Dancing With the Software: Selecting Your Imputation Partner

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

Abstract
The United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS), in conjunction with the USDA Economic Research Service (ERS), conducts the three-phase Agricultural Resource Management Survey (ARMS) to study the economic well-being of farm households. Due to item nonresponse, some of the ARMS data are missing. Prior to 2015, a complete data set for use by NASS was formed by a mixture of conditional mean imputation and manual imputation. Since 2015, Iterative Sequential Regression (ISR), a multivariate imputation methodology, has been used for ARMS's third phase (ARMS 3). ISR is an in-house developed software program that requires a significant amount of support to maintain. Also, ISR has been developed for use on continuous and semi-continuous data, and NASS needs to impute other data types including categorical and ordinal data. Hence, NASS is exploring alternative commercial off-the-shelf (COTS) imputation approaches, specifically, IVEware, a product of the University of Michigan, and SAS® PROC MI. ISR, IVEware, and PROC MI are empirically compared for use in the ARMS 3 survey with attention not only given to data quality but also to ease of implementation and maintainability.

Key Words: Imputation, SAS® PROC MI, IVEware

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 of NASS and the USDA 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 allow 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 on the ARMS Phase 3. One potential source of non-response on the ARMS 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 question that is commonly refused due to its sensitive nature surrounding the personal finances of respondents.

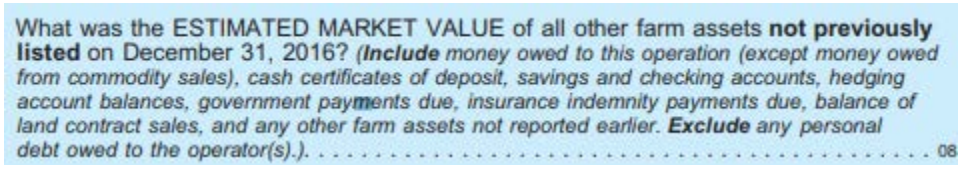
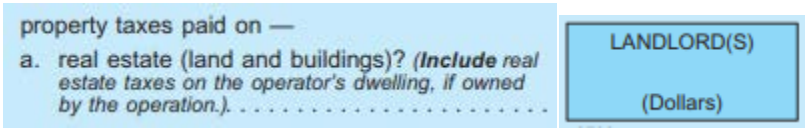

What was the ESTIMATED MARKET VALUE of all other farm assets **not previously listed** on December 31, 2016? (***Include*** *money owed to this operation (except money owed from commodity sales), cash certificates of deposit, savings and checking accounts, hedging account balances, government payments due, insurance indemnity payments due, balance of land contract sales, and any other farm assets not reported earlier.* ***Exclude*** *any personal debt owed to the operator(s).).* 08

Figure 1. Question asking for personal financial info 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.


property taxes paid on —

a. real estate (land and buildings)? (***Include*** *real estate taxes on the operator's dwelling, if owned by the operation.).* . . . . . LANDLORD(S) (Dollars)

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 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 executes between January and August annually.

