# Comparative Survey Imputation Methods for Farm Household Income

Mary Ahearn, David Banker, Dawn Marie Clay, and Daniel Milkove

Farm household well-being is an important indicator used by public decision-makers when formulating policies, as well as by private sector decision-makers in the agricultural marketplace. In 2008, the mean household income of the 2.1 million farm operator households was $78,803 (USDA, ERS 2010). Although it varies greatly by farm size, the average farm component of the operator households' 2008 household income was only 11%. An example of this indicator's importance in the policy arena was exemplified in the April 21, 2010 testimony of Secretary of Agriculture Vilsack before the House Agriculture Committee when opening debate on the 2012 Farm Bill (USDA, OC 2010). The indicator of farm financial well-being featured by the Secretary in his more than two hours of testimony was the decline in farm household's share of income from farming (from 47% in 1960).

The Agricultural Resource Management Survey (ARMS) is the primary source of information on the economic well-being of America's farm operator households. The ARMS is conducted jointly by the Economic Research Service (ERS) and the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA). Currently, the USDA uses a relatively simple conditional mean imputation approach to deal with some missing household data. This approach generates estimates of means and totals that satisfy the objective of predictive accuracy. However, the data generated through simple imputation methods may yield less than satisfactory results when estimates of the population distribution are important, e.g. the share of households with high or low incomes, or when tests of statistical significance relying on unbiased standard errors are important, e.g. whether farm management practices are statistically significantly related to high or low incomes. These concerns are of particular importance in policy-related uses of the data. Moreover, accurate imputations for household variables in the ARMS is arguably more urgent than for farm business variables because: (1) survey weights are based on farm characteristics (principally farm size, type, and location), and the post calibration of survey weights is based exclusively on farm characteristics, not household characteristics; and (2) for many of the household variables, when a respondent refuses, data reported elsewhere in the survey that are used to guide an analyst in the imputation process are based on weaker assumptions of relationships than for refused farm production or business variables.

Recent advances in the statistical literature (Little and Rubin 2002), followed by the development of user-friendly software, have made it possible to consider more complex multivariate approaches than imputing for missing data. In a recent review of the ARMS program, the review panel recommended that, "NASS and ERS should consider approaches for imputation of missing data that would be appropriate when analyzing the data using multivariate models. Methods for accounting for the variability due to using imputed values should be investigated," (National Research Council 2010, p. 167). This article summarizes the results from a formal study of alternative imputation methods for selected household variables, and considers implications for improvements to current procedures.

Mary Ahearn and David Banker are senior economists, and Daniel Milkove is an economist, respectively, at the Economic Research Service (ERS), U.S. Department of Agriculture. Dawn Marie Clay is an intern at ERS and a Ph.D. student at North Carolina State University. Senior authorship is not assigned. The authors are grateful for the useful comments of Kirk White. The views expressed are those of the authors and do not necessarily represent the views of USDA. This article was presented during an invited paper session at the 2010 AAEA annual meeting in Denver, CO. The articles in these sessions are not subject to the journal's standard refereeing process.

*Amer. J. Agr. Econ.* 93(2): 613–618; doi: 10.1093/ajae/aaq167
Published by Oxford University Press on behalf of the Agricultural and Applied Economics Association 2011

## Current Imputation Methods in ARMS

There are approximately 2,100 variables on the ARMS survey in a typical year, and 1,500 are allowed to have data missing. In 2008, NASS imputed for 147 of the variables where refusals were allowed. The only household-specific variables which are imputed by NASS are the family living expenditure variables, and ERS does not use these imputations, but prefers to use its alternative imputation approach.[1] In 2008, ERS imputed for 47 household-specific variables using an approach similar to the general approach currently used by NASS in its ARMS imputations. The ERS approach is to use weighted conditional mean values to replace missing data items. The conditional means are matched based on characteristic variables from ARMS assessed to be closely related to individual missing data items. While preserving the overall mean calculations for imputed variables, a consequence of this method is to systematically understate variances for variables that are imputed.

