Dancing with a New Partner: Imputing New Demographic Questions on the Census of Agriculture Using Commercial-off-the-Shelf (COTS) Software

Darcy Miller[1],
[1]National Agricultural Statistics Service, 1400 Independence Avenue SW, Washington, DC 20250

Abstract
The census of agriculture (COA) is the only source of uniform, comprehensive agricultural data for every state and county in the United States. The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) conducts the COA every five years, in years ending in 2 and 7. In 2015, a panel of experts recommended that the COA update information collected about women and new or beginning farmers. Subsequently, NASS redesigned the demographics section of the 2017 COA. Some of the updates included allowing multiple principal operators and adding more than a dozen detailed farm operation decision-making questions to the 2017 COA. This major redesign to the questionnaire required changes/updates to downstream processes such as editing and imputation. We describe the changes made to the development of donor pool values for the new decision-making questions and updated methods for editing and imputation of some of the demographic variables.

Key Words: Imputation, Commercial-off-the-Shelf (COTS) Software, Editing, Agriculture, Survey Methods

1. Introduction

The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) provides timely, accurate, and useful statistics in service to U.S. agriculture. NASS has two primary programs: the census of agriculture (COA) and the agricultural estimates program. The COA is conducted every five years, in years ending in 2 and 7. Census data provide a foundation for farm policy. They are used to make decisions about community planning, company locations, availability of operational loans, staffing at service centers, and farm programs and policies. The agricultural estimates program provides reports on virtually every aspect of U.S. agriculture. Many estimates provide market-sensitive information. Both the census and agricultural estimates reports provide all market participants accurate supply/demand information for the agricultural sector simultaneously, which promotes efficiency and fairness in competitive markets.

The COA is the only source of uniform, comprehensive agricultural data for every state and county in the United States. The COA has a list frame with approximately 3 million records. The COA is the leading source of information on the characteristics of people operating U.S. farms and ranches. By USDA's definition, a farm is any place from which $1000 or more of agricultural products were produced and sold or normally would have been sold during the Census year. Understanding changes in farm structure and the

demographics of farm operators over time is important in assessing how well USDA programs serve the farm population.

After a panel of experts recommended that the COA update information collected about women and new or beginning farmers, NASS redesigned the demographics section and added more than a dozen detailed farm operation decision-making questions (referred to as the decision-making matrix) to the 2017 COA. Data collected from the new questions are unique and similar data are not collected elsewhere in the hundreds of surveys NASS conducts. This major redesign to the questionnaire required changes/updates to downstream processes such as editing and imputation.

The COA is imputed using a nearest neighbor methodology. Before each COA, an initial donor pool is formed based on records from previous COAs as well as a content test conducted for the upcoming COA. As data are collected, edited, and imputed for the current census, those records are added to the donor pool, with a preference for records to be used from the current COA. For the decision-making matrix, data from the 2012 COA or similar data from other NASS survey programs do not exist to use in the initial donor pool or imputation. This necessitated the development and implementation of a different method to create values for the initial donor pool and imputation for the decision-making matrix. With limited time to develop and update the edits and imputation, NASS selected Commercial-off-the-Shelf (COTS) software to implement an imputation methodology.

2. History of Editing and Imputation in the Census of Agriculture

The COA moved from the U.S. Census Bureau to NASS in 1997, though the Census Bureau collaborated with NASS for the 1997 COA. Until it accepted full responsibility for the data editing of the 2002 COA, NASS handled nearly all of its imputations manually. The size of the census of agriculture brought the need for automated (statistical) imputation to NASS and introduced NASS to a broader understanding of statistical data editing. The NASS Prism system was developed in-house to continue the use of decision logic tables (DLTs) for COA processing, as had been done previously at the Census Bureau. However, the Census Bureau's imputation strategy was modified in the NASS implementation of the DLTs. Editing and imputation systems are integrated for both manual imputation and statistical imputation so that editing and imputation happen as data are collected and entered into the system. The imputation does not occur at the end of the process, after all of the records are collected.

