United States Department of Agriculture

National Agricultural Statistics Service

Research and
Development Division
Washington DC 20250

RDD Research Report
RDD-07-03

April 2007

Results from the 2002 Classification Error Study

Denise A. Abreu

EXECUTIVE SUMMARY

In 1997, a real-time study known as the Classification Error Study (CES) was conducted to evaluate the June Area Survey (JAS) and Fall Area Survey (FAS) as a replacement for the Classification Error Survey (re-interview approach). The 1997 CES (area based approach) was conducted for the 11 Western States comprising the West Census Region. The study intended to measure those farms incorrectly classified as nonfarms, nonfarms incorrectly classified as farms, and duplication of farms. Since the Area Surveys were already being used to measure the Not-on-Mail-List (NML) component of coverage error, using them to account for classification errors would eliminate the need to recontact farm operators. Additionally, the approach would focus on specific tracts of land instead of an entire operation either being misclassified or duplicated.

The study's results compared favorably to the survey estimates. Recommendations were to replace the Classification Error Survey (re-interview approach) with the Classification Error Study (area frame approach) for the 2002 Census of Agriculture.

This report outlines the results of the 2002 CES (area frame based approach). The 2002 CES was implemented operationally for the 48 conterminous states. Its main objective was to determine the relative size and likelihood of classification errors to warrant future CES studies. The reported commodity inventories were not used during calibration of the census aggregates.

The results of the 2002 study indicated that although the CES comprises a small portion of the overall coverage number, it needs to be addressed further. The 2002 CES found an overall misclassification overcount of farms, while in the 1997 CES there was a net misclassification undercount of farms. Due to this inconsistency from census to census, the study should be conducted again.

RECOMMENDATIONS

The objective of the 2002 CES was to determine the relative size and likelihood of classification errors to warrant future CES studies. The results show that although the CES comprises a small portion of the overall coverage number, it needs to be addressed further. In 1997, a net misclassification undercount was found, while in 2002 we note an overall misclassification overcount. The results are not consistent from census to census; this is one reason to conduct this study again. From the results we see that we disproportionately missed farms for all types of errors in the various characteristics: size of farm, type of farms, and total value of production (TVP). In all these cases, records with different characteristics were disproportionately undercounted, overcounted or duplicated. This is another reason to conduct this study again. As a result of these findings, the CES should be repeated again in 2007.

If this recommendation is adopted, the following items should be taken into consideration:

1. The primary focus of the CES should be on addressing discrepancies between the sources: census and area. The reviewers noted there were errors observed in some instances in both census and area reports.

2. More time and effort should be devoted to addressing non-ag tracts in the area frame. A large number of the records classified as overcount had notes from the reviewers indicating that the area non-ag tract was actually erroneously classified. For example, the area survey missed hay, horses, CRP acres etc. However, since they were instructed to hold the area as 'truth' they had to code the record as overcount, when in fact it wasn't. If we are to hold a source as 'truth' these types of comments should not be there.

3. Conduct an overhaul of the IDAS instrument used during the second review.
    a. Design sections to look for discrepancies between the two sources (census and area). Have reviewers comment as to which source showed an error in reporting.
    b. A sort function or "go to" option should be included in the application. On average it took about 5 minutes to resolve a link group during this review. Most of the concerns expressed by the FOs were that they had to scroll down to the next link group upon completion of the review of a link group. A sort function that will move the reviewed link group to the bottom and return the main screen to show the next link group(s) requiring review will expedite this process.

4. To adequately evaluate the usefulness of the names and addresses collected, a link group identification number should be designed to attach the matched census record to the correct area record it matched initially. This should be done as part of the record linkage review. This will allow for later evaluation as to which area record (target, additional name, additional address, or landlord name) the census record matched to. This should be an analysis of ALL the information collected and not only of those in the scope of the CES.

5. Clearly defined procedures to apply for misclassification rules should be written. There were many comments provided by the FO reviewers noting non-ag tracts matched to census records reporting CRP acres. Perhaps, this should not have been coded as an

error. One suggestion to address this issue would be to include a question on the area screener questionnaire asking for CRP acres in the tract. Also, if the area survey considers certain records as ag or non-ag perhaps similar rules should be applied to the census records. In the census, this is the case for those records that were computer or analyst out-of-scoped. If these records had not changed status, they would have correctly matched the ag tract to which they were linked.

Results from the 2002 Classification Error Study

Denise A. Abreu[1/]

Abstract

The 1997 Coverage Evaluation program was designed to measure four components of error in the census farm counts: 1) undercount due to farms not-on-the-mail list (NML); 2) overcount due to farms duplicated or enumerated more than once; 3) undercount due to farms incorrectly classified as nonfarms; and, 4) overcount due to nonfarms incorrectly classified as farms. The NML component is by far the largest contributor to coverage error. It utilizes the NASS area frame to assess the number of farms not on the list. The other three components account for misclassification error.

In 1997, the Classification Error Study (CES) was conducted to evaluate the feasibility of using the NASS area frame to measure misclassification error and replace the Classification Error Survey (re-interview approach) in place at the time. The results of the 1997 CES, conducted on the 11 Western states, compared favorably to the re-interview approach estimates. Recommendations were to replace the classification error survey (re-interview approach) with the Classification Error Study (area frame approach) for the 2002 Census of Agriculture.

This report outlines the results of the 2002 Classification Error Study (area frame based approach). The 2002 CES was implemented operationally for the 48 conterminous states. Its main objective was to determine the relative size and likelihood of classification errors to warrant future CES studies.

Key Words: Classification Error; Coverage Error; Probabilistic Record Linkage; Classification Resolution

[1/] Denise A Abreu is a Mathematical Statistician with the National Agricultural Statistics Service – Research & Development Division, located at 3251 Old Lee Highway, Room 305, Fairfax, VA 22030. A special thanks to Dale Atkinson and Bill Iwig for their guidance in the development of this report, Jay Johnson for his tremendous assistance throughout the entire project, Mike Hogye for his programming assistance with variance estimation formulas, and Phil Kott for providing the standard error methodology. Also thanks to Lindsay Drunasky for making available various census estimates, Kara Daniel and Tom Pordugal for their record linkage processing, Mark Apodaca for this superb assistance in developing the phase II resolution application, and the Maryland FO for their help in testing the review instrument.

