Nonresponse Bias Adjustment in Establishment Surveys: A Comparison of Weighting Methods using the Agricultural Resource Management Survey (ARMS)

Morgan Earp[1], Melissa Mitchell[2], Phil Kott[3], Frauke Kreuter[4], & Eric Porter[5]

[1]Bureau of Labor Statistics, PSB Suite 1950, 2 Massachusetts Avenue NE, Washington, DC 20212
[2]National Agricultural Statistics Service (NASS), 3251 Old Lee Highway Room 305, Fairfax, VA 22030
[3]Research Triangle Institute, 6110 Executive Blvd., Suite 902, Rockville, MD 20852
[4]University of Maryland, 1218 LeFrak Hall, College Park, MD 20742
[5]NASS, 1400 Independence Avenue SW, Washington, DC 20004

Abstract

There are numerous ways to address nonresponse bias adjustment in surveys; two such methods are calibration weighting and propensity score models. Calibration is a viable technique when good external benchmarks exist; however, good external benchmarks are not always available. An alternative method to calibration is to use propensity scores to adjust for nonresponse. There are at least three main modeling techniques used to create propensity scores, but little if any research has focused on which methods provide the best propensity scores in terms of nonresponse adjustment. This paper compares calibration weights with three propensity score adjustment methods. One propensity weight is based on logistic regression models; the other two are based on classification trees (using either a single or an ensemble tree approach).

This research focused on the Agricultural Resource Management Survey Phase III (ARMS III), which adjusts for potential bias resulting from unit nonresponse by calibrating weights so that estimates equal published benchmarks from other sources. Using Census of Agriculture (COA) data, we were able to compare the effectiveness of using calibration weights versus propensity score weights to reduce (unit) nonresponse bias. Bias comparisons were done by using COA data as proxy data for the 2000-2008 ARMS III samples, since the COA includes items surveyed on ARMS III as well as a number of items pertaining to operational characteristics. Nonresponse bias of the mean was compared across 30 production and demographic type items. The results indicate that tree weights outperform logistic regression weights, and that calibration weighting reduces nonresponse bias of the mean to the lowest levels. The results also suggest that tree weighting is the next best option when calibration targets are not available.

Key Words: Nonresponse Adjustment; Nonresponse Weighting; Nonresponse Bias; Propensity Scores; Calibration; Classification Trees; Ensemble Trees; Logistic Regression

1. Introduction

Survey nonresponse is concerning to both statisticians and survey methodologists; however, each has a different perspective on how to address nonresponse bias. According to Singer (2006), statisticians mainly focus on adjusting for nonresponse, where survey methodologists are more interested in understanding reasons for nonresponse and increasing response rates.

We will concentrate here on unit (whole-record) nonresponse. Both statisticians and survey methodologists use propensity scores to manage unit nonresponse. Traditionally, propensity scores are developed using logistic regression (Little and Vartivarian, 2005; Little, 1986; Rosenbaum and Rubin, 1983); however, in cases where there are large amounts of auxiliary data, using logistic regression is not the best approach for two main reasons: 1) due to the fact that we have to hypothesize specific variables up front, we are forced to assume that these are the only causes of nonresponse; and 2) the more auxiliary variables we include in our model the more problematic it will be to specify interaction terms, account for missing data, avoid issues of multicollinearity, and interpret the results (Phipps and Toth, 2012). Alternatively, if we use a data mining approach such as classification trees, we are able to include a large number of auxiliary variables, automatically detect significant interaction effects worth exploring, automatically include item missingness as an indicator of nonresponse, and use multicollinearity to our benefit by allowing variables that are highly correlated to work as surrogates when other variables are missing.