Edit logic is written by subject-matter experts and are applied in coherent “modules” of the census of agriculture report. The "conditions" portion of DLT processing identifies each data inconsistency, allowing an "action" chosen from a hierarchy of three imputation strategies. First, any value that can be determined through DLT evaluation of relevant responses (such as a missing total) is imputed. As its next choice for imputation, DLT logic makes use of previously-reported data. For COA purposes, previously-reported data are assembled from a variety of NASS surveys, as well as the previous COA, and are maintained in their own database. Donor imputation is invoked as the third option.

Donor imputation requires a pool of donors who provide values to recipients needing imputation. The donor pool membership begins with a mixture of data from the previous census and preliminary COA test data. As editing proceeds over a period of several months, recently-edited records that have passed all of the edits are used to incrementally update the donor pool. Donor data are maintained separately for each "module," which

roughly correspond to sections of the COA questionnaire. Many of the distinct donor pools function together to provide imputation during the editing of an entire COA record. Each time donor records are added or updated, all donor records are stratified using a data-driven algorithm. This algorithm groups farms by type, size and income, according to a strategy developed for each edit module and its respective donor pool. Early in the editing schedule, newer donors are favored over similar donors with older data since the initial donor pool is composed of records from the previous census and preliminary census test data.

During editing, each recipient is classified into an appropriate stratum, and the ensuing search is limited to donors in its stratum. Donor selection employs Euclidean distance computations, which are normalized across values within each stratum. The distance computation during the donor search always includes an estimated mileage between the respective county centroids. When appropriate, the donor value may be scaled before imputing the value into the recipient's record. When a recipient falls outside all current strata definitions, or when none of the donors in the recipient's stratum meet the DLT selection criteria, a backup automated strategy using donor averages may be applied. Otherwise, the record may be referred to an analyst for manual resolution.

3. Changes to the Census of Agriculture Form

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. Recognizing that farm structure has evolved into complex entities where responsibilities are often divided amongst several individuals, the panel made recommendations to change how data are collected in the demographic section of the COA.

First, the panel recommended defining operators in terms of function rather than titles. Titles such as "operator" and "principal operator" do not have universal definitions and are open to interpretation. Deference is often given to the oldest male family member, regardless of that individual's involvement on the farm or in farm related decisions. As a result, the panel recommended that NASS define operators based on function. NASS elected to use the term "person" instead of "operator" (see Figure 1).