1. INTRODUCTION

The 1997 Coverage Evaluation program was designed to measure four components of error in the census farm counts: 1) undercount due to farms not-on-the-mail list (NML); 2) overcount due to farms duplicated or enumerated more than once; 3) undercount due to farms incorrectly classified as nonfarms; and, 4) overcount due to nonfarms incorrectly classified as farms. The first component, NML, is by far the largest contributor to coverage error. It utilizes the NASS area frame to assess the number of farms not on the list. Although the duplication component does not occur as often, it tends to involve large operations that have a greater impact on acreage and total value of production estimates. Misclassification of either farms or nonfarms can arise in either reporting or processing the information reported. Historically, the assessment of duplication and misclassification has been evaluated through a re-interview of sampled census respondents, known as the Classification Error Survey.

In 1997, a real-time study known as the Classification Error Study (CES) was conducted to evaluate the feasibility of using the June Area Survey (JAS) and Fall Area Survey (FAS) as replacements for the Classification Error Survey (re-interview approach). The 1997 CES (area based approach) was conducted for the 11 Western States comprising the West Census Region. The study intended to measure those farms incorrectly classified as nonfarms (undercount), nonfarms incorrectly classified as farms (overcount), and duplication of farms (overcount).

Since the area surveys were already being used to measure the NML component of coverage error, using them to account for classification errors would eliminate the need to recontact farm operators. Additionally, the approach would focus on specific tracts of land instead of an entire operation being misclassified or duplicated, thus reducing nonsampling errors. To facilitate this approach, additional information on the farm operations was collected on both 1997 June and Fall Area questionnaires. The study's results compared favorably to the survey estimates. Recommendations were to replace the Classification Error Survey (re-interview approach) with the Classification Error Study (area frame approach) for the 2002 Census of Agriculture.

Due to budgetary constraints, the FAS was discontinued in 2000. For the 2002 Census of Agriculture, a combination of the 2002 JAS and a supplemental area sample, referred to as the 2002 Agricultural Coverage Evaluation Survey (ACES) segments were used to measure coverage error (i.e. NML, classification error and duplication) in the 48 contiguous states. The ACES tract questionnaire was a shortened version of the JAS tract questionnaire and contained only those sections needed to collect data to measure coverage error.

This report outlines the results of the 2002 CES (area frame based approach). Further mentions of classification error will refer to the area frame based approach only. The 2002 CES was implemented operationally for the 48 conterminous states. Its main objective was to determine the relative size and likelihood of classification errors to warrant future CES studies. The reported commodity inventories were not used during calibration of the census aggregates.

The underlying basis for the analysis was to hold the area frame survey as 'truth'. Significant reliability is expected from face-to-face interviewing and personal observations by the interviewers. Also, the fact that the area frame survey is based on a

specific piece of land associated with each tract allowed misclassification of farm status to focus on a known acreage. However, even though the interviewer physically goes out to the field and accounts for every piece of land in the sampled area, the data are not immune to errors. Area frame data are sometimes coupled with incomplete name and address information or lack thereof (respondent refusals). Similarly, inadequate name and address data can be also a problem with the information collected on the census questionnaires. The name matching approach used for the study and the set of rules applied to records to determine which type of misclassification has occurred depends solely on which source is assumed to be correct. Ultimately, if we assume the area to be correct we arrive at one conclusion; and if we assume the census to be correct we arrive at a different one.

2. AREA FRAME DATA COLLECTION

For the evaluation, additional name, address, and telephone information were collected on both the JAS and ACES through the addition of the following three questions to the area survey instruments:

i. 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?

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

iii. 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).

These were the same questions asked during the re-interview version of the CES conducted along with the 1997 Census of Agriculture in the 11 Western States. This information was primarily used to identify duplication on the census mail list. However, it proved valuable in linking area records to census records that would not have been linked otherwise. The following section outlines in more detail the matching process. Determining whether an area tract was misclassified or duplicated was the final step in the process.

The data collection effort yielded 18,967 additional names from both area surveys. There were 15,467 landlord names, 1,678 additional names (i.e., spouse, partners) and 1,822 additional addresses for the operation. See Attachment A for counts by state. These additional names were used to help identify misclassification of farm operations in the 2002 CML. Nevada respondents provided no new names, whereas Texas operators reported over 2,000 additional names and/or addresses. The additional name and address information collected was matched against the 2002 Census Mail List (CML) in a two phase review process.

3. CES FIELD OFFICE REVIEWS – PROBABILISTIC RECORD LINKAGE & FARM CLASSIFICATION REVIEW

The first step of the CES analysis involved the use of Probabilistic Record Linkage (PRL) to match the additional information collected on the area surveys to the names and addresses on the 2002 CML. PRL is a technique used to identify records that are believed to correspond to a CML record. Records are brought together into groups which possibly represent the same operation. These groups are called link groups. Each link group is classified into three distinct

types: matches, possible matches and non-matches. The non-matches were part of the NML component and are not in the scope of this report. The CES is primarily concerned with the definite matches and possible matches.

The PRL process involved matching the target names and additional information collected from the area surveys against the names on the 2002 CML. The JAS survey had 110,072 names and the ACES had 22,341. It was necessary to match all the operator names along with all the additional information collected for them to the names on the CML. In doing so, it was guaranteed that a census record that would not have matched to the target operator could match to an additional partner or spouse name collected on the area questionnaire. For example, an area respondent lists two partners on the questionnaire, while the census questionnaire is completed by one of the partners. The census respondent is not the one listed as the target operator on the area side. If we match the target operator name to the name on the census questionnaire (one-to-one matching) they would not have matched. However, allowing the program to also match to the additional names reported on the area questionnaire would have picked up the partner name as a valid match. If the matching process had involved solely matching to the operator names, matches from partners, landlord, spouses, etc. not reported on the census could be missed. This would lead to improper classification of records as overcount, undercount or duplication. It may even lead to considering two records as non-matches when they are in-fact a valid match.

