Assessment of an Imputation Process Used in the 2017 Census of Agriculture

Tara Murphy, Habtamu Benecha, Denise A. Abreu, Darcy Miller
National Agricultural Statistics Service, USDA,
1400 Independence Avenue, Washington DC 20250

Abstract

The National Agricultural Statistics Service (NASS) conducts a Census of Agriculture (COA) every five years using a list frame. The 2017 COA used capture-recapture methods to adjust the COA for undercoverage, nonresponse and misclassification of farms/non-farms. NASS's June Area Survey (JAS) was used as the independent survey in the capture-recapture approach. The JAS uses an area frame and the data are collected via in-person interviews. For capture-recapture, a matched dataset consisting of all matches of a COA record to a JAS record is formed. This dataset is the foundation for modeling the probabilities of coverage, response and correct classification of farms/non-farms for the COA. These probabilities are estimated through a series of weighted logistic regression models. Demographic characteristics are crucial covariates considered in the models' variable selection. In 2017, NASS redesigned the demographics section of the COA questionnaire to allow up to four principal operators per farm. The JAS questionnaire gathers information on only one principal operator. Multivariate imputation was used to address this missing-data problem. This paper evaluates the effectiveness of the imputation.

Key words: Capture-recapture; Imputation; List frame; Area frame; Logistic regression

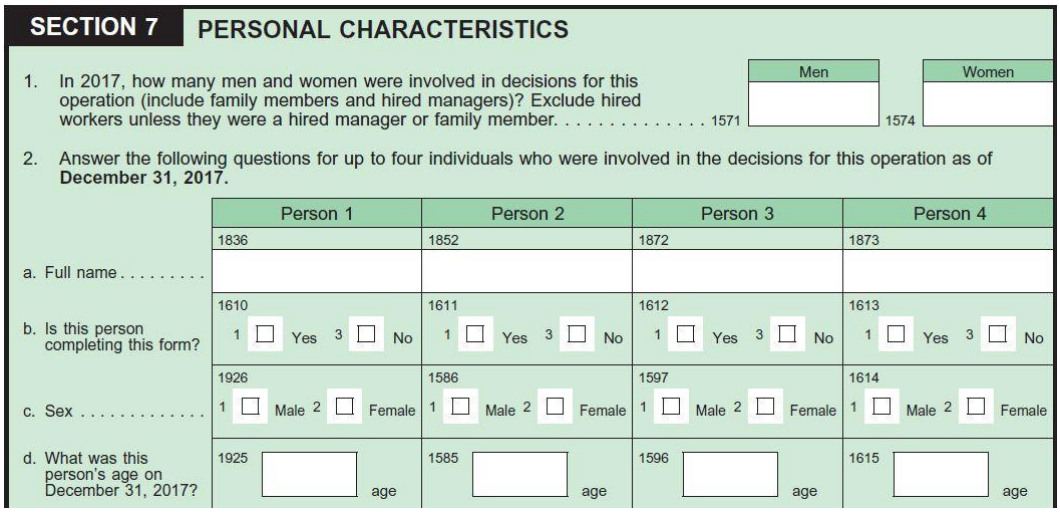
1. Introduction and Background

The United States Department of Agriculture's National Agricultural Statistics Service (NASS) conducts over one hundred surveys each year and prepares more than 500 reports annually that cover every facet of U.S. agriculture. The majority of the reports provide estimates that impact U.S. markets and the prices of commodities. Some examples of these include corn, soybeans, wheat, and upland cotton estimates of acreage and forecasts of yield. NASS conducts the Census of Agriculture (COA) every 5 years, in years ending in 2 and 7. The COA provides information on characteristics of U.S. farms and ranches and the people who operate them. A farm is defined to be any place from which $1,000 or more of agricultural products were produced and sold or normally would have been sold during the year. During the COA, data are collected on land use and ownership, operator characteristics, production practices, income and expenditures, and numerous other characteristics. The COA is the leading source of information on characteristics of the people operating farms and provides the most uniform comprehensive agricultural data for every county in the nation. It is used by federal, state, and local governments and others who provide services to farms and rural communities. COA estimates are published at the national, state, and county levels. The estimates impact community planning, availability of operational loans and other funding, location and staffing of service centers, and farm programs and policies.

