Evaluating Imputation Methods for the Agricultural Resource Management Survey

Darcy Miller[1] , Andrew Dau[1] and Audra Zakzeski[1]
[1]United States Department of Agriculture – National Agricultural Statistics Service, 1400 Independence Avenue, Washington, DC 20250

Abstract
The National Agricultural Statistics Service (NASS), in conjunction with the Economic Research Service (ERS), conducts the three-phase Agricultural Resource Management Survey (ARMS) to study the economic well-being of farm households. Since 2015, Iterative Sequential Regression (ISR), a multivariate imputation methodology, has been used to address item nonresponse in the third phase of the survey (ARMS 3). ISR is an in-house developed software program that requires a significant amount of support to maintain. Also, ISR was developed for use on continuous and semi-continuous data, and NASS wants to impute other data types, including categorical and ordinal data. Hence, NASS is exploring alternative "off-the-shelf" imputation approaches, specifically, IVEware, a product of the University of Michigan, and the Fully Conditional Specification Option in SAS® PROC MI. A 2018 JSM paper empirically compared ISR to these two alternatives using a subset of ARMS 3 data. This paper builds on that simulation work and culminates in an impact assessment of a change to one of the alternatives on reported estimates and operational resources through an application to the full ARMS 3 dataset.

Key Words: Imputation, SAS® PROC MI, Agriculture

1. Background

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) is responsible for the publication of over 400 agricultural statistical publications annually. Production and supplies of food and fiber, prices paid and received by farmers, farm labor and wages, farm finances, chemical use, and changes in the demographics of U.S. producers are only a few examples of the many publications produced by NASS (USDA, 2018).

A majority of NASS publications are driven by data collected via survey. The Agricultural Resource Management Survey (ARMS) is conducted annually through a joint effort between NASS and the Economic Research Service (ERS). The ARMS provides an annual snapshot of the financial health of the farm sector and farm household finances. The ARMS is the only source of information available for objective evaluation of many critical policy issues related to agriculture and the rural economy (Farm, 2018).

NASS conducts the ARMS in three phases. The initial phase (ARMS Phase 1) screens a large sample of farms and ranches to determine which farms qualify for subsequent phases of ARMS. Subsamples of qualifying farms are selected for the other two phases. The second phase (ARMS Phase 2) collects data on agricultural production practices, chemical use, and costs of production for designated commodities. ERS determines the commodity

rotation and is responsible for estimating the cost of production for major commodities from the data NASS collects (Farm, 2018).

The third phase (ARMS Phase 3) collects whole farm finance and operator characteristics for a calendar year. Respondents from the second phase are included in the third phase to obtain financial and farm production expenditure data for the operation. It is vital that both the ARMS Phase 2 and the ARMS Phase 3 be completed for these designated crop commodity operations. Data from both phases provide the link between agricultural resource use and farm financial conditions, and allows for economic impact analysis of regulation and policy. This is a cornerstone of the ARMS design. In addition, costs of production, and farm production-expenditure data for designated livestock commodities are collected in one interview during the third phase (Farm, 2018).

NASS has worked in recent years to increase awareness of the importance of the ARMS, while also taking measures to reduce respondent burden. Despite those efforts, unit and item level non-response still remain high on the ARMS Phase 3. One potential source of non-response on the ARMS 3 comes from its 24 page length (Roszkowski, 1990). Another source of non-response stems from the nature of questions that are asked in order for the ARMS Phase 3 to successfully fulfill its goals. Some of those questions ask about potentially sensitive personal and financial information in order to properly assess the financial health of farms. Figure 1 below shows an example of a question that is commonly refused due to its sensitive nature surrounding the personal finances of respondents.

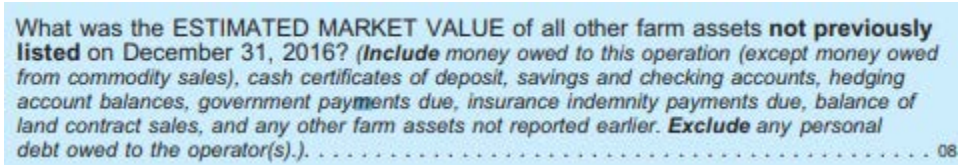

Figure 1. Question asking for personal financial information on the ARMS Phase 3.

