Examining the Challenges of Missing Data Analysis in Phase Three of the Agricultural Resource Management Survey

Darcy Miller[1], Michael Robbins[2], Josh Habiger[3]
[1]National Agricultural Statistics Service, 3251 Old Lee Highway Rm 305, Fairfax, VA 22030
[2]Michael Robbins, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709
[3]Josh Habiger, Oklahoma State University, Department of Statistics, Stillwater, OK; National Institute of Statistical Sciences, 19 T.W. Alexander Drive Research Triangle Park, NC 27709

Abstract
The National Agricultural Statistics Service (NASS) collaborates with the Economic Research Service (ERS) to conduct the Agricultural Resource Management Survey (ARMS), which provides a source of information for addressing issues relating to agriculture and the rural economy. ARMS is a multi-phase, multi-mode, and dual frame survey. The third phase, ARMS III, collects data which are critical to assessing the relationship of the over-all financial health of the farm household and the farm operation with production practices. Like many surveys, ARMS III is subject to item nonresponse and utilizes imputation to mitigate the effect of the missing information on statistical analysis. Examination of the ARMS III data set and the current imputation method has revealed a rich potential for fruitful investigations into multivariate imputation techniques for a large, semi-continuous data set.

Key Words: Missing Data, Imputation, Item Nonresponse, Complex Surveys, Nonresponse, Agriculture

1. INTRODUCTION

The Agricultural Resource Management Survey, administered by the National Agricultural Statistics Service (NASS), is an invaluable source of information on the current state of agriculture. Data users of the ARMS survey include Congress, USDA, NASS, Economic Research Service (ERS), Bureau of Economic Analysis (BEA), academic researchers and agri-business officials. Some data users develop forecasts for private enterprise which have implications for the food supply and prices. Others use ARMS data as part of an analysis to establish and review policy or assess standards in an array of areas such as food production, rural economics, bioenergy, and the environment.

The ARMS is administered in three phases. The first phase is a screening phase for in-scope and in-business farms as well as presence of the targeted commodities for that year. Targeted commodities change from year to year. The second phase asks for detailed field-level data for the targeted commodity for that year. The third phase (ARMS III) is a multiple frame survey of about 35,000 farms and ranches conducted annually in February and March in all states except Alaska and Hawaii. NASS utilizes a list frame, which contains most of the farms with the largest expenditures, and an area frame which compensates for the incompleteness of the list frame. The ARMS III survey instrument is mixed mode and has as many as five versions in a given year.

Because the resulting data are used for in-depth, detailed analyses of the link between policy, operation profitability, and operator household financial health, the ARMS III survey instrument is necessarily long and complex. Some versions of the survey are 51 pages long, with data collected on sometimes more than 800 variables. Moreover, survey questions encompass the characteristics, management, income, and expenses of both the farm operation and the farm household. Collecting full responses on all of the items is a challenge. Details concerning expenses of a contractor or landlord are also collected from the operator (respondent) and these items are often, not surprisingly, the most problematic, sometimes with over half of the observations missing. (Miller, D. & O'Connor, 2010). NASS has taken many steps to increase awareness of the benefits of the survey and to reduce respondent burden. Many operators, however, still may not find utility in responding due to the magnitude of perceived personal costs (time, privacy, etc.), confusing questionnaire design or possibly anti-government sentiment (Dillman, 2007).

Both unit nonresponse (units sampled do not respond to any items), and item nonresponse (units sampled only answer some of the items) create gaps in data that require special handling throughout the analysis and estimation processes. Missing data can be problematic at the very least through the reduction in efficiency due to the decrease in sample size. Moreover, the possible systematic differences between respondents and nonrespondents can lead to biased estimators of a particular parameter of interest. The quality of estimators in the presence of missing data is subject to the ability of the data analyst to account for the missingness. Techniques used to mitigate issues inherent in the analysis of missing data cost money, manpower, time, and potentially affect quality, largely depending upon the data user's needs.

Despite the difficulty in obtaining full responses from all operations that are sampled, the need remains to perform effective statistical analysis with the data available. To fill this need, NASS imputes generated values into missing items for those variables for which summary statistics are published. The current NASS procedure involves the use of conditional averages following the elimination of outliers. Most often, the imputation is carried out by conditioning on three categorical variables; however, if there is little information, NASS may use a national average.