The Census is a list-based endeavor. The list contains both agricultural operations that are in the target population (farms) and agricultural operations that are not in the target population (non-farms). The Census Mailing List (CML) is incomplete; not all farms are on the list. To account for farming operations not on the CML, NASS uses the June Area Survey (JAS). The JAS uses an area frame and, during pre-screening, tracts of land are classified as agricultural or non-agricultural based on the agricultural activity of the area. The JAS is conducted annually and also provides an estimate of the number of farms. In 2007, the difference in the estimated number of farms from the COA and from the JAS was larger than could be attributed to sampling error alone (Abreu et al., 2010). This led to the decision to use capture–recapture or dual system estimation (DSE) methodology as the foundation for adjusting the 2012 Census of Agriculture, and future censuses, for undercoverage, nonresponse, and misclassification (Young et al., 2017).

To implement capture-recapture methodology, a matched dataset consisting of all matches of a COA record to a JAS tract is formed. The matching is performed using probabilistic record linkage. This dataset is the foundation for modeling the probabilities of coverage, response, and correct classification of farms/non-farms for the COA. These probabilities are estimated through a series of weighted logistic regression models. Demographic characteristics are crucial covariates considered in the models' variable selection.

In 2015, a panel of experts reviewed the COA to determine improvements that could be made to allow data users to better understand the role and effectiveness of USDA programs directed at women and beginning farmers. The panel recommended several updates to the COA questionnaire to achieve this goal. In response to one of the recommendations, NASS redesigned the demographics section of the 2017 COA questionnaire to allow up to four principal producers per farm ("Report of the Expert Panel" 2015) (Figure 1). A principal producer is defined as an individual on the operation who is involved in decision making (Figure 2).



SECTION 7 **PERSONAL CHARACTERISTICS**

1. In 2017, how many men and women were involved in decisions for this operation (include family members and hired managers)? Exclude hired workers unless they were a hired manager or family member. . . . . 1571 — Men: 1574 — Women

2. Answer the following questions for up to four individuals who were involved in the decisions for this operation as of **December 31, 2017**.

	Person 1	Person 2	Person 3	Person 4
	1836	1852	1872	1873
a. Full name				
b. Is this person completing this form?	1610 1 ☐ Yes 3 ☐ No	1611 1 ☐ Yes 3 ☐ No	1612 1 ☐ Yes 3 ☐ No	1613 1 ☐ Yes 3 ☐ No
c. Sex	1926 1 ☐ Male 2 ☐ Female	1586 1 ☐ Male 2 ☐ Female	1597 1 ☐ Male 2 ☐ Female	1614 1 ☐ Male 2 ☐ Female
d. What was this person's age on December 31, 2017?	1925 ☐ age	1585 ☐ age	1596 ☐ age	1615 ☐ age

Figure 1: 2017 COA demographics section snapshot

3. Was this person involved in these specific decisions as of **December 31, 2017**? For each person and for each item, **mark all that apply**.

	Person 1	Person 2	Person 3	Person 4
a. Day-to-day decisions	1642 1 ☐	1643 1 ☐	1644 1 ☐	1645 1 ☐
b. Land use and/or crop decisions, including planting, crop spraying, or other, e.g., grazing	1650 1 ☐	1651 1 ☐	1652 1 ☐	1653 1 ☐
c. Livestock decisions, including purchases, sales, breeding, and pasturing.	1654 1 ☐	1655 1 ☐	1656 1 ☐	1657 1 ☐
d. Record keeping and/or financial management.	1776 1 ☐	1777 1 ☐	1778 1 ☐	1779 1 ☐
e. Estate planning or succession planning . .	1757 1 ☐	1758 1 ☐	1759 1 ☐	1760 1 ☐

	Person 1	Person 2	Person 3	Person 4
4. Is this person a Principal Operator or Senior Partner?	1765 1 ☐ Yes 3 ☐ No	1766 1 ☐ Yes 3 ☐ No	1767 1 ☐ Yes 3 ☐ No	1768 1 ☐ Yes 3 ☐ No