Lastly, the ARMS asks questions about information that may not be directly available to the respondent. Figure 2 below shows an example of a question asked on the ARMS that is difficult for a respondent to answer. The question asks about an expense paid by their landlord, which is often times unknown to the respondent.

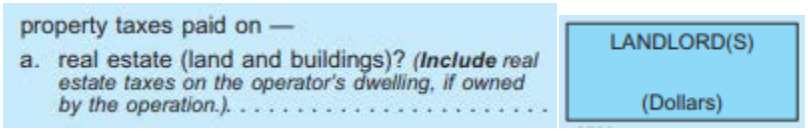

Figure 2. Question asking about landlord information on the ARMS Phase 3.

2. ARMS Phase 3 Survey Process

The ARMS Phase 3 survey process has many steps that can affect the operational viability of any new process that is implemented. An understanding of the timing and necessity of each process will impact decision making that is presented later in this paper. Figure 3 below shows the abbreviated survey process and how it is executed between January and August annually.

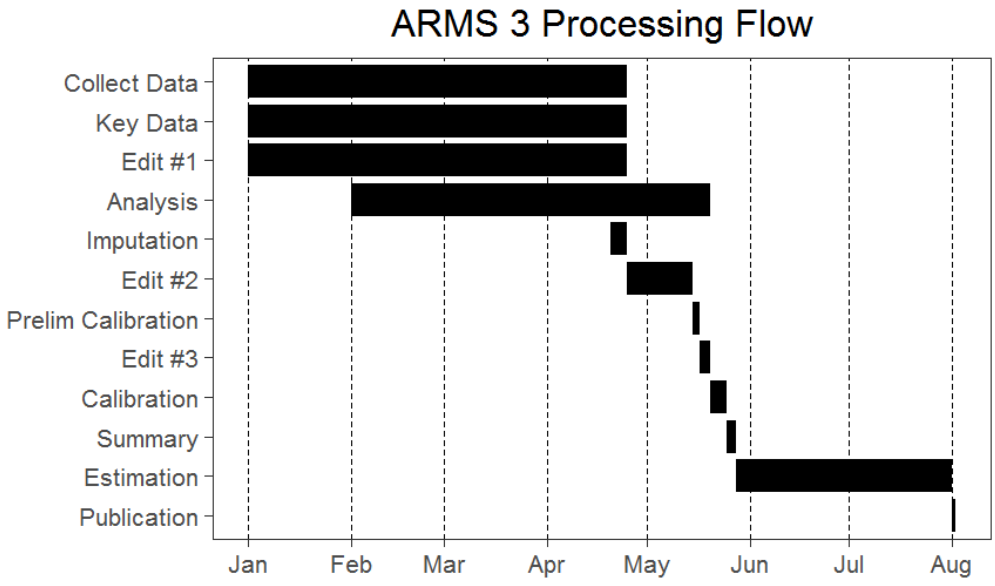

Figure 3. Gantt chart of the ARMS 3 survey processing flow.

2.1 ARMS Pre-Imputation Processes

Prior to imputation, several steps occur that have an impact on the resulting imputation procedures. In particular, the "Edit #1" phase really sets the table for future imputation work. The edit phase involves a complex computer edit system, which flags different levels of errors and either uses pre-programmed methodology to fix the errors, or asks the analyst for manual intervention to resolve the errors. In addition to resolving errors, the computer edit also flags missing variables that require imputation. A combination of the edit and editors have determined that flagged variables must be non-zero. This is a very important step because in the NASS imputation process for ARMS 3, a value of zero should rarely (if ever) be returned from any imputation module.

2.2 ARMS Imputation

Once an initial edit has been performed, imputation is required for missing data in selected variables such as landlord tax expenses, contractor's marketing expenses and more. Prior to 2014, missing data on the ARMS Phase 3 was imputed using a conditional mean approach. Data were subset into similar groups using a combination of farm type, size of farm, and location. Then a mean was computed for each group and imputed for the missing data values.

Because ARMS Phase 3 has many complex multivariate relationships, the conditional mean imputation methodology used prior to 2014 could not generally condition on all variables that might be in a multivariate imputation. Therefore, some important relationships between variables were not used in these imputations. To incorporate more information when conducting imputation, NASS collaborated with the National Institute of Statistical Sciences (NISS) to develop an alternative imputation methodology. Iterative sequential regression (ISR) was adapted to ARMS Phase 3 and implemented for the 2014 survey year.