Continually reviewing imputation procedures for potential improvement is an integral part of maintaining the highest level of quality in this influential data set. The complexity and importance of the survey led NASS to look for partners to bring fresh ideas to a research team. NASS entered a cooperative agreement with the National Institute of Statistical Sciences (NISS). NISS is an organization which works to facilitate and promote innovative statistical research. NISS contracted appropriate academic researchers to join a team of researchers from NASS and ERS in continuing research towards improving current imputation methods. Specifically, the methods developed should be appropriate for data users who utilize multivariate models.

2. METHODS

2.1 Organizational Structure of Research

2.1.1 NISS-NASS Research Agreement

Research is being conducted through a cooperative agreement between NASS and the National Institute of Statistical Sciences (NISS). The objective of the research is to develop a comprehensive, multivariate imputation scheme that produces results reflecting the distribution of agricultural data, that supports both economic modeling and direct estimates, and that provides for an estimable impact of imputation on mean squared error.

NISS is a private organization developed to foster high-impact cross-disciplinary research involving the statistical sciences. NISS' experience in cross-disciplinary research was invaluable in the appropriate assignment of academic researchers to collaborate with NASS and ERS researchers and staff. Statistics and Economics Faculty guide a post-doctoral student and graduate student in technical research components. Researchers and staff from the government agencies that are working in the area of imputation research or the administration of the ARMS facilitate the research process to ensure that the academic members understand the ARMS and that the results of the research are implementable.

The research program spans two full years. Several weeks out of each summer are spent in residence at NISS by team members. When the faculty and graduate student return to their universities for the fall and spring semesters, the postdoctoral researcher works at NASS' Research and Development Division. Weekly teleconferences, a series of webconferences, and professional conferences allow the team members to continue collaboration while working at different locations.

2.1.2 Privacy and Technical Considerations

Record level data are protected by law. NASS is governed by specific statutes that explicitly prohibit the agency from releasing any information that is collected under our pledge of confidentiality and that could be traced back to an individual producer or farm operation. These statutes include Title 7, U.S. Code, Section 2276 and the Confidential Information Protection and Statistical Efficiency Act. Every person working for or in cooperation with NASS from the Agency Administrator to the person collecting the information signs a confidentiality form which states that no confidential reported information will be compromised. Any offender is subject to a jail term (five years), a fine ($250,000), or both. This includes sworn agents who are authorized by NASS to provide data collection support or statistical research. For the academic researchers to have access to the data necessary to perform the research successfully, they would need to obtain access to information about individuals that are collected on the survey. The academic researchers took the confidentiality pledge, became sworn agents, and then were authorized to utilize the record level data necessary for the specific statistical research at hand.

The off-site location for the summer of 2009 posed another security challenge. Research required a data lab at NISS with security standards equivalent to the security standards

currently in place for the operation of the NASS secure data analysis facilities in the USDA's South Building. Since the NISS/NASS Data Lab is physically separated from NASS, certain operational and physical aspects of the security plan implementation differed. NASS developed and implemented a security plan with guidance from:

- NIST Special Publication 800-18: "Guide for Developing Security Plans for Federal Information Systems"
- USDA DM 3550-002, Chapter 10, part 2 "Sensitive but Unclassified (SBU) Information Protection"
- USDA DM 3535-002, Chapter 7, Part 1 "USDA's C2 Level of Trust"

The fundamental operational requirement for the lab was to keep data secure inside of it. That is, no data would be exiting the lab for public release or review. The data lab was disconnected from the NISS LAN and was subject to unannounced security audits. To add another layer of protection, the data sets used off-site were at least five years old and truncated to only variables that were necessary for exploration.

For the post-doctoral and graduate student to continue work at NISS when faculty returned to their universities, results such as graphs and summary tables from some analyses would need to be shared. A formal written data exit authorization request was submitted to the point of contact at NASS who would approve or disapprove of any data related items leaving the lab.
(Barboza et. al., 2009)

For the summer of 2010 in residence at NISS, NASS addressed the challenge of working off-site differently. NASS contracted with the National Opinion Research Center (NORC) to provide access to the NASS data sets through the NORC Data Enclave. The NORC Data Enclave provides a confidential, protected environment within which authorized researchers from organizations such as the National Science Foundation, National Institute of Standards and Technology, USDA, and more, can access sensitive micro data remotely. The Data Enclave is fully compliant with DOC IT Security Program Policy, Section 6.5.2, the Federal Information Security Management Act, provisions of mandatory Federal Information Processing Standards and all other applicable NIST Data IT system and physical security requirements.

