United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research and
Development Division
Washington DC 20250

RDD Research Report
Number RDD-11-04

November 2011

Adjusting the June Area Survey Estimate of the Number of U.S. Farms for Misclassification and Non-response

Kenneth K. Lopiano
Andrea C. Lamas
Denise A. Abreu
Pam Arroway
Linda J. Young

This report has been prepared for limited distribution to the research community outside the U.S. Department of Agriculture (USDA). The views expressed herein are not necessarily those of NASS or USDA.

EXECUTIVE SUMMARY

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 conducted every five years (in years ending in 2 and 7). The Census of Agriculture employs the area frame from the JAS as well as a list frame composed of all known agricultural operations. Both surveys provide an estimate of the number of farms in the United States. Following each census, previous annual number of farms estimates are revised, if necessary, based on intercensal trends. The JAS annual estimate showed a decline in number of farms from 2003-2006 prior to the 2007 Census. In addition, results from the 2007 Census indicated that the 2007 JAS was underestimating the number of farms. This led to an intercensal trend adjustment to the number of farms estimates that was larger than could be attributed to sampling error alone.

Previous studies conducted by NASS indicated that one possible source of the underestimate in the JAS is misclassification (Abreu, Dickey and McCarthy, 2009; Johnson 2000). Misclassification occurs when an operating arrangement with agricultural activity present is incorrectly identified as a non-farm, or when a non-farm arrangement is incorrectly identified as a farm.

Another potential factor associated with the JAS undercount is the estimation of agricultural activity for sampled tracts. When a tract operator is either inaccessible for a JAS interview or refuses to participate in the JAS, enumerators are instructed to estimate the tract-level agricultural items. As a result, farm-level items are left to be imputed. When calculating the total number of farms for the JAS, the tract-to-farm ratio (the tract acreage divided by the total farm acreage) is used to represent the proportion of a farm that is present in a tract. For agricultural tracts that are estimated, the tract-to-farm ratio is imputed. For non-agricultural tracts, the tract-to-farm ratio is 0.

Recent research has identified misclassification and estimation as two sources of error in the JAS (Abreu et. al, 2010; Lamas et. al, 2010; Lopiano et. al, 2010; Appendix A). This research report presents methodologies to adjust the JAS number of farms indication for both misclassification and non-response.

In years when a census is conducted, JAS records can be matched to the census respondents list and misclassification can be adjusted for directly. In this context, the census information is considered a follow-up. More broadly, if the JAS can be matched to any validation source, then misclassification can be accounted for directly. When matching to another source is not possible, the effect of misclassification can be estimated using data from a previous year for which follow-up was conducted. Here, generalized linear models are used to model the processes associated with misclassification and to obtain an estimated tract-to-farm ratio for the non-farm tracts. Because the information available for non-agricultural tracts is limited, only covariates that were observed for all non-agricultural tracts (land-use stratum and a description of the tract) were used in the model.

The misclassification model assumes that the misclassification process (i.e., rates and behavior of misclassification) are independent of time. Another implicit assumption of the model is the tract-to-farm ratio is 0 when no follow-up was done.

The resulting adjusted estimator based on these modeled tract-to-farm ratios includes a design-based portion (the traditional JAS estimator) and a model-based portion (the adjustment for misclassification). Although estimates of the variances associated with each portion of the estimator are derived, the two portions are correlated. An estimator of the variance that accounts for this correlation merits further research.

When the agricultural activity in a tract is estimated for the JAS, the tract-to-farm ratio is imputed using either previously reported/administrative data or a median imputation approach. Because the JAS does not currently identify the imputation method used to complete estimated records, the quality of the imputed values cannot be assessed. Thus, each estimated tract was treated as a non-respondent. The probability of obtaining a response for a tract is modeled as a function of covariates and design variables, and observations are reweighted based on their response probability. Given the response model, the response weight is the inverse of the estimated probability of response.

By combining the methodologies for misclassification and non-response, an estimator of the number of farms adjusted for both non-response and misclassification is constructed when a follow-up is possible. In addition, an estimator of the variance associated with this estimator is provided. The estimator still potentially represents an undercount because it is possible that some of the JAS non-farm records that did not match to a census record could be farms. Combining non-response and misclassification when misclassification is modeled merits further research.

RECOMMENDATIONS

1. **Thoroughly evaluate current JAS imputation procedures and develop appropriate imputation methodology**. Currently, the quality of imputed values for estimated tracts cannot be determined. The quality of the imputed data for total farm acreage is likely related to the method of imputation. The quality could be better assessed if the information regarding the source or method of imputation was retained. *This recommendation is currently being addressed. An office use box has been added to the 2011 JAS survey instrument which will collect the source of the farm acreage item reported on the questionnaire. Upon completion of the 2011 data collection processes, the data will be analyzed and various imputation approaches should be tested as per this recommendation.*

2. **Develop non-response methodology that reflects a combination of a revised imputation methodology (noted in the first recommendation) and a rigorous non-response methodology for estimated tracts that have no quality information available for imputation**.

3. **A final JAS survey indication should include adjustments for non-response, imputation, and misclassification.** In addition, future research is needed to develop a methodology that accounts for these three sources of error in the farm number indication and provides an appropriate measure of uncertainty associated with the final JAS indication.

