Working Paper.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (ii): New and emerging methods

Assessing the Impact of a New Imputation Methodology for the Agricultural Resource Management Survey

Prepared by Wendy Barboza, Darcy Miller, and Nathan Cruze, National Agricultural Statistics Service, United States Department of Agriculture, USA

I. Introduction

1. The National Agricultural Statistics Service (NASS) is a statistical agency located under the United States Department of Agriculture (USDA). NASS's mission is to provide timely, accurate, and useful statistics in service to U.S. agriculture. In order to successfully accomplish the agency's mission, NASS conducts hundreds of surveys every year and publishes numerous reports covering virtually every aspect of U.S. agriculture. Some examples of areas covered in NASS's reports are production and supplies of food and fiber, prices paid and received by farmers, farm labor and wages, farm income and finances, chemical use, and rural development. A wide variety of topics are covered within these different areas. The subject matter ranges from traditional crops, such as corn and wheat, to specialty commodities, such as mushrooms and flowers; from agricultural prices to land in farms; from once-a-week publication of cheddar cheese prices to detailed census of agriculture reports every five years. In order to publish these reports, the size of the target population varies from fewer than 50 for a survey to nearly 3 million for the census of agriculture.

2. The Agricultural Resource Management Survey (ARMS) is conducted by NASS and cosponsored by the USDA's Economic Research Service (ERS). The ARMS provides an annual snapshot of the financial health of the farm sector and farm household finances, and it is the only source of information available for objective evaluation of many critical policy issues related to agriculture and the rural economy. Its data are essential to USDA and other federal administrative, congressional, and private-sector decision makers when they must weigh alternative policies and programs or business strategies that touch the farm sector or affect farm families.

3. 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, which changes 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 multi-mode, dual frame survey conducted annually in all states except Alaska and Hawaii. The sample consists of approximately 35,000 farms and ranches and is selected from NASS's list frame that attempts to cover all agricultural establishments within the U.S. and an area frame which compensates for the incompleteness of the list frame. The survey questionnaire is mailed to the entire sample, but additional modes of data collection include web, face-to-face, and computer-assisted telephone, although telephone interviews are rare for this survey.

4. Based on data collected from the ARMS III, NASS publishes estimates of farm production expenditures for the U.S. (except Alaska and Hawaii) in addition to five regions. The regional estimates are broken down by the fifteen leading cash receipt states and then all other states within the region. Farm production expenditures are also estimated for eight economic sales classes and two

farm type categories. In addition to farm production expenditures, the ARMS III also collects data on production practices and costs of production for one to three targeted crop and livestock commodities each year, selected on a rotational basis. The production practices and cost of production data for these designated commodities are collected in the top producing states while the farm production expenditures data are collected in all states (except Alaska and Hawaii).

II. Issues with Missing Data

5. Because the survey data are utilized for in-depth analyses of critical policy issues related to agriculture and the rural economy, the ARMS III survey questionnaire is long and complex. Some versions of the survey are 51 pages long, with data collected on more than 800 variables. The 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 respondent and these items are often the most problematic, sometimes with over half of the observations missing. NASS has taken extensive steps to increase awareness of the importance of the survey as well as to reduce respondent burden by utilizing a sampling procedure that reduces the probability of an operation being selected two years in a row.

6. Response to the ARMS III is voluntary. Like many surveys, the ARMS III is subject to both unit nonresponse (the sampled record does not respond to the entire questionnaire) and item nonresponse (the respondent does not answer an item(s) on the questionnaire). Both unit and item nonresponse create gaps in the data that need to be addressed prior to the estimation process. 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 item of interest. The quality of estimators in the presence of missing data is subject to the ability of the nonresponse adjustment or imputation procedure to adequately account for the missingness.

III. Nonresponse and Imputation Methodology

7. The first step in the editing procedures is for a statistician to perform a cursory manual edit of the ARMS III questionnaire. The questionnaire is then processed through a computer edit that checks the consistency of the data and verifies data values fall within a certain range. After this, a statistician reviews all questionnaire items that fail any of the edits. The statistician has the option of manually imputing the data item or letting the computer program impute the item. A manual imputation is typically performed over a machine imputation when the statistical analyst has knowledge about the questionnaire item for that operation.