The NORC data enclave uses virtual private network (VPN) technology to prevent outsiders from reading data transmitted between the researcher's computer and NORC's network. All applications and data run on the server at the data enclave (i.e. researchers cannot utilize the PC or internet while in the data enclave). Data files cannot be downloaded to the researcher's PC, the "cut and paste" feature in Windows cannot move data from the Citrix session, and the researcher cannot print data from the local computer. Sessions in the data enclave are tracked through audit logs and audit trails. Results can be downloaded from the NORC Data Enclave through approval of the designated NASS agents and NORC.

2.2 Complexity of ARMS

2.2.1 Survey Design Overview

ARMS has three phases of survey administration and several different questionnaires at each phase. Different modes of administration include face-to-face, mail, telephone and internet; however, ARMS III uses primarily face-to-face and mail with a telephone follow-up.

The ARMS III survey instrument usually has between three and five versions. The Cost and Returns Report (CRR) collects detailed economic data and production practices for the whole farm operation and farm household and is administered every year. Also administered annually is the Core, which is a condensed version of the CRR that is mailed to a subset of states for NASS to publish state level estimates. Several other versions of the questionnaire will target specific commodities. Targeted commodity version content will essentially cover the same items as the Core, but with some additional, commodity specific detail.

The ARMS sample is drawn from a list frame and an area frame. The current sampling design utilizes a Sequential Interval Poisson sampling procedure to reduce respondent burden by decreasing the probability of selection for operations that were sampled for ARMS the previous year or for other major NASS surveys.

NASS contracts with the National Association of State Department of Agriculture (NASDA) for data collection. For ARMS III, NASDA enumerators visit operations (sampled units) for CRR and targeted commodity versions. The National Processing Center in Jeffersonville, Indiana mails and receives the Core. Collected data are keyed at the NPC. Some surveys that are returned very late in the data collection cycle are keyed in NASS' Field Offices (FO's). Processing and analysis are completed in FO's and Headquarters.

2.2.2 ARMS Data Collection and Processing

Several data sets are available that act as snapshots through the survey editing and imputation process. The first set is the original cell values from NPC keyed data. There is also a pre-NASS machine imputation data set, a final post-NASS imputation data set used for estimation, and a final post-ERS imputation data set available to the research team. Figure 1 illustrates ARMS data processing.

Data collected from the respondent may be changed by a field enumerator or a survey statistician before it is keyed at NPC. In the Virginia FO, most of these are deterministic and/or administrative changes. For example, a missing enumerator identification code may be manually imputed, or the sum of parts is manually imputed into a total value that is missing.

NASS collects the data, keys them, and hand imputes questionnaires throughout the editing process, but predominantly before machine imputation. Initially, a batch edit sweeps through the data flagging possible errors within the records. Analysts have several weeks to make sure that the items within a record (except those eligible for machine imputation) will pass all of the edits. Next, the approximately 150 items eligible

for NASS machine imputation with positive values are put into similar groups, analyzed using ad hoc outlier tools, and extreme values are dropped as an analyst sees fit. The remainder of the values in each group are averaged and imputed into cells that are determined to need non-zero values by the edit. The edit is run again to ensure that the machine imputations are valid as determined by edit specifications. If a record does not pass the edits, an analyst determines what values should be changed within a record. Then, the sample weights are calibrated to totals for number of farms, acreage, and inventory. Data are summarized. More outlier analysis is done on the summarized data and more changes are made to the weights, if necessary. After the NASS editing process is complete, NASS has a final dataset that is used for NASS estimation purposes and shared with ERS. The remaining items that are eligible for imputation in the data set are available to ERS to impute by a separate ERS process.