Adjusting the June Area Survey Estimate of the Number of U.S. Farms for Misclassification and Non-response

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

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, which uses both a list and the JAS area frame, also produces an estimate of the number of U.S. farms. In 2007, the two estimates were further apart than could be attributed to sampling error alone. Previous studies of the JAS identified misclassification of JAS sampled units as a source leading to an undercount in the number of farms in the U.S. Using data from the 2007 JAS and the 2007 Census, misclassification of tracts as agricultural or non-agricultural were identified. Research has also identified the estimation of agricultural activities for sampled tracts as another factor that contributes to the discrepancy in the JAS number of farms estimate. This research report presents methodology that adjusts for two known sources of error on the JAS: misclassification and estimation (which later will be addressed as non-response).

KEY WORDS: June Area Survey, Misclassification, Non-response, Generalized Linear Models

[1] Department of Statistics, University of Florida, Gainesville, FL 32611
[2] National Agricultural Statistics Service, USDA, 3251 Old Lee Hwy, Fairfax VA 22030
[3] Department of Statistics, North Carolina State University, Raleigh, NC 27695

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 conducted every five years (in years ending in 2 and 7). The Census of Agriculture employs the JAS area frame as well as a list frame composed of all known agricultural operations. Both surveys provide an estimate of the number of farms in the United States. 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. Any government agricultural payments received are included in determining whether an operation is a farm. Following each census, previous JAS 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 a large intercensal trend adjustment to the number of farms estimates.

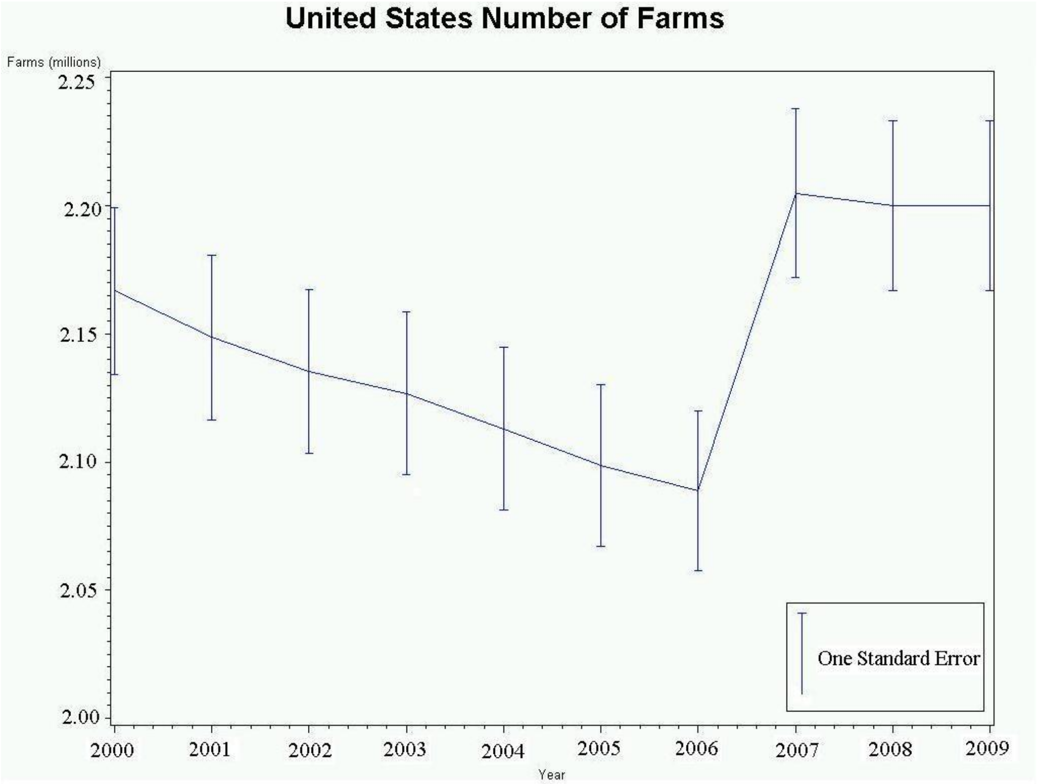

Figure 1: Published estimates of the number of U.S. farms from 2000 to 2009 and bars with length of one standard error (in either direction).

Previous studies conducted by NASS indicated that a possible source of this underestimate is misclassification. Misclassification occurs when an operating arrangement that meets the definition of a farm is incorrectly classified as a non-farm, or when a non-farm arrangement is incorrectly classified as a farm. One such study is the Classification Error Survey (CES) conducted in 2007, which was based on a final set of 67 respondents (Abreu, Dickey and McCarthy, 2009). The CES results suggested that, during the screening procedures of the JAS, some agricultural operations were incorrectly classified as non-agricultural, leading to more intensive efforts to understand the source and extent of misclassification in the JAS. The Farm Numbers Research Project (FNRP), based on an intensive post-June survey re-screening was conducted in 2009 (Abreu, McCarthy and Colburn, 2010) to address misclassification as it relates to the farm numbers indication. Concurrently, through a collaborative agreement with the National Institute of Statistical Sciences (NISS), a team of researchers was formed to review the methodology associated with the JAS and to recommend improvements. The team consisted of two NASS researchers, two university faculty members, a post-doctoral fellow, and a graduate student. The team evaluated several measures to address misclassification on the JAS. By matching the 2007 JAS to the 2007 Census of Agriculture list frame, the team evaluated misclassification on the JAS (Abreu, et al. 2010). In addition, the team identified the estimation of agricultural activities for sampled tracts as another factor contributing to the discrepancy in the JAS number of farms estimate (See Appendix A). Note a tract is a unique land operating arrangement. All land in sampled areas is divided into tracts. When a tract operator is either inaccessible for a JAS interview or refuses to participate in the JAS, enumerators are instructed to estimate the tract-level agricultural items whenever possible. This research report presents methodology that adjusts for two known sources of error on the JAS: misclassification and estimation.