A. Current Imputation Procedure

8. For the ARMS III, item-level nonresponse is accounted for by imputing data where there are missing values. Imputed values are calculated through an automated imputation system that calculates an unweighted mean for an imputation group based on locality, farm type, and value of sales class. These groups of homogeneous farms exclude extreme outliers (both high and low) so that the imputed values are not biased as a result of a few large/small or unique operations. An imputation group must have a minimum of ten or more positive responses. When a group lacks a sufficient number of responses, groups are collapsed by value of sales class, locality, and farm type according to a defined hierarchy preserving as much of the homogeneity as possible. About 75 percent of the necessary imputations are completed at the first level of imputation groups. The imputation algorithm fails to deliver an acceptable value less than one-half of one percent of the time.

9. After the imputation routine is complete, the records with imputed data are re-edited to ensure the imputed values are acceptable. Relationships between data items on the current survey are verified and in certain situations, items are compared to data from earlier surveys to make certain specific relationships are logical. A statistician is required to manually impute any item that fails an edit or

could not be imputed. The edit logic also ensures administrative coding follows the methodological rules associated with the survey design. In this case, a statistician is required to manually impute any item that fails an edit.

.

B. Unit Nonresponse Adjustment

10. Calibration is a weighting technique used in survey sampling to adjust the survey weights for sampled elements so that the weighted sum of a set of benchmark variables equals a pre-determined set of values for the population. For the ARMS III, the weights generated from the sampling procedures are used as input into the calibration algorithm. Sampling weights are calculated based on numerous factors so that the sample allocation can be representative of the entire population of farms at the state level for the fifteen leading cash receipts states and the five regions for all other states.

11. Due to survey nonresponse and the possibility of disproportionate responses across different farm types and economic sales classes, weights are adjusted through a calibration algorithm. Calibration adjusts the sampling weights so that the expanded data will match several known commodity, livestock and farm number published totals. This ensures that the expense data collected will accurately represent the expense breakdowns for all farm types and farm sizes as well as cover the expenses for the entire target population.

C. Proposed New Imputation Procedure

12. There are some disadvantages with the mean imputation methodology, especially in the ARMS III. The methodology relies on the use of conditional means as estimates of missing values. For survey estimates of univariate-level statistics or statistics cross-classified by several variables, this methodology should be adequate in general. However, estimates of variability in the data will typically be artificially reduced. When more complex multivariate relationships are estimated, conditional mean imputation generally cannot condition on a sufficiently large set of variables to maintain relationships between the variables imputed and all variables that might be included as related variables in a multivariate analysis. A multivariate imputation approach preserves important relationships, the distribution of the respondents' data, and provides a better estimate of uncertainty.

13. To develop methodology that would incorporate more information (covariates) when making imputations, NASS entered into a cooperative agreement with the National Institute of Statistical Sciences (NISS). The new methodology requires data to be transformed marginally and 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. Various criteria are used to select the covariates which allow 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 sampling method. Consequently, this new method developed is labelled iterative sequential regression (ISR).

D. Calibration and Imputation Interaction

14. Economic class information used for both imputation methods is based on information collected for the operation up to the point of the imputation. The calibration routine is run after the imputations are made and the economic class of each operation and determination of an operation being in-scope is updated after incorporating the additional information provided from the imputations. In-scope means that the operation is in the target population. Some of the imputed variables are used in the updating process. So, although estimated totals of imputed variables are not directly compared to the published totals in calibration, imputed variables can affect the calibration through the updated economic class assignment and determination of being in-scope.

15. For the conditional mean imputation methodology, it is possible for an operation to change from one economic class to another or change in-scope status after imputation. However, due to the nature of utilizing a mean for imputation, this occurrence is rare. Contrarily, in the case of ISR, imputations are being drawn from a distribution, and there is a possibility to draw an extreme value from the tail

of the distribution for one of the variables used in updating the economic class and in-scope status. Operations are more likely to change economic class or occasionally change in-scope assignment using ISR. Hence, the calibrated weights for a dataset imputed using the conditional mean methodology may differ from the calibrated weights when ISR is used to impute the same dataset.