Figure 2: 2017 COA demographics section snapshot

The 2017 JAS questionnaire collected demographic information on only one principal operator (Figure 3), the person who makes most of the day-to-day decisions (one of the decision-making questions on the COA). For purposes of simplicity, the JAS operator will henceforth be referred to as a "producer". Ideally, the demographic information on the 2017 JAS would have been collected in the same manner as the 2017 COA to complete the DSE weighting process for the 2017 COA. When the COA and the JAS matched dataset was created, the demographic variables associated with producers 2, 3, and 4 were missing for the JAS records. JAS records are a crucial element for modeling coverage of the CML. Because COA publications include demographic estimates at the county level, it is essential for the demographic variables to be included in the model.

SECTION P - OPERATOR CHARACTERISTICS

1. **Age of operator as of December 31, 2016?**
[*Check* (√) *age of operator and enter code.*]

		Code
☐ Less than 25 years.	= 1	
☐ 25 - 34 years.	= 2	
☐ 35 - 44 years.	= 3	821
☐ 45 - 54 years.	= 4	
☐ 55 - 64 years.	= 5	
☐ 65 years and over.	= 6	

2. **Ethnicity of operator?** [*Check* (√) *one and enter code.*]

☐ Hispanic or Latino	= 1	822
☐ Not Hispanic or Latino	= 3	

3. **Race of operator?** [Check (√) one or more and enter code.]

☐ White.	= 1	
☐ Black or African American	= 2	823
☐ American Indian or Alaska Native (*Specify tribe*:)	= 3	
☐ Asian.	= 4	
☐ Native Hawaiian or Other Pacific Islander	= 5	

4. **Sex of operator?** [Complete from observation and enter code.]

☐ Male	= 1	824
☐ Female	= 2	

	Year
5. **In what year did the operator begin to operate any part of this operation?**	896

Figure 3: 2017 JAS demographics section snapshot

2. Imputation Methodology Overview

In order to combat the missing information issue, hot deck imputation was used to impute demographic information for up to three additional producers on the JAS form using donors from the COA administered in the same year. Hot deck imputation often describes a general class of imputation methods that utilize the current survey data (the 'hot' data) to model and impute data. It has also evolved to be used as a term for a specific imputation process where groups of 'like' records are formed and a respondent value is drawn from the same group as the recipient to provide an imputed value for the recipient.

In this implementation of the hot deck method, no donors were available in the current JAS survey to use to impute demographic items for more than one producer for other JAS records. So, 2017 Census demographic data was added to the pool of donors for more than one producer on the JAS records. Groups were formed based on the values of the producer collected on the JAS form and the producer listed in the first column of the COA (most often a principal producer). Demographic variables used to form similar groups included age, race, and sex of the first producer listed. An entire COA record was drawn from the group to impute producers 2, 3, and 4 on the JAS. Using the entire COA record as a donor, the distributions of the number of producers and joint demographics of producers were maintained. The distribution of the number of producers was preserved since records drawn would have zeros as placeholders for variables collecting information on producers beyond the number of producers on the farm. For example, the demographic values on the COA record drawn could all be 0, meaning that the census record only had one producer, ensuring that the distribution of single producer farms was still preserved in the JAS. Demographic values drawn for producers 2, 3, and 4 could all be zero except for values corresponding to the second producer, preserving the distribution of two producer farms, and similarly for three and four or more producer farms. Any items requiring imputation in the data for the one producer that was collected on the JAS form was imputed using other JAS records where all of the items were reported before imputing data for potential additional producers. Imputation was implemented using the PROC SURVEYIMPUTE procedure available in the SAS software.

2.1 PROC SURVEYIMPUTE

PROC SURVEYIMPUTE is a SAS procedure that implements imputation techniques that do not use explicit models. Hot-deck imputation is the most commonly used imputation technique for survey data. A donor is selected for a recipient unit, and the observed values of the donor are imputed for the missing items of the recipient. Although the imputation method is straightforward, the variance estimator that accounts for imputation variance might not be simple and is often ignored in practice. PROC SURVEYIMPUTE does not create imputation-adjusted replicate weights for hot-deck imputation. Available donor selection techniques include simple random selection with or without replacement, probability proportional to weights selection (Rao and Shao, 1992), and approximate Bayesian bootstrap selection (Rubin and Schenker, 1986). For the JAS imputation process, simple random selection with replacement was used.

