AN ALTERNATIVE TO THE REINTERVIEW SURVEY USED TO MEASURE COVERAGE ERROR IN THE UNITED STATES CENSUS OF AGRICULTURE

Jay V. Johnson
National Agricultural Statistics Service

Keywords: Coverage error, misclassified, tract.

Abstract

In 1997, the National Agricultural Statistics Service (NASS) was granted the responsibility of conducting the Census of Agriculture. This census was previously conducted by the U.S. Department of Commerce. Since NASS conducts many ongoing sample surveys on the Census of Agriculture's target population, the opportunity arose to investigate alternative data sources to measure various aspects of the Census of Agriculture. This paper will examine the methodology and results of utilizing an existing area frame survey to replace the independent reinterview survey designed to measure classification error and duplication. The area frame survey successfully measured classification error and duplication comparable to results from the operational CES while utilizing less resources and reducing response burden.

Introduction

The goal of a census is to contact every member of a given population, but this is often not a realistic goal to attain. The coverage evaluation program for the United States Census of Agriculture is designed to measure the accuracy and completeness of farm counts and other selected characteristics at the U.S., regional and State levels. For more than 150 years, the U.S. Department of Commerce, Bureau of the Census, conducted the Census of Agriculture. However, the 1997 Appropriations Act transferred the responsibility from the Bureau of the Census to the United States Department of Agriculture (USDA), National Agricultural Statistics Service (NASS). Since NASS conducts many ongoing sample surveys on the Census of Agriculture's target population, the opportunity arose to investigate alternative data sources to measure various aspects of the Census of Agriculture.

Historically, two independent components have been used to measure and detect duplication and classification errors in the U.S. Census of Agriculture. In the past, the Department of Commerce contracted to obtain information from NASS's area frame survey of farm land and it's operators. A sample from this independent area frame survey is used to estimate the number of farms not on the mail list (NML). This is one form of undercount. Additionally, a reinterview of a sample of census respondents is used to measure those farms incorrectly classified as nonfarms (undercount), nonfarms incorrectly classified as farms (overcount), and duplication of farms (overcount). This reinterview survey is called the Classification Error Survey (CES). Finding a way to replace the reinterview survey with an alternative method could reduce respondent burden and potentially result in significant cost savings. It was also hoped that the use of the area frame survey would increase the reliability of estimating coverage error for individual commodities. This was believed to be a possibility because the area frame had a larger sample than the operational CES.

NASS conducts area frame surveys each June and December. The sample consists of stratified, randomly selected areas of land. Each parcel of selected land is carefully and completely enumerated for any presence of agriculture or agriculture producers. Data is collected not only on the land inside the selected land unit, but also for the entire operation of any farm or ranch operator who lives in the selected land unit. The information from the area sample is used heavily by NASS in its annual estimating procedures of livestock inventory, crop acreage and crop inventory in storage. As mentioned previously, the Department of Commerce utilized information from NASS's area frame surveys to measure the NML component of the overall coverage evaluation. However, the area frame surveys would have needed modifications to replace the CES. The Department of Commerce did not have control over the content of the information collected and therefore the area sample was not a possible alternative to the CES. NASS, having the ability to alter the content of the data collected from the area sample, is now in position to study the feasibility and reliability of using the area frame survey to replace the CES. This would create a uniform approach to obtain all components of the overall coverage error, reduce respondent burden and reduce/eliminate the costs associated with conducting the additional CES interview.

The motivation for using the area frame surveys to replace the CES were:

1. Reduce respondent burden by utilizing existing survey processes to replace the existing CES, a reinterview survey.

2. Save valuable resources by eliminating the need for a separate survey to estimate misclassification and duplication in the census of agriculture.

3. Increase the ability to estimate coverage error for individual commodities.

4. Provide for the inclusion of all farms in the misclassification/duplication processing. The current CES makes assumptions that large farms, multiunits, abnormals (Indian Reservations, university research farms, etc) and records tagged for special handling are accurately accounted for during census data collection. This allows these operations to be excluded from the CES and reduce respondent burden.

This paper compares the CES reinterview approach currently used and the use of the independent area frame survey which was tested. Implications of using an independent, existing survey to replace the reinterview, including associated costs, respondent burden issues and reductions in sampling error will be the primary focus.

Methodology

Questionnaire

Three questions (see Figure 1) were added to the area frame survey questionnaire. These questions were designed to pick up additional names and addresses an operation may operate under and to identify any landlords for the acreage in the sampled area of land. The questions were very similar

to those asked on the operational CES. These names were used to aid in identification of duplicated census farms.