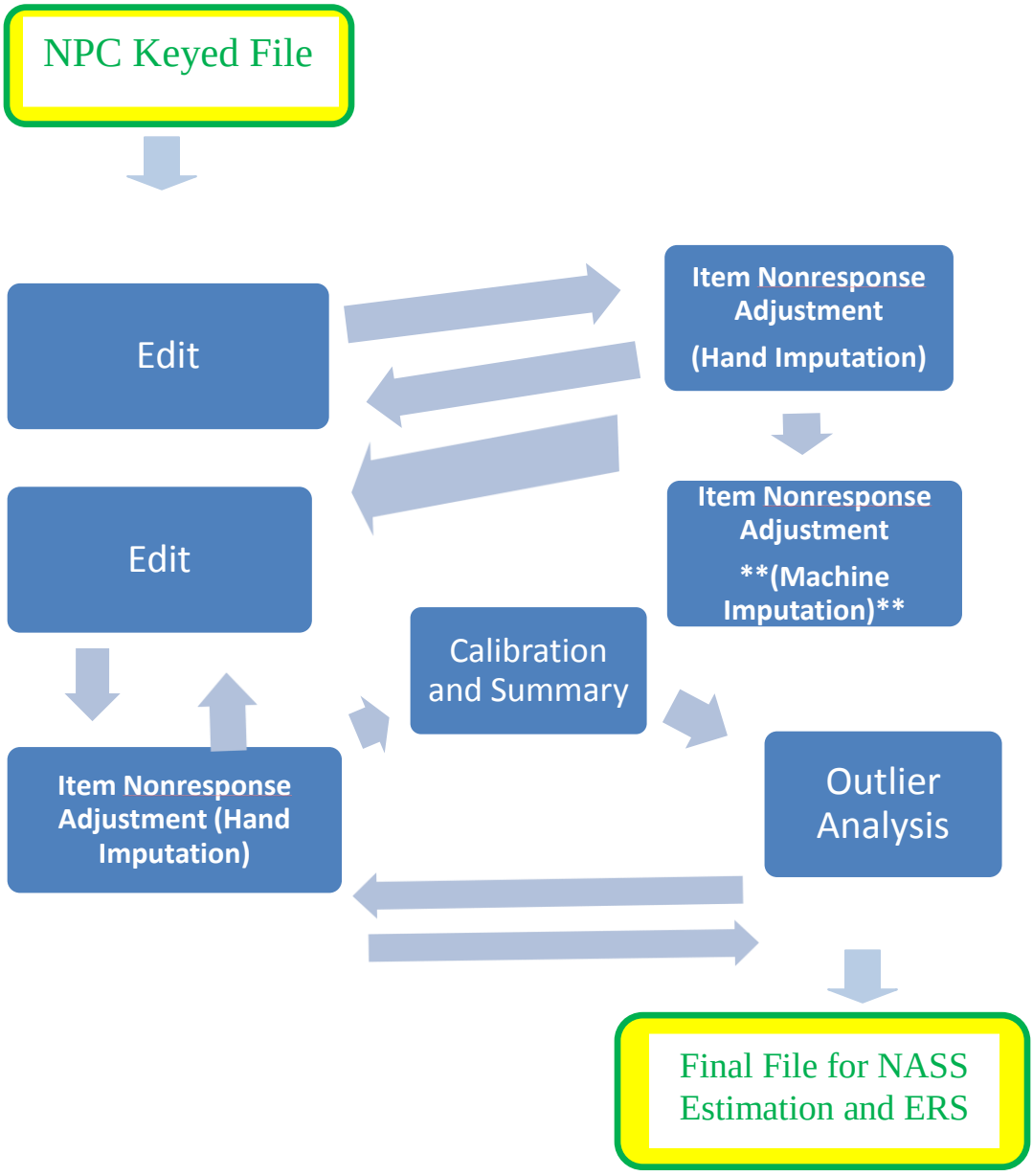

Figure 1: Processing flow overview after NPC keyed data are put into a batch edit to Final File.

The project uses three data sets: one that has been edited and hand imputed, just before the machine imputation part of the process; the final NASS data set used for estimation and sent to ERS; and the final dataset after ERS imputations.

2.2.3 Current NASS Imputation Procedure

Approximately 150 variables are machine imputed by NASS, and ERS machine imputes close to 50 more. The current NASS procedure for imputing missing data involves the use of conditional averages in place of missing data. This is equivalent to a regression on categorical variables. The categorical variables used come from a predetermined hierarchy, with a broader set of variables used when information needed for imputation is insufficient at a finer level (See Tables 1 & 2).

1	Sales Class, Farm Type, State
2	Farm Type, State
3	Sales Class, State
4	State
5	Sales Class, Farm Type, Region
6	Farm Type, Region
7	Sales Class, Region
8	Region
9	U.S.

Table 1: Imputation Groupings for ARMS III: Value in Land and Buildings

1	Region, Sales Class, Farm Type
2	Region, Pooled Sales Class, Farm Type
3	Sales Class, Farm Type
4	Pooled Sales Class, Farm Type
5	Region, Sales Class, Pooled Farm Type
6	Region, Pooled Sales Class, Pooled Farm Type
7	Sales Class, Pooled Farm Type
8	Pooled Sales Class, Pooled Farm Type
9	Region, Farm Type
10	Region, Sales Class
11	Region
12	U.S.