2. JUNE AREA SURVEY & THE CENSUS OF AGRICULTURE

The June Area Survey (JAS) is conducted annually utilizing an area frame. It collects information on U.S. crops, livestock, grain storage capacity and type and size of farms. Land within the JAS area frame is divided into homogeneous 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, providing 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 the total operating arrangement, which includes land both inside and outside of the JAS-selected tract, has qualifying agricultural activity. Otherwise, the tract is defined as non-agricultural.

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 enumeration 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 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 to assess potential misclassification of the JAS tracts defined as non-agricultural during the JAS.

3. METHODS

NASS's area frame is complete because the population of interest (land in the U.S.) is entirely covered by the sampling frame with no overlaps or gaps. Therefore, it has long been assumed that estimates derived from the JAS, using the area frame, are unbiased. However, recent research conducted by the NISS-NASS team (Abreu, et al. 2010) indicated two sources of error in the JAS: misclassification and estimation. Misclassification occurs when a tract, which has some portion of a farming operation inside it, is identified as a non-farm or when a non-farm tract is classified as a farm. Agricultural activity in the tract is estimated when the tract operator is either inaccessible for or refuses an interview. The failure to adjust for these sources of error contributes to the undercount of the number of farms in the JAS. This research report considers methodologies to adjust the JAS number of farms indication for misclassification and estimation.

3.1 Misclassification

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. If the tract was identified as a farm in the JAS and a non-farm in the census, then the tract was considered a farm. If the tract was identified as a non-farm in the JAS and a farm in the census, then the tract was considered a farm. In other words, if the tract was identified as a farm in either the JAS or the census, then the tract was considered a farm. The assumption ignores the potential overcount in the JAS that can arise from non-farm tracts being identified as farms. Historically, the overcount, although important, is known to be negligible. As a result, the focus here is only on the undercount.

3.1.1 Quantifying Misclassification

In years when a census is conducted (i.e., years ending in 2 and 7), the JAS records can be matched to the records of census respondents, allowing a direct adjustment for misclassification. More broadly, if the JAS can be matched to any validation source, then misclassification can be accounted for directly. However, when matching to another source is not possible, the effect of misclassification can be estimated if it is reasonable to assume that misclassification behaves similarly in years when a follow-up is conducted. Under that assumption, a model of misclassification can be developed from the follow-up year's matched data and used to adjust for misclassification in the year for which no follow-up information is available. The process of developing a model is described in the next section.

3.1.2 Modeling Misclassification of Non-Farms

Because the focus here is in adjusting the JAS indication for an *undercount*, misclassification of JAS non-farm tracts is modeled.

The current NASS estimate for the number of farms is defined as

$$\sum_{i \in F} \pi_i^{-1} t_i$$

where π_i and π_i^{-1} are the inclusion probability and the expansion factor associated with farm *i*, respectively, t_i is the tract-to-farm ratio (tract acres divided by total farm acres) and *F* is the set of sampled farm tracts. However, to adjust for misclassification, consider the following estimate

$$\sum_{F} \pi_i^{-1} t_i + \sum_{NF} \pi_i^{-1} t_i$$

where *NF* is the set of sampled non-farm tracts.

The tract-to-farm ratio is unobserved in non-farm tracts; that is, t_i is missing in the second sum. If the tract is correctly classified as a non-farm the tract-to-farm ratio is zero. If it is incorrectly classified as a non-farm, then the tract-to-farm ratio should be greater than zero but less than or equal to one. Because some tracts are misclassified, an estimate of the tract-to-farm ratio for tracts misclassified as non-farms is needed. Here, t_i is estimated with a modeled estimate defined as

$$\hat{t}_i = E(t_i)$$

where $\hat{t}_i$ is the estimated tract-to-farm ratio of a misclassified tract. The challenge is to obtain a good estimate of t_i for all non-farm tracts. To do this, a hierarchical model was developed that accounts for the process used to identify misclassification.

Consider a tract that was identified as a non-farm. Let **X** be a set of covariates. Let *u* be an indicator of whether or not a tract had census follow-up. Furthermore, suppose

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

where π_u depends on **X**. Let f be an indicator that a farm is present in the tract. Conditional on u being 1, let

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

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

$$f_1(f|u) = \pi_f^f\left(1-\pi_f\right)^{1-f} I(u=1) + I(u=0)I(f=0),$$

where $I()$ is an indicator function. 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 0 if the tract-to-farm ratio is 1). Thus, conditional on u and f being 1,

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

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

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

Finally, let t denote the true 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,

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

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

$$\begin{gathered} f_3(t|z,f,u) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(1\text{-}\mu)}\, t^{\mu\phi\text{-}1}(1-t)^{(1-\mu)\phi-1}I(z=1)I(f=1)I(u=1) \\ +\,1I(z=0)I(f=1)I(u=1) \\ +I(u=0)I(t=0) + I(u=1)I(f=0)I(t=0) \end{gathered}$$

The first term in the above sum corresponds to tracts with a tract-to-farm ratio less than 1 (i.e., $z = 1$), while the second part of the sum corresponds to when the tract to farm ratio is 1 (i.e., $z = 0$).