Like household surveys, establishment surveys have (unit) nonresponse, but with establishment surveys nonresponse bias can be a more serious threat to estimates since individual establishments can have a large effect on the final estimates even after larger establishments are given higher probabilities of selection. According to the 2007 Census of Agriculture (COA), 0.3 percent of farms with total annual sales of five million dollars or more accounted for 27.9 percent of total sales in the US; thus, the impact of nonresponse on the estimate of total sales is much greater for these operations than for other operations (US. Department of Agriculture, 2007, Table 2). To adjust for possible nonresponse bias, NASS weights the Agricultural Resource Management Survey Phase III (ARMS III) respondent sample so that estimated variable totals for a large set of items match "target" figures from other sources. This is done through a weighting process called "calibration" (Deville and Sarndal, 1992; Kott and Chang, 2010). Calibration weighting adjusts the survey weights so that certain targets, typically estimates from sources outside the survey, are met. NASS uses official estimates as calibration targets since they are correlated with the economic activity of farm operations; calibration targets include estimates of total number of farms, total number of farms by state and by economic class; corn, soybeans, wheat, cotton, hay, rice, peanuts, sugar (sugarcane/sugar beets), tobacco, fruits, and vegetable acreage; egg and milk production; cattle, hog, broiler, and turkey inventories; and nursery and floriculture Using some form of calibration weighting assures that the calibration-weighted sum of the survey data will equal the NASS official estimate produced using sources other than ARMS. NASS uses a truncated linear version of calibration (Kott 2009, p 74, 75). In this version no adjusted weight is allowed to fall below one Sometimes, however, calibration targets cannot be reached and need to be dropped.

In addition to reducing the confusion in the user community that might result if NASS released alternative estimates for the same totals, calibration weighting produces ARMS

Phase III estimates that generally have less bias than the unadjusted estimates; however, in 2008 a third of the 30 variables assessed still exhibited nonresponse bias levels that were significantly different from zero even after calibration; thus, leading in part to the research of alternative weighting methods discussed in this paper (Earp, McCarthy, Porter, & Kott, 2010).

Nonresponse bias is very difficult to evaluate directly, since data are lacking for the nonrespondents. Fortunately, data similar to those collected in ARMS III are available from the Census of Agriculture (COA). We can measure the difference between the average ARMS III respondent and the average of the full sample without any nonresponse adjustment, after calibration weighting, and after propensity score adjustments using logistic regression and classification tree procedures.

Propensity methods have been developed to reduce a large set of covariates to one single variable with which adjustment is done (Rosenbaum and Rubin, 1983). A propensity score is the fitted probability that a given case will become a survey respondent. Both logistic regression and classification tree models can be used to create the propensity scores. However, given the lack of knowledge about the nonresponse mechanism in ARMS III in particular and establishment surveys more generally, data-driven methods like classification trees might be more suitable. Classification trees offer a number of advantages over logistic regression: 1) classification trees automatically detect significant relationships and interaction effects without pre-specification, reducing the risk of selecting the wrong variables or other model specification errors; 2) the classification tree models identify both the variables that are correlated with the target variable, but also the optimal breakpoints within these variables for maximizing their correlation; 3) they identify hierarchical interaction effects across numerous variables and summarize them using a series of simple rules; 4) they incorporate missing data into the model and assess whether missingness on a given variable is related to the target; 5) they create a series of simple rules that are easy to interpret and use for identifying subgroups with higher propensities; and 6) they reduce the subjectivity of selecting variables to include in the model. This paper will compare nonresponse adjustment as currently done using calibration, with four sets of propensity score adjustments, one derived directly from a logistic regression, one using the logistic regression to adjust the base weights within 10 classes (n/r), one using a classification single-tree, and one using an ensemble of trees.

Other nonresponse models have been developed using auxiliary data, but most begin with hypotheses about a small set of relevant predictor variables and have generated response propensity scores based on logistic regression or similar models (Abraham, Maitland, and Bianchi, 2006; Johansson and Klevmarken, 2008; Johnson, Cho, Campbell, and Holbrook, 2006; Lepkowski and Couper, 2002; Nicoletti and Peracchi, 2005). These types of models may accurately predict which cases become nonrespondents, but they do not typically include large sets of auxiliary variables. Furthermore, as Groves (2006) states, this approach assumes that these "…variables are the only possible 'common causes' of response propensity and survey variables." (p. 654)

While other propensity score models have been built using decision trees, most typically use a single tree to predict nonresponse (Phipps and Toth, 2012). According to Phipps and Toth (2012), by taking a more conservative approach and only modeling the variables that are strongly associated with response, they are able to avoid over fitting and thus produce more stable estimates; however, by limiting the analysis to only the variables with strong associations to nonresponse, they also admit possibly limiting the

accuracy of their estimates. According to Dietterich (1999), bagging can be used to exploit model instability and improve classification accuracy. Bagging involves creating multiple trees with varying criteria and then taking the average propensity score across all of the trees (Brieman, 1998). This approach results in an ensemble of classification trees, which is more stable and powerful than a single classification tree (Brieman, 1998).