Table 2: Imputation Groupings for ARMS III: Other Items

An imputation donor is defined as a record with non-zero data for the item of interest. Donors for each item of interest are placed in imputation groups based on the first step in the table (e.g. in Table 1: region, sales class, and farm type). If fewer than ten suitable donors are available, a broader set of classification variables found in the next step in the table is used. The process continues until the donor pool consists of at least ten suitable donors. For most items, the initial group is used (Schauer, 2008).

Before the average of each group is calculated for an item, an analysis to determine the skewness of the distribution of donors within each group is conducted. A distribution is deemed too skewed if the mean is more than one quartile length away from the upper or lower quartiles. The most extreme values in the groupings may be excluded by the analyst. The remaining donor values in each group are averaged for most items and imputed into recipient records. For a few items where it is appropriate, NASS utilizes information from another variable within the survey (e.g. owned acreage to impute value of owned land). Otherwise, information utilized is limited to the groupings in Table 1 & Table 2.

After outlier analysis, the values are imputed. The data set is then run through a batch edit. The processing continues at a summary level beyond this point and revisions can be made by hand.

2.3 Data Exploration

The research team selected approximately 100 variables from the ARMS III survey on which to focus initial analysis. Of these variables, approximately 22 may have missing values and 85 are must, meaning they will not have any missing values to be machine imputed. The focus was on 22 policy variables related to farm program receipts, specifically, farm payment receipts. The 85 non-missing variables related to the farm payment receipt variables were selected by the economic experts on the team.

After defining the variables to focus on, the research team conducted a detailed evaluation of the general distributional properties and other characteristics of the variables. This included detailed study of data plots, considerations of discrete/continuous relationships, and the categorical, censored, or continuous nature of the variables. For farm program payments, a large number of values are zero. However, during the ARMS III edit, which is completed before machine imputation, cells are marked as needing a positive value. So, exploration continued with positive parts of the variable distributions.

Most of the ARMS III items imputed by NASS are continuous (specifically, the nonzero part of the distribution). In accordance with expectation, the farm program payment receipt variables are non-normal and exhibit positive skewness. After a log-transformation, variables studied overall best fit a skew normal distribution (See Figure 2).

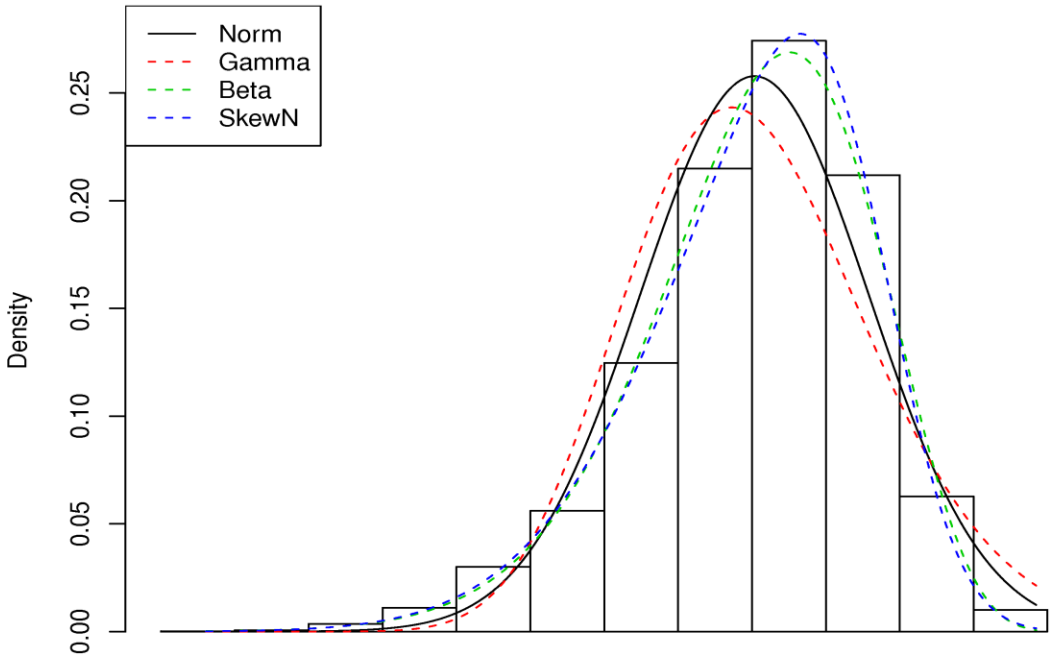

Figure 2: P525 Modeled using the Normal (pvalue = 0), Gamma (pvalue = 0), Beta (pvalue = .01), and Skew Normal (pvalue = .17).