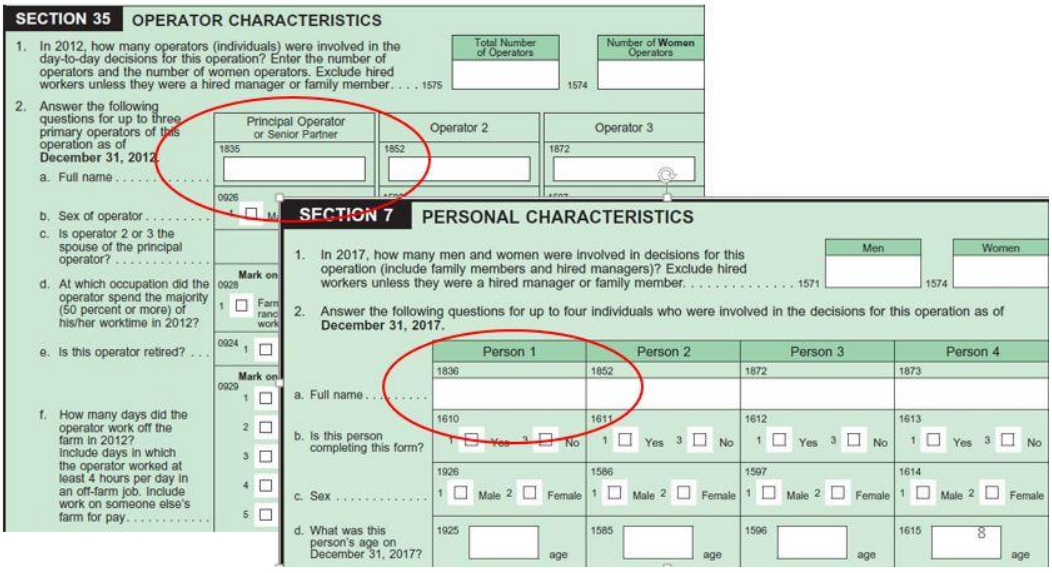


Figure 1: Select demographic questions from the 2012 Census of Agriculture (top left) and 2017 Census of Agriculture (bottom right) comparing the use of the term "operator" and "person".

Furthermore, the panel recommended removing the restriction from one "principal operator" on a farm operation to allow for joint decision making. NASS needed a "bridging" cycle so placed the "principal operator" question on the form and allowed for multiple "principal operators" to be selected (see Figure 2).

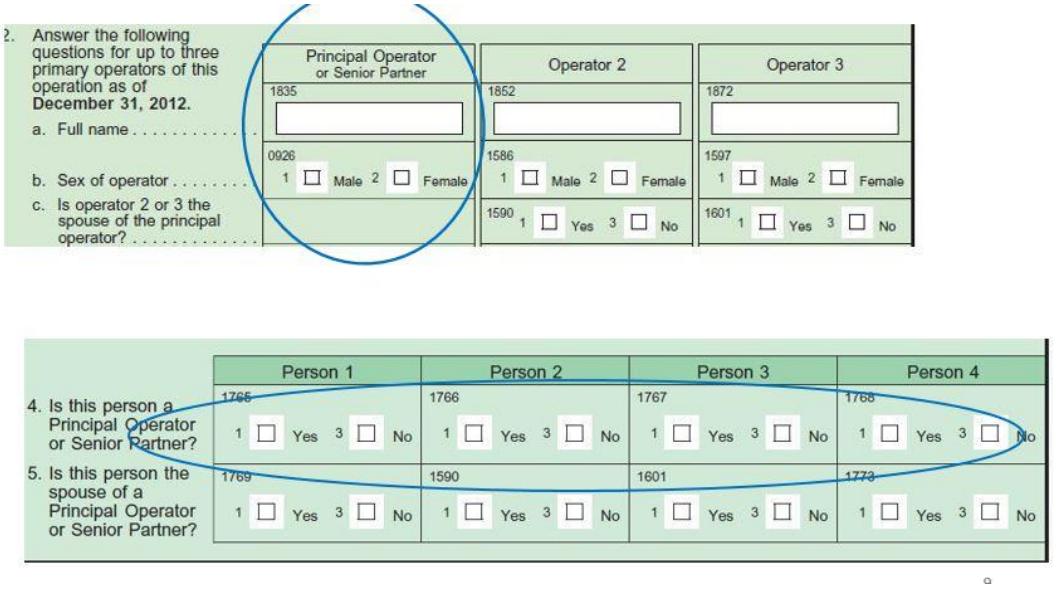
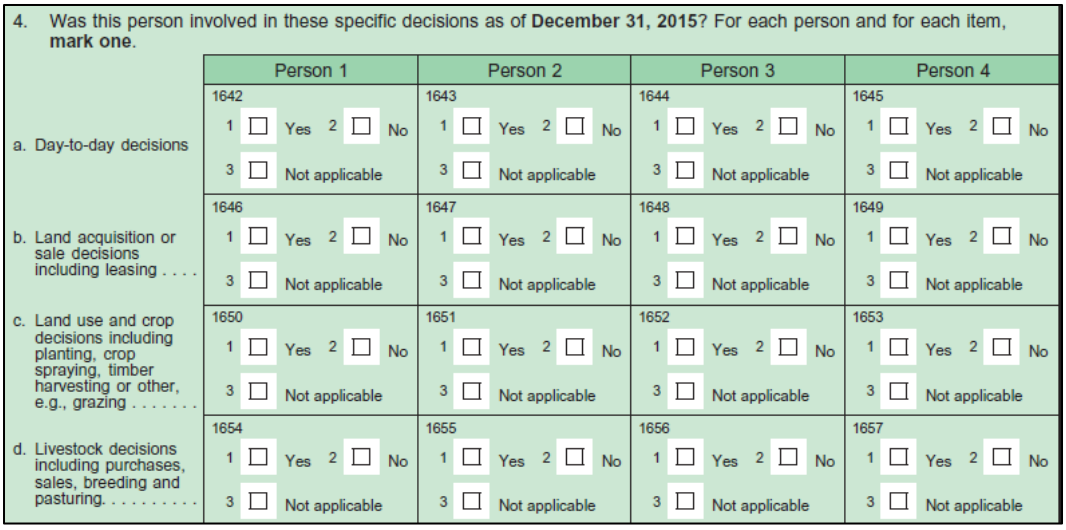