The unobserved tract-to-farm ratio of a non-agricultural tract, t, is of primary interest. Here, E(t) is used to estimate a tract's unobserved tract-to-farm ratio, t. Based on the hierarchy defined above, the expected value of t is calculated as follows:

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

An implicit 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 result from that tract not being a farm. Thus, the unobserved tract-to-farm ratio would be 0. If all JAS tracts had a census follow up (π_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 of E(t) would therefore be,

$$\widehat{E(t)} = \hat{\pi}_f\hat{\pi}_u\big((\hat{\mu}-1)\hat{\pi}_z+1\big) \qquad (1)$$

Based on the distributional assumptions, generalized linear models are used to estimate each of the unknown parameters. Data for all non-farm tracts are used to develop the model, but only data available for all types of non-farm tracts can be used. The information available for non-agricultural tracts is limited; other non-farm tracts have additional information available. Thus, only covariates that were collected for non-agricultural tracts can be considered in model development. The two covariates included were land-use stratum and tract description. Cultivated land is divided into several land-use strata based on the distribution of cultivation in a state. The strata take on one of four values indicating whether or not the tract falls into a land use 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). Tract description is a variable identifying the tract as 1. Agricultural (i.e., an agricultural tract that did not qualify as a farm); 2. Non-Agricultural with Potential; 3. Non-Agricultural with Potential Unknown; or 4. Non-Agricultural with No Potential. Note, i indexes the tract's stratum and j indexes the tract's description.

To estimate μ, the following beta regression model with a logit link was used

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

To estimate π_f, π_u and π_z, the following logistic regression models were used

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

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

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

In all levels of each model, the parameters were estimated using maximum likelihood estimation under the constraint that $\beta_4 = 0, \beta_4^z = 0, \beta_4^u = 0,$ and $\beta_4^f = 0$ (Ferrari and Cribari 2004, McCullagh and Nelder 1989). The estimated parameters were used to estimate the probabilities π_z, π_u, π_f, and μ which in turn are used to estimate $\hat{t} = \mathrm{E}(t)$ as in formula (1) above.

The model-adjusted indication for the total number of farms is

$$\sum_F \pi_i^{-1} t_i + \sum_{NF} \pi_i^{-1} \hat{t}_i\,;$$

that is, it is the sum of expanded observed tract-to-farm ratios for farm tracts plus the sum of expanded estimated tract-to-farm ratios for tracts identified as non-farms. The first term is the traditional JAS indication. The second term compensates for the undercount resulting from the misclassification of some tracts identified as non-farms during the JAS. Let

$$Y_O = \sum_F \pi_i^{-1} t_i,$$

which is the part of the indication based on the JAS farm tracts, and let

$$Y_m = \sum_{NF} \pi_i^{-1} \hat{t}_i,$$

which is the modeled part of the model-adjusted indication.

3.1.3 Uncertainty

A measure of uncertainty provides a measure of the error associated with a given estimator. Here, the variances of estimators are the uncertainty measure and in practice, the variances of estimators are often estimated as well. Here, an estimate of variance for the model-adjusted indication is considered.

The estimator $Y_O = \sum_F \pi_i^{-1} t_i$ is design-based. Thus, the estimator of its variance is also based on the design. In contrast, $Y_m = \sum_{NF} \pi_i^{-1} \hat{t}_i$ is a model-based estimator, and the estimator of its variance is based on the asymptotic normality of the parameter estimate. From maximum likelihood theory,

$$\hat{\mathbf{p}} = \begin{pmatrix} \hat{\alpha}_1^u \\ \hat{\alpha}_2^u \\ \hat{\alpha}_3^u \\ \hat{\alpha}_4^u \\ \hat{\beta}_1^u \\ \hat{\beta}_2^u \\ \hat{\beta}_3^u \\ \hat{\alpha}_1^f \\ \hat{\alpha}_2^f \\ \hat{\alpha}_3^f \\ \hat{\alpha}_4^f \\ \hat{\beta}_1^f \\ \hat{\beta}_2^f \\ \hat{\beta}_3^f \\ \hat{\alpha}_1^z \\ \hat{\alpha}_2^z \\ \hat{\alpha}_3^z \\ \hat{\alpha}_4^z \\ \hat{\beta}_1^z \\ \hat{\beta}_2^z \\ \hat{\beta}_3^z \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \\ \hat{\alpha}_4 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} \dot{\sim} N \left(\mathbf{p} = \begin{pmatrix} \alpha_1^u \\ \alpha_2^u \\ \alpha_3^u \\ \alpha_4^u \\ \beta_1^u \\ \beta_2^u \\ \beta_3^u \\ \alpha_1^f \\ \alpha_2^f \\ \alpha_3^f \\ \alpha_4^f \\ \beta_1^f \\ \beta_2^f \\ \beta_3^f \\ \alpha_1^z \\ \alpha_2^z \\ \alpha_3^z \\ \alpha_4^z \\ \beta_1^z \\ \beta_2^z \\ \beta_3^z \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_u & 0 & 0 & 0 \\ 0 & \hat{\Sigma}_f & 0 & 0 \\ 0 & 0 & \hat{\Sigma}_z & 0 \\ 0 & 0 & 0 & \hat{\Sigma}_t \end{pmatrix} \right)$$

where $\widehat{\sum}_u, \widehat{\sum}_f, \widehat{\sum}_z, \widehat{\sum}_t$ are the asymptotic covariance matrices of the parameters associated with estimating *u*, *f*, *z*, and *t*, respectively.