The matching resulted in 17,567 definite matches and 35,434 possible matches – 53,001 total matches. Each field office (FO) reviewed the possible matches in their state and determined the match vs. non-match status. See Attachment B for FO workloads. On average it took an FO staff member about 1 to 1 ½ minutes to review a link group. The definite matches consisted of one-to-one (one census record matching one area record) matches where the records coincided in all the fields (i.e., name, address, city, state, etc) involved in the matching process. Additionally, area and census records which had a valid area-to-list link were also considered definite matches.

Upon completion of the PRL review, 8,444 link groups were considered non-matches (dropped from the scope of the study) by the FO reviewer and 44,557 link groups remained for further review, spawning Phase II of the CES review process (Farm Resolution Classification). HQ reviewed 30,893 of these link groups, and the remaining 13,664 were sent back to the FOs for additional review. See Attachment C for workloads by state. These link groups were broken into two groups. The first group consisted of area records matching two or more census records. Reviewing these records helped identify duplication on the CML. The second group included records (area and census) where the reported acreage differed by more than 25%. A SAS/AF application similar to an IDAS survey instrument was developed to handle the review of these cases in Phase II.

4. CLASSIFICATION ERROR RESOLUTION AND METHODOLOGY

Once all the reviews were completed, area tract status was matched against the census current status code to determine the type of classification error. The underlying philosophy for the CES project was to hold the area frame survey as 'truth'. The fact that the area frame survey identifies a specific piece of land associated with each tract allowed the resolution to focus on a known

acreage. Classification decisions for the study were based on the land inside the tract. Any deviations from the ag/non-ag tract status were identified as classification error.

The possibility of multiple misclassifications or duplications existed for each area record (see Table 4.1). Therefore an outcome table was used to determine final results.

Table 4.1 Classification Decision Outcome Table

AREA FRAME STATUS	# OF TIMES (i) TRACT ACRES RPTD ON IN-SCOPE CENSUS RECORDS	RESULT
Non-Ag	i=0	OK
Non-Ag	i > 0	i misclassified overcount(s)
Ag	i=0	i misclassified undercount(s)
Ag	i=1	OK
Ag	i > 0	(i-1) Duplications

Misclassification overcount occurred when a non-ag tract matched a census in-scope record or a census nonrespondent. A match to a census non-respondent was deemed a valid match, since a non-response adjustment was used to account for these records. When a non-ag tract matched a census out-of-scope, it was determined that both agreed in status and no further investigation was necessary. Misclassification undercount occurred when an ag tract matched a census out-of-scope record. If an ag tract matched exactly one census in-scope record it was determined that both records agreed in status and no further review was necessary. List duplication was accounted for whenever an ag tract matched two or more census in-scope records.

Once the resolution was completed for a tract, the results were coded as indicator variables on all operations identified as misclassified overcount, misclassified undercount or duplication. Tract to farm ratios were calculated for each ag and non-ag tract. For estimated ag tracts, the 2002 median tract to farm ratios at the stratum or land use level, as provided by the Statistical Methods Branch, were used to calculate the total land acres. Since non-ag tracts only have tract acres available, the total farm acres as reported on the census in-scope record were used. Whenever the tract-to-farm ratio for a non-ag tract was greater than 1, it was rounded to 1. When a non-ag tract matched a census non-respondent, 2002 classified control data were used for the non-respondent record to calculate a tract to farm ratio. Additionally, counting all non-respondents as overcount misclassification would have inflated the results, since it is believed that approximately half of all non-respondents are likely to be nonfarms. To address this issue, a nonresponse adjustment as calculated for census estimates was applied to each nonrespondent's tract to farm ratio.

The traditional NASS weighted estimator was used to summarize misclassification and duplication.

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 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 the value 1 or 0, 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:

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

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 Reference section.

5. RESULTS

The 2002 CES looked at three types of errors – overcount, duplication and undercount. The study intended to measure nonfarms incorrectly classified as farms (overcount), duplication of farms, and farms incorrectly classified as nonfarms (undercount). There were 4,526 tracts in the scope of the CES. Of these, 1,645 were undercounted tracts, 2,311 were overcounted and 570 duplicated. See Attachment D for counts by state. The US level net misclassification error is calculated as:

Overcounted farms + Duplicated farms - Undercounted farms

Table 5.1 Expanded Number of Farms for Each Type of Error & US Level Net Total

Type of Error	Number of Tracts	Expanded Number of Farms	Standard Error
Overcount	2,311	141,069	4,975
Duplication	570	16,760	1,191
Undercount	1,645	106,484	4,147
US Net CES Error	4,526	51,345	6,456

Table 5.1 provides the expanded number of farms for each type of error and the overall net error for the US. The results show that there was a significant net misclassification overcount of 51,345 farms at the US level.

The CV or level of precision associated with this indication was 12.5%.

Table 5.2 Total Number Farms Misclassified for Each Type of Error by TVP

Type of Error	TVP							
	Less than $1,000	$1,000-$2,499	$2,500-$9,999	$10,000-$24,499	$25,000-$49,999	$50,000-$99,999	$100,000+	Total
Overcount	31,412	39,815	41,987	14,160	4,965	2,703	6,028	141,070
Duplication	384	641	2,655	2,458	2,239	2,106	6,275	16,758
Undercount	27,955	24,168	23,935	10,313	6,873	5,087	8,153	106,484
Net Totals	**3,841**	**16,288**	**20,707**	**6,305**	**331**	**(278)**	**4,150**	**51,344**

- Data may not add to the total due to rounding.