ISR is founded on the normal distribution. Many of the ARMS Phase 3 data are semi-continuous with a large proportion of records reporting zero, and a normal distribution among the remaining positive records. For example, for an item such as number of cattle, a large number of records may not have any cattle (i.e. report zero) and the remaining records reporting cattle will follow a normal distribution. Thus, the semi-continuous nature of many of the variables in the ARMS Phase 3 requires special handling. To handle the probability mass at zero, an indicator variable is constructed for each item to denote whether a value of the item is non-zero or zero. Marginal transformations of the non-zero, continuous portion of each variable are then joined to form a multivariate normal joint density. The multivariate joint density is decomposed into a series of conditional linear models, and a regression-based technique is used to produce values to impute.

Subject-matter experts select the covariates, which allows for flexibility in the selection of the covariates while still providing a valid joint distribution. Parameter estimates for the sequence of linear models and imputations are obtained in an iterative fashion using a Markov-chain-Monte-Carlo (MCMC) sampling method. The ISR method is a blend of data augmentation (DA) and fully conditionally specified (FCS) models, having the covariate choice flexibility of the FCS methods but the theoretical background of the DA methods (See Robbins, et al. 2013 for more details). A similar joint modeling approach was used for the EM algorithm in Lipsitz and Abrahim (1996) and Abrahim, Lipsitz, and Chen (1999).

2.3 ARMS Post-Imputation Processes

Following imputation, the data are processed again through a computer edit (Edit #2). It is due to this edit process and the need for a singular dataset for researchers, that NASS uses a singular imputation approach for the ARMS Phase 3. This second edit examines reasonableness of multivariate relationships within the imputed data at a record level. Also, now that the imputed data are present, the additional edit checks and analyses are executed. After the computer/analyst resolves all the errors, the data are considered clean and continues into the calibration and summary phases. During the calibration and summary phases, an outlier board is held where outlying weighted values are reviewed and weights may be adjusted.

3. Motivation

Since 2014, ISR has served NASS well for the purposes of the ARMS imputation. However, commercial off the shelf (COTS) approaches to imputation may reduce ongoing program maintenance and provide expanded flexibility in imputation.

First, ISR currently lacks the flexibility to impute categorical or ordinal data. Recently, ERS has examined methods to extend ISR to impute ordinal data using the Anderson-Darling Method to fit an estimate density to the observed data (Burns, 2015). It is possible a similar extension could be developed to focus on imputing categorical values as well. However, extending ISR in this way would require substantial capital investment in software development and maintenance. This leads to a second motivation to examine COTS solutions.

Resources to maintain the ISR program are limited. Currently, ISR is housed on aging hardware, and it will eventually need to be migrated to a different platform. In addition, as staffing changes occur, there is a growing disconnect from the original developers of ISR. This could create issues if any debugging is needed, as the person debugging will have to spend a significant amount of time and resources to understand the current methodology and how it is programmed. For NASS purposes, it is ideal if more time could be spent on the imputation models themselves, and less time on the underlying imputation code. COTS solutions would provide that opportunity.

Lastly, the ARMS program at NASS is not the only survey program that requires imputation. Currently, a variety of different methodologies are applied on a survey by survey basis. COTS would potentially provide the ability to standardize imputation processes across survey platforms. The capital investment to extend ISR methodology to other surveys would be quite large, and in some cases may not be the most viable solution.

4. Goal

The goal of this research is to examine two COTS solutions that use multivariate approaches for imputation and evaluate the impact of changing the imputation method.

Most of NASS production work is executed using SAS, so to study the first part of this goal, the focus was primarily on two COTS solutions that could be executed in SAS: IVEware and PROC MI. The predictive mean matching (PMM) method within PROC MI's fully conditional specification (FCS) option was particularly appealing since it would allow NASS to relax assumptions of normality, specify models for each variable imputed, and use respondent values as imputed values. The last reason listed eases the transition to a new method, since NASS has more experience with imputation methods that use reported values as the imputed values. A simulation study was conducted in 2018 to reach the first part of the goal. The conclusion of the study was that PMM implemented using SAS PROC MI was the best replacement for ISR in terms of balancing data quality and operational wants/needs (Dau, A. et al., 2018).