3. Evaluating the Impact of the JAS Imputation

To assess impacts of the imputation on COA estimates, comparisons were made between model estimates with and without imputation for the demographic variables associated with producers 2, 3, and 4 on the JAS. The characteristics evaluated were age, gender, race and

ethnicity. For age, six groups were formed (less than 25, 25-34, 35-44, 45-54, 55-64, and greater than or equal to 65). For each characteristic, two types of variables were created: "farms with at least one…" and "farms with any…". "Farms with at least one" indicated that at least one producer met the characteristic and was a principal producer. "Farms with any" indicated that any of the producers had the characteristic but may NOT necessarily be a principal producer. For example, a farm with at least one male principal producer indicates at least one producer on the farm is male and that same producer is considered a principal producer. A farm with any male producer indicates the farm has one or more male producers, regardless of whether they are designated as a principal producer or not. Remember that not all producers are principal producers. In other words, not all persons associated with the operation are involved in decision making. A total of 28 demographic characteristic variables were created (see Appendix A).

Estimates based on the imputed data are already available from the official 2017 Census of Agriculture publications. Estimates from the matched data with missing variables (i.e., without imputation) were obtained by applying the same procedures as the COA estimation on the incomplete data. The DSE modeling process, including variable selection, was applied and produced estimates based on the incomplete data. The resulting estimates are henceforth referred to as the study DSE estimates. The study DSE estimates (i.e., estimates from the incomplete matched dataset) and the published DSE estimates (i.e., estimates from the complete matched dataset) were compared for several demographic variables by using paired t-tests and graphical means.

4. Findings

T-tests performed to compare the study DSE estimates and the published DSE estimates showed that estimates from the two approaches are significantly different ($p < 0.01$) at the national level for eleven of the 28 demographic characteristics. Table 1 shows these variables.

Table 1: Significantly different variables at $p < 0.01$

Farms with...	
at least one male principal producer	any producer less than 25 years of age
any male producer	at least one principal producer between the ages of 35 & 44
at least one female principal producer	any producer between the ages of 35 & 44
any female producer	at least one principal producer between the ages of 45 & 54
at least one Hispanic principal producer	any producer between the ages of 45 & 54
	at least one principal producer between the ages of 55 & 64

Based on research used to redesign the 2017 COA demographics section, there was an expectation to capture more young (less than 25 years of age) and female producers (Ridolfo et al., 2016). Particular attention was paid to these demographic variables to determine how the imputation efforts for the JAS producers may have better reflected those with these characteristics.

4.1 Young Producers

Nationally, the study DSE and the published DSE estimates were different for farms with at least one principal producer aged less than 25 at $p = 0.11$. Regional graphical analysis was done to review the mean percent difference between the study DSE estimates and the published DSE estimates by region. Agricultural regions were defined by subject matter experts based on similar agricultural activity (Figure 4).