Models can be combined in various ways. In the current study, we calculated the average of the propensity scores produced across all of our classification trees, which, according to Bauer and Kohavi (1999), performs better than the other methods used to combine trees.

Although the 2007 COA data do not perfectly match the 2008 ARMS III data, they are moderately to highly correlated (Earp, McCarthy, Porter, & Kott, 2010). This paper will compare 2008 ARMS III survey respondents to nonrespondents using their 2007 COA data. One of the weaknesses of this approach according to Groves is that not all of the variables of interest have auxiliary data available (2006). While it is true that not everything collected on ARMS is available on the COA, we were able to assess bias across 30 estimates including both household and establishment type items that are of particular interest to both NASS and the Economic Research Service (ERS).

2. Method

The ARMS III is an annual survey conducted by NASS and ERS. ARMS III is one of the most complex and detailed sample survey data collections conducted by NASS. It collects calendar year economic data from agricultural producers nationwide. The ARMS is conducted in three phases. Phase I screens for potential samples for Phases II and III. Phase II collects data on cropping practices and agricultural chemical usage, and Phase III collects detailed economic information about the agricultural operation, as well as the operator's household. Phase III is the only phase of the ARMS with unit response rates lower than 80 percent, which falls below the Office of Management and Budget requirement. Surveys with less than an 80 percent unit level response rate are required to complete an analysis of nonresponse bias (United States, 2006). This paper focuses on unit nonresponse and in part addresses a recommendation made by the Committee on National Statistics report (2007) to identify characteristics of ARMS nonrespondents. This paper focuses on unit nonresponse and in part addresses the recommendation to identify characteristics of ARMS nonrespondents. This paper also considers how this information could be used to create adjustment weights.

The COA is a mandatory collection of data from all known agricultural operations. NASS has data from the COA on items of interest for many of the ARMS nonrespondents; however, the COA itself is incomplete. An estimated 16.2 percent of all farms were missing from the 2007 COA Mailing List, and 14.6 percent of farms on the Mailing List failed to respond to the COA (USDA, 2007, Table A). Moreover, 5.7 percent of the operations sampled for ARMS III could not be matched to 2007 COA records. Nevertheless, by comparing the 2007 COA values of the ARMS III respondents to the full sample of ARMS III cases (including nonrespondents), data from the COA were matched to both respondents and nonrespondents in the ARMS III 2000-2008 samples to create response propensities, as was done by Groves and Couper (1998). Matching 2007 COA data were available for 71.3 percent of the ARMS III 2000-2008 sample. Nonresponse bias was assessed using just the 2008 sample; for that year matching 2007 COA data were available for 94.3 percent of the records. The match rates

for the 2008 ARMS III respondents (94.5 percent) and nonrespondents (94.1 percent) were approximately equal. Our analysis is based just on these matching cases.

Sixty-nine COA variables were used to model ARMS III nonresponse using logistic regression and classification trees (Earp, Mitchell, McCarthy, & Frauke, 2012, Table 2). For both models, we used variables thought to be related to unit nonresponse as predictors. These variables include operator demographics, farm type, size, commodities raised, expenses, and location. Our analysis of nonresponse bias focuses on 30 specific production and demographic variables collected on both the ARMS and the COA. These variables were selected by NASS and ERS subject area experts as the variables that are common to both ARMs and COA. Nineteen of these 30 variables were included in the logistic regression and tree models. The majority of the variables not included as predictors are not collected during data collection, but are calculated after data collection.

The logistic regression and tree approaches are quite different from calibration. Calibration does not include any household attributes as targets, since they are not considered to be related to physical farm attributes as much as business economic information. The logistic regression and tree models do include both business unit and household attributes as predictors of unit nonresponse.

Unit nonresponse propensity scores were created using logistic regression and classification trees. Both the logistic regression and the classification tree models were set to predict the probability that operations were unit respondents.