The second part of the goal is to evaluate the impact of changing the imputation method using the full ARMS 3 data. The purpose is to project what change in estimates during the next cycle will be industry change and what part of the change is due to using a different imputation method. PROC MI using the FCS option and PMM where appropriate was selected as the best alternative, so research into the second part of the goal continued with this imputation method. This paper presents initial results from an impact assessment when moving from using ISR to PROC MI.

4.1 SAS PROC MI

As an alternative to IVEware, PROC MI has been developed by and is available in SAS. The MI procedure is a multiple imputation procedure that creates multiple imputed data sets for incomplete p-dimensional multivariate data. It uses methods that incorporate appropriate variability across the *m* imputations. The imputation method of choice depends on the patterns of missingness in the data and the type of the imputed variable.

Flexibility is a strength of the MI procedure as it can handle both monotone and arbitrary missing patterns. The data for a continuous variable with a monotone missing pattern can be imputed using a regression method (Rubin 1987), a predictive mean matching method (Heitjan and Little, 1991), or a propensity score method (Rubin, 1987; Lavori, Dawson, and Shera 1995). For a categorical variable, a logistic regression method or a discriminant function method can be used depending on whether the variable is binary, nominal, or ordinal.

Data sets that have an arbitrary missing data pattern, similar to ARMS Phase 3, can use either a Markov-chain Monte-Carlo (MCMC) method (Shafer 1997) or a fully conditional specification (FCS) method (Brand 1999; Van Buuren 2007). Similar to data with a monotone missing pattern, continuous variables can be imputed using a regression method or a predictive mean matching method. Furthermore, categorical variables can be imputed using a logistic regression method or a discriminant method depending on whether the variable is binary, nominal, or ordinal.

Empirically, FCS methods, like those implemented in PROC MI (FCS option), have produced good 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, convergence may not be reached due to a potential lack of a valid joint distribution (Miller 2015) and uses diagnostics to assess if the process has converged.

Several options are built into the MI procedure. The SAS MI procedure user guide details these. A few options that were explored during this ARMS Phase 3 research included TRANSFORM, ROUND, MINIMUM, and MAXIMUM. The TRANSFORM statement allows the user to transform variables prior to the imputation process and automatically reverse transforms the data back. The ROUND option allows the user to specify the magnitude for which the resulting imputed data should be rounded. Lastly, MINIMUM and MAXIMUM allows the user to set bounds for the imputed data. SAS deploys PROC MI within its SAS/STAT product and for this research SAS 9.4 with SAS/STAT 14.1 was used (SAS, 2015).

5. Impact Assessment

For this study, analysis was conducted on the impact of using PROC MI vs ISR (the current method), as the imputation process. Currently, a single imputation is used for ARMS 3, so the evaluation begins under this condition. Confidence intervals for the differences in estimates from ARMS 3 when using PROC MI and ISR across multiple years of ARMS 3 data will be examined. Due to the partial government shutdown earlier in the year, the resources to complete the assessment, were significantly reduced. Results are shown for the 2013 ARMS 3 survey year.

4.1 Methods

The 2013 ARMS Phase 3 dataset was imputed using both ISR and PMM. The data were calibrated and summarized using operational standards. Data were not passed through the post-imputation edit and analysis routines where outliers or records with significant impacts would be further scrutinized and values or weights adjusted accordingly. Replicating this part of the process was not possible.

After calculating a 95% confidence interval for the difference in the estimates (PMM estimate – ISR estimate), confidence intervals for the difference in estimates as a percentage change from ISR (the method currently used) are provided. Estimates at the national level, regional level and state level (where published) are examined.

4.2 Preliminary Results

The following charts display 95% confidence intervals for the differences in select estimates from ARMS 3 when using PROC MI (PMM method) and ISR on 2013 ARMS 3 data before post-imputation processing (editing and analysis). The vertical axis denotes the estimate level (i.e. U.S., Region, State). The horizontal axis represents the percent change from the current method (ISR). Intervals that overlap zero indicate no significant change in the estimate; intervals entirely above zero indicate larger estimates using PMM. Keep in mind that without post-imputation editing and analysis, these results are the same as results would be had PMM replaced ISR within the full NASS process. Estimates that NASS produce that contain imputed values are included in the results as well as an estimate used by ERS. ARMS 3 uses a calibration technique to reweight the data, and some of the variables used to calibrate will have components that are imputed. An estimate that NASS produces that does not contain imputed values is also included to demonstrate the effect imputation can have on calibration. Selected estimates provided in this section are listed in the table below.