Figure 2: Select demographic questions from the 2012 Census of Agriculture (top) and 2017 Census of Agriculture (bottom) which highlight the change in the form allowing joint principal operators.

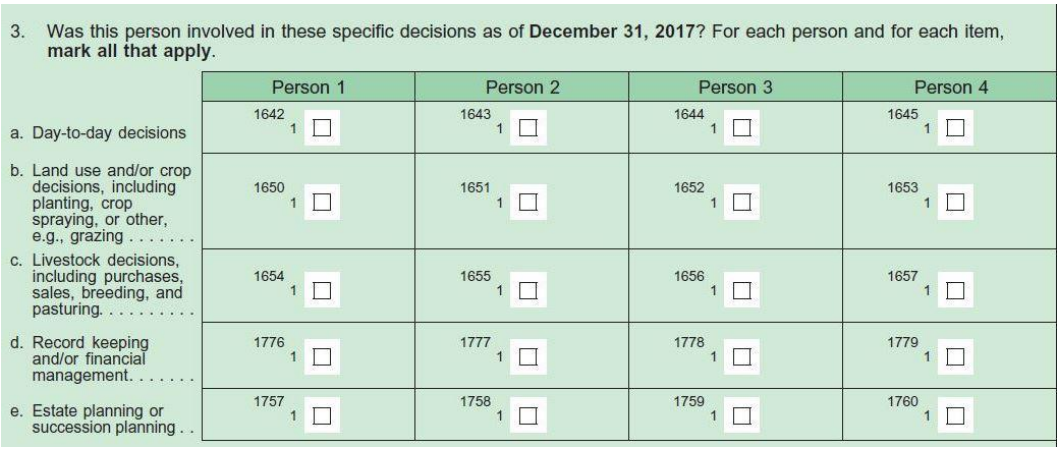
Finally, the panel recommended collecting additional data on the types of decisions made by individuals contributing to decisions on the farm. With this new content in mind, NASS designed a new set of questions to add to the 2017 COA, referred to as the decision-making matrix. The questions were tested in multiple rounds before the language was confirmed for the 2017 COA form (see Figures 3 & 4). Additional information on the panel recommendations can be found in the *Report of the Expert Panel on Statistics on Women and Beginning Farmers in the USDA Census of Agriculture* (2015). These data are not available from previous COAs or other surveys at NASS. Hence, an initial donor pool for the decision-making matrix needed to be constructed differently than the other 2017 COA questions.



4. Was this person involved in these specific decisions as of **December 31, 2015**? For each person and for each item, **mark one**.