E. Review of Outliers

16. Outliers may be caused by aging control data resulting in misstratification, data errors, or the nonresponse and calibration adjustments to the sampling weight. A preliminary calibration and summary are run and any individual record accounting for 0.5 percent of the national estimate for total expenses or 2.5 percent of a regional estimate for total expenses is tagged as an outlier. After verifying the data have not been misrecorded or mishandled, background information on these outliers is compiled and presented to a National Outlier Board. This Board is a team of NASS and ERS analysts that meet to discuss the national outliers and form a consensus on a course of action.

17. Most outliers trace back to unique situations that do not exist in the target population as often as a large calibrated sample weight indicates. The Board looks at other respondents of the same locality, farm type, and sales class as the reported data on the outlier. The Board examines the weights of the comparable respondents and most often overrides the outlier's weight with the median weight of the comparable respondents. After the most extreme outliers have been addressed, the Board reviews the national totals by expense category following the same methodology and, when necessary, overrides the outlier's weight with the median weight of the comparable reports. Finally, staff within NASS examine outliers found at the state level for the published expense categories. A determination is made as to whether a weight adjustment is justified. Adjustments are not made to all outliers, but they are reviewed closely for accuracy. It is important to note that the calibration algorithm is implemented after each stage of the outlier review process.

IV. Impact of Methodology Change on the Estimates

18. The new multivariate imputation technique (i.e., ISR) has been thoroughly researched and is in the process of being evaluated on 2011 and 2012 data. The results of this evaluation will be utilized to access the impact of the new imputation methodology on the survey estimates along with a parallel test in 2013. To date, several operational hurtles have prevented the authors from presenting a more comprehensive analyses comparing the survey estimates using the two methodologies. A more detailed paper will be presented at the Joint Statistical Meetings in August 2014, which will be held in Boston, Massachusetts, USA.

A. Summary Estimates

19. As stated earlier, the ARMS III survey collects data on more than 800 variables. In 2011 and 2012, nearly 150 of these questionnaire items were eligible for potential machine imputation. After the final calibration is performed, estimates are generated during a process referred to as the summary. During this process, there are more than 400 estimates generated (typically as sums and ratios) and an estimate may be defined in terms of one or more imputed component variables. NASS publishes a set of 18 key variables in the annual Farm Production Expenditures report, which is issued annually in August. These key variables have been the initial focus of comparisons between the conditional mean imputation and ISR methodologies. The variables cover 17 broad categories of farm spending (e.g., seeds, livestock and feed expenditures, fertilizer, chemicals, labor, taxes, and machinery and capital expenditures) which are further summed to obtain total farm production expenditures. Thus, it is important to measure the impact of the change in estimates between both imputation methodologies after the summary. Note that the outlier review process discussed earlier was <u>not</u> performed prior to producing these estimates.

B. Comparison of Select Estimates after the Summary

20. Ultimately, the performance of ISR and conditional mean imputation must be assessed for a large set of estimates at the state, regional, and national levels. For brevity, comparisons of the resulting expenditure estimates from ISR and conditional mean imputation are shown for the 15 leading cash receipts states. In addition, comparisons on just four of the 18 aforementioned key variables are presented, including agricultural chemical expenditures, farm services expenditures, all tax expenditures, and total expenditures. In 2011 and 2012, the first three variables represented approximately 4, 11, and 3.5 percent of national total expenditures, respectively. For each state, two corresponding post-imputation datasets were produced using ISR and conditional mean imputation (referred to as mean in the upcoming formula). The comparison lends itself to a matched pairs t-test. Tests of the null hypothesis that the difference in estimates (ISR-mean) is equal to zero were conducted with a 5 percent level of significance. These differences are depicted by confidence intervals in Figures 1-4 below. Since the sample sizes within the 15 estimating states are quite large (the smallest being Indiana (2011), n=641), the half-width of each 95% confidence interval is approximately two standard errors wide. The midpoints and endpoints of the intervals were converted to show percent change in the estimate between the ISR and conditional mean imputations (i.e., percent change = 100*(ISR-mean)/mean). Thus, a positive value indicates that the estimate using ISR is greater than the estimate using the conditional mean imputation.