Table 1. List of estimates and corresponding imputation level and user.

Estimate	Imputation Status	User
Figure 4. Fuel Expenditures	Zero Imputed Values	NASS
Figure 5. Assets	Contains Greater than 30% Imputed Values	ERS
Figure 6. Farm Services Expenditures	Contains 10-30% Imputed Values	NASS
Figure 7. Tax Expenditures	Contains Greater than 30% Imputed Values	NASS
Figure 8. Total Expenditures (sum of all of the expenditure estimates)	Contains Less than 10% Imputed Values	NASS

Figure 4 examines the impact of changing the imputation method from ISR to PMM (implemented using PROC MI) for the published estimate, fuel expenditures. The e estimate for Fuel Expenditure is fully observed and does not require imputation; however, it is evident that different estimates are obtained when imputing using each method. Imputed values for other variables affect the economic class of a unit and the data are calibrated within economic class domains. Any unit being assigned to one economic class for data imputed using ISR and another economic class for data imputed using PMM (implemented using PROC MI) leads to different estimates. Calibration tends to make larger changes to weights in the higher economic classes, so domains where units are assigned to different higher economic classes (larger farm operations) can see even larger changes due to calibration than when units are in different lower economic classes. The impact due to calibration is relatively small; other estimates without imputed values produced similar results.

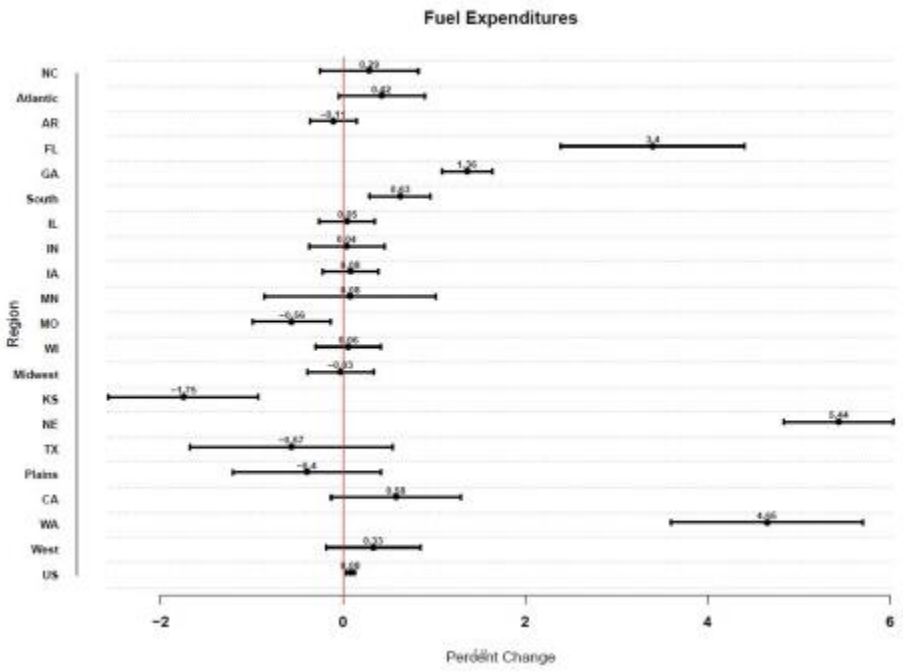

Figure 4. 95% confidence intervals for the difference in fuel expenditure estimates (PMM – ISR) expressed as a percent of the ISR estimate.

Results from comparing assets and tax expenditures estimates are shown in figures 5 and 6, respectively. The assets estimate is used in ERS research; the tax expenditures estimate is one of the key estimates NASS publishes in its Farm Production Expenditures Report. Both estimates contain a large number of imputed values. Changes are larger than for estimates without imputation, such as fuel expenditures (see figure 4), especially in domains with larger farms such as Florida. It is expected that some of this change to be muted by the post-imputation editing and analysis, especially in Florida where a small number of big farms have relatively large weight changes between the ISR imputed dataset and PMM imputed dataset.