Table 5.2 shows the distribution of the total number misclassified of farms by the type of misclassification error and TVP. The results show that there is an overcount of farms for most TVP ranges, with the exception of an undercount of operations with TVP between $50,000 and $99,999. One reason to explain this is that the 2002 Census picked up a large number of small farms that were point farms at the time of the Area Survey interview, but that became or qualified to be farm operations at the time of the census. The above table shows that the majority of the overcounted farms are those with TVP between $1 and $9,999. By a small margin these farms could have been in-scope or out-of-scope at the time of the 2002 Census. This together with the fact that close to 20% of the farms in the undercount category were computer or analyst out-of-scope helps justify this finding.

Of the 48 states participating in this study, 33 showed non-significant misclassification net error whether overcount, duplication or undercount had occurred. Alabama, Arkansas, California, Georgia, Illinois, Indiana, Iowa, Kansas, Maryland, Minnesota, Mississippi, New York, Pennsylvania, Washington and Wisconsin showed statistically significant overcounts of farms.

5.1 Characteristics of Overcounted Farms

Overcounted tracts consisted of a non-agricultural tract (including records with less than 1,000 points or $1,000 in sales) matching a census in-scope or a census nonrespondent. There were 2,311 tracts overcounted. The expanded number of overcounted farms was 141,069. The only data available for these farms were those collected on the in-scope census report forms. Based on these data, 78.5% of the total overcount estimate came from matching to a census in-scope record. The remaining 21.5% came from non-ag tracts matching to census non-respondents. These matches were considered valid since we account for nonfarms among the non-respondents through the non-response adjustment. Data from the 2002 Classify cycle were used to estimate for the nonrespondents.

The overcount comprised 6.6% of the US total number of farms. When looking at the characteristics by type of farm, we find the census disproportionately overcounted tobacco operating arrangements. About 40% of overcounted records had this information missing and thus estimates were not available. It is important to note that if data were collected on these nonrespondents the percentages shown below could easily shift upwards for some of these groups.

The results indicate that most of the farms disproportionately overcounted by the census had TVP between $1,000 and $9,999. Large arrangements (TVP >= $100,000) were disproportionately overcounted at a much smaller rate. Results by size of farm show

that we disproportionately overcounted small sized farms (1-9 acres). It was surprising to find out that we disproportionately overcounted farms with acreage between 10 and 49 acres. This could be due to the operation being at a different location. See Tables 5.3-5.5 in Attachment E for the results presented in this section.

5.2 Characteristics of Duplicated Farms

Duplicated tracts consisted of an agricultural tract matching two or more census in-scope records. There were 570 duplicated tracts which expanded to 16,760 duplicated farms. The expanded duplication estimate represented less than 1% of the US published number of farms.

Although in small percentages, duplication was disproportionately high for grains, tobacco, cotton & cottonseed, fruit & nut trees, and milk and other dairy products types of farm. It also was disproportionately higher for all operations with $10,000 or more in TVP. The results by size of farm show that mostly farms with acreage between 50 and 499 were disproportionately duplicated. Tables 5.6-5.8 in Attachment E, provide the results presented in this section

5.3 Characteristics of Undercounted Farms

Undercounted tracts were coded as out-of-scope on the census, while the area frame shows these as having adequate data reported for them (agricultural tracts) to qualify as farms. These records had over 1,000 points, $1,000 or more in sales or more than 100 acres of pasture. There were 1,645 undercounted tracts which expanded to 106,484 farms. Close to 99% of these farms consisted of one-to-one matches during the review process. This indicates that there was a high level of agreement in the matching fields (name, address, SSN, EIN, etc). In other words, the majority of these farms did not require further field office review after the PRL phase. For the area frame (still considered as truth), about 11% of the undercounted farms had data estimated for them. To estimate for these records, the 2002 median tract-to-farm ratios at the stratum or land use level were applied. Over 20% of the records had some data reported for them but were deemed non-ag since they did not have at least $1,000 in sales, 1,000 points, or 100 acres of pasture as measured by the MFARMEDT variable.

The distribution of farms by current status code shows that 52% of the farms undercounted were from the 'Other' type of out-of-scope records on the census. Nearly all these records are grouped together due to lack of information. Another 18% of the records were computer or data review out-of-scoped. It is important to note that if the scope status of these records had not been changed by the computer or data analysts, they would have been matched correctly to their respective ag tract area records. Additionally, 14% of the farms were reported as landlord only operations on the census.

The undercount estimate represented 5% of the US total number of farms. When compared to the 2002 Census published estimates, the census disproportionately undercounted Tobacco, Christmas Tree, Other crops & Hay, Sheep & Goat, and Horse & Pony.

Similarly for TVP, the census disproportionately undercounted farms with TVP between $1,000 and $4,999. Records with zero TVP and point farms ($1-$999) comprised about 4.9% of the US total number of farms. Large farm operations ($250,000 or more) comprised about 6.1%. It is important to keep in mind that smaller operations are more likely to go out of business in a shorter period of time than are larger farm arrangements. Thus, the results

are not that surprising.

When looking at farms by size, the results seem consistent with the findings by TVP and type of farm that small or specialty type operations are being missed. The results show that small farm arrangements (1-9 acres) are disproportionately missed by the census. It was a bit surprising to see that operations with more than 10 acres would be missed. This could be due to those arrangements where the operation is at a different location. See Tables 5.9-5.12 in Attachment E for the results presented in this section.

5.4 Additional Information Collected

The data collection effort for the area surveys produced 18,967 additional pieces of information (names, addresses, and/or landlords). A significant percentage of the tracts in this study contained at least one piece of additional information on one of the area questionnaires (804 of 4,526 tracts or 18%). These 804 tracts produced 996 names and/or addresses. Some of the area tracts had information collected of all three types (name, address and landlords) and in some instances one tract had multiple names listed (i.e., landlords, partners). Table 5.13 provides the number of additional names (i.e., spouse, partners), addresses and landlord names collected for each type of error for the tracts involved in the study. In this study, there were 996 out of 18,967 additional pieces of information collected from both area surveys. Most of the information collected was landlord names.