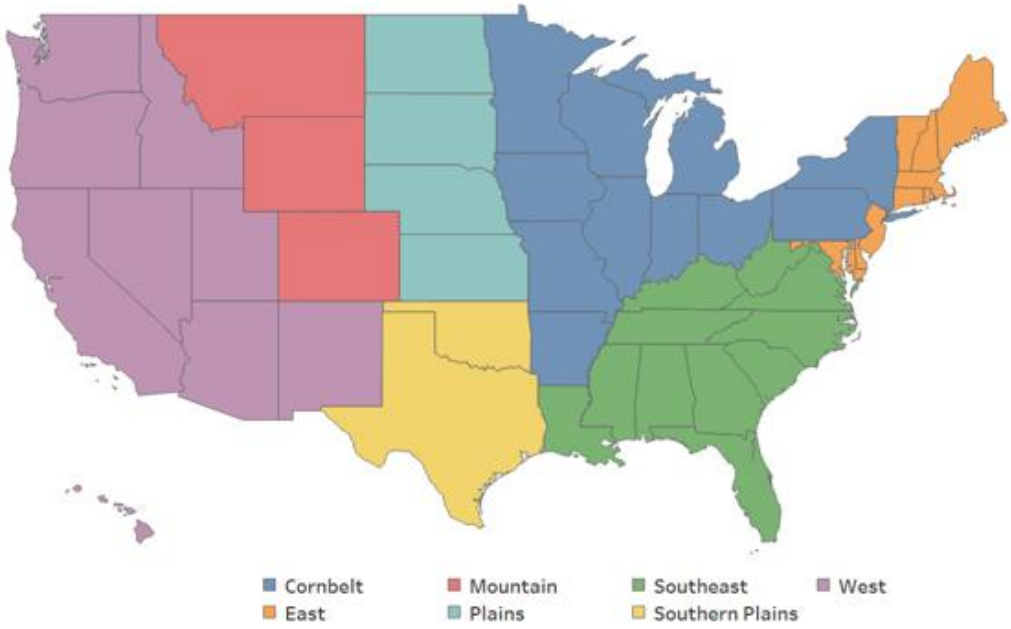


Figure 4: Agricultural regions

For the majority of the regions, the study DSE estimates were less than the published DSE estimates, as shown in Figure 5 identified by the bars below the zero line. In the Southern Plains region, the study DSE estimate was greater than the published DSE estimate (identified by the bar above the zero line), meaning the DSE estimate calculated with the data where potential producers 2, 3, and 4 on the JAS were not imputed was larger than the published DSE estimate for that region. One possible reason for this could be that different variables were selected for the model; however, this requires further research.