21. Agricultural Chemical Expenditures: This estimate takes into account both the costs of materials and the cost of application of agricultural chemicals. None of the questionnaire-level components of the agricultural chemical expenditures estimate is imputed and this is reflected in the small number of significant differences shown in Figure 1. Note that while the definition of the agricultural chemical expenditures does not include any imputed component variables, both imputation methodologies have an interaction with the determination of calibration weights. This interaction accounts for the significant differences observed in Kansas and North Carolina in 2011 and in Texas in 2012. The cause for the difference in Kansas in 2011 merits further exploration to understand why the calibration routine yielded such a large change. Although significant differences were observed in North Carolina and Texas, the percent changes are small (less than one percent in absolute terms) and are of little practical difference.

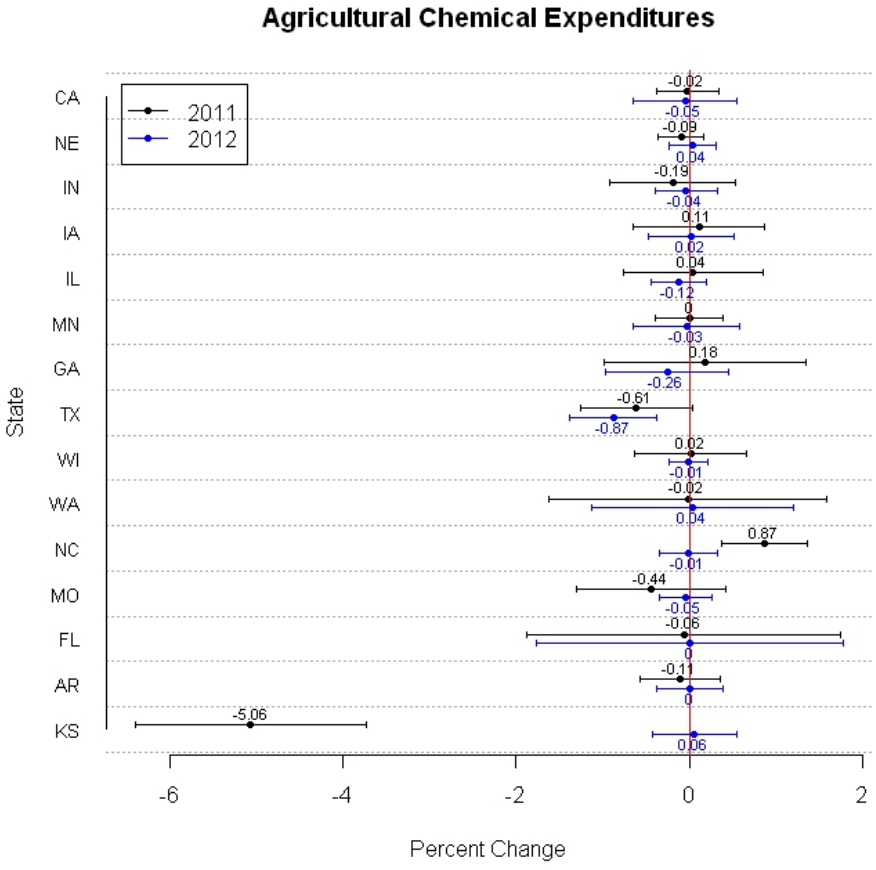

Figure 1: Percent change in estimates for agricultural chemical expenditures

22. Farm Services Expenditures: Farm services expenditures is the sum of 36 component variables which describe crop and veterinary custom services, transportation, marketing, storage, utilities, and a variety of other business expenditures. In particular, three variables related to farm marketing and storage costs are imputed. This contributes to the scatter of intervals shown in Figure 2. In four cases (Kansas and California in 2011 and Georgia and Washington in 2012), the ISR methodology results in an absolute change in the estimate of 10 percent or more relative to the conditional mean imputation. California shows an especially large increase in the difference between the estimates. It was determined that this is largely due to a large contractor's marketing and storage expenditures imputed for one farm. If the outlier review process was performed prior to producing these estimates, this large difference would not have occurred.