Table 5.13 Number of Additional Names, Addresses & Landlord Names by Type of Error

TYPE OF ADDITIONAL INFORMATION	UNDERCOUNT	OVERCOUNT	DUPLICATION	TOTAL
Additional Names	51	14	24	89
Additional Addresses	53	13	33	99
Landlord Names	381	35	392	808
Total	485	62	449	996

There was more information collected to identify misclassification undercount and duplication errors than there was for overcount. It is expected that the number of additional names and addresses collected for the misclassification overcount was low since these were the non-ag tracts. Non-ag tracts don't always contain complete name and address information. However, the collection of this additional information may assist in addressing the overcount problem in the future.

It is possible that the remaining names not directly used in the analysis for this report helped to correctly match area and census records which would have not been matched otherwise. For example, an area respondent lists two partners on the questionnaire, while the census questionnaire is completed by one of the partners. The census respondent is not the one listed as the target operator on the area side. If we only matched the target operator name to the name on the census questionnaire (one-to-one matching) they would not have matched. However, matching to the additional names reported on the area questionnaire would have picked up the partner name as a valid match. If the matching process had involved solely matching to the operator names we could miss matches from partners, landlord, spouses, etc. that were not reported on the census. This would lead to improper classification of records as overcount, undercount or duplication. It may even lead

to considering two records as non-matches when they are in fact a valid match.

5.5 1997 & 2002 CES Eleven State Comparison

In 1997, the CES was conducted for the 11 Western States which comprised the West Census Region. The states participating in the 1997 study were Arizona, California, Colorado, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington and Wyoming. The 1997 CES reported a net 27,971 misclassified undercounted farms. The 2002 CES resulted in an estimated net misclassification overcount of 5,438 farms for these eleven states, with a standard error of 3,245. The results are not consistent across censuses. Note that there is a correctable positive bias in the area frame estimators due to the lack of data for the non-ag tracts for which only the available census reported data were used.

6. REFERENCES

2002 Census of Agriculture, United States Summary and State Data, Volume 1, Geographic Area Series, Part 51, National Agricultural Statistics Service, USDA.

Abreu, Denise (2002). Review of Possible Matches. Unpublished User Instructions. Available from the author upon request.

Abreu, Denise (2002). Farm Classification Resolution Review. Unpublished User Instructions. Available from the author upon request.

Allen, J. (1998). 1997 Census Classification Error Survey. Internal Document. National Agricultural Statistics Service, USDA.

Johnson, J.V. (2000). Agricultural Census – Classification Error Estimation Using An Area Frame Approach. Data Quality Research Section Unpublished Manuscript. 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.

ATTACHMENT A

Additional Names and Address Information Collected during the Area Surveys

State Name	JAS Names	ACES Names	Additional Addresses	Additional Names	Landlord Names	Total Add'l Info	Total Names JAS + ACES	Additional Names %
Alabama(1)	2,686	0	10	7	176	193	2,686	7.2%
Arizona(4)	733	179	16	6	61	83	912	9.1%
Arkansas(5)	4,153	167	38	31	627	696	4,320	16.1%
California(6)	3,832	944	138	63	364	565	4,776	11.8%
Colorado(8)	1,621	378	27	12	148	187	1,999	9.4%
Connecticut(9)	27	65	1	2	14	17	92	18.5%
Delaware(10)	258	9	0	0	26	26	267	9.7%
Florida(12)	1,200	916	20	10	139	169	2,116	8.0%
Georgia(13)	3,045	1,016	23	9	295	327	4,061	8.1%
Idaho(16)	1,478	991	85	223	259	567	2,469	23.0%
Illinois(17)	5,105	595	23	13	793	829	5,700	14.5%
Indiana(18)	4,451	400	22	197	640	859	4,851	17.7%
Iowa(19)	4,923	145	70	72	1,054	1,196	5,068	23.6%
Kansas(20)	3,142	180	30	4	552	586	3,322	17.6%
Kentucky(21)	1,996	882	3	0	166	169	2,878	5.9%
Louisiana(22)	2,266	267	26	5	433	464	2,533	18.3%
Maine(23)	169	35	13	3	15	31	204	15.2%
Maryland(24)	953	118	16	20	103	139	1,071	13.0%
Massachusetts(25)	68	147	8	2	26	36	215	16.7%
Michigan(26)	1,747	760	23	35	420	478	2,507	19.1%
Minnesota(27)	4,545	298	109	55	745	909	4,843	18.8%
Mississippi(28)	3,655	0	19	18	455	492	3,655	13.5%
Missouri(29)	3,350	2,328	59	19	451	529	5,678	9.3%
Montana(30)	2,219	77	66	84	328	478	2,296	20.8%
Nebraska(31)	4,199	0	85	183	697	965	4,199	23.0%
Nevada(32)	66	76	0	0	0	0	142	0.0%
New Hampshire(33)	72	440	16	39	53	108	512	21.1%
New Jersey(34)	826	139	0	0	37	37	965	3.8%
New Mexico(35)	676	135	16	3	29	48	811	5.9%