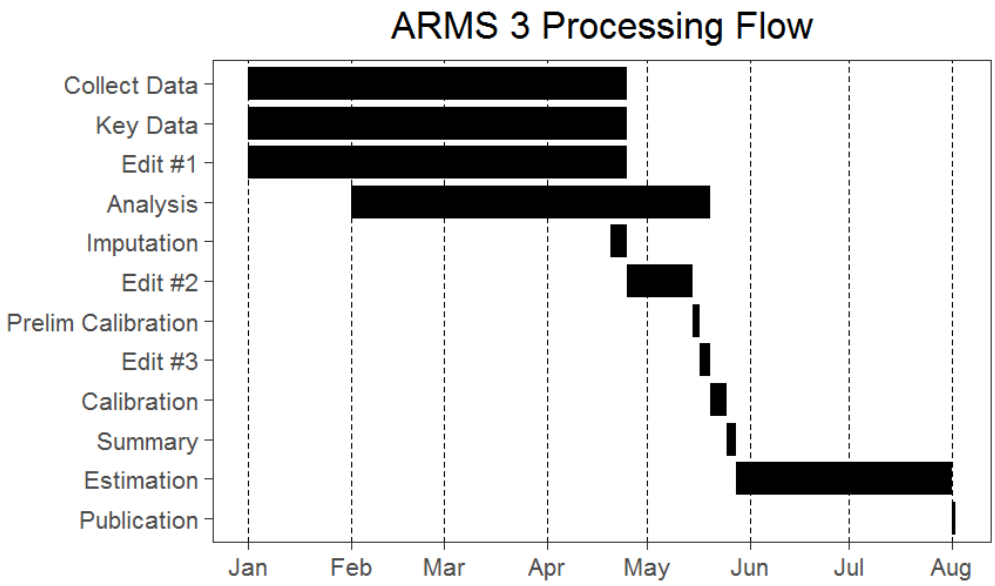


Figure 3. Gantt chart of the ARMS 3 survey processing flow.

2.1 ARMS Pre-Imputation Processes

Prior to imputation, several steps that have an impact on the resulting imputation procedures occur. In particular, the "Edit #1" phase prepares the dataset 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. These missing flagged variables must be non-zero. This is an important step because in the NASS imputation process for the ARMS Phase 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. Prior to 2014, missing data on the ARMS Phase 3 was imputed using a conditional mean approach. Data were subsetted into similar groups using a combination of farm type, size and location and a mean was computed for each group. The mean calculated was then imputed for the missing data values.

Because ARMS Phase 3 has 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 among variables were not used. 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)

(Robins, et al. 2013) was adapted to ARMS Phase 3 and implemented for the 2014 survey year.

ISR is founded on the normal distribution. The ARMS Phase 3 data often have a probability mass at zero. For example, for an item such as feed expense, a large number of records may not have any feed expense (i.e. report zero). 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 in the imputation process.

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 described as 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).

2.3 ARMS Post-Imputation Processes

Following imputation, the data are processed again through a computer edit. 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, additional edit checks that were previously skipped due to the missing data are executed. After the computer/analyst resolves all the errors, the data are considered clean and continue into the calibration and summary phases.

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, making COTS solutions more attractive.

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, involvement by original developers of ISR decreases. This results in a substantial learning curve for those tasked with maintaining or improving the software. For NASS purposes, it is ideal for more time to be spent on the imputation models themselves, and less time on the underlying imputation code. COTS solutions provides 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 may not always be a viable solution.

4. Goals

The goal of this research is to examine two COTS solutions that use multivariate approaches for imputation. Most of NASS production work is executed using SAS, so for this initial study we focused on two COTS solutions that could be executed in SAS: IVEware and PROC MI.

4.1 IVEware

IVEware is software created by survey researchers at the Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan, to produce single or multiple imputations using Sequential Regression Multiple Imputation (SRMI) as described in Raghunathan, et al., 2001. SRMI is a popular and well understood methodology; a brief overview of the process follows to allow for comparison to ISR.

As with ISR, the joint conditional distribution can be factored as a series of univariate conditional distributions. SRMI methodology uses a Gibbs sampling algorithm (Geman and Geman 1984; Gelfand and Smith 1990). After initialization, iterative draws of parameters and imputations can be made, where each conditional model may be linear or nonlinear (e.g. generalized logit) in nature and a diffuse prior is used for the parameters (Miller, 2015).

IVEware is available as a stand-alone program, or it can be run in SAS (SAS callable). For this research, the SAS callable version of the software was used. IVEware has several modules. The primary module used for this research was the IMPUTE module, that has several features that may benefit NASS work.

For example, within the IMPUTE module, the type of regression depends on the variable type. Variable types that can be imputed include continuous, binary, categorical (polytomous with more than two categories), counts, and semi-continuous. All variables in the dataset are potentially used in each conditional model, unless indicated in the transfer statement. Hence, variables may not take on all of the roles allowed in the ISR program; some of the relationships preserved by the conditional models may not be preserved using IVEware. IVEware allows for model selection, such as step-wise regression, minimum R-squared, and maximum number of predictors. It can also incorporate some types of edits, such as restrictions on variables to be imputed based on the value of other variables and bounded imputations. Data can also be transformed before imputing (Miller, 2015).

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 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, the user accepts that convergence may not be reached due to a potential lack of a valid joint distribution (Miller, 2015).

4.2. SAS PROC MI

As an alternative to IVEware, PROC MI is available in SAS. The MI procedure is a multiple imputation procedure that creates multiply 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 huge 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 (SAS, 2015).