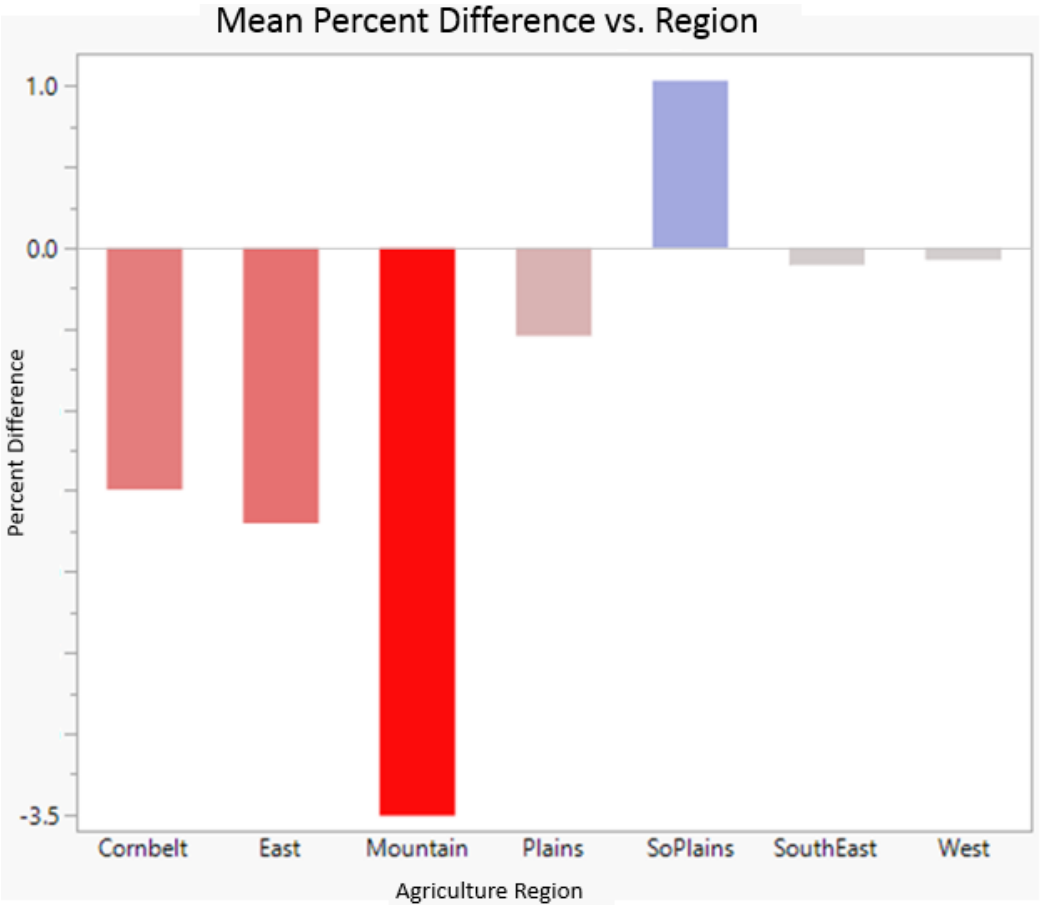


Figure 5: Mean percent difference by region for farms with at least one principal producer less than 25 years of age

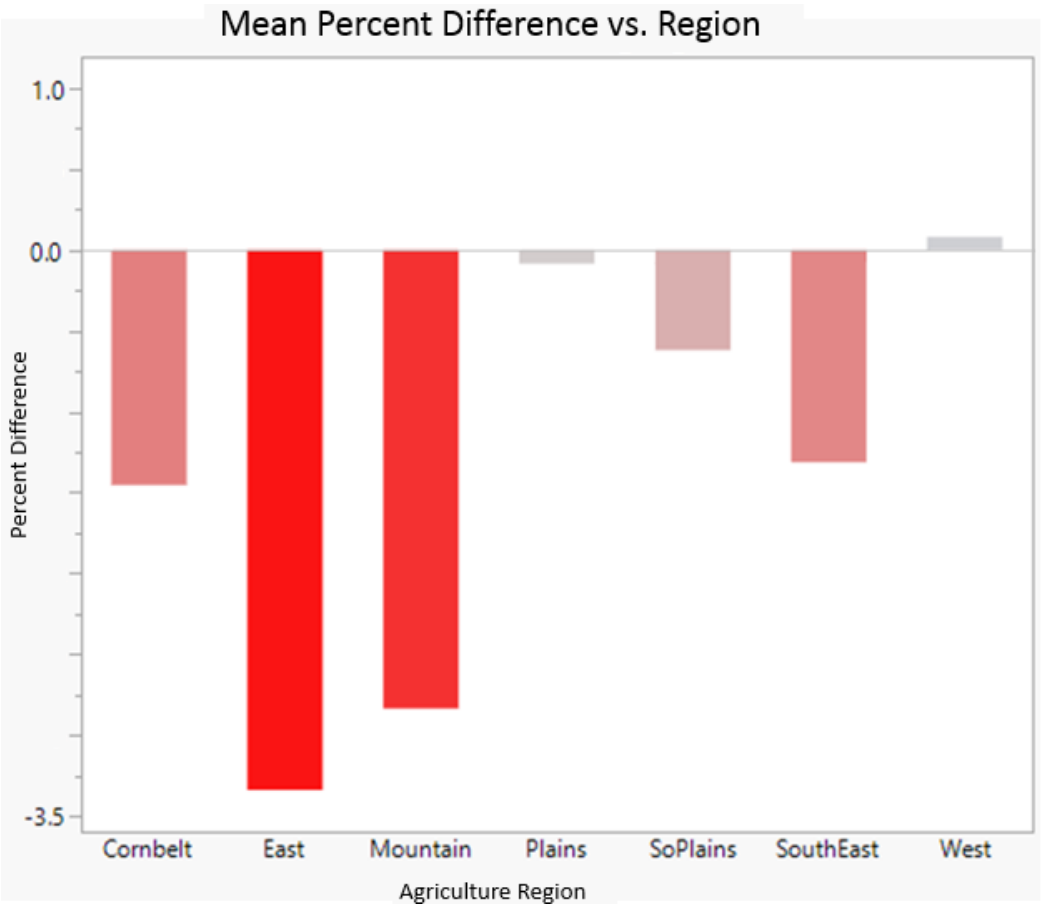


Figure 6: Mean percent difference by region for farms with any producer less than 25 years of age

Figure 6 shows mean percent difference for farms with any producer, regardless of 'principal' designation, less than 25 years of age on the same scale as Figure 5. For this variable, the study DSE estimates were less than the published DSE estimates for most of the regions; however, the West region showed the study DSE estimate to be slightly larger than the published DSE estimate. The difference in farms with any producer less than 25 years of age was found to be significant at $p < 0.01$ nationally.

