)}{1-\pi_u(i,j)}\right) = \alpha_i^u + \beta_j^u, \text{ and}$$

$$\log\left(\frac{\pi_f(i,j)}{1-\pi_f(i,j)}\right) = \alpha_i^f + \beta_j^f$$

In all levels of the model, the parameters were estimated using maximum likelihood estimation under the constraint that $\beta_4^z, \beta_4^u, \beta_4^f, \beta_4 = 0$ respectively for each of the models (Ferrari and Cribari 2004, McCullagh and Nelder 1989). The estimated parameters are used to construct

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

Note that $\hat{E}$ is calculated for each tract and depends on the tract's stratum and substratum. The segment level estimate is

$$\hat{y}_{ijk} = \sum_{m=1}^{z_{ij1}} \hat{E}\left(t_{ijkm}\right).$$

Finally, the model-based indication for the total number of farms is

$$\sum_{i=1}^{l}\sum_{j=1}^{s_i}\sum_{k=1}^{n_{ij}} e_{ijk}\, a_{ijk} + \sum_{i=1}^{l}\sum_{j=1}^{s_i}\sum_{k=1}^{n_{ij}} e_{ijk}\, \hat{y}_{ijk}.$$

That is, the sum of expanded, observed tract-to-farm ratios plus the sum of expanded, estimated tract-to-farm ratios for tracts initially identified as non-farm in the JAS. This second term compensates for the undercount resulting from the misclassification of some portion of these tracts.

4. Results and Conclusions

Using the 2007 JAS and follow-up information, a direct estimate can be obtained for the number of farms in the United States. The updated estimate is obtained by summing the JAS design-based estimate for farm tracts, and the direct estimate from the non-farm tracts. This estimate is 91.7% from farm tracts and 8.3% from non-farm tracts.

Using the 2007 follow-up information to develop a model, a modeled estimate can be obtained for 2007. The updated estimate is obtained by summing the JAS design-based estimate for farm tracts and the modeled estimate for non-farm tracts. This estimate is 91.5% from farm tracts and 8.5% from non-farm tracts. The modeled and direct estimates are close, indicating that the model captures the misclassification behavior well.

Using the model developed based on the 2007 follow-up data, a 2009 modeled estimate was calculated. The updated estimate is obtained by summing the JAS design-based estimate from farm tracts and the modeled estimate from non-farm tracts. This estimate is 91.2% from farm tracts and 8.8% from non-farm tracts. However, this assumes that misclassification rates and behavior are independent of time.

5. References

Ferrari, S., Cribari-Neto, F., (2004). Beta regression for modeling rates and proportions. Journal of Applied Statistics 31, 799815

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd ed. London: Chapman and Hall

Agresti, A. Categorical Data Analysis. Wiley, New York, NY, 2002.

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

Abreu, Denise A., Pam Arroway, Andrea C. Lamas, Kenneth K. Lopiano, and Linda J. Young. 2010. "Using the Census of Agriculture List Frame to Assess Misclassification in the June Area Survey" *Proceedings of the Joint Statistical Meetings*.

Davies, C. (2009). “Area Frame Design for Agricultural Surveys”. RDD Research Report. Washington, DC: USDA, National Agricultural Statistics Service.