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 (SAS, 2015).

Several options which come built into the MI procedure. The SAS MI procedure user guide details these. A few options 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. For this research SAS 9.4 with SAS/STAT 14.1 was used (SAS, 2015).

4. Simulation Study

For this study, empirical analysis on the feasibility of using IVEware or PROC MI for our imputation process was conducted. The goal of the study was to compare means and frequencies of imputed datasets relative to fully reported datasets of the same type and to evaluate which method produced data closest to the "true" full dataset values.

4.1 Methods

The 2013 ARMS Phase 3 dataset was subsetted to 10 variables that were fully reported and required no computer imputation. These variables were a combination of categorical, continuous, and semi-continuous variables.

Variable	***Variable Description***	***Variable Type***
FARMTYPE*	Type of Farm	Categorical (Crop = 1, Livestock = 2)
P864*	End of Year Breeding Livestock Value	Semi-Continuous
FERTSEXP*	Fertilizer Expense	Semi-Continuous
LVSTKEXP	Livestock Expense	Semi-Continuous
SEEDSEXP	Seed Expense	Semi-Continuous
P889	End of Year Crop Value	Semi-Continuous
P63	Cropland Acres	Semi-Continuous
P26	Total Acres on the Operation	Continuous
Region	ARMS Region	Categorical
GVCLS	Gross Value of Sales	Categorical

*indicates variable imputed in study

Table 1. List of variables used in simulation study.

Once the full datasets were created, three types of missingness were imposed, all at a rate of 30% missing: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). A seed was used and altered to create 250 different datasets of each type of missingness.

Missingness	***Method***
MCAR	RANUNI* to generate a number between 0 and 1. Values where RANUNI returned a value of <= .30 were set to missing.
MAR	Missingness of response variable is dependent on GVCLS. Rates of missingness differ between GVCLS.
MNAR	Missingness of the response variable is dependent on the response variable itself. Generally, speaking, larger values had more missingness imposed than smaller values

* The RANUNI function returns a number that is generated from the uniform distribution on the interval (0,1) using a prime modulus multiplicative generator with modulus 2^{31} and multiplier 397204094 (Fishman and Moore, 1982)

Table 2. Detailed breakdown on how missingness was imposed.

After the missingness was imposed, there were 750 datasets with missing fields that needed to be imputed. Seven different methods to impute the missing data, some of which were minor variations on how the mass of zeroes should be handled, were used. All methods used all the variables presented in table 1 as covariates. In the end, the four imputation mechanisms which did the best job for each type of imputation method were examined.

Method
Iterative Sequential Regression (ISR)
IVEware
PROC MI Regression
PROC MI Predictive Mean Matching

Table 3. Imputation methods examined.

For each of the 750 imputed datasets, the means of FERTSEXP and P864 were computed. The frequency of crop farms based on the FARMTYPE variable was calculated. These means and frequency contained a combination of the 30% imputed data and the 70% reported data. Next, the calculated means and frequencies with imputation to the original, fully completed dataset, the "truth" dataset were compared. For each dataset comparison, a relative difference was computed.

$$\boldsymbol{relative\_difference} = \frac{(\boldsymbol{Imputed} - \boldsymbol{True})}{\boldsymbol{True}}$$

The value of a relative difference shows the magnitude and direction for which the imputation shifted the mean or frequency.

4.2 Results

The following charts display box plots of the relative difference for each variable by type of missingness.

4.2.1 MCAR

Under a MCAR missingness pattern, both ISR and PROC MI PMM performed well on continuous variables as evidenced in figure 4. For the categorical variable Farm Type, ISR performed the worst. However, ISR was not developed to impute categorical variables, so this result was expected. All of the other imputation methodologies worked well for the categorical variable based on the relative difference output. Overall, under an MCAR missingness pattern, PROC MI PMM performed the best when evaluating the relative difference output.