If the parameter values are known, the modeled-based adjustment to the JAS estimator of farms, Y_m, written as a function of *p* is

$$f(p) = \sum_{i=1}^{4}\sum_{j=1}^{4} N_{i,j}\, \pi_u(i,j)\pi_f(i,j)((\mu(i,j) - 1)\pi_z(i,j) + 1)$$

where

$$\pi_u(i,j) = \frac{\exp(\alpha_i^u + \beta_j^u)}{1 + \exp(\alpha_i^u + \beta_j^u)}$$

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

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

$$\mu(i,j) = \frac{\exp(\alpha_i + \beta_j)}{1 + \exp(\alpha_i + \beta_j)}$$

$$N_{i,j} = \sum_{k|\text{Stratum}=i,\ \text{Tract Status}=j} e_{ijk}$$

Note e_{ijk} is the expansion factor for the k^{th} tract with stratum *i* and tract status *j*. Thus, based on the multivariate delta method, the estimated variance of this function is

$$\widehat{\text{var}}(Y_m) = \nabla f(\hat{\mathbf{p}})'\widehat{\sum}\nabla f(\hat{\mathbf{p}})$$

where $\nabla f(\hat{\mathbf{p}}) = \left[\frac{\partial f(\mathbf{p})}{\mathbf{p}_1}, \dots, \frac{\partial f(\mathbf{p})}{\mathbf{p}_{28}}\right]|_{\{\mathbf{p}=\hat{\mathbf{p}}\}}$ is the gradient function evaluated at $\hat{\mathbf{p}}$. An overall estimate of the variance of $\sum_F \pi_i^{-1} t_i + \sum_{NF} \pi_i^{-1} \hat{t}_i$ is difficult to obtain. The modeled and design-based portions of the estimator are correlated, and this correlation is not accounted for by simply adding the two variances of the two terms. With this in mind, a bootstrap/multiple imputation procedure could potentially be used to estimate the variability associated with the model-based components and the covariance between the design-based and the model-based components. Additional research is needed to fully assess the viability of this approach.

3.2 Estimation

The estimation of agricultural activity for sampled tracts contributes to the discrepancy between the JAS design-based and the model-adjusted estimators. During the sample selection process, tracts of land are selected to be surveyed for agricultural activity. When a tract operator is either inaccessible for a JAS interview or refuses to participate in the JAS, enumerators are instructed to estimate the tract-level agricultural items based on a physical observation of the tract. Consequently, farm-level items are left to be imputed using other sources (other NASS surveys, previous year JAS, Farm Service Agency information, etc.) or imputation methodologies.

One farm-level item that is imputed is total farm acreage, which together with the estimated tract acreage, is used to compute the tract-to-farm ratio. When calculating the total number of farms, the tract-to-farm ratio (the tract acreage divided by the total farm acreage) is used to represent the proportion of a farm that is present in a tract. When the agricultural activity in a tract is estimated, enumerators accurately calculate the tract acreage in person and Field Office (FO) staff are instructed to hand impute the total farm acreage using either previously reported or administrative data. If this information is not available, they are instructed to use strata-level median tract-to-farm ratios calculated for each state. FOs multiply these state/strata median tract-to-farm ratios by the tract acres to estimate the total farm acreage. Although median imputation was a common solution when this problem was first addressed, more recent research has illustrated its limitations. Therefore, estimation of total farm acreage for non-response tracts is a potential area of improvement in the process of estimating the number of farms in the United States.

In 2009, NASS conducted the Farm Numbers Research Project (FNRP). In this study, 595 estimated tracts' farm-level items from the JAS and the FNRP were compared. Substantial discordance was observed for a number of variables, including total farm acreage and consequently tract-to-farm ratio (See Appendix A). The quality of the imputed data for total farm acreage is likely related to the method of imputation. However, prior to 2011, the source used for imputation was not recorded. The quality of the imputed farm-level values could be assessed if the information was known. Here, quality is an overall measure of the validity and/or properties of the imputed value based on either the imputation source or the imputation methodology. The specific definition of quality and subsequent quantification merits further research. Because the quality of imputed values for estimated tracts cannot currently be determined, an intermediate solution is to treat each estimated tract as a unit non-respondent. Then, the JAS-based estimate of the number of farms can be adjusted using unit non-response methodologies. Such an approach, although statistically viable, is not able to fully utilize the information collected from estimated tracts, but it is used here. Note: For 2011, the JAS survey instrument has been amended so that the quality of the sources used for farm-level imputation can be assessed. Thus, in the future, tracts with quality information can be treated as respondents, and those remaining will be treated as unit non-respondents.

3.2.1 Non-Response

Non-response Model

The current estimate for the number of farms based on the JAS can be simplified to the following expression,

$$T = \sum_{i \in R} \pi_i^{-1} y_i t_i,$$

where R denotes the set of respondents, π_i denotes the inclusion probability of respondent i, y_i=1 if the tract contains a farm and is 0 otherwise, and t_i = tract-to-farm ratio. If ϕ_i denotes the probability of response for unit i, then the non-response weighted estimate for the total number of farms would be

$$T_{NR} = \sum_{i \in R} \pi_i^{-1} \phi_i^{-1} y_i t_i.$$

where ϕ_i is the probability the ith tract responds. In practice, ϕ_i is unknown and must be estimated. ϕ_i can be estimated in several ways.

Although sampling weights have often been incorporated in non-response methodologies (Platek and Gray, 1983), Little and Vartivarian (2003) show that “weighting response rates by sampling weights to adjust for design variables is either incorrect or unnecessary." (pp. 1589) Further, they recommend modeling non-response as a function of covariates and design variables. Given the model, the response weight is the inverse of the estimated probability from this model.

3.2.2 Estimating the Probability of Response: Logistic Regression

A logistic regression model was developed to estimate the probability of responding to the JAS. The model is based on the assumption that each tract has a probability ϕ_i of having a response recorded during the JAS. Further, the probability a tract has a response is independent of the probability that a response is obtained for any other JAS tract. Finally, the probability a tract has a response can be predicted using available tract, state and land-use stratum items.