The logistic probability of response, p_i^{lr} , has the form

$$\text{logit}\left(p_i^{lr}\right) = \log\left(\frac{p_i^{lr}}{1-p_i^{lr}}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK},$$

where the subscript i denotes a farm, x_{ik} the k^{th} ($k = 1, \ldots, K$) explanatory variable associated with farm i, and β_k is the k^{th} logistic regression coefficient. The response probability is estimated using a logistic regression routine by

$$\hat{p}_i^{lr} = \frac{exp(b + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_K x_{iK})}{1 + \exp\left(\alpha + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_K x_{iK}\right)},$$

where b_k is a large-sample estimator for β_k. This leads to initial nonresponse weights of

$$\widetilde{w}_i^{lr} = w_i^0 \frac{1}{1-\hat{p}_i^{lr}} \ , \qquad (1)$$

where w_i^0 was the base weight for farm i. The base weights are each farm's ARMS III sample weight before calibration multiplied by its COA sample weight (the latter to account for the COA's undercoverage).

In addition to using the $\widetilde{w}_i^{lr}$, we also followed the advice of Eltinge and Yansaneh (1997) and sorted the sample by the farms' $\hat{p}_i^{lr}$ values. Farm classes were created based on sorted quintiles ($C = 5$) and deciles ($C = 10$). A pooled inverse probability of response was estimated within each class c ($c = 1, \ldots, C$; $C = 5$ or 10) as

$$a_c = n_{cs}/n_{cr}, \quad (2)$$

where n_{cs} is the sample size of class c, and n_{cr} the respondent size.

The alternative logistic regression weight for a respondent in c was then

$$w_i^{lr} = w_i^0 a_c. \quad (3)$$

Eltinge and Yansaneh argued that the final weights from equation (3) are often less variable than those from equation (1), which in turn leads to more stable estimates. Nevertheless, the use of equation (3) should still remove much of the nonresponse bias. Moreover, by employing the estimates of the actual (weighted) response rates within classes (the $1/a_c$) in equation (2), w_i^{lr} may be more reflective of the true shape of the nonresponse function than $\widetilde{w}_i^{lr}$.

The logistic regression model was used to predict likely nonrespondents, where the classification trees were used to model characteristics of likely nonrespondents. Therefore, while the logistic regression analysis included all operations sampled for ARMS III between 2000 and 2008 with matching 2007 Census of Agriculture Data, the classification trees were created using only a randomly selected subset of the data to avoid over-fitting. We randomly partitioned the data using simple random sampling into subsets to be used for training (40%), validation (30%), and testing (30%). The training dataset was used to construct tree models that identified subsets of records that responded at lower rates than the overall sample. This model was then applied to the validation and the test datasets, and the average squared error was compared across results from all three datasets; this procedure helps prevent generating a model that would not fit other data or that would be unreliable (i.e., overfitted).

The classification tree nonresponse propensities were calculated using a single tree and an ensemble of classification trees. A classification tree model is constructed by segmenting the data using the application of a series of simple rules (SAS, 2009). Each rule assigns an observation to a subgroup, or "segment," based on the value of one predictor variable. The rules are applied sequentially, resulting in a hierarchy of segments within segments. The rules are chosen to subdivide cases into segments that have the largest difference with respect to the target variable, in this case, nonresponse rates. Thus, the rule selects both the variable and the best breakpoint to separate the resulting subgroups maximally. Variables can be used more than once to further segment groups, and thus may appear multiple times throughout a tree.

The hierarchy of segments is called a tree, and each segment in it is called a node. The original segment contains the entire set of cases and is called the root node of the tree. A node with all its successors is termed a branch of the node that created it. The final nodes are called leaves. In our analysis, we are ultimately interested in the leaves that contain a higher proportion of nonrespondent cases.

The optimal splits of cases are found using significance testing or reduction in variance criteria. Significance tests (based on F or chi square tests) use the p value as the stopping rule. Interval variables were assessed using an F test criteria, and nominal level variables

were assessed using a chi square test, where the best split is the one with the smallest *p* value (Enterprise Miner, 2009). Bonferroni adjustments were applied to the *p* value before the split was selected to "…mitigate the bias towards inputs with many values." (Neville, 1999, p. 18) Ordinal variables were assessed using entropy. Splitting rules were selected by measuring the reduction in entropy, after adjusting for ordinal differences.

$$entropy = -\sum_{all\ branches} P_b \log_2 P_b$$

where,

P_b = proportion of observations in the node assigned to branch b

Like other data mining techniques, classification trees describe subsets of data and are constructed without any theoretical guidance. Variables are chosen to create maximally different segments, so if variables are correlated, only one or a few of these (which individually might be related to the target) may appear in the tree.