## Approach for Evaluating the Quality of Imputations

An improved imputation approach would at least maintain the accuracy of the means, but produce more accurate variance estimates, avoiding the downward bias in the variance estimates of the current ERS approach. Our ultimate goal is to determine which of the two imputation methods, the current ERS approach or a more complex multivariate approach, would yield datasets with sample means and standard errors that were closer to a representation of the population. We use the 2007 ARMS dataset to evaluate the representativeness of the means and standard errors generated from the two imputation approaches. We first created a dataset (a subset of the actual 2007 dataset) without missing values for the variables of interest, referred to as *full dataset*. The variables of interest included in this dataset are the focus variables to be imputed, and the variables used in the development of conditional means in the ERS approach (age, education, metropolitan status, state, and region).[2] In recognition that the study results could differ depending on the frequency of valid nonzero responses, another aspect of our study design was to evaluate two focus variables, household income from retirement and disability sources, and operator wage and salary income (table 1), which differed in the extent of valid zero responses.

**Table 1. Distribution of Sample Survey Responses**

| Response | Farm operator wages off-farm and salaries | Farm operator household private retirement income and disability payments |
|---|---|---|
| | Percent | |
| Missing (refused) | 15.7 | 13.8 |
| Valid zero | 59.0 | 79.5 |
| Valid non-zero | 25.3 | 6.7 |

Source: USDA, ERS/NASS, 2007 Agricultural Resource Management Survey, Phase III

Since a pragmatic requirement of the study was to draw conclusions about a preferred imputation method that would be robust over time (and eliminate the need to restudy the issue annually with each new ARMS), we created and analyzed distributions of datasets for each approach. The distributions of datasets were created from the *full dataset*. First, to provide a basis for comparison, we created a set of 500 synthetic replications of the full dataset.[3] The 500 synthetic datasets used for both the ERS and the complex multivariate imputation approach were created from the 500 full synthetic datasets by creating missing values in the same percentages as existed in their respective variables in the actual ARMS data. The missingness mechanism we used to

[1] ERS' approach is to impute for missing values in total household expenditures based on the age of the principal farm operator, and the total income of the household. NASS' approach is based on farm size, commodity specialization, and region.

[2] In practice, ERS imputation approaches use different variables when calculating conditional means, depending on the focus variable and the year. The 5 variables used in this study were chosen because there has been recent discussion to expand the set of variables used to calculate conditional means to the 5 identified.

[3] When creating the 500 synthetic datasets, we regressed each of the focus variables on their coimputation variables (age, education, metropolitan status, state and region). This process of running logistic regressions was repeated where each variable of the imputation set became the dependent variable. The vector of probabilities for each level of the six respective dependent variables was captured from each polytomous or binary logistic regression. These level-probabilities were used to generate the synthetic data variable counterparts to the six variables from the original full dataset. This procedure was run 500 times using a random number generator with the level-probabilities as parameters to create the 500 synthetic full datasets.

create missing values was a missing at random (MAR) mechanism.

We determined the probabilities of missingness used for creating missing values by running a logistic regression of a binary (0-1) missingness variable for principal operator off-farm wages on five predictors of missingness. These five covariates were fully reported, so this satisfies the condition in Little and Rubin (2002, p. 16) regarding predictor-dependent MCAR being the same as MAR. These missingness probabilities were used to generate a binary (0-1) variable in each dataset that formed the criteria for data-deletion. If a data-deleting variable used in the process was equal to one, then the data point for that simulated observation was recoded to missing. Otherwise, the data value was unchanged. Each data-deleting variable was used in this process to create missing data for each respective full synthetic imputation set. We then used each imputation approach to impute for missing values in each of the synthetic datasets resulting in 500 *completed datasets* for both the ERS approach and the multivariate approach.

The final step in the study design compares distributions of focus variables for the alternative imputation approaches relative to distributions from the *full dataset*. Box plots and summary statistics were used to evaluate the performance of the alternative imputation approaches. Box plots provide useful visual depictions of distributions. Note that the comparisons were made with unweighted data since the synthetic datasets are no longer linked to the actual survey weights.