During the JAS, tract-level items are recorded for both respondents and non-respondents. Here, for tract-level items, a simple binary indicator of the presence or absence is used as a covariate. For example, if an enumerator observes corn in a tract, then the corn indicator for that tract is 1. In addition, state and land-use strata are used as covariates. State and land-use strata are common to both respondents and non-respondents, and are design variables used in the sample selection procedure. The land-use strata may be defined slightly differently from state to state. Here, strata were combined as necessary to form a land-use strata variable that takes one of five values: greater than or equal to 50% cultivated, 15 to 50% cultivated, agricultural urban/commercial, less than 15% agricultural or non-agricultural. The final logistic regression model can be expressed as follows. For a given tract,

$$Z_i \sim \text{Bernoulli}(\phi_i)$$
$$\text{logit}(\phi_i) = X_i \beta$$

where ϕ_i is the probability that a JAS response is obtained for tract i, Z_i is 1 if a response was obtained from the i^{th} tract and is 0 otherwise, X_i is the vector of covariates for the i^{th} tract and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients.

Based on the logistic regression model, ϕ is estimated for each respondent and incorporated in the non-response model. That is, the estimated non-response adjusted estimate is given by

$$\hat{T}_{NR} = \sum_{i \in R} \pi_i^{-1} \hat{\phi}_i^{-1} y_i t_i,$$

where $\hat{\phi}_i$is the estimated response probability for tract i.

3.2.3 Variance Estimation

The additional uncertainty due to non-response must be accounted for when estimating the variance of the estimator. Moreover, the error in estimating the response propensity must be accounted for in the variance calculations. The methodology of Kim and Kim (2007) provides a framework for estimating the variance of design-based estimates adjusted for non-response.

Recall, the usual estimate of the number of farms is given by

$$T = \sum_{i \in R} \pi_i^{-1} y_i t_i.$$

The non-response adjusted estimate of the number of farms is given by,

$$T_{NR} = \sum_{i \in R} \pi_i^{-1} \phi_i^{-1} y_i t_i.$$

Finally, the estimated non-response adjusted estimate is given by,

$$\hat{T}_{NR} = \sum_{i \in R} \pi_i^{-1} \widehat{\phi_\iota}^{-1} y_i t_i.$$

where $\hat{\phi}_i$ is the estimated response probability for tract i. Kim and Kim (2007) show that under certain assumptions, the variance of this estimate is estimated using the following formula,

$$V\widehat{\left(\hat{T}_{NR}\right)} = \sum_{i \in R} \frac{1-\pi_i}{\pi_i^2} \hat{\phi}_i^{-1} (y_i t_i)^2 + \sum_{i \neq j \in R} \sum_{j \in R} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} \hat{\phi}_i^{-1} \hat{\phi}_j^{-1} \left(y_i t_i y_j t_j\right)$$
$$+ \sum_{i \in R} \pi_i^{-2} \frac{\left(1-\hat{\phi}_i\right)}{\hat{\phi}_i^2} \left(y_i t_i - \pi_i \hat{\phi}_i \hat{\mathbf{h}}_i^T \hat{\gamma}\right)^2.$$

where π_{ij} is the joint inclusion probability of tract i and tract j, $\hat{\mathbf{h}}_{\boldsymbol{i}} = \frac{\partial logit(\phi_i)}{\partial\beta}|_{(\beta=\hat{\beta})} = X_i'$ and $\hat{\gamma} = \left(\sum_{i\,\epsilon R}\left(1-\hat{\phi}_i\right)\hat{\mathbf{h}}_{\boldsymbol{i}}\hat{\mathbf{h}}_{\boldsymbol{i}}^{\boldsymbol{T}}\right)^{-1}\sum_{i\epsilon R}\pi_i^{-1}\,\hat{\phi}_i^{-1}\hat{\mathbf{h}}_i y_i t_i$

3.3 Adjusting for Both Non-response and Misclassification

The JAS farm numbers estimate can be adjusted for both non-response and misclassification when a follow-up to the JAS is conducted using census records. Let U denote the set of respondents to the JAS after the records are updated using the information obtained from matching to census records. That is, U contains records that were classified as non-farms in the JAS but were identified to be part of a farming operation in the census. Note: Records identified as non-respondents are not considered in the matching process; they are accounted for using the non-response weights.

The response probability for each tract in U is estimated under the modeling framework as described in the previous section. With $\hat{\phi}_i$ for all records in U, the non-response, misclassification adjusted estimate for the number of farms in the U.S. is given by

$$\hat{T}_{NR.M} = \sum_{i\in U}\pi_i^{-1}\,\hat{\phi}_i^{-1}y_i t_i.$$

Note: This estimate still potentially represents an undercount because it is possible that some of the JAS non-farm records that did not match to a census record could be farms. Recall here it is assumed that a tract identified as a farm in either the JAS or the census is a farm. If misclassification of JAS farms were considered, a different framework would need to be developed.

The variance of this estimator is (Kim and Kim, 2007),

$$V\left(\widehat{\hat{T}_{NR.M}}\right) = \sum_{i\epsilon U}\frac{1-\pi_i}{\pi_i^2}\hat{\phi}_i^{-1}(y_i t_i)^2 + \sum_{i\neq j\in U}\sum_{j\in U}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\hat{\phi}_i^{-1}\hat{\phi}_j^{-1}\left(y_i t_i y_j t_j\right)$$
$$+\sum_{i\in U}\pi_i^{-2}\frac{\left(1-\hat{\phi}_i\right)}{\hat{\phi}_i^2}\left(y_i t_i - \pi_i\hat{\phi}_i\hat{\mathbf{h}}_i{}'\hat{\gamma}\right)^2.$$

Further research is needed to develop the variance of a JAS estimator of farm numbers that accounts for both non-response and misclassification when misclassification is modeled as in Section 3.1.2.