There are multiple stopping criteria that can be used to decide how large to grow a classification tree. After the initial split, the resulting nodes are considered for splitting using a recursive process that ends when no nodes can be split further (SAS, 2009). A node can no longer be split when there are too few observations, the maximum depth (hierarchy of the tree) has been reached, or no significant split can be identified. For purposes of our research, the minimum number of observations for a node was set to five, the maximum depth was set to six, and the significance level was set to 0.20.

The characteristics associated with nonrespondents were first identified using the training data set (with n = 72,954 records). This model was then validated using 30 percent of the data (n = 54,446), and finally tested using final 30 percent of the data (n = 54,447). We compared the average squared error to determine that the model performed nearly as well in all three data sets.

A decision tree split can be selected automatically to maximize the dichotomy or it can be forced. When variables are automatically selected to maximize the dichotomy of the outcome, the selection is done looking at a single level of the tree, automatic selections do not consider the effect of subsequent splits. Due to the sheer number of nodes involved in the single tree (116), it would be impractical to try to display them. This approach was used to create the single-tree propensity scores. As a result, while a variable may initially provide the most optimal split for maximizing the dichotomy of the target, it may not ultimately result in the best model after subsequent splits are applied. For example, a model using the worst initial split for maximizing the dichotomy may in fact identify more of the target with less misclassification error than the model using the best initial split due to the effect of subsequent splits. In our case we had 69 COA variables we were using to model characteristics of nonrespondents, and thus we created 69 separate trees, with each tree using one of the 69 classification tree predictor variables for the initial split. This allowed us to identify more nonrespondents by forcing the trees to consider all 69 predictor variables in relation to nonresponse. After the initial split was forced, the following splits were selected using the automated methods described above. Separate propensity scores were created for each tree. Within each tree, propensity

scores were calculated first using the training data and then using the validation data. The propensity score for the entire tree was calculated by taking the mean of the training and validation propensity scores. Propensity scores are not calculated for the test data set to avoid over-fitting the models. The overall nonresponse propensity score for the ensemble of decision trees was calculated by averaging all 69 tree propensity scores. The single-tree propensity scores were derived from the first tree, which started with the most optimal split.

The ensemble tree nonresponse propensities were calculated as

$$\hat{p}_i^{et} = \sum_{t=1}^{69} \frac{\hat{p}_i^{tr(t)} + \hat{p}_i^{val(t)}}{2T}.$$

This leads to the following nonresponse weights for the ensemble classification tree model. They were calculated using:

$$w_i^{et} = w_i^0 \frac{1}{1 - \hat{p}_i^{et}}.$$

Here, *t* denotes the tree, $\hat{p}_i^{tr(t)}$ the tree-*t* nonresponse propensity score of farm *i* using the training data, and $\hat{p}_i^{val(t)}$ is the tree-*t* nonresponse propensity score of farm *i* using the validation data.

Calibration weights, w_i^{cal}, were created by taking the base weights for the subset of farms responding to the ARMS III and calibrating them using a truncated linear routine so that no final weight ever fell below one and the calibration equation held. This means that final weighted totals from the 2008 respondents equaled the weighted total computed from the entire 2008 matched sample for the following list of calibration variables: acreages for corn, soybean, wheat, cotton, hay, rice, peanuts, sugar , tobacco, fruits, and vegetables; production of egg and milk; inventories of cattle, hogs, broilers, and turkeys; indicators for nursery and floriculture; number of farms by economic classes; number of farms by non-estimated states; and total number of farms. Each of these target variables were used operationally to calibrate the ARMS III data. Targets initially selected for calibration, but not used operationally were excluded from the list.

Mathematically, in linear calibration:

$$w_i^{cal} = w_i^0(1 + g_1 z_{i1} + g_2 x z_{i2} + \cdots + g_P z_{iP}),$$

(4)

where the z_{ip} ($p = 1, \ldots, P$) are the calibration variables, and the g_p are chosen so that the calibration equation holds. In truncated linear calibration, when the right-hand side of equation (4) is outside the permissible range (e.g., below 1) , w_i^{cal} is set at the boundary of the range (e.g., 1). The g_p are recalculated (if possible) so that even with some w_i^{cal} set at boundary values, the calibration equation holds.