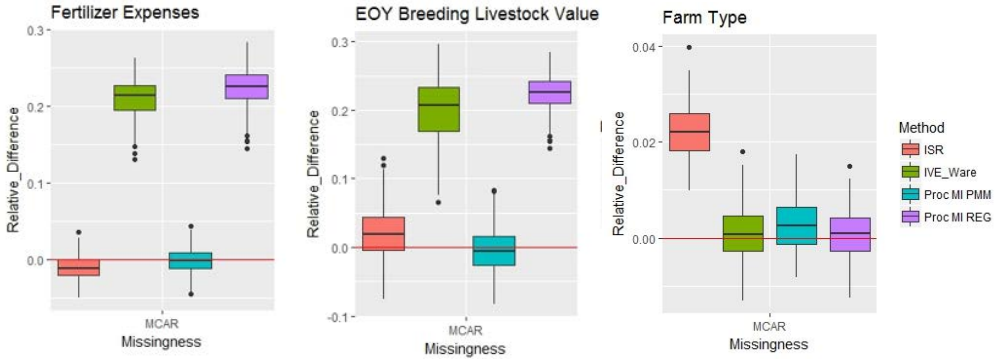


Figure 4. Relative difference boxplots of imputed data under MCAR missingness pattern.

4.2.2 MAR

Under a MAR missingness pattern with missing rates related to the value of sales, the results tended to be similar to the MCAR output. Both ISR and PROC MI PMM performed well for continuous variables. However, ISR performed worst on the categorical variable. Overall, under a MAR missingness pattern, which is what is generally assumed with the ARMS Phase 3 data, PROC MI PMM performed the best when evaluating the relative difference plots in figure 5.

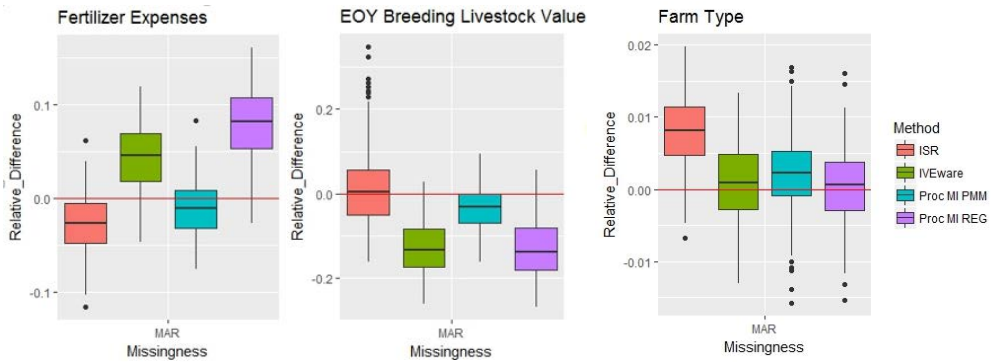


Figure 5. Relative difference boxplots of imputed data under MAR missingness pattern.

4.2.3 MNAR

Under a MNAR missing pattern with missing values related to the size of the value, all imputation models tend to perform poorly when evaluating the relative difference output. Despite this, PROC MI does offer the user some flexibility in imputing MNAR data via the MNAR option (SAS, 2015). However, for this to perform correctly, the user must have knowledge of how the data is MNAR, which may not always be viable in a production setting.