4. RESULTS AND CONCLUSIONS

Recent research identified misclassification and estimation as two sources of error in the June Area Survey (JAS). Three methods have been developed in this report: (1) an adjustment for JAS misclassification when a follow-up is conducted and in the case for which a follow-up is not possible, when the effect of misclassification is modeled, (2) an adjustment for JAS non-response, and (3) an adjustment for both misclassification and non-response.

The adjustment for JAS for misclassification requires relevant follow-up data. When a follow-up is conducted, such as when census records are matched to JAS records, misclassification on the JAS can be adjusted for directly. If, as in non-census years, a follow-up is not conducted, the modeling framework described in Section 3.1 provides an approach to estimating misclassification. For this framework, the final JAS estimate of the number of farms consists of a design-based portion for farm tracts and a model-based portion for non-farm tracts. An estimator of the variance for the adjusted number of farms estimator for this approach has yet to be determined. The modeled and design-based terms of the estimator are correlated, and this correlation impacts the variance of the estimator. A bootstrap procedure could potentially provide an estimate of the variability associated with the model-based component and the covariance with the design-based components. This and alternative methodologies for estimating the variance of the farm numbers estimator merit further research.

For the second method, a framework for adjusting the JAS for non-response was developed by assuming that each tract has a certain probability of responding to the survey. The probabilities were estimated using logistic regression. The estimated probabilities were used to calculate farm number estimates with appropriate measures of uncertainty.

Because misclassification and non-response are both concerns for the JAS, a unified framework was developed to account for misclassification and non-response. The effect of misclassification is quantified based on a follow-up, and the probability of response is modeled and used to adjust for non-response. An estimator of the variance using the methods of Kim and Kim (2007) was presented. Combining non-response and misclassification where both components are modeled merits future research.

5. RECOMMENDATIONS

1. **Thoroughly evaluate current JAS imputation procedures and develop appropriate imputation methodology**. Currently, the quality of imputed values for estimated tracts cannot be determined. The quality of the imputed data for total farm acreage is likely related to the method of imputation. The quality could be better assessed if the information regarding the source or method of imputation was retained. *This recommendation is currently being addressed. An office use box has been added to the 2011 JAS survey instrument which will collect the source of the farm acreage item reported on the questionnaire. Upon completion of the 2011 data collection processes, the data will be analyzed and various imputation approaches should be tested as per this recommendation.*

2. **Develop non-response methodology that reflects a combination of a revised imputation methodology (noted in the first recommendation) and a rigorous non-**

response methodology for estimated tracts that have no quality information available for imputation.

3. **A final JAS survey indication should include adjustments for non-response, imputation, and misclassification.** In addition, future research is needed to develop a methodology that accounts for these three sources of error in the farm number indication and provides an appropriate measure of uncertainty associated with the final JAS indication.

6. REFERENCES

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, 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, 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.

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

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

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.

Kim, J.K. and J.J. Kim (2007). Non-response weighting adjustment using estimated response probability. The Canadian Journal of Statistics. Vol. 35, No. 4, 2007, Pages 501-514.

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.

Little, R.J.A. and S. Vartivarian (2003). On weighting the rates in non-response weights. Statistics in Medicine. 2003; 22:1589-1599.

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

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

Platek, R. and G.B. Gray (1983). Imputation methodology. In Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies, Madow WG, Olkin I, Rubin DB (eds). Academic Press: New York, 1983; 255294.

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.

APPENDIX A

In 2009, NASS conducted the Farm Numbers Research Project (FNRP). As a result, 595 tracts estimated in the June Area Survey were rescreened, yielding a dataset that contained both the estimated farm level information and the actual farm level information. The estimated June values were compared to the actual values obtained during FNRP.

For brevity, the results are summarized for three variables that play an important role in estimating the number of farms: the tract-to-farm ratio, the total land, and the edited value of sales. A scatter plot of the June tract-to-farm ratio versus the FNRP tract-to-farm ratio for 595 tracts indicates substantial discordance between the two (Figure 2). In addition, a scatter plot of the June total land versus the FNRP total land illustrates similar discordance (Figure 3). The results indicate the estimation procedure does not accurately estimate the two variables that are needed to calculate the number of farms.

Finally, the edited value of sales were compared (Table A). The number of off-diagonal elements indicates discordance between the FNRP and estimated June values. A large number of off-diagonal values confirm the inaccuracy of the estimation procedure. Due to the inability to determine the quality of estimated values, it is assumed that estimated tracts are non-respondents. The methodology in the report describes the consequences of this assumption and provides a framework for estimating the number of farms in the presence of non-response.