State Name	JAS Names	ACES Names	Additional Addresses	Additional Names	Landlord Names	Total Add'l Info	Total Names JAS + ACES	Additional Names %
New York(36)	852	455	4	9	41	54	1,307	4.1%
North Carolina(37)	2,661	167	28	24	349	401	2,828	14.2%
North Dakota(38)	3,108	0	42	29	698	769	3,108	24.7%
Ohio(39)	2,354	829	31	37	547	615	3,183	19.3%
Oklahoma(40)	2,994	438	48	33	255	336	3,432	9.8%
Oregon(41)	1,637	890	71	95	169	335	2,527	13.3%
Pennsylvania(42)	3,037	1,251	77	84	540	701	4,288	16.3%
Rhode Island(44)	43	3	1	2	6	9	46	19.6%
South Carolina(45)	1,200	179	0	0	138	138	1,379	10.0%
South Dakota(46)	2,418	0	35	16	426	477	2,418	19.7%
Tennessee(47)	4,014	900	6	7	87	100	4,914	2.0%
Texas(48)	12,528	1,924	373	115	1,795	2,283	14,452	15.8%
Utah(49)	604	489	14	20	105	139	1,093	12.7%
Vermont(50)	232	82	10	2	64	76	314	24.2%
Virginia(51)	974	587	0	0	157	157	1,561	10.1%
Washington(53)	2,312	664	20	11	200	231	2,976	7.8%
West Virginia(54)	1,379	641	46	41	200	287	2,020	14.2%
Wisconsin(55)	3,688	961	24	32	539	595	4,649	12.8%
Wyoming(56)	576	194	30	6	45	81	770	10.5%
US Totals	*110,072*	*22,341*	*1,822*	*1,678*	*15,467*	*18,967*	*132,413*	14.3%

ATTACHMENT B

CES Possible Matches Resolution Review – Workloads by State

State	Area Records	Census Records	Possible Matches	Definite Matches	Total Link Groups
AL-1	2,686	54,464	579	361	940
AZ-4	912	9,999	149	58	207
AR-5	4,320	55,443	1,173	400	1,573
CA-6	4,776	96,415	1,108	531	1,639
CO-8	1,999	37,710	448	347	795
CT-9	92	5,733	10	18	28
DE-10	267	2,802	61	25	86
FL-12	2,116	49,184	452	337	789
GA-13	4,061	53,721	987	437	1,424
ID-16	2,469	28,932	536	318	854
IL-17	5,700	93,928	1,464	947	2,411
IN-18	4,851	75,491	1,190	615	1,805
IA-19	5,068	114,973	1,769	854	2,623
KS-20	3,322	77,824	1,160	693	1,853
KY-21	2,878	110,659	786	399	1,185
LA-22	2,533	34,198	632	226	858
ME-23	204	10,320	24	32	56
MD-24	1,071	16,964	228	79	307
MA-25	215	7,776	44	31	75
MI-26	2,507	69,904	655	331	986
MN-27	4,843	99,629	1,569	659	2,228
MS-28	3,655	48,405	982	287	1,269
MO-29	5,678	126,901	1,591	982	2,573
MT-30	2,296	38,668	683	256	939
NE-31	4,199	59,639	1,302	604	1,906
NV-32	142	3,949	15	11	26
NH-33	512	4,314	48	32	80
NJ-34	965	13,212	116	84	200
NM-35	811	18,185	202	156	358
NY-36	1,307	42,148	170	245	415
NC-37	2,828	63,052	765	449	1,214
ND-38	3,108	37,876	1,026	475	1,501
OH-39	3,183	92,397	842	618	1,460
OK-40	3,432	94,782	1,087	581	1,668
OR-41	2,527	46,230	472	321	793
PA-42	4,288	79,720	1,157	280	1,437
RI-44	46	1,436	4	8	12
SC-45	1,379	42,102	327	160	487
SD-46	2,418	42,811	905	393	1,298
TN-47	4,914	94,149	1,184	765	1,949
TX-48	14,452	271,825	4,528	1,922	6,450
UT-49	1,093	20,048	284	129	413
VT-50	314	7,026	56	34	90
VA-51	1,561	54,545	380	213	593
WA-53	2,976	39,470	532	221	753
WV-54	2,020	22,295	392	118	510
WI-55	4,649	81,084	1,255	423	1,678
WY-56	770	11,240	105	102	207
US-Totals	**132,413**	**2,563,578**	**35,434**	**17,567**	**53,001**

ATTACHMENT C

CES FARM CLASSIFICATION RESOLUTION REVIEW – WORKLOADS BY STATE

State Alpha	Fips	Total Link Groups	HQ Link Groups (No SSO Review)	SSO Workloads		Total Link Groups for SSO Review
				AG Link Groups for Review	Non-Ag & Out-of-Business Link Groups for Review	
AL	1	742	540	126	76	202
AZ	4	194	106	71	17	88
AR	5	1,375	820	403	152	555
CA	6	1,366	870	419	77	496
CO	8	719	473	220	26	246
CT	9	25	20	5	0	5
DE	10	64	44	14	6	20
FL	12	673	478	160	35	195
GA	13	1,124	797	225	102	327
ID	16	718	491	194	33	227
IL	17	2,112	1,446	593	73	666
IN	18	1,442	1,124	272	46	318
IA	19	2,314	1,636	630	48	678
KS	20	1,691	1,207	447	37	484
KY	21	934	692	194	48	242
LA	22	682	453	185	44	229
ME	23	45	34	9	2	11
MD	24	237	166	55	16	71
MA	25	64	51	13	0	13
MI	26	865	701	145	19	164
MN	27	1,861	1,404	396	61	457
MS	28	960	655	234	71	305
MO	29	2,158	1,526	487	145	632
MT	30	979	540	318	121	439
NE	31	1,763	1,259	482	22	504
NV	32	25	16	7	2	9
NH	33	56	47	9	0	9
NJ	34	170	127	31	12	43
NM	35	320	209	72	39	111
NY	36	364	310	34	20	54
NC	37	1,124	559	480	85	565
ND	38	1,411	1,016	381	14	395
OH	39	1,313	1,069	235	9	244
OK	40	1,399	1,014	315	70	385
OR	41	666	479	164	23	187
PA	42	1,002	780	200	22	222
RI	44	11	9	2	0	2
SC	45	376	265	90	21	111
SD	46	1,212	785	371	56	427
TN	47	1,498	1,116	281	101	382
TX	48	5,204	3,246	1,737	221	1,958
UT	49	364	147	191	26	217
VT	50	79	62	16	1	17
VA	51	484	370	100	14	114
WA	53	653	453	141	59	200