All other data of interest collected on the current CES, e.g. commodity acreage, livestock inventories, total acres operated, gross value of sales, etc., were already asked on the area frame survey questionnaire.

Figure 1: Questions Added to June/Fall Area frame survey

Section A

X. During the past two years, has the operator received mail for this operation, at any address other than the one shown on the face page?

Xa. Please provide the other address. . . . . .

Name:
Address:
City: State: Zip:
Phone:

X. Excluding partners and landlords, were any other names associated with this operation in the past year? (For example, other business names, spouses names, etc.) [Enter Code]

Xa. Please provide the other name. . . . . .

Name:
Address:
City: State: Zip:
Phone:

Section D

X. Is any of the land inside the blue tract boundary rented from others? (Include land for which you paid cash rent, land used rent free, or land rented on shares.)

Name:

Address:

City: State: Zip:

Phone:

Data Collection

Additional data collection activities were minimal for this project. The majority of the data used were already collected in the area frame survey. The only additional data collection necessary was the three (3) questions added to the questionnaire (see previous sub-section). By utilizing existing data, it was hoped that respondent burden could be reduced and resources conserved.

Once the additional names were collected and merged with other names associated with the operation (i.e. operator and partners), they could be matched to the census mail list (CML).

Record Linkage

Automatch, the official record linkage system of NASS, was utilized to match all names from the area frame survey to the names on the CML. These names included farm operators, partners,

additional names collected for this study (e.g. landlords), and names of those who do not farm, but live within a sampled unit of land. Using an automated matching system results in a high level of accuracy and substantially minimizes the manual review process.

A large portion of this work was done during the operational NML processing. The procedures of the NML matches farm operators and partners to the CML. Those names from the area sample which cannot be found on the CML constitute the NML component of the coverage error. Those names which did match were operationally discarded, but were used by this study to determine potential misclassification or duplication.

Since the matching process performed for the operational NML was utilized, only the additional names which were collected (see previous sub-section) and the names of those who do not farm but live within a sampled unit of land needed to be processed through Automatch.

Upon a final determination for all potential matches, a database was constructed containing all matching records for each farm and nonfarm in the sampled unit of land. This database served as the mechanism used to determine misclassification or duplication on the CML.

Classification Resolution

The database containing all matching records for each operation was resolved. Resolution for this project involved reviewing each operation and it's matching names and each non-farm residence with any matching CML names. The fact that an area frame survey gives a specific piece of land associated with each operation (or non-farm residence) allowed the resolution to focus on a known acreage. This specific piece of land will be referred to as the 'tract'. This study based decisions on the land inside the tract. Decisions resulting from the resolution were based on the flow diagram shown in Figure 2.

Figure 2: Resolution Decision Diagram

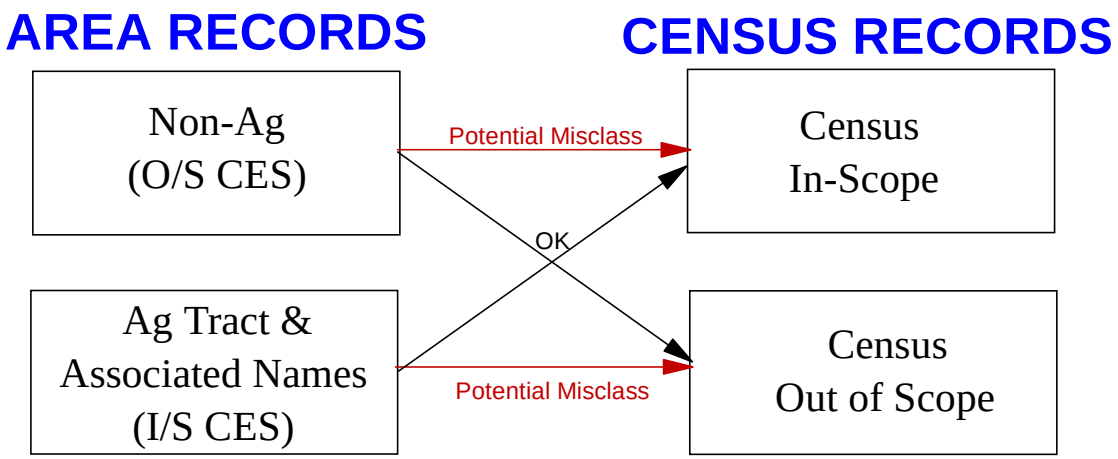


Figure 2 shows that if the area frame reported a non-farm residence that matched an in-scope census record(s) (i.e. farm) then potential misclassification(s) was identified and the 'truth' of whether these were actually the same place and if the land inside the tract was included on the Census was

determined through other readily available sources. If the area frame reported a non-ag residence which matched an out-of-scope census record (i.e. non-farm) then the determination was made that both agreed in status and no further investigation was necessary. A similar review was conducted for agricultural tracts.