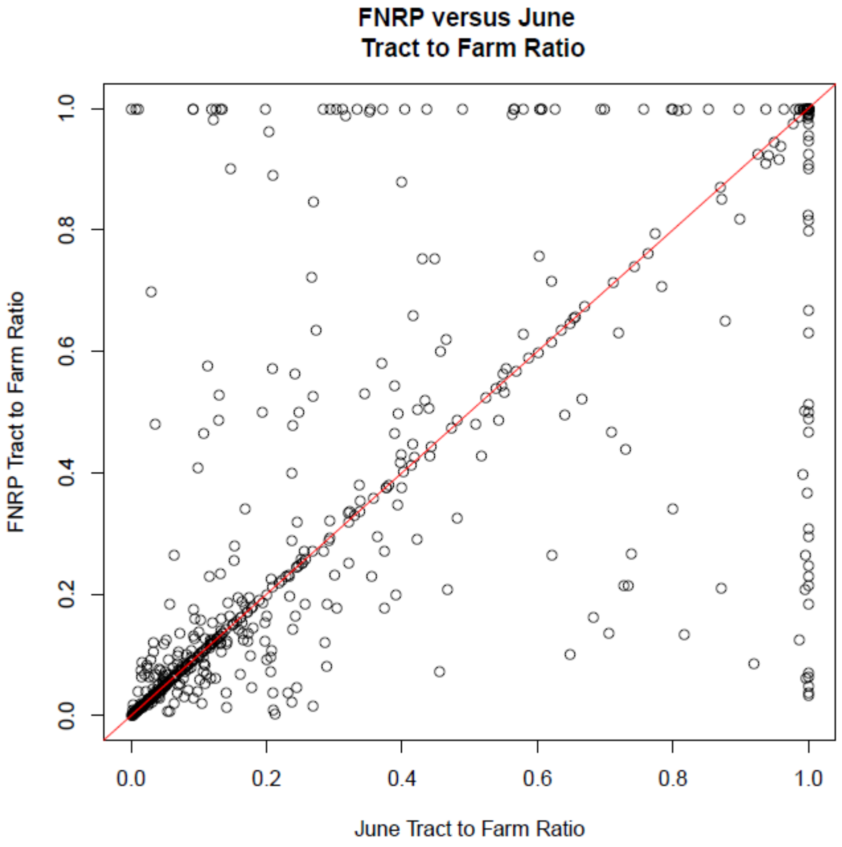

Figure 2

APPENDIX A

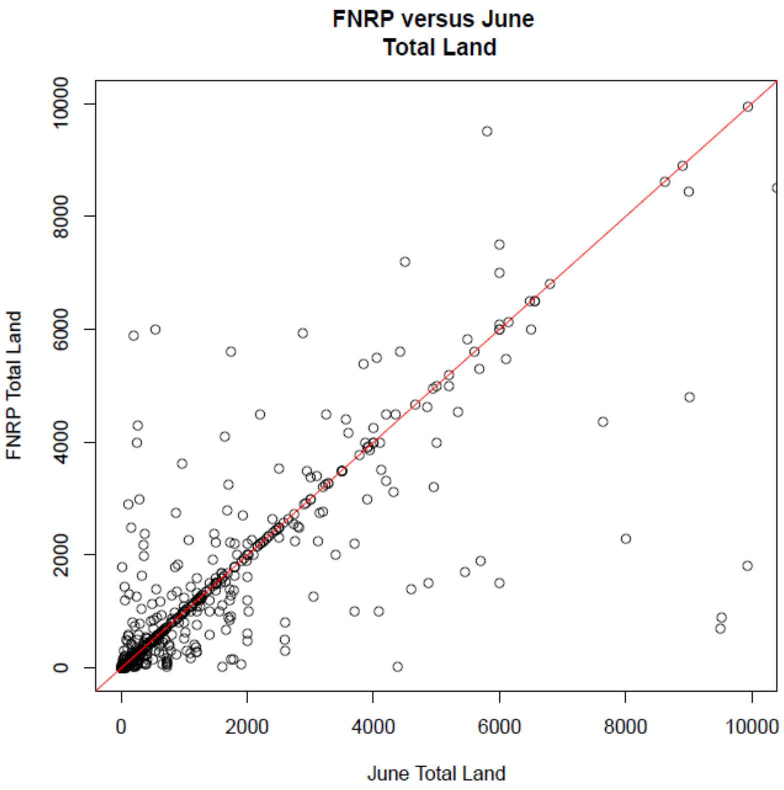

Figure 3

Table A. A Comparison of Sales Class Values for Matched FNRP and JAS Frame Records

JAS Sales Class	FNRP Sales Class												
	$1,000-	$2,500-	$5,000-	$10,000-	$25,000-	$50,000-	$100,000-	$250,000-	$500,000-	$1,000,000-	$2,500,000-		
	$2,499	$4,999	$9,999	$24,999	$49,999	$99,999	$249,999	$499,999	$999,999	$2,499,999	$4,999,999	$5M+	Total
$1,000-$2,499	37	15	10	5	5	2	1	2	0	0	0	0	77
$2,500-$4,999	6	11	5	5	1	0	1	1	0	0	0	0	30
$5,000-$9,999	5	8	13	5	2	2	1	2	0	1	0	0	39
$10,000-$24,999	1	3	7	23	5	2	1	1	1	1	0	0	45
$25,000-$49,999	1	0	1	11	9	3	3	2	3	0	2	0	35
$50,000-$99,999	1	1	1	3	12	28	8	2	3	2	0	0	61
$100,000-$249,999	3	5	2	5	4	11	49	8	5	1	0	1	94
$250,000-$499,999	1	1	1	2	1	6	17	32	12	5	1	0	79
$500,000-$999,999	0	0	0	0	0	0	6	12	37	9	1	2	67
$1,000,000-$2,499,999	0	0	1	1	1	2	3	2	3	25	0	0	38
$2,500,000-$4,999,999	0	0	0	0	0	1	1	1	1	2	8	1	15
$5M+	0	0	0	0	0	1	0	0	0	2	1	11	15
Total	55	44	41	60	40	58	91	65	65	48	13	15	595