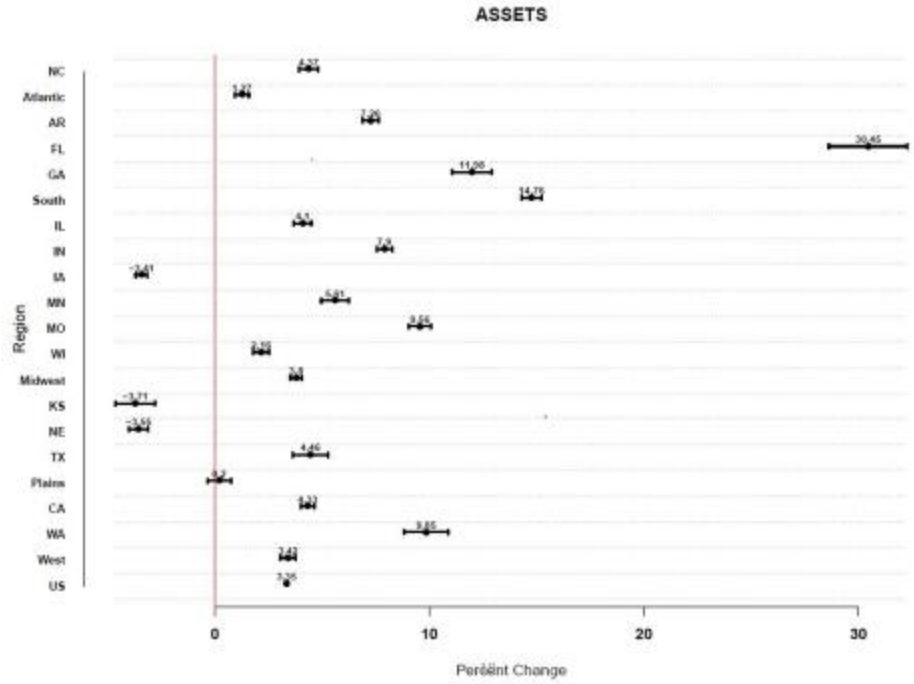

Figure 5. 95% confidence intervals for the difference in the assets estimates (PMM – ISR) expressed as a percent of the ISR estimate.

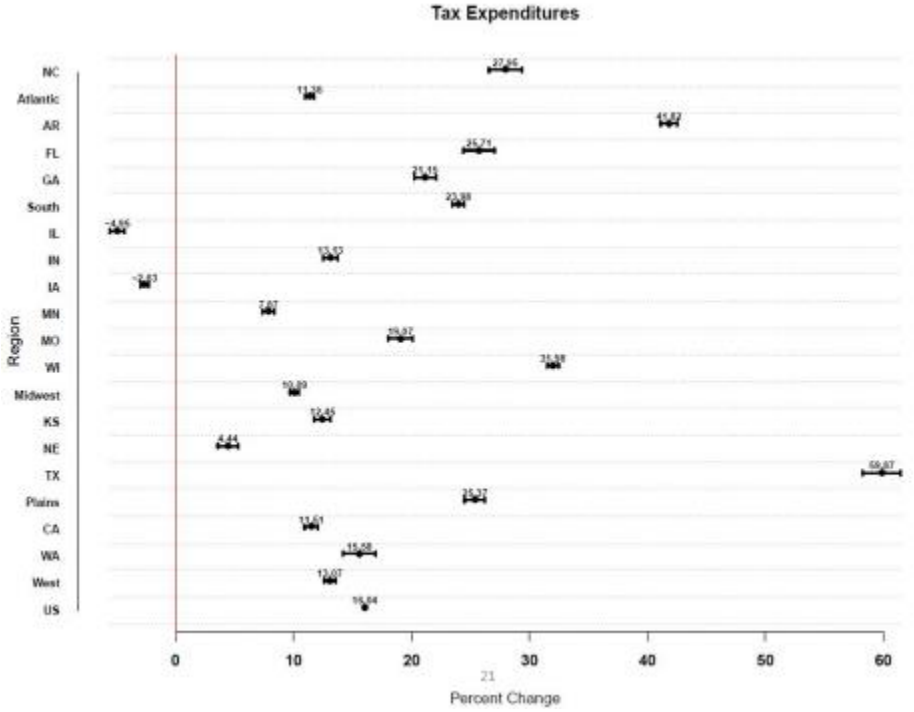

Figure 6. 95% confidence intervals for the difference in the tax expenditures estimates (PMM – ISR) expressed as a percent of the ISR estimate.