4.2 Female Producers

Nationally, the study DSE estimates and the published DSE estimates were significantly different for farms with at least one female principal producer and for farms with any female producer ($p < 0.01$). Reviewing these estimates at the regional level, Figure 7 shows the study DSE estimates were less than the published DSE estimates for all of the regions for farms with at least one female principal producer.

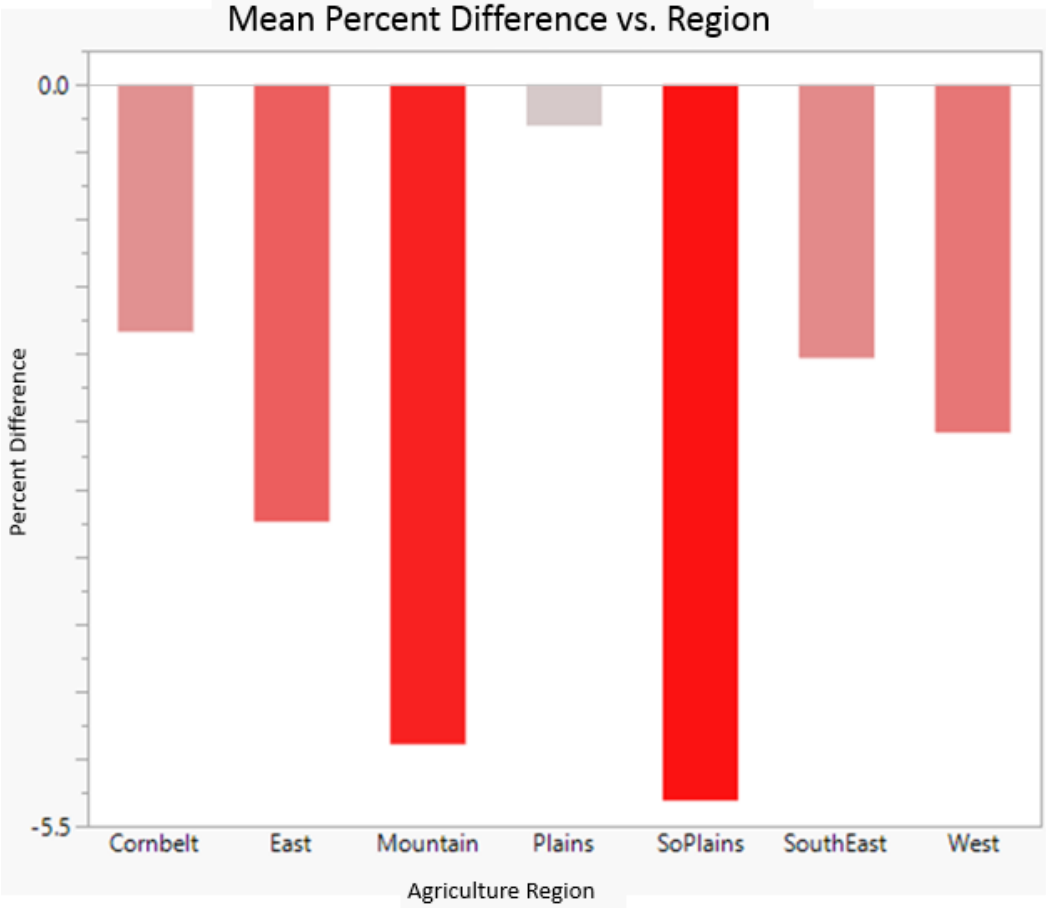


Figure 7: Mean percent difference by region for farms with at least one female principal producer

Figure 8 shows in the Plains region the study DSE estimate was greater than the published DSE estimate for farms with any female producer.