Summary of p-values from goodness of fit tests for several candidate distributions confirmed that the data are clearly non-normal (See Table 3). In addition, a histogram of p-values fitting the skew normal to the log-transformed data was relatively flat and much more uniformly distributed on (0,1) than the Normal, Gamma, or Beta models. This is another indication that the Skew Normal is the best overall fit.

Model	*Normal*	*Gamma*	*Beta*	*Skew Normal*
# P-vals < .01	29	52	11	7
# P-vals < .05	40	61	18	15
# P-vals < .1	46	66	22	19

Table 3: Summary of goodness of fit tests applied to each variable for each imputation model.

Bivariate examination of the variables through scatterplots revealed that the variables with missing values were linearly related to the continuous "must" variables on the log scale. (See Figure 3). This suggested that linear imputation models may be appropriate.

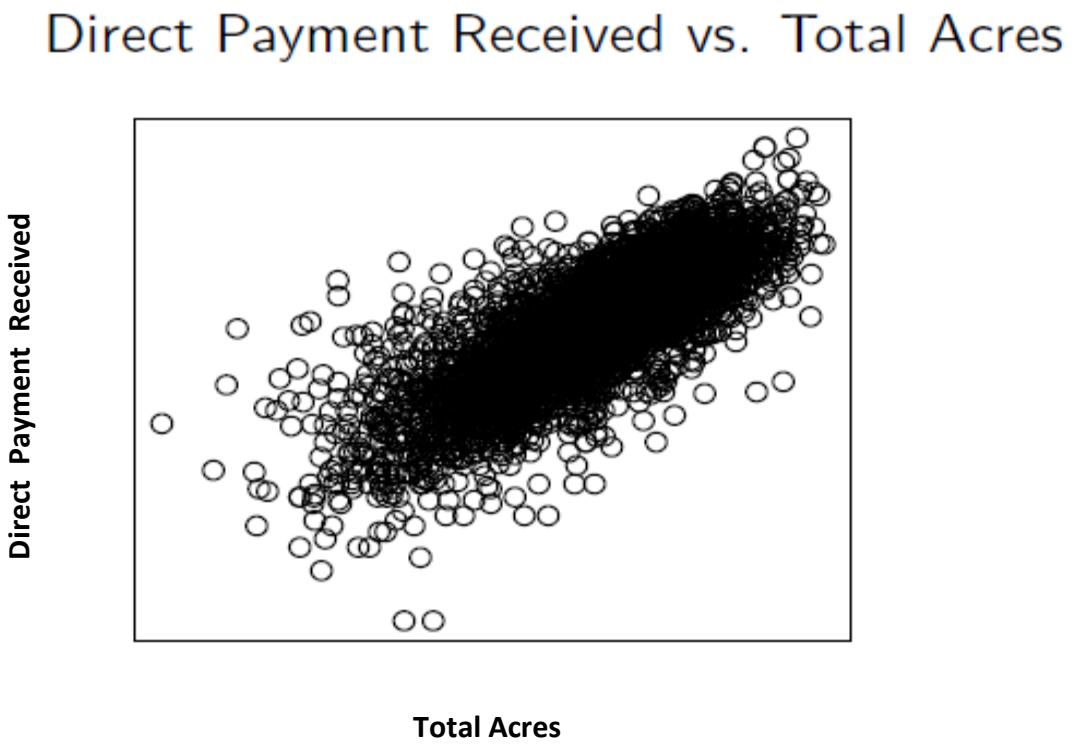

Figure 3: Scatter plot of Direct Payments Received and Total Acres Operated, both on the log-scale.

3. RESULTS

3.1 Effects of the NASS Imputation Procedure

Mean imputation disturbs the shape of the distribution by reducing the variation in the data set. The magnitude of variation depends upon the amount of data imputed. Although most items have small rates of imputation, NASS sometimes imputes more than fifty percent of the positive values for some items (Miller, D. & O'Connor, T., 2010). Imputation rates of this magnitude can cause a large downward bias in the variance and disturb the shape of the distribution. (See Table 4).