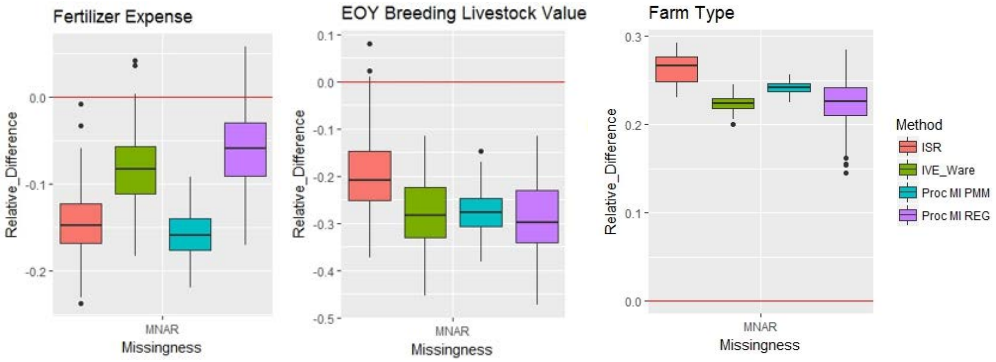
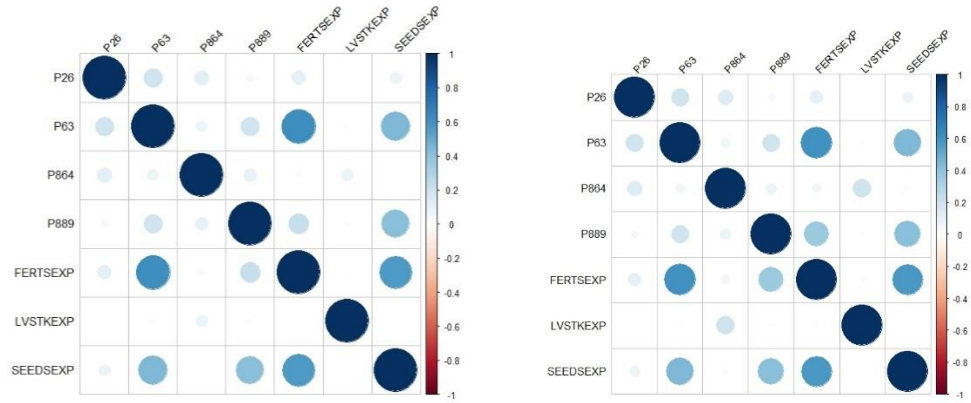


Figure 6. Relative difference boxplots of imputed data under MNAR missingness pattern.

While relative difference is an effective way to initially evaluate the imputed data, other components of analysis should be completed. Advanced multivariate analysis should be conducted after an initial analysis of the bivariate relationships between variables. Figure 7 below shows the bivariate relationships that are present between variables in the true dataset and also the bivariate relationships that are present in a randomly selected MAR dataset imputed using PROC MI PMM. Evaluating this output shows that the bivariate relationships between variables appear to be holding on this randomly selected MAR imputed dataset.



True Imputed MAR

Figure 7. Correlation plot output of bivariate relationships in true and imputed MAR datasets using PROC MI PMM.

5. Conclusion

While individual in-house developed imputation systems may perform well for individual surveys, these often require extensive resources to develop and maintain. In addition, these systems are often difficult to use on other surveys. If COTS software can perform as well or better than these systems, they may significantly reduce the need for staff resources. Based on initial output, promising results from COTS software were obtained and encourage future research. In particular, further research analyzing the sensitivity to missingness models, incorporating additional variance from imputation while still using a singular dataset, assessing pointwise accuracy using the full ARMS Phase 3 datasets with all imputation eligible variables included, testing the scalability of the software to see how long execution takes on larger datasets, and improving model selection efforts which include expert opinion along with statistical evaluation of models should be done.

References

Brand, J. P. L. (1999). Development, Implementation, and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets. Ph.D. thesis, Erasmus University.

Burns C., Prager, D. Ghosh, S. Goodwin, B. (2015) "Imputing for Missing Data in the ARMS Household Section: A Multivariate Approach". 2015 AAEA & WAEA Joint Annual Meeting

Gefland, A. E. and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," Journalof the American Statistical Association, 85, 398-409.

Geman, D. and Geman, S. (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Reconstruction of Images". IEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

Heitjan, F., and Little, R. J. A. (1991). "Multiple Imputation for the Fatal Accident Reporting System." Journal of the Royal Statistical Society, Series C 40:13–29.

Lavori, P. W., Dawson, R., and Shera, D. (1995). "A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data." Statistics in Medicine 14:1913–1925.

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.

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.

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.

Robbins, M., Gosh, S., and Habiger, J. (2013). "Imputation in high-Dimensional Economic Data as Applied to the Agricultural Resource Management Survey". Journal of the American Statistical Association, 108:501, 81-95, DOI: 10.1080/01621459.2012.734158.

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

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

SAS Institute Inc. (2015). SAS/STAT® 14.1 User's Guide. Cary, NC: SAS Institute Inc.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. New York: Chapman & Hall.

USDA - 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

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

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

Van Buuren, S. (2007). "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification." Statistical Methods in Medical Research 16:219–242.

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

White, T.K., and Reiter, J.P. (2008). "Multiple Imputation in the Annual Survey of Manufacturers". 2007 Research Conference Papers. Federal Committee on Statistical Methodology, Office of Management and Budget, Washington D.C.