WV	54	332	203	93	36	129
WI	55	1,204	948	184	72	256
WY	56	183	130	48	5	53
US	**99**	**44,557**	**30,893**	**11,479**	**2,185**	**13,664**

ATTACHMENT D

Total Number of Undercounted, Overcounted and Duplicated Tracts by State

State Name	Fips Code	Total Undercounted Tracts	Total Overcounted Tracts	Total Duplicated Tracts	Total Misclassified Tracts
Alabama	1	29	97	1	127
Arizona	4	9	24	0	33
Arkansas	5	44	107	29	180
California	6	53	74	31	158
Colorado	8	31	32	9	72
Connecticut	9	1	1	1	3
Delaware	10	5	3	2	10
Florida	12	36	53	8	97
Georgia	13	70	150	3	223
Idaho	16	26	18	4	48
Illinois	17	39	100	5	144
Indiana	18	44	75	21	140
Iowa	19	55	86	39	180
Kansas	20	37	47	24	108
Kentucky	21	54	65	15	134
Louisiana	22	35	51	3	89
Maine	23	2	3	1	6
Maryland	24	6	23	11	40
Massachusetts	25	0	3	0	3
Michigan	26	29	33	11	73
Minnesota	27	50	102	8	160
Mississippi	28	22	78	7	107
Missouri	29	79	60	11	150
Montana	30	20	10	6	36
Nebraska	31	38	35	23	96
Nevada	32	3	2	0	5
New Hampshire	33	2	4	1	7
New Jersey	34	8	20	0	28
New Mexico	35	16	15	10	41
New York	36	14	25	3	42
North Carolina	37	38	24	25	87
North Dakota	38	39	23	12	71
Ohio	39	58	28	20	106
Oklahoma	40	71	103	2	176
Oregon	41	26	44	5	75
Pennsylvania	42	52	56	29	140
Rhode Island	44	1	0	0	1
South Carolina	45	19	34	3	56
South Dakota	46	52	41	18	111
Tennessee	47	86	143	4	233
Texas	48	233	221	119	573
Utah	49	11	23	4	38
Vermont	50	2	2	0	4
Virginia	51	31	18	6	55
Washington	53	18	48	0	66
West Virginia	54	11	22	3	36
Wisconsin	55	29	78	33	140
Wyoming	56	11	7	0	18
Totals		**1,645**	**2,311**	**570**	**4,526**

ATTACHMENT E

Table 5.3: Characteristics of Overcounted Census Records by Type of Farm

DESCRIPTION	NUMBER OF FARMS OVERCOUNTED 2002 CES	TOTAL NUMBER OF FARMS 2002 CENSUS	PERCENT
Grains, oilseeds, dry beans, and dry peas	4,452	324,812	1.4%
Tobacco	2,767	36,782	7.5%
Cotton and cottonseed	373	14,498	2.6%
Vegetables, melons, potatoes and sweet potatoes	1,267	34,293	3.7%
Fruit, tree nuts, and berries	4,839	97,502	5.0%
Nursery, greenhouse, floriculture & sod	1,109	47,483	2.3%
Cut Christmas trees & short rotation woody crops	818	16,954	4.8%
Other crops & hay	21,382	390,421	5.5%
Cattle & calves	30,078	712,831	4.2%
Milk & other dairy products	2,153	71,230	3.0%
Hogs & pigs	2,261	87,213	2.6%
Sheep, goats, and their products	1,550	39,628	3.9%
Horses, Ponies, etc	9,163	174,412	5.3%
Poultry and eggs	1,313	29,912	4.4%
Aquaculture	339	3,284	10.3%
Other animals & other animal prods	871	47,729	1.8%
None/Unknown	56,337	--	--
Total	**141,072**	**2,128,982**	**6.6%**

- Data may not add to the total due to rounding.

Table 5.4: Characteristics of Overcounted Census Records by Total Value of Production

DESCRIPTION	NUMBER OF FARMS OVERCOUNTED 2002 CES	TOTAL NUMBER OF FARMS 2002 CENSUS	PERCENT
Less than $1,000	31,412	570,919	5.5%
$1,000-$2,499	39,815	255,639	15.6%
$2,500-$4,999	25,546	213,436	12.0%
$5,000-$9,999	16,441	223,168	7.4%
$10,000-$24,499	14,160	256,157	5.5%
$25,000-$49,999	4,965	157,906	3.1%
$50,000-$99,999	2,703	140,479	1.9%
$100,000-$249,999	2,411	159,052	1.5%
$250,000-$499,999	1,123	81,694	1.4%
$500,000-$999,999	1,072	41,469	2.6%
$1,000,000 and over	1,422	28,673	5.0%
Total	**141,070**	**2,128,982**	**6.6%**

- Data may not add to the total due to rounding.

Table 5.5: Characteristics of Overcounted Records by Size of Farm

DESCRIPTION	NUMBER OF FARMS 2002 CES	NUMBER OF FARMS 2002 CENSUS	PERCENT
1-9 Acres	35,762	179,346	19.9%
10 – 49 Acres	55,645	563,772	9.9%
50 – 179 Acres	37,310	658,705	5.7%
180 – 499 Acres	7,686	388,617	2.0%
500 – 999 Acres	2,634	161,552	1.6%
1,000 – 1,999 Acres	1,870	99,020	1.9%
2,000 or more Acres	163	77,970	0.2%
Total	**141,070**	**2,128,982**	**6.6%**

- Data may not add to the total due to rounding.