Variable	P525	P526	P527	P528	P529	P530	P531
%Missing	3.567	8.279	18.61	4.304	4.769	22.67	11.47
%Change	0.456	-1.841	-4.919	-0.927	-0.542	-7.442	-4.358

Variable	P532	P533	P534	P535	P536	P537	P538
%Missing	30.901	42.735	26.754	17.000	0.857	6.050	3.879
%Change	-6.487	-23.911	-12.729	-6.352	-1.121	-2.397	-1.512

Variable	P539	P540	P541	P542	P543	P544	P545	P557
%Missing	30.901	1.949	3.979	4.530	36.075	6.131	4.680	38.402
%Change	-16.415	-0.622	-1.407	-1.784	-13.775	-2.354	4.109	-19.256

Table 4: Percent of data missing by variable and percent of change in variance from before to after imputation by variable. For variable listing, see Appendix 1.

Examination of the distributions before and after machine imputation provided further confirmation that the ARMS III imputation process was disturbing the distribution (See Figure 4).

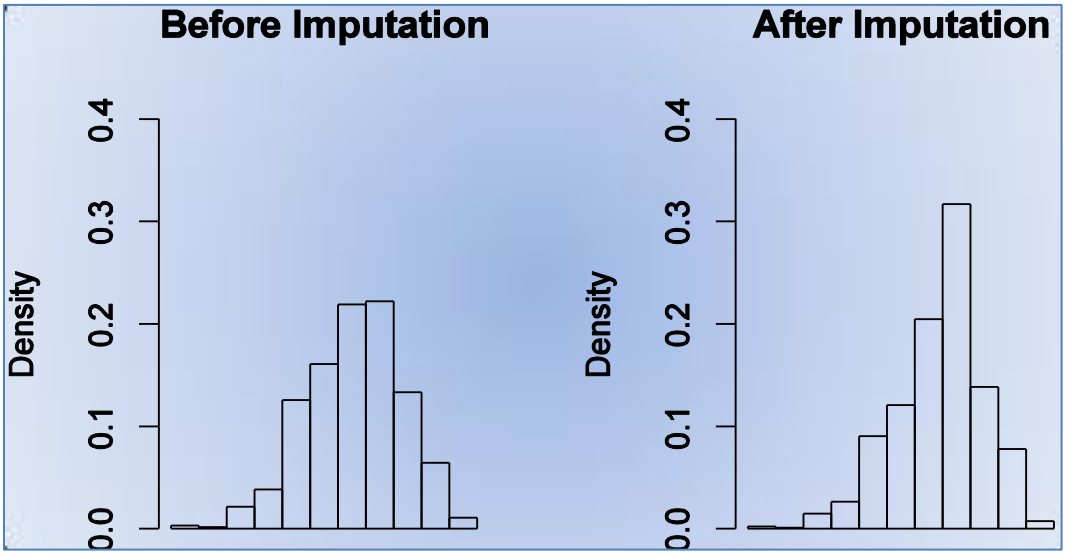

Figure 4: Distribution of Counter Cyclical Payments Before and After Imputation (Robbins, M. et al, 2009)

In addition, if non-respondents are not like respondents with respect to an item's value within each of the NASS groupings, the mean will also be biased.

3.2 Evidence for Improvement

Since initial analysis of correlation patterns among variables suggest that the interrelationships appear to be linear on the log-scale, linear imputation models may be appropriate. Economic experts further structured the ARMS III variables selected for study into groups to compare alternative methods of imputation which use more than the three variables used by NASS. Four farm program payment receipt variables eligible for imputation were selected: Direct Payments Received, Counter Cyclical Payments Received, Target Commodity Payments Received, and Milk Income Loss Payments. All of these variables were modeled on three variables currently used: Farm Type, Region, and Sales Class. Then, economic experts selected additional "must" variables in the survey on which to model the four farm program payment receipt variables. Significant

improvements in adjusted R^2 terms are apparent when multivariate relationships are used to expand the set of variables used to impute missing values. (See Table 4).

Variable to Impute	*DPR*	*CPR*	*TCPR*	*Milk*
Three Variables	.16	.18	.15	.12
More Than Three	.72	.58	.69	.33

Table 4: Adjusted R^2 values for modeling Direct Payments Received, Counter Cyclical Payments Received, Target Commodity Payments Received, and Milk Income Loss Payments on the three variables Sales Class, Region, Farm Type, then on those three variables and the additional 19, 20, 10, and 6 variables selected by the economic experts.

4. CONCLUSION AND RECOMMENDATIONS

The size of the ARMS III data set, high missingness rates and the semi-continuous nature of the survey variables all come together to present a missing data research problem which proves to be simultaneously challenging and ripe for innovative investigations. The NASS-NISS research team has been fruitful in uncovering possibility through data exploration and is excited with the potential of applying new imputation methodologies.