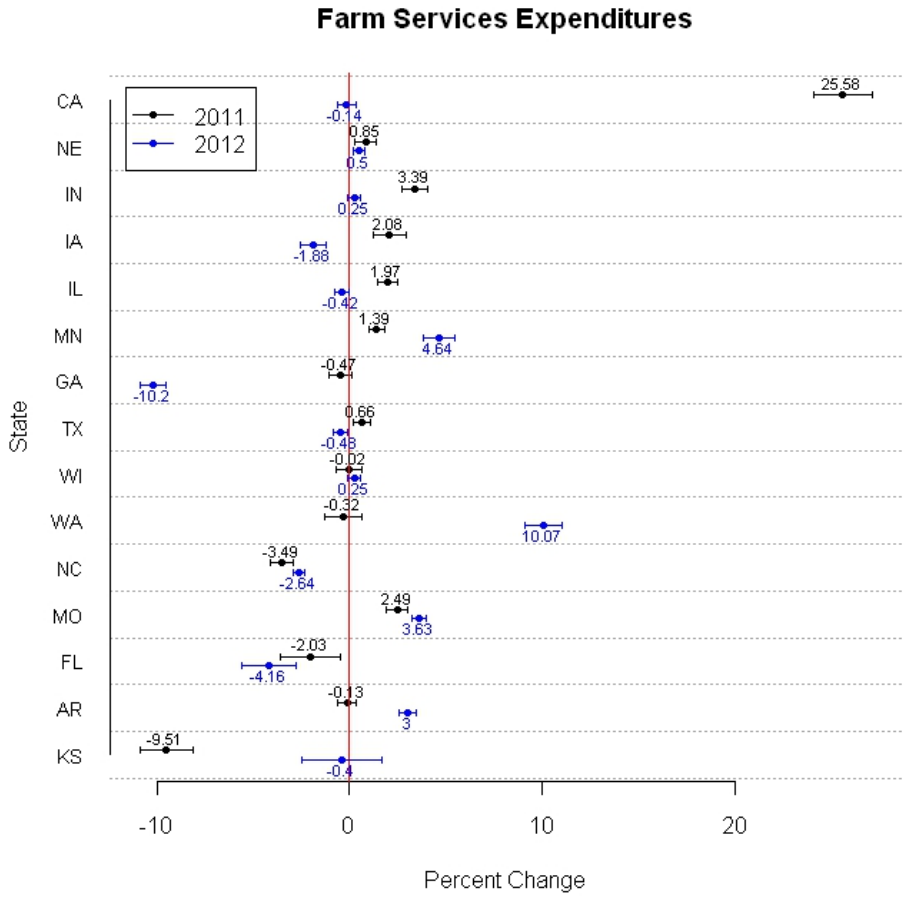

Figure 2: Percent change in estimates for farm services expenditures

23. All Tax Expenditures: All tax expenditures is the sum of six component variables related to the real estate and personal property taxes paid by operators, landlords, and contractors. Two of these components are imputed variables related to real estate taxes paid by landlords and paid by contractors; these two variables have the highest imputation rate in the ARMS. A further analysis showed that there was little or no change between ISR and conditional mean imputation in terms of contractor real estate taxes, and that typically only a dozen observations nation-wide even contained non-zero values. In contrast, the real estate taxes paid by the landlord were imputed in more than half of all cases, contributing to the high degree of scatter shown for both years in Figure 3. In general, the direction of change seemed to be positive, indicating that ISR may be imputing higher landlord real estate taxes than the conditional mean imputation and possibly determining higher calibration weights. Missouri (2011) and Kansas (2012) were the only exceptions observed among the largest cash receipt states.