	Person 1	Person 2	Person 3	Person 4
a. Day-to-day decisions	1642 1 Yes 2 No 3 Not applicable	1643 1 Yes 2 No 3 Not applicable	1644 1 Yes 2 No 3 Not applicable	1645 1 Yes 2 No 3 Not applicable
b. Land acquisition or sale decisions including leasing	1646 1 Yes 2 No 3 Not applicable	1647 1 Yes 2 No 3 Not applicable	1648 1 Yes 2 No 3 Not applicable	1649 1 Yes 2 No 3 Not applicable
c. Land use and crop decisions including planting, crop spraying, timber harvesting or other, e.g., grazing	1650 1 Yes 2 No 3 Not applicable	1651 1 Yes 2 No 3 Not applicable	1652 1 Yes 2 No 3 Not applicable	1653 1 Yes 2 No 3 Not applicable
d. Livestock decisions including purchases, sales, breeding and pasturing	1654 1 Yes 2 No 3 Not applicable	1655 1 Yes 2 No 3 Not applicable	1656 1 Yes 2 No 3 Not applicable	1657 1 Yes 2 No 3 Not applicable

Figure 3: Decision-making questions from one of the rounds of testing -- 2016 Census of Agriculture electronic data reporting test.



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

Figure 4: Final Decision-making questions used on the 2017 Census of Agriculture form.

4. COTS Software Selected

NASS forms a team each census cycle to lead the effort to implement the editing and imputation system for the COA, which includes individuals from several divisions at NASS. This team discussed options to create an initial donor pool for the decision matrix. They decided to make the decision-making matrix a separate module from the other demographic module and impute the COA content test using a multivariate model to create an initial donor pool for the decision-making matrix module. The Prism system continued to flag cells that required values. After the initial donor pool was formed using the imputed census content test data, the nearest neighbor methodology was utilized for the remainder of the COA records with two or less persons making decisions and a separate multivariate model was used for records with three or more persons making decisions.

To form imputed values for the initial donor pool and for farm operations with a larger number of decision makers, NASS selected COTS software to implement a multivariate

imputation method. Benefits to using COTS software such as fast development relative to scripting custom code, ease of use, ability to quickly make changes to models, reproducibility, and generalized code were appealing to NASS. These benefits outweighed the challenges to implementing the COTS software, such as meeting specific edit logic requirements and implementing the program in a process that is not modularized.

4.1 Creating the Initial Donor Pool (IVEware)

To implement the imputation for the subset of demographic variables in the COA content test and electronic data reporting test data, the team selected IVEware, an iterative multivariate imputation approach recently implemented in other NASS surveys. IVEware is a flexible imputation program developed by the University of Michigan and based on the Fully Conditional Specification (FCS) method described in Ragunathan (2001). The joint distribution is induced from a conditional specification. Parameter estimates and deviates used for imputation are generated through a Gibbs sampling routine (Geman and Geman 1984; Gelfand and Smith 1990). After initialization of this routine, sets of parameter values are drawn iteratively. For each set of parameter values, missing data are imputed based on a conditional model, where each conditional model may be linear or non-linear (e.g. generalized logit) in nature and a diffuse prior is used for the parameters. IVEware is available as a stand-alone program, or it can be run in SAS (SAS callable). It is easy to implement and NASS's familiarity with SAS led to the decision to run it in SAS.

IVEware has several modules available to perform imputation and to conduct analysis of the data. For the purpose of creating an initial donor pool using census content test data, the IMPUTE module was used. The IMPUTE module defines the model and also contains a host of other appealing features. Within the IMPUTE module, the type of regression used can be determined by defining the variable type. Variable types that can be imputed include continuous, semi-continuous, binary, categorical (polytomous with more than two categories), and counts. All variables in the dataset are potentially used in each conditional model, unless indicated in the transfer statement. The imputation programmer has options to utilize statements for model selection, such as for step-wise regression. The user also has the option to incorporate some types of edits, such as restrictions on variables to be imputed based on the value of other variables and bounded imputations.