**Table 2. Distribution Values for Mean Private Retirement Income and Disability Payments**

| | Distribution | | | Difference From Full Distribution | |
|---|---|---|---|---|---|
| | ERS | Full | SRMI | ERS | SRMI |
| | | USD | | Percent | |
| Minimum | 1,544 | 1,548 | 1,543 | −0.3 | −0.3 |
| Mean | 1,929 | 1,887 | 1,948 | 2.2 | 3.2 |
| Maximum | 2,295 | 2,225 | 2,322 | 3.2 | 4.4 |

### Imputations with Sequential Regression Multivariate Imputation

The complex multivariate imputation method used in this study is the Sequential Regression Multivariate Imputation (SRMI) method implemented in a software application called IVEware (Raghunathan, Solenberger, and Van Hoewyk 2000; Raghunathan, et al. 2001). SRMI considers imputation on a variable-by-variable basis, but conditioned on all selected variables. The basic strategy is to create imputations through a sequence of multiple regressions, varying the type of regression used to model the missingness of a variable by the classification of the variable being imputed. Therefore, a continuous variable would have a different regression regarding its missingness compared to a regression model for missingness of a discrete variable. The IVEware SRMI method of imputation also allows for variables to be entered as mixed (partially discrete and partially continuous), which was the selection used in our study due to the relatively high proportions of zero values. In this case, the calculations occur in two stages.

The initial stage fits a binary logistic regression of one for all nonzero values and zero otherwise for the dependent variable (the variable to be imputed). The final stage uses the results of the logistic regression to then fit a linear regression to determine the imputation values for this mixed variable. Each iterative sequence of multiple regressions yields a dataset (implicate) which initially draws from a flat prior distribution that is updated after every iteration of imputations for all specified variables that have missing data values. Several iterations between recorded copies of imputations allow for independent multiple imputations of missing values of variables. The implicates yield distributions for missing values that provide imputation variance measures across implicates that are not possible with the ERS approach.

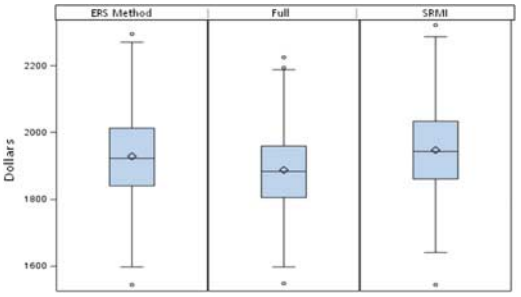


**Figure 1. Box plots of the distribution of mean values for the full dataset and by imputation method for private retirement and disability income**

## Results

The box plots contain the middle 50% of the data, and the line indicates the median value of the distribution. The mean of the distribution is shown as the diamond shape in the box. Figure 1 presents the box plot comparisons for the distributions of mean values for private retirement income and disability payments for operator households. For this focus variable, the ERS imputation method yielded a distribution from the 500 means that was closer to the respective *full dataset* distribution than that generated by the alternative SRMI imputation method. The ERS method of imputation yielded a distributional (unweighted) mean of $1,929, compared to the *full dataset* distributional mean of $1,887 (table 2). The counterpart measure from the SRMI imputation method was $1,948. The maximum mean values of the distributions from the ERS method were also closer to those of the *full dataset* distribution than those from the SRMI method. The extreme and central values of the standard error distribution from the ERS method (figure 2, table 3) were lower than those from the *full dataset* distribution, while those from the SRMI method were higher.

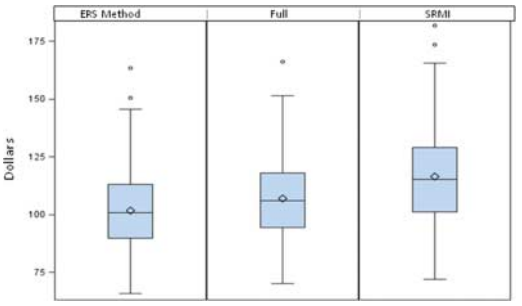


**Figure 2. Box plots of the distribution of the standard errors for the full dataset and by imputation method for private retirement and disability income**

**Table 3. Distribution Values for Standard Errors for Mean Private Retirement Income and Disability Payments**

| | Distribution | | | Difference From Full Distribution | |
|---|---|---|---|---|---|
| | ERS | Full | SRMI | ERS | SRMI |
| | | USD | | Percent | |
| Minimum | 65.8 | 70.2 | 72.0 | −6.3 | 2.6 |
| Mean | 101.7 | 107.0 | 116.3 | −5.0 | 8.7 |
| Maximum | 163.4 | 166.1 | 181.9 | −1.6 | 9.5 |