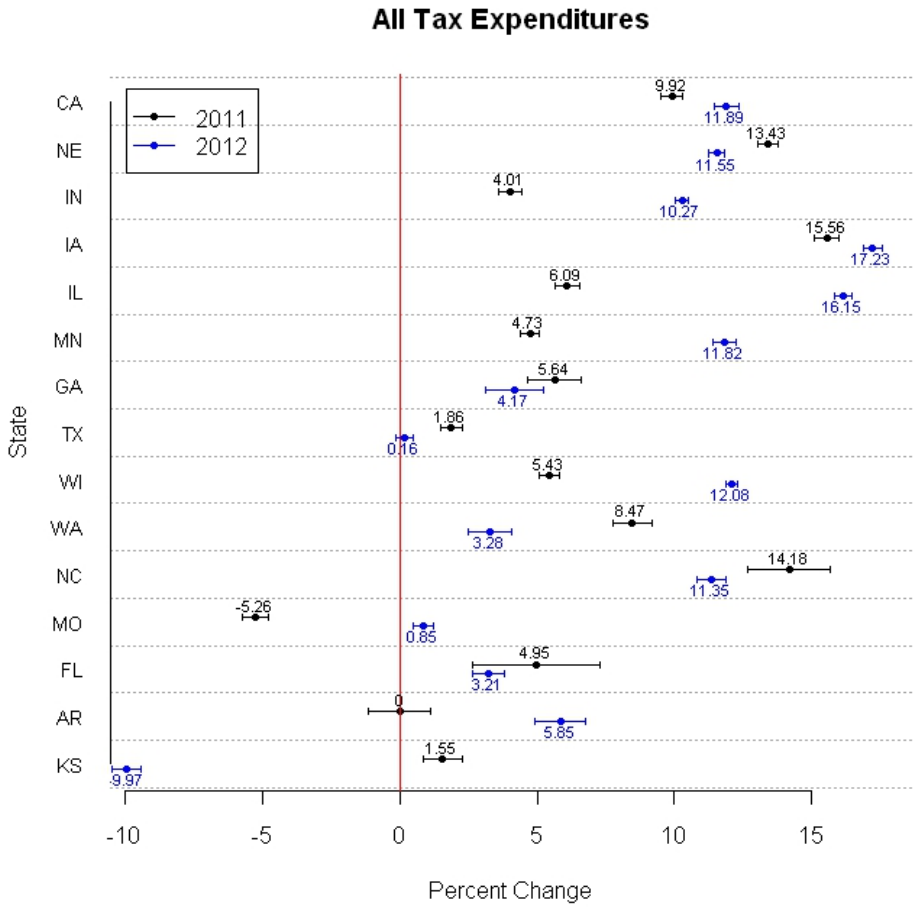

Figure 3: Percent change in estimates for all tax expenditures

24. Total Expenditures: Total farm production expenditures is the sum of 17 other variables, including agricultural chemical expenditures, farm services expenditures, and all tax expenditures. The total expenditures estimate depends on the three imputed variables contained in the farm services estimate and the two imputed variables contained within the all tax expenditures estimate (i.e., the other 15 variables which go into total expenditures do not contain any imputed components). Figure 4 shows that the overall impact on the total expenditures estimates between the two methodologies is generally small, usually resulting in less than a 1 percent change. Obvious exceptions are California and Kansas in 2011, and Georgia and Washington in 2012; note that these differences correspond exactly to the four most extreme differences observed in the farm services expenditures estimates as depicted in Figure 2. Outliers are determined by an individual record's share of total expenditures; if the differences shown in Figure 4 are being driven by a few unusual observations, these would likely be detected and corrected in the outlier review process.