IVEware is free, user-friendly, and easy to apply on a variety of data sources. Empirically, FCS methods, like those implemented in IVEware, have produced reasonable results (see Ragunathan et al., 2001; van Buuren et al., 2006; White and Reiter, 2008) with a high degree of variable flexibility and other desirable features for implementation by a statistical agency. However, the user accepts that convergence may not be reached due to a potential lack of a valid joint distribution. NASS has implemented IVEware for the 2014 Tenure, Ownership, and Transition of Agricultural Land (TOTAL) survey, 2016 Local Food Marketing Practices Survey, 2017 Organic Food Survey, and development of the initial donor pool for the 2017 COA decision-making questions.

4.2 Imputing Values During Production (PROC MI)

PROC MI is a procedure offered in SAS with the FCS option to implement fully conditional specification added in the launch of SAS 9.3. The FCS statement implements a two-step iterative process to impute values sequentially over the variables taken one at a time at each iteration for the number of iterations specified (van Buuren, 2006). Each

variable being imputed is specified by a different imputation model according to the data type of the variable being imputed. For continuous variables, the REG method, using a regression model, and the REGPMM method, using a predictive mean matching method, are available. For categorical variables, the DISCRIM method, which uses a discriminant function to implement imputation, and the LOGISTIC method, which uses logistic regression modeling can be used. The PROPENSITY method is also available for both continuous and categorical variables and implements a propensity score to impute values. The data being imputed for the COA using PROC MI are binary, so the LOGISTIC method was selected.

Features available in SAS PROC MI include transformation and back-transformation of variables, specifying minimum and maximum values for imputed values, rounding, and explicit selection of covariates for each variable to be imputed. The last feature of PROC MI and NASS's update from SAS 9.2 to SAS 9.4 was the primary reason for NASS to move from IVEware to PROC MI for use in production.

5. Changes to the Census of Agriculture Imputation

5.1 Creating the Initial Donor Pool (IVEware)

5.1.1 Stress Test

Outcomes from imputing NASS surveys in the past along with data types similar to the type of data found in the decision-making matrix using IVEware were assessed by NASS's operational units and research division and deemed successful. Additional stress testing of the software included running IVEware for 20 iterations on up to 500,000 units with a similar number of variables and data types as the decision-matrix. Up to 50% missing values were missing in the test data. Given that the number of records to be imputed to form the donor pool is approximately 15,000 records, IVEware can handle the task.

5.1.2 Model Development

Often, one of the most time consuming steps to implement imputation methods, including this method, is developing the models. NASS developed imputation models for the target variables (principal operator, principal operator spouse, and decision-making variables) through a combination of statistical analysis in the research and methodology divisions as well as the farm operator expertise of a NASS subject matter expert. The 2017 COA edit and imputation team met bi-monthly and additional individual communications occured between those directly involved with creating this initial donor pool. Both verbal communication as well as visual tools such as spreadsheet grids describing the models tested were used to create a successful imputation model to produce the initial donor pool for the decision-making matrix.

5.1.3 Implementation

The COA questionnaire and instrument testing occurs before the COA is administered, including processing the data (editing and imputing). For these rounds of testing, the editing and imputation are primarily done after data collection. Hence, NASS edited and imputed all of the data except for the target variables using the standard methods. These

questions were edited, but markers were placed where data would need to be imputed. Then, target variables were imputed outside of the COA processing system using IVEware. The imputed values were loaded into the database and were prepared for use in the initial donor pool.

5.2 Imputing Values During Production (PROC MI)

NASS elected to continue to impute the target variables for operations with more than three operators largely due to the complexity in updating customized code (and developing customized code) to identify the nearest neighbor which met the edit logic. With multiple changes to the questionnaire and edit logic between the testing instruments and the 2017 COA form, updates to code needed to be made relatively swiftly. NASS selected SAS PROC MI with FCS Option over IVEware due to the flexibility in building the models.