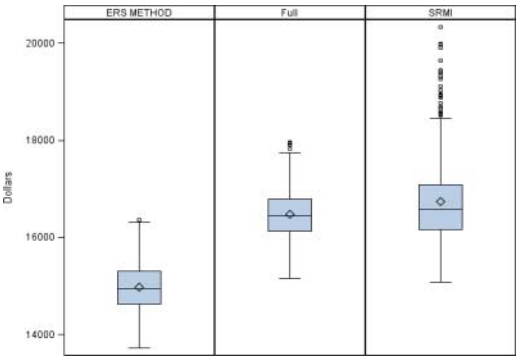


**Figure 3. Box plots of the distribution of mean values for the full dataset and by imputation method for operator wage and salary income**

**Table 4. Distribution Values for Mean Farm Operator Household Off-Farm Wages and Salaries**

| | Distribution | | | Difference From Full Distribution | |
|---|---|---|---|---|---|
| | ERS | Full | SRMI | ERS | SRMI |
| Minimum | 13,726 | 15,162 | 15,083 | −9.5 | −0.5 |
| Mean | 14,974 | 16,473 | 16,742 | −9.1 | 1.6 |
| Maximum | 16,359 | 17,964 | 20,333 | −8.9 | 13.2 |

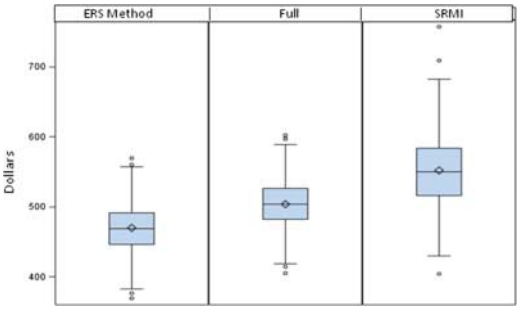


**Figure 4. Box plots of the distribution of standard errors for the full dataset and by imputation method for operator wage and salary income**

In contrast, for the other focus variable, operator off-farm wages and salaries, the SRMI method yielded mean and minimum mean values that were closer to the *full dataset* distribution than those generated by the ERS method (figure 3, table 4). The ERS method's standard errors (figure 4, table 5) were numerically closer, but lower than those from the *full dataset*.

**Table 5. Distribution Values for Standard Errors for Mean Farm Operator Household Off-farm Wages and Salaries**

| | Distribution | | | Difference From Full Distribution | |
|---|---|---|---|---|---|
| | ERS | Full | SRMI | ERS | SRMI |
| Minimum | 369.2 | 404.7 | 456.8 | −8.8 | 12.9 |
| Mean | 469.6 | 504.2 | 552.0 | −6.9 | 9.5 |
| Maximum | 568.9 | 602.5 | 682.3 | −5.6 | 13.2 |

## Discussion

The mean income of farm operator households is relatively high, exceeding the U.S. mean household income since 1996. The off-farm income component alone has exceeded the average U.S. household income since 1998 (USDA, ERS 2010). The major source of household income for farming households are the wages and salaries of the farm operator ($25,875 in 2008), one of the focus variables in our study.[4] It is only on the 10% of all farms that are large (gross farms sales are $250,000 or more) that farm income and the off-farm wages and salaries of the farm spouse exceeded the off-farm wages and salaries of the principal operator.

It is reasonable to question the role of the imputation approach in the growth of farm household income. In particular, could imputation methodology be a contributing factor to the high levels of estimated off-farm income? One conclusion from this analysis is that the ERS imputation approach has likely not been a factor in rising off-farm income estimates as reported by USDA. In fact, our analysis provides some evidence that the current imputation approach may reduce the estimated mean income of farm operator households.

An advantage of our study design was that it provided a comparison of methods to a *full dataset*, which can be considered as a benchmark or gold standard dataset. For imputing farm operator household private retirement income and disability payments, a variable with a large share of responses equal to zero, we found that the ERS method was closer to the *full dataset* than the SRMI method. Perhaps this is due to the difficulty of fitting a polytomous logistic regression to a variable with a very small proportion of valid non-zero values. For variables like this with a truncated distribution, it might be easier to model a joint relationship less directly (using weighted conditional mean imputation by a few characteristic variables) than to model with regressions in two stages. In contrast, for imputing farm operators' off-farm wages, a variable where more of the data were non-zero, the SRMI method was closer to the *full dataset* than the ERS method. This could indicate that more complex multivariate imputation approaches for the actual data set would yield more reliable means and standard errors for variables that have lower percentages of zero observations.