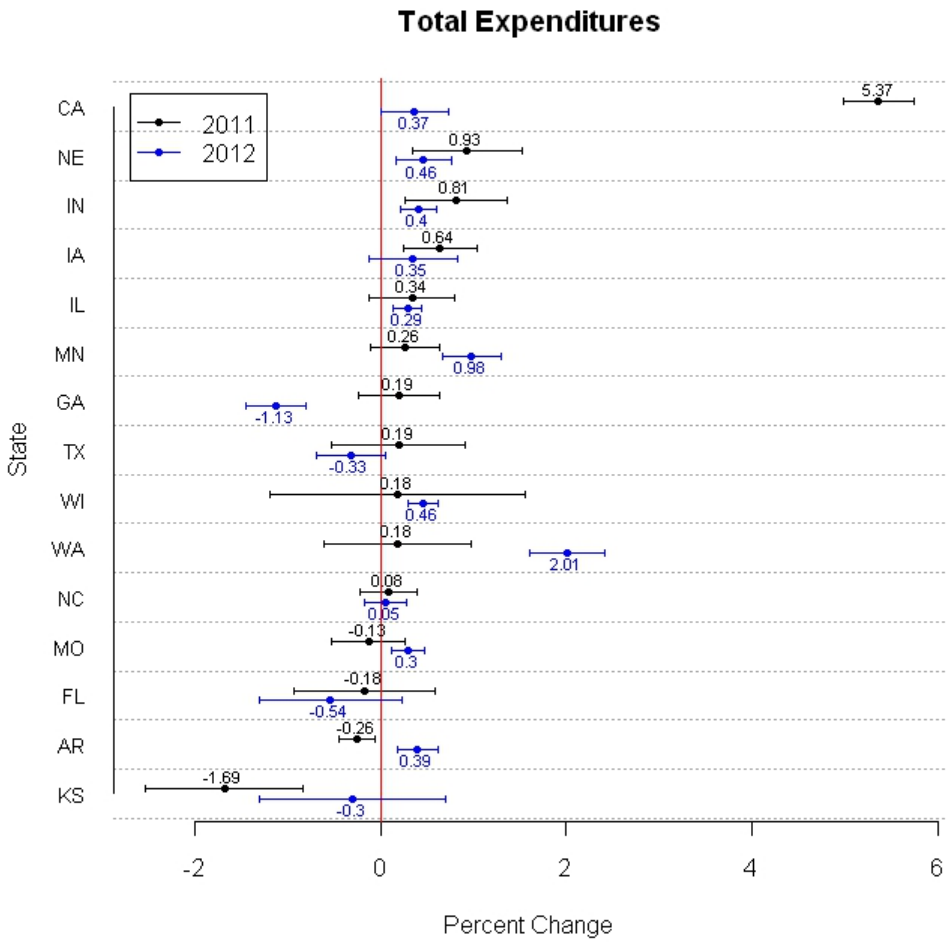

Figure 4: Percent change in estimates for total expenditures

V. Conclusion

25. As discussed earlier, there are some major disadvantages with using the conditional mean imputation methodology. Since there are many complex multivariate relationships in the ARMS III, conditional mean imputation generally cannot condition on a sufficiently large set of variables to maintain relationships between the variables imputed and all variables that might be included as related variables in a multivariate analysis. ISR will preserve important relationships, the distribution of the respondents' data, and provide a better estimate of uncertainty. Preliminary analysis has shown that there is a significant difference in some estimates between the two methodologies. However, it is difficult to compare the two methodologies directly because of the interaction between calibration and imputation as well as the lack of an outlier review process for this evaluation. The preliminary results are promising and the differences will be explored in more detail in the near future.

VI. References

National Agricultural Statistics Service (2014). "Farm Production Expenditures Methodology and Quality Measures".
http://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Farm_Production_Expenditures/08_2013/fpxq0813.pdf

National Agricultural Statistics Service (2014). "Farm Production Expenditures 2012 Summary".
http://usda01.library.cornell.edu/usda/current/FarmProdEx/FarmProdEx-08-02-2013.pdf

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.

Robbins, M., Ghosh, S., and Habiger, J. (2010). “Innovative Imputation Techniques Designed for the Agricultural Resource Management Survey”. Proceedings of the 2010 Joint Statistical Meetings, pages 634-641.