Duplication was detected when an agricultural tract (i.e. farm) was accounted for on multiple in-scope census records (i.e. farm) which were deemed to be the same operation. Again, the match was determined on whether the area frame survey and census were both accounting for the acreage recorded inside the tract.

The possibility of multiple misclassifications or duplications existed for each area record, therefore an outcome table (see Figure 3) was used to determine final results.

The results were then coded as indicator variables on any operations which were identified as misclassified overcount, misclassified undercount or duplicated.

Figure 3: Resolution Outcome Table

Area Frame Status	# of Times Tract Acres Rptd on In-Scope Census Records	Result
Non-Ag	0	OK
Non-Ag	$i > 0$	i Misclassified Overcount(s)
Ag	0	1 Misclassified Undercount
Ag	1	OK
Ag	$i > 0$	(i-1) Duplication(s)

Summary Procedures

After the entire database was reviewed and determinations made about misclassification and duplication of each census record, the final summary and analysis was performed. This study compared the operational CES against the use of the area frame surveys to estimate classification errors for the number of farms, land in farms, number of farms by size category, number of farms by value of sales, number of farms by type of organizational structure, acreage and production levels of many different commodities and head of livestock items. The primary focus of this presentation will be on the number of farms and land in farms.

Results and Discussion

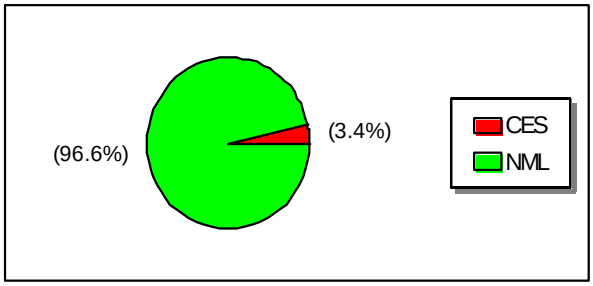
Complete enumeration of agriculture operations satisfying the farm definition of $1,000 or more in agricultural sales is complicated by the variety of arrangements under which farms are operated, the multiplicity of names used for an operation, the number of operations in which an operator participates and the difficulty in classifying those operations just around the $1,000 sales range.

The coverage evaluation program is designed to measure four components of error in the census farm counts. The Not on the Mail List (NML) component of coverage error estimates the number of farm operations not on the census mail list - a form of undercount. The Classification Error Survey (CES) measures the other three components, specifically:

- farms incorrectly classified as nonfarms (undercount)
- nonfarms incorrectly classified as farms (overcount)
- farms duplicated in the census (overcount).

The NML component is by far the largest component of coverage error. In 1997 it accounted for over 96 percent of the total coverage error in farm count for the 11 Western States which comprise this study.

Figure 4: Component Breakdown of Total Coverage Error
Farm Count - 11 State Western Region



Since the CES accounts for such a small portion of the total coverage error, but is conducted as a stand alone survey, it was identified as a potential area for using alternative sources of data. Additionally, NASS envisioned being able to report coverage error for major commodities within each state. Traditionally, the CES only estimated coverage error for number of farms and land in farms. All of these factors led to the motivation to investigate the feasibility of studying an alternative data source to replace the CES.

The NASS area frame was already being used to measure the NML component of Net Coverage Error and can be slightly adapted to pick up the missing information necessary to make it a useable tool for measuring misclassification and duplication, thus replacing the CES. Previously, the Bureau of the Census did not have the authority or the resources to alter NASS' data collection activities and use the area frame for all components of Net Coverage Error.

This study was conducted in the Western Region, constituting 11 States (Arizona, California, Colorado, Idaho, Montana, New Mexico, Nevada, Oregon, Utah, Washington, and Wyoming). This region was chosen because it historically has one of the highest classification/duplication error rates and the states which compose the region are diverse in the agriculture commodities present.

The 1997 Classification Error Survey used a detailed questionnaire designed to gather information to help determine the true farm status (identify potential misclassification/duplication) for the Census of Agriculture. The questionnaire asked for demographic information, landlord names, additional

names the operation may operate by (other than what was on the mailing label) and commodity information. The area frame survey questionnaire collects commodity information and for this study collected landlord and additional names. Demographic information was not added to the area frame survey questionnaire for this study.