5. REFERENCES

Allison, Paul D. (2001) "Missing Data," Sage Publications, Inc., Thousand Oaks, CA.

Dillman, Don A. (2007) "Mail and Internet Surveys: The Tailored Design Method," 2nd Ed. John Wiley & Sons, Inc., Hoboken, NJ.

Earp, M. et al (2008) "Asssessing the Effect of Calibration on Nonresponse bias in the 2006 ARMS Phase III Sample Using Census 2002 Data," U.S. Department of Agriculture, National Agricultural Statistics Service, RDD Research Report, RDD-08-06, Fairfax, VA.

Ford, B. (1976) "Missing Data Procedures: A Comparative Study," U.S. Department of Agriculture, Statistical Reporting Service, Sampling Studies Section, SF 76-02, Washington, DC.

Ford, B. (1978) "Missing Data Procedures: A Comparative Study (Part 2)," U.S. Department of Agriculture, Economics, Statistics, and Cooperatives Service, Washington DC.

Gardner et. al. (2008) "Understanding American Agriculture: Challenges for the Agricultural Resource Management Survey," The National Academies Press, Washington, DC.

Groves et. al. (2004) "Survey Methodology," John Wiley & Sons, Inc., Hoboken, NJ.

Hoge, S. & Willmack, D. (1991) "Analysis of Item Nonresponse, Imputation and Editing in the 1989 Farm Costs and Returns Survey for Iowa and North Carolina," U.S. Department of Agriculture, National Agricultural Statistics Service, SRB Research Report, SRB-91-09, Washington, DC.

Hogye, M. (2009) "ARMS III SAS Imputation Programs: Overview," U.S. Department of Agriculture, National Agricultural Statistics Service, Fairfax, VA.

Hopper, R. (2006) "2006 Agricultural Resource Management Survey: Phase III Cost and Returns Report - Survey Administration Manual," U.S. Department of Agriculture, National Agricultural Statistics Service, Washington, DC.

Hopper, R. (2007) "2006 Agricultural Resource Management Survey: Phase III Cost and Returns Report – Survey Administration Manual," U.S. Department of Agriculture, National Agricultural Statistics Service, Washington, DC.

Kott, P. & Bailey J. (2000) "The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling," Proceedings of the Second International Conference on Establishment Surveys, Invited Papers, 269-278.

Kott, P. (2000) "Poisson Sampling, Regression Estimation, and the Delete-a-Group Jackknife," Joint Statistical Meetings, Indianapolis, Indiana.

Little, R. & Rubin, D. (1987) "Statistical Analysis with Missing Data," John Wiley & Sons, Inc. New York, NY.

Miller, D. & O'Connor, T. (2010) "ARMS III Item Nonresponse Overview," Presentation to NASS RDD, April 8, 2010, U.S. Department of Agriculture, Washington, D.C.

Office of Management and Budget (2006) "Standards and Guidelines for Statistical Surveys," Office of Management and Budget, Washington, DC.

Robbins, M. et al. (2009) "ARMS III Multivariate Imputation," Presentation to NASS Operations, December 4, 2009, U.S. Department of Agriculture, Washington, D.C.

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

United States House of Representatives, Office of the Law Revision Counsel, "U.S. Code, Title 7, Chapter 55, Section 2266a" Washington, D.C.

104th Congress, "Food Quality Protection Act of 1996," Public Law 140-170, Washington, D.C.

6. Appendix 1

Item Code	Variable Description
P525	Direct Payments Received
P526	Counter-Cyclical Payments Received
P527	Direct Payments For Target Commodity
P528	Loan Deficiency Payments Received
P529	Loan Deficiency Payments, etc. Received
P530	CC/LDP Payments For Target Commodity
P531	Marketing Loan Gains Received
P532	Counter-Cyclical Payments for Target Crop
P533	Value of Commodity Certificates
P534	Government payments Received Through Coops
P535	Peanut Quota Buyout Payments
P536	Milk Income Loss Payments
P537	Agriculture Disaster Payments
P538	Cropland Reserve Program (CRP) Payments
P539	Wetland Reserve Program (WRP) Payments
P540	Environmental Quality Incentives Program (EQIP) Payments
P541	Other Federal Agriculture Program Payments
P542	Other State/Local Agriculture Program Payments
P543	Landlord Received in Government Payments
P544	CRP WRP EQIP Payments
P545	Other Federal/State Agriculture Program Payments
P557	Other Federal/State Payments for Target Commodity