Figure 8: Mean percent difference by region for farms with any female producer

5. Discussion and Future Work

A preliminary evaluation of the effectiveness of imputing JAS producers for DSE modeling was performed. Study DSE estimates were calculated after disregarding the imputation conducted for potential producers 2, 3, and 4 on the JAS. These estimates were compared against the published DSE estimates for 28 demographic variables. In a few cases, the study DSE estimates were found to be greater than the published DSE estimates. Simulation studies are planned and further analysis will be done to determine why this occurred. Further, the demographics section of the JAS will be reviewed and potentially redesigned to allow reporting for up to four producers, which would provide consistency in the demographic data collected for the COA publications.

Acknowledgements

The findings and conclusions in this report are those of the author(s) and should not be construed to represent any official USDA or U.S. Government determination or policy.

References

Abreu, D. A., J. S. McCarthy, and 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.

Miller, Darcy. (2018). "Dancing with a New Partner: Imputing New Demographic Questions on the Census of Agriculture Using Commercial-off-the-Shelf (COTS) Software," 2018 Joint Statistical Meetings Proceedings.

N. K. Rao, J & Shao, J. (1992). Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. Biometrika. 79. 10.2307/2337236.

Pick, Kenneth, Kathy Ott, Heather Ridolfo, Jaki McCarthy, Rachel Sloan, Jeremy Beach, and Shirley Samson. (2016). *2015 Census of Agriculture Content Test: Results from Round 2 Cognitive Testing.* Washington, DC: National Agricultural Statistics Service.

Report of the Expert Panel on Statistics on Women and Beginning Farmers in the USDA Census of Agriculture. 2015. Unpublished report.

Ridolfo, Heather. (2015). *Census of Agriculture Section 6 Personal Characteristics: Results from Cognitive Testing.* Fairfax, VA: National Agricultural Statistics Service.

Ridolfo, Heather, Emilola Abayomi and Virginia Harris. (2016). "Challenges to Developing New Survey Questions: When Cultural Norms Run Counter to Survey Questions." Paper to be presented at the International Conference on Questionnaire Design, Development, Evaluation, and Testing (QDET2), Miami, FL.

Rubin, Donald and Nathaniel Schenker (1986). "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." Journal of the American Statistical Association. Vol. 81, No. 394, Pp. 366-374.

Sloan, Rachel, Heather Ridolfo, Zulma Riberas, Jaki McCarthy, Kathy Ott, and Shirley Samson. (2015). *2015 Census of Agriculture Content Test – Results from Round 1 Cognitive Testing.* Fairfax, VA: National Agricultural Statistics Service.

Young, Linda J., Lamas, Andrea C. and Abreu, Denise A. (2017). "2012 Census of Agriculture: A Capture-Recapture Analysis." Journal of Agricultural, Biological, and Environmental Statistics. 22:523-539.

Appendix

Appendix A Demographic Characteristic Variables
Farms with at least one Principal Producer less than 25 years of age
Farms with any Producer less than 25 years of age
Farms with at least one Principal Producer between the ages of 25 and 34
Farms with any Producer between the ages of 25 and 34
Farms with at least one Principal Producer between the ages of 35 and 44
Farms with any Producer between the ages of 35 and 44
Farms with at least one Principal Producer between the ages of 45 and 54
Farms with any Producer between the ages of 45 and 54
Farms with at least one Principal Producer between the ages of 55 and 64
Farms with any Producer between the ages of 55 and 64
Farms with at least one Principal Producer aged 65 or older
Farms with any Producer aged 65 or older
Farms with at least one Black or African American Principal Producer
Farms with any Black or African American Producer
Farms with at least one American Indian or Alaska Native Principal Producer
Farms with any American Indian or Alaska Native Producer
Farms with at least one Asian Principal Producer
Farms with any Asian Producer
Farms with at least one Native Hawaiian or Other Pacific Islander Principal Producer
Farms with any Native Hawaiian or Other Pacific Islander Producer
Farms with at least one White Principal Producer
Farms with any White Producer
Farms with at least one Hispanic, Latino, or Spanish origin Principal Producer
Farms with any Hispanic, Latino, or Spanish origin Producer
Farms with at least one Female Principal Producer
Farms with any Female Producer
Farms with at least one Male Principal Producer
Farms with any Male Producer