While the 1997 CES questionnaire made great strides towards efficiencies, it was still eight pages long and was being asked to respondents who had just completed the census report form a few weeks earlier. Many field staff and interviewers made comments about the dissatisfaction and burden that this created for those contacted. Response burden issues are always a concern for NASS since we conduct many surveys each year on the same population.

As mentioned earlier, only three (3) additional questions were asked during both the 1997 June Area Survey and the 1997 Fall Area Survey. These questions resulted in about one additional name or landlord being collected on every third operation interviewed (see Table 1).

Table 1: Additional Names Collected During 1997 June & Fall Area Surveys

State	Operations Interviewed	Additional Names	%
AZ	365	98	27%
CA	2490	791	32%
CO	801	170	21%
ID	885	370	42%
MT	1442	985	68%
NV	53	9	17%
NM	702	221	31%
OR	687	172	25%
UT	477	158	33%
WA	713	242	34%
WY	299	92	31%
Total	**8914**	**3308**	**37%**

Table 1 shows that all the States (except Montana) collected a proportionately similar number of additional names. Montana operators had a large number of lease agreements with the State of Montana.

The cost associated with the data collection and processing of the additional names are considered to be minimal. Very few staff hours were used to update the questionnaire and process the names after data collection. The data collection also resulted in minimal additional cost. Since the June and Fall Area Frame Surveys are existing surveys, the only additional cost was the time associated

with asking the three (3) additional questions and recording the answers. While this cost is certainly not zero, it is minimal in the overall cost of conducting these surveys.

In contrast, the operational CES involved considerable staff time for the design of the survey instrument, both paper and CATI, the testing of the instruments, printing of the paper questionnaires, presurvey activities, including reviewing and coordinating the sample, preparing the mailing, training the enumerator staff and finally processing the completed questionnaires. Considerably more time was spent on the operational CES preparation than was necessary for the additional questions added to the area frame study. The data collection costs associated with the operational CES included training, telephone interviewing and some field data collection. These data collection costs are documented at over $300,000.

The potential gains of using an existing survey instrument to replace the current reinterview methodology are not only seen in cost reduction, but also the ability to reduce respondent burden. The benefits of eliminating a survey contact, especially one which so closely follows the census contact, are unmeasurable.

Recall that during the classification resolution phase, decisions of misclassification and duplication were based solely on the land inside the tract boundaries. This restriction resulted in the use of a weighted estimator that prorates whole farm values. The proration technique utilized by NASS for weighted estimators of area frame surveys is the land inside the tract boundaries for an operation divided by the reported whole farm acreage. Therefore, the resolution actually only determined that a piece of farm was misclassified or duplicated.

The weighted estimator has the form:

$$Y_{state} = \sum_{h \in A_L} Y_h$$

where A_L is the set of all land-use strata in the state. Each Y_h is calculated as:

$$Y_h = \sum_{j \in B_h} \sum_{k \in G_{hj}} e_{hjk} \sum_{m \in T_{hjk}} w_{hjkm}\, a_{hjkm}\, y_{hjkm}$$

where h is the land-use stratum,

- B_h is the set of all substrata in h,
- G_{hj} is the set of all segments in substratum j of land-use stratum h,
- T_{hjk} is the set of all tracts in segment k of substratum j of land-use stratum h,
- e_{hjk} is the expansion factor for all tracts in segment T_{hjk},
- w_{hjkm} is an indicator variable, i.e. it takes on either the value 0 or 1 depending on whether the tract was misclassified/duplicated.
- a_{hjkm} is the weight used to prorate Y_{hjkm} (presently the tract acres divided by the farm acres),
- y_{hjkm} is the entire farm value associated with the tract.

The corresponding variance estimator is:

$$V_{state} = \sum_{h \in A_L} V_h$$

Each V_h is calculated as:

$$V_h = \sum_{j \in B_h} [(n_{hj}/[n_{hj}-1]) * \{ \sum_{k \in G_{hj}} (Y_{hjk.}^{e})^2 - (\sum_{k \in G_{hj}} Y_{hjk.}^{e})^2 / n_{hj}\}].$$

Each subscript is defined above. For further information regarding this weighted estimator and the associated variance see Mathematical Formulae for the 1989 Survey Processing System (SPS) Summary listed in the References section.