Table 5.6: Characteristics of Duplicated Records by Type of Farm

DESCRIPTION	NUMBER OF FARMS DUPLICATED 2002 CES	TOTAL NUMBER OF FARMS 2002 CENSUS	PERCENT
Grains, oilseeds, dry beans, and dry peas	4,332	324,812	1.3%
Tobacco	736	36,782	2.0%
Cotton and cottonseed	537	14,498	3.7%
Vegetables, melons, potatoes and sweet potatoes	67	34,293	0.2%
Fruit, tree nuts, and berries	1,785	97,502	1.8%
Nursery, greenhouse, floriculture & sod	183	47,483	0.4%
Cut Christmas trees & short rotation woody crops	158	16,954	0.9%
Other crops & hay	1,917	390,421	0.5%
Cattle & calves	4,353	712,831	0.6%
Milk & other dairy products	1,668	71,230	2.3%
Hogs & pigs	143	87,213	0.2%
Sheep, goats, and their products	117	39,628	0.3%
Horses, Ponies, etc	516	174,412	0.3%
Poultry and eggs	26	29,912	0.1%
Aquaculture	0	3,284	0.0%
Other animals & other animal prods	220	47,729	0.5%
Total	**16,758**	**2,128,982**	**0.8%**

- Data may not add to the total due to rounding.

Table 5.7: Characteristics of Duplicated Records by Total Value of Production

DESCRIPTION	NUMBER OF FARMS DUPLICATED 2002 CES	TOTAL NUMBER OF FARMS 2002 CENSUS	PERCENT
Less than $1,000	384	570,919	0.1%
$1,000-$2,599	641	255,639	0.3%
$2,500-$4,999	1,178	213,436	0.6%
$5,000-$9,999	1,477	223,168	0.7%
$10,000-$24,999	2,458	256,157	1.0%
$25,000-$49,999	2,239	157,906	1.4%
$50,000-$99,999	2,106	140,479	1.5%
$100,000-$249,999	2,673	159,052	1.7%
$250,000-$499,999	1,580	81,694	1.9%
$500,000-$999,999	1,243	41,469	3.0%
$1,000,000 and over	779	28,673	2.7%
Total	**16,758**	**2,128,982**	**0.8%**

- Data may not add to the total due to rounding.

Table 5.8: Characteristics of Duplicated Records by Size of Farm

DESCRIPTION	NUMBER OF FARMS DUPLICATED 2002 CES	TOTAL NUMBER OF FARMS 2002 CENSUS	PERCENT
1-9 Acres	586	179,346	0.3%
10 – 49 Acres	2,779	563,772	0.5%
50 – 179 Acres	7,649	658,705	1.2%
180 – 499 Acres	4,504	388,617	1.2%
500 – 999 Acres	957	161,552	0.6%
1,000 – 1,999 Acres	206	99,020	0.2%
2,000 or more Acres	80	77,970	0.1%
Total	**16,758**	**2,128,982**	**0.8%**

- Data may not add to the total due to rounding.

Table 5.9: Characteristics of Undercounted Records by Current Status Code

TYPE OF OUT-OF-SCOPE RECORD	NUMBER OF FARMS	PERCENT
Deceased	9,012	8.5%
Landlord Only	14,507	13.6%
Non-Ag Never Farmed	7,571	7.1%
Retired OR Disabled	253	0.2%
Other	55,499	52.1%
Computer	17,185	16.1%
Data Review	2,458	2.3%
Total	**106,484**	**100.0%**

Table 5.10: Characteristics of Undercounted Records by Type of Farm

DESCRIPTION	NUMBER OF FARMS UNDERCOUNTED 2002 CES	TOTAL NUMBER OF FARMS 2002 CENSUS	PERCENT
Grains, oilseeds, dry beans, and dry peas	11,499	324,812	3.5%
Tobacco	2,229	36,782	6.1%
Cotton and cottonseed	373	14,498	2.67%
Vegetables, melons, potatoes and sweet potatoes	1,592	34,293	4.6%
Fruit, tree nuts, and berries	2,640	97,502	2.7%
Nursery, greenhouse, floriculture & sod	1,319	47,483	2.8%
Cut Christmas trees & short rotation woody crops	1,639	16,954	9.7%
Other crops & hay	25,159	390,421	6.4%
Cattle & calves	37,295	712,831	5.2%
Milk & other dairy products	1,382	71,230	1.9%
Hogs & pigs	1,247	87,213	1.4%
Sheep, goats, and their products	2,383	39,628	6.0%
Horses, Ponies, etc	15,470	174,412	8.9%
Poultry and eggs	1,138	29,912	3.8%
Aquaculture	121	3,284	3.7%
Other animals & other animal prods	998	47,729	2.1%
Total	**106,484**	**2,128,982**	**5.0%**

- Data may not add to the total due to rounding.

Table 5.11: Characteristics of Undercounted Records by Total Value of Production

DESCRIPTION	NUMBER OF FARMS UNDERCOUNTED 2002 CES	TOTAL NUMBER OF FARMS 2002 CENSUS	PERCENT
Less than $1,000	27,955	570,919	4.9%
$1,000-$2,499	24,168	255,639	9.5%
$2,500-$4,999	12,607	213,436	5.9%
$5,000-$9,999	11,328	223,168	5.1%
$10,000-$24,999	10,313	256,157	4.0%
$25,000-$49,999	6,873	157,906	4.4%
$50,000-$99,999	5,087	140,479	3.6%
$100,000-$249,999	4,964	159,052	3.1%
$250,000-$499,999	1,887	81,694	2.3%
$500,000-$999,999	731	41,469	1.8%
$1,000,000 and over	571	28,673	2.0%
Total	**106,484**	**2,128,592**	**5.0%**

- Data may not add to the total due to rounding.

Table 5.12: Characteristics of Undercounted Records by Size of Farm

DESCRIPTION	NUMBER OF FARMS UNDERCOUNTED 2002 CES	TOTAL NUMBER OF FARMS 2002 CENSUS	PERCENT
1-9 Acres	15,770	179,346	8.8%
10 – 49 Acres	38,576	563,772	6.8%
50 – 179 Acres	37,049	658,705	5.6%
180 – 499 Acres	12,154	388,617	3.1%
500 – 999 Acres	1,980	161,552	1.2%
1,000 – 1,999 Acres	453	99,020	0.5%
2,000 or more Acres	502	77,970	0.6%
Total	**106,484**	**2,128,982**	**5.0%**

- Data may not add to the total due to rounding.