Figure 7 examines the impact of changing the imputation method from ISR to PMM (implemented using PROC MI) for an estimate with a medium number of imputed values, farm services expenditures. As expected, changes are smaller relative to the assets and tax expenditures. Some of the larger, but still small, changes are seen where the interaction with calibration is greater.

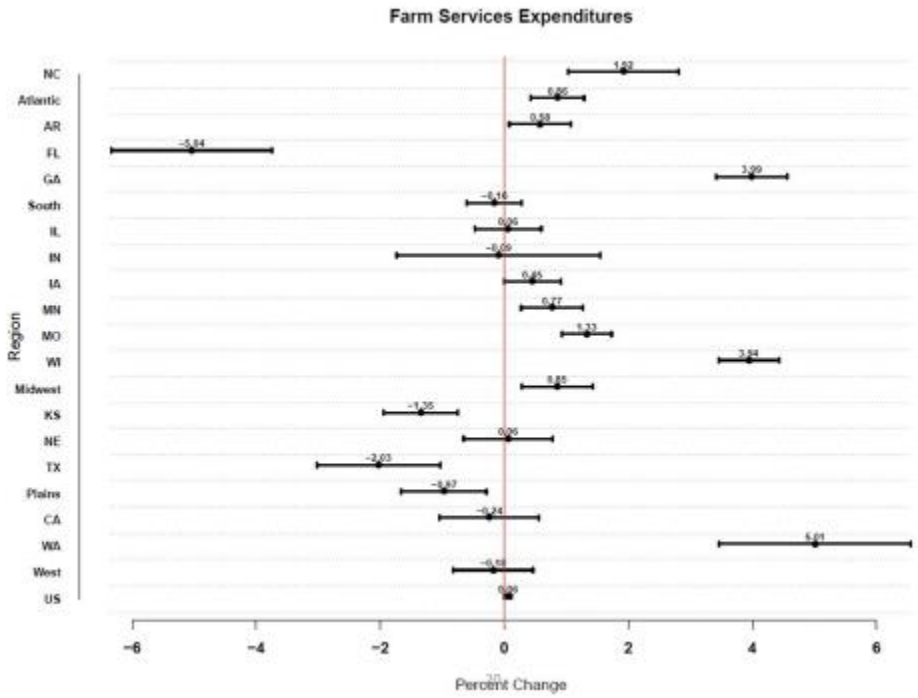

Figure 7. 95% confidence intervals for the difference in the farm services expenditures estimates (PMM – ISR) expressed as a percent of the ISR estimate.

Results from comparing total expenditures estimates are shown in figure 8. Total expenditures is the sum of underlying published expenditures, such as farm services expenditures and tax expenditures, and is also published in NASS's annual Farm Production Expenditures report. Since the estimate is a combination of other estimates, most of which do not contain any imputed values, the number of imputed values contained in the total expenditures estimate is small. Results show that the expected change in the estimate when moving from ISR to PMM would be small. Analysis of the weighted values revealed thatost of the change seen will be due to an interaction with the calibration routine (e.g. records changing economic classes in the larger economic classes). According to NASS expert operational staff, changes seen in figure 8 would be dampened by the post-imputation edit and analysis where weights and values can still be changed.