5.2.1 Model Development

First, code to impute variables needed to be adapted to accommodate changes to the form. For example, questions were collapsed or deleted and the response option to the decision-making question was changed from "Yes," "No", "N/A" to a simple presence/absence (see Figure 3 above). Most of the questions that were collapsed were easy to map from one questionnaire version to the next. However, the change in response options required more time to evaluate. Moving to presence/absence meant that deciphering between nonresponse and absence of decision-making was not as clear. Ultimately, the team concluded that if any decision on the form was checked (presence), then any absence was just absence of decision-making (i.e. the respondent read all of the decision-making questions and absence meant "No" or "N/A"). If there were not any decision-making boxes checked for any person, the entire matrix was considered as a nonresponse and required imputation. To develop the models to be used in PROC MI, NASS began with the models selected in IVEware. Then, subject matter experts provided input and added or deleted variables used in each model. Finally, a statistical assessment was conducted of the models and feedback was provided to the subject matter experts. The final model was selected using a combination of the statistical analysis and subject matter expertise. One of the most time consuming steps to implement most imputation methods, including this method, is developing the models. NASS developed imputation models for the target variables (principal operator, principal operator spouse, and decision-making variables) through a combination of statistical analysis in the research and methodology divisions as well as the farm operator expertise of a NASS subject matter expert.

5.2.2 Implementation

PROC MI, using the FCS Option, was implemented for the target variables where the record had more than two persons making decisions on the farm. This comprised 6 -7% of farms in 2012 (U.S. Census of Agriculture Full Report, 2012). Results from 2017 are not yet available for public release. The processing flow was similar to what was used for developing the initial donor pool, except there were several rounds of imputation of the target variables throughout COA data collection and processing. Since the COA is processed on a record by record basis during data collection, a set of records needed to be collected before applying the COTS software. Once a certain threshold of records had run through the other edit and imputation modules, the records were set aside and quarantined from other areas of the process until imputation of the target variables using the COTS

software were completed. At each round all records that had passed edit logic up to that point and any records that needed imputed values for target variables were used to impute records still requiring imputation. After each round of imputation, data needed to be loaded into the database, and the record needed to be removed from quarantine and allowed to be used in other COA processing steps.

Unlike records imputed using the traditional imputation method (nearest neighbor at the time of record processing), the target data were not necessarily passed through the COA edit. This was because this imputation had to be implemented into the workflow of COA edit and imputation processes. Hence, post-imputation edits were required for the set of target variables. In addition to post-imputation edits, additional variables needed to be calculated that would normally be calculated as a part of the regular COA edit and imputation process, such as, the number of women listed on the form. These post-edits and calculations comprised most of the code required to implement this COTS software into a customized process run on a record by record basis with intertwined editing and imputation steps. The required code to format the data for imputation, impute the data, and reformat the data to load into the database was small. Implementing COTS software was simpler than altering and adding to customized code, but it would have been much easier if editing and imputation were modularized.

6. Conclusion

NASS has been working to implement COTS software in its survey programs to replace customized code. So far, the use of COTS software has been primarily focused on new survey programs or where new survey data are collected in current survey programs. These are places where the benefits of agility and ease of implementation of COTS software are most realized. Otherwise, it has been used in one instance as an overall upgrade to an imputation methodology. One major challenge in implementing COTS software is fitting it in current survey processes. In this application, the COTS software was applied in a separate step from editing; the editing and imputation were modularized. For many of the smaller surveys, this requires modularizing the editing and imputation processes. In a time where technology is available with modern methodologies implemented, the benefits to using COTS software generally outweigh the work needed to address the challenges. Moving forward, NASS plans to continue to identify ways to meet this processing challenge and use COTS software instead of customized code where it is appropriate.

References

Dau, A. and Miller, D. (2018). “Dancing with the Software: Selecting Your Imputation Partner,” 2018 Joint Statistical Meetings Proceedings.