Perhaps a greater concern with the ERS imputation approach is its effect on estimates of standard errors. Our analysis showed that the ERS approach resulted in lower standard error estimates than SRMI approaches for both focus variables. Standard errors that are biased downwards may incorrectly lead researchers to conclude that relationships are statistically significant, and to understate the extremes in income levels of farm households. For some policy analysis purposes, such as capping farm subsidies based on high levels of adjusted gross income, this understatement is potentially of concern.

It is not possible to generalize our findings to reach a definitive conclusion about a preferred imputation approach for household variables in ARMS, but observations can be made about future research that will better support definitive recommendations. In particular, the data in our study were unweighted and the observations in the initial *full dataset* necessarily included only respondents who provided complete data. To the extent that those who refused to respond to the items or the full survey were not representative of the population at large, our results are not conclusive. In addition, we know that large farms have higher sampling rates, lower sampling weights, and lower off-farm income than smaller farms. Future work also needs to consider placing reasonable boundaries on the imputed values, for example, by constraining imputed values to avoid negative income values where appropriate, and eliminating or otherwise dampening the undue effects of outlier values on the imputations. The survey weights are based exclusively on farm characteristics, not household characteristics, and until this issue is directly addressed in the survey design, constraining the distribution of imputed values for household-related

[4] This value is much larger than reported in our analysis. We used unweighted data in our study, thus resulting in substantially lower mean values, compared to weighted data, for both operator off-farm wages and private retirement income and disability payments.

variables may be a pragmatic solution. An interesting next step will be to compare means and standard errors between the ERS and the SRMI imputation approaches over a period of years to determine if the general results of this study are upheld when all observations are included, and if the data are statistically weighted to reflect the underlying complex sample design.

In recent years, the nine off-farm sources of income have been collected on the ARMS through the use of thirty-one value codes (and some, like nonfarm business income, are allowed to be negative), rather than collecting actual dollar amounts. In the earlier years of the survey preceding ARMS, the Farm Costs and Returns Survey, actual dollar values of off-farm income were collected. The switch was made to a value codes collection approach in an attempt to increase item response rates. This raises the question about differences in data quality between the two data collection approaches, and the role of the current value code collection approach in our analysis of the alternative imputation approaches. Tests of alternative data collection approaches, in conjunction with alternative imputation approaches, can be performed to shed some light on this question.

Ultimately, the ARMS-based household indicators must be corroborated or consistent with evidence from other information, past and present, and in relation to the conditions in the farm and general economy. As stated in the recent review of the ARMS program, when judging the imputation procedures, one criterion that should be considered is the "imputation plausibility" of the resulting values (National Research Council 2008, p. 121). This is a special challenge involving expert judgment, because the timeliness of these indicators, as much as accuracy, is critically important.

## References

National Research Council. 2008. Understanding American Agriculture: Challenges for the Agricultural Resource Management Survey. Panel to Review USDA's Agricultural Resource Management Survey. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Little, R.J.A., and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*, 2nd ed. New Jersey: John Wiley & Sons.

Raghunathan T.E., P. Solenberger, and J. van Hoewyk. 2000. *IVEware: Imputation and Variance Estimation Software (User's Guide)*. Ann Arbor, University of Michigan, Institute for Social Research.

Raghunathan, T. E., J.M. Lepkowski, J. van Hoewyk, and P. Solenberger. 2001. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology* 27, 85–95.

U.S. Department of Agriculture (USDA, ERS). 2010. Farm Household Economics and Well-Being Briefing Room. Washington, D.C. Economic Research Service. Available at: http://www.ers.usda.gov/Briefing/WellBeing/. Accessed on April 1, 2010.

U.S. Department of Agriculture (USDA, OC). 2010. Agriculture Secretary Vilsack Makes Case for Stronger Rural America: Vilsack testifies before House Committee on Agriculture on the 2012 Farm Bill. Washington, D.C. Office of Communications. USDA News Release No. 0198.10.