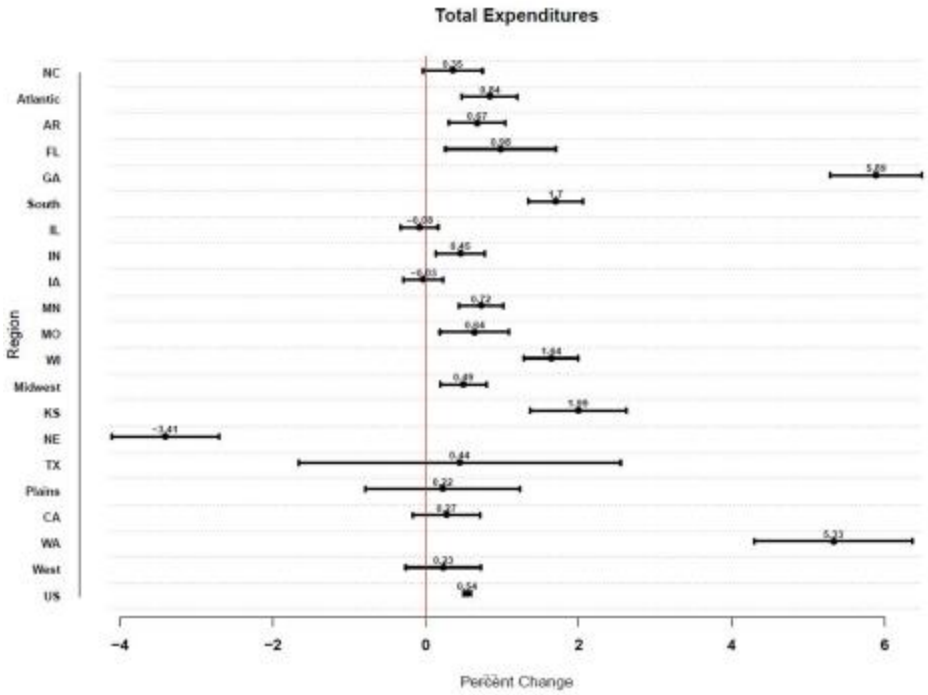

Figure 8. 95% confidence intervals for the difference in the total expenditures estimates (PMM – ISR) expressed as a percent of the ISR estimate.

6. Conclusions

A study comparing ISR (current imputation method) to IVEware and PROC MI as alternatives along with known operational preferences and needs pointed to utilizing PROC MI (PMM option) to impute ARMS 3 data. Additional work was requested to assess the impact of making a change in imputation method and program from ISR to PMM implemented with PROC MI. Although this type of study with ARMS 3 data has its challenges, a couple of conclusions bubble to the surface. Much of the empirical evidence from application to the 2013 ARMS data confirm intuitive thoughts. Examining 2013 ARMS data, results showed that the more data that are imputed, the larger the change. In addition, the interaction between imputation and the calibration routine has potential to be much larger in states that have large farm operations. Weight adjustments are proportionally larger in higher economic classes, because response rates are lower. So, any imputation that changes the economic class at the higher economic class level between the two methods would have a relatively large impact on the change. Determining what the difference actually would have been if PMM was used instead of ISR is not possible due to the inability to replicate the manual post-edit and analysis phase. However, continuing analysis to additional years of ARMS 3 data provide will provide some insights into the magnitude and direction of the impact when changing the imputation method which will assist staff in explaining changes seen in the time series.

References

USDA – National Agricultural Statistics Service – About NASS – Agency Overview. (2018). https://www.nass.usda.gov/About_NASS/

Farm Production Expenditures Methodology and Quality Measures. (2018). https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Farm_Production_Expenditures/08_2018/fpxq0818.pdf

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", New Jersey: John Wiley & Sons, 2nd ed.

Miller, D., Robbins, M., and Habiger, J. (2010). "Examining the Challenges of Missing Data Analysis in Phase Three of the Agricultural Resource Management Survey". Proceedings of the 2010 Joint Statistical Meetings, pages 816-829

Miller, D. and Dau, A. (2015). "Capturing Additional Variability Introduced by Imputation within the Agricultural Resource Management Survey". 2015 Joint Statistical Meetings Proceedings.

Robbins, M., Ghosh, S., Goodwin, B., Habiger, J., Kosler, J., Miller, D., and White, K. (2011). "ARMSimpute: A Computation Algorithm for Imputation in ARMS III." Tech. re., National Institute of Statistical Sciences/National Agricultural Statistics Service.

Roszkowski, M., & Bean, A. (1990). Believe It or Not! Longer Questionnaires Have Lower Response Rates. Journal of Business and Psychology, 4(4), 495-509. Retrieved from http://www.jstor.org/stable/25092255

Vizcarra, B. and Sukasih, A. (2013). "Comparing SAS PROC MI and IVEware Callable Software". 2013 SouthEast SAS Users Group Conference Proceedings.