Little, R. J. A. and Rubin, D.B. (2002. “Statistical Analysis with Missing Data,” Second Edition, New York: John Wiley & Sons.

Liu, Yang et al. (2015). "Multiple Imputation by Fully Conditional Specification for Dealing with Missing Data in a Large Epidemiologic Study," International Journal of Statistics in Medical Research, August 2015, 4(3): 287-295.

Manning, A. and Atkinson, D. (2009). "Toward a Comprehensive Editing and Imputation Structure for NASS – Integrating the Parts," *USDA NASS RDD*. United Nations Statistical Commission and Economic Commission for Europe, Conference for European Statisticians, Work Session on Statistical Data Editing. Neuchatel, Switzerland, 5-7 October 2009.

Miller, D. (2017). "Creating an Initial Donor Pool for New Questions in the Census of Agriculture," United Nations Statistical Commission and Economic Commission for Europe, Conference for European Statisticians, Work Session on Statistical Data Editing. The Hague, Netherlands, 24-26 April 2017.

Miller, D., Dau, A., and Lisic, J. (2016). "Imputation's Reaction to Data: Exploring the Boundaries and Utility of IVEware and Iterative Sequential Regression (ISR)," Fifth International Conference on Establishment Surveys. Geneva, Switzerland, 20-23 June 2016.

Miller, D. and Young, Linda (2015). "Imputation at the National Agricultural Statistics Service," United Nations Statistical Commission and Economic Commission for Europe, Conference for European Statisticians, Work Session on Statistical Data Editing. Budapest, Hungary, 14-16 September 2015.

Miller, D., Ridolfo, H., Harris, V., McCarthy, J., and Young, L. (2015). "Expert Panel on Federal Statistics on Women and Beginning Farmers in U.S. Agriculture," Documentation for the Expert Panel on Federal Statistics on Women and Beginning Farmers in U.S. Agriculture. Washington, DC. 2-3, April 2015. Unpublished Report.

Oudshoorn, C.G.M., van Buuren, S., and Rijckevorsel, J. L. A. (1999). "Flexible Multiple Imputation by Chained Equations of the AVO-95 Survey," TNO Prevention and Health, TNO Report PG/VGZ/99.045.

Raghunathan, T.E., Lepkowski, J.M., Hoewyk, J.V. and Solenberger, P. (2001). "A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models," Survey Methodology, 27, 85-95.

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

Report from the 2012 Census of Agriculture, Volume 1, Chapter 2, Table 45. https://www.agcensus.usda.gov/Publications/2012/Full_Report/Volume_1,_Chapter_2_US_State_Level/st99_2_045_045.pdf

Ridolfo, H., Harris, V., McCarthy, J., Miller, D., Sedransk, N., and Young, L. (2016). "Developing and Testing New Survey Questions: The Example of New Questions on the Role of Women and New/Beginning Farm Operators," Fifth International Conference on Establishment Surveys. Geneva, Switzerland, 20-23 June 2016.

Rubin, D. B. (1987). "Multiple Imputation for Nonresponse in Surveys,"New York: John Wiley & Sons.

Schafer, J.L. (1997). "Analysis of Incomplete Multivariate Data," Chapman & Hall/CRC.

van Buuren , S., Brand , J. P.L., Groothuis-Oudshoorn, C. G.M., and Rubin, D.B. (2006). "Fully conditional specification in multivariate imputation," Journal of Statistical Computation and Simulation, 76:12, 1049-1064, DOI: 10.1080/10629360600810434

Vizcarra, B. and Sukasih, A. (2013). "Comparing SAS PROC MI and IVEware Callable Software," Proceedings from 2013 Southeast SAS Users Group (SESUG) Conference.
de Waal, T., Pannekoek, J., and Scholtus, S. (2011). "Handbook of Statistical Data Editing and Imputation". Wiley Handbooks in Survey Methodology. John Wiley & Sons, Inc.