The accuracy of the operational CES is unknown. However, it is known that the CES component represents a very small portion of total coverage error. Table 2 shows the results of the research study when compared to the operational CES. The table shows the aggregate data for the 11 State Western Region. The precision of both techniques is very similar. However, the estimator levels vary. There appears to be a positive bias in the research study. The net error for each state, except one, was positive. This indicates that the area frame may have done a better job in detecting undercount than it did in finding overcount. This phenomenon is due to the fact that many of the residential (non-farm) tracts had insufficient name and addresses. This can be rectified through targeted training of the enumerators who conduct the survey. Correcting this bias will result in a lower area frame estimator which is closer to the operational CES results. Recall that the accuracy of the operational CES is unknown and that the CES component constitutes only about 3 percent of the total coverage error (see Figure 4). After correcting the bias in the area frame estimator, it will also have a small effect on overall coverage error. Therefore, the area frame approach should be seriously considered as a replacement to the current procedure. However, the area frame approach did not significantly improve on the ability to estimate coverage error for individual commodities. The occurrences of misclassification or duplication are rare and resulted in very thin data at the individual commodity level for both the area frame study and the operational CES. The lack of observations at the commodity level resulted in estimates which varied widely.

Conclusion

The results of this study have shown that using the area frame survey to replace the Classification Error Survey will achieve three of the four goals of the study. First, use of the area frame survey would reduce respondent burden by using an existing survey instrument to replace the reinterview approach. Second, it would save valuable resources by eliminating the need to plan, develop, coordinate and conduct a separate survey. These resources can be measured in both staff time and monetary savings. Third, the study successfully allowed all farm operations inclusion into the population. This removes any assumptions which were made to reduce respondent burden.

This study was unsuccessful in achieving the fourth goal of estimating coverage error for individual commodities. The fact that misclassification and duplication remain a rare finding causes the precision and accuracy of indications for specific commodities to become very unstable.

Table 2: Comparison of Results - 11 State Total 1/

Item of Interest	Operational CES Estimate	Standard Error	Area Frame Estimate 2/	Standard Error
# of Farms	**1,628** 3/	**1,441**	**22,121**	**2,730**
Land in Farms	**(2,514,100)**	**1,655,713**	**7,335,906**	**2,123,186**
Farms by Size:				
Less than 10 acres	**82**	**727**	**9,480**	**1,926**
10 to 49 acres	**1,579**	**777**	**7,635**	**1,318**
50 to 179 acres	**891**	**807**	**2,131**	**875**
180 acres or more	**(923)**	**683**	**2,878**	**738**
Farms by Value of Sales:				
Less than $2,500	**307**	**1,201**	**12,345**	**1,993**
$2,500 to $9,999	**826**	**675**	**8,350**	**1,249**
$10,000 or more	**496**	**857**	**1,428**	**544**
Farms by Type of Organization:				
Individual or Family	**2,336**	**1,342**	**25,521**	**2,469**
Partners, Corp., or other	**(707)**	**525**	**(3,399)**	**1,083**

1/ Total may not add due to rounding.
2/ A correctable positive bias exists in the area frame estimators.
3/ The CES constitutes only 3 percent of the total coverage error for # of farms. Other items show similar results.

While one goal was unattainable with the current area sample, the study proved the use of the area frame survey to be as reliable as the stand alone CES, while being more efficient and less burdensome. Therefore, the use of the area frame to measure both the NML and CES components of the coverage error is a realistic alternative to the current procedure.

The results from this study are specific to one survey, but are hoped to foster discussion and studies by other federal agencies in the use of existing data sources. Response burden and monetary constraints are key issues that every data collection agency faces on a regular basis. Maximizing the use of existing data sources is one way to ease these concerns.

References

1997 Census of Agriculture, Classification Error Survey Interviewer's Manual (1998), National Agricultural Statistics Service, USDA.

1997 Census of Agriculture, United States Summary and State Data, Volume 2, Geographic Area Series, Part 51, National Agricultural Statistics Service, USDA.

BROADBENT, K. (1996). Record Linkage III: Experience Using AUTOMATCH in a State Office Setting. Survey Technology Branch Research Report, STB-96-02, National Agricultural Statistics Service, USDA.

KOTT, P.S. (1990). Mathematical Formulae for the 1989 Survey Processing System (SPS) Summary. NASS Staff Report, SRB-90-08, National Agricultural Statistics Service, USDA.

LEWIS, P. and IMEL, J. (1994). NML Estimation Processing Documentation, Bureau of the Census, US Department of Commerce.

MATTHEWS, R.V. (1988). Screening Residential Tracts for Agricultural Activity. NASS Staff Report, SSB-88-05, National Agricultural Statistics Service, USDA.