Calibration weighting is unbiased in some sense for the estimated mean of a survey variable if it behaves like a linear function of the calibration variables whether or not the

farm responds (Kott 2006). Linear calibration will also return large-sample, unbiased estimators in some sense if a farm's probability of response, which is implicitly estimated by w_i^0/w_i^{cal}, is the inverse of a linear function of the calibration varibles (*Fuller et al.* 1994). Truncated linear calibration will share this property when the only weight restriction is that no w_i^{cal} be lower than one.

As in the operational program, the ARMS III respondent subset was calibrated independently in 20 regions. These included the 15 leading cash receipts states (Arkansas, California, Florida, Georgia, Illinois, Indiana, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Carolina, Texas, Washington, and Wisconsin). The remaining 35 states (Alaska and Hawaii are not sampled for ARMS) were grouped using the five production regions: 1) Atlantic, 2) South, 3) Midwest, 4) Plains, and 5) West.

Using the base weights in combination with either logistic regression weights, the classification tree weights, or the calibration weights, we calculated the nonresponse bias of 30 estimates collected on both the ARMS III and the COA. We then compared the amount of the remaining nonresponse bias under each of the weighting schemes. We compared the weighted estimates for the ARMS III responding operations with the weighted estimates for the entire ARMS III sample; we used COA data, which were available for both the ARMS III respondents and the ARMS III nonrespondents. The inverse of the fitted propensities from the logistic regression model and from the classification tree model were multiplied by the base weights to give the final weight for these estimates. We computed estimates of the relative nonresponse bias for means under each weighting scheme. Because it treats upward and downward biases symmetrically, we used the log scale to compare average nonresponse biases across production and demographic type items (i.e., $\log(a/A) \times 100\%$ was our measure for the relative bias of a as an estimate of A).

3. Results

Nonresponse bias was assessed under the base weights and under a variety of adjustment schemes. These included the methods described earlier in the text: using fitted logistic probabilities of response directly, creating five weight-adjustment classes based on the sorted logistic fit, 10 reweighted-adjustment classes based on the sorted logistic fit, using a single classification tree, using an ensemble tree, and (truncated linear) calibration weighting. We also created and employed alternative single and ensemble tree probabilities based on sorting the initial tree probabilities into five and 10 classes as we did the fitted logistic probabilities (see equations 2 and 3).

Before we could compare the weighting methods against each other, we had to assess which adjustment method involving logistic regression (and the base weights) worked best on our data. We also had to compare the single tree weighting methods and the three ensemble tree weighting methods. Through these analyses we found that the logistic regression weighting using 10 classes appeared to result in the least amount of bias across the production type items while all weights do well when adjusting for bias of demographic type items with the exception of race. Moving forward, we will only examine the logistic regression weighting method using 10 classes, since that method results in the least amount of bias for both production and demographic items. The single tree method using a direct approach appears to result in the least amount of bias across the production type items while all weights do well when adjusting for bias of

demographic type items with the exception of race. Again, moving forward, we will only examine the single tree direct method because that method results in the least amount of bias for both production and demographic items. Finally, examining the ensemble tree method, we found the same results as for the single tree weight method. Therefore, we will only examine the ensemble tree direct method since that method results in the least amount of bias for both production and demographic items. For more in-depth information, including figures that display these findings, contact the first author.

After we identified the best overall weighting method using the logistic regression, single tree, and ensemble tree models, we then compared the best weighting methods from each model with calibration weighting. According to Figure 1, it appears that the best weighting method varies across the production items. Both of the best tree methods and calibration outperform the best logistic regression weighting method for all of the production variables. According to Figure 2, the weights perform relatively the same for the demographic variables, except for race. According to Figure 3, the single tree resulted in the least amount of bias on average for the production type items, and calibration resulted in the least amount of bias on average for the demographic type items. Overall, according to Figure 4, when we looked across all of the items including both production and demographic, calibration resulted in the least amount of bias on average.

In conclusion, we found that overall logistic regression weights perform better using classes, tree weights perform better when used directly, and that tree weights performed better than logistic regression weights. When we looked specifically at production versus demographic type variables, we found that while the 10 class approach was the best method regardless of variable type when using a logistic regression model; the best method varied depending on the type of item for both the tree type models. With both of the tree models, it appeared that the direct method worked best for the production type items, and the five class approach worked best for the demographic type items.

When we looked across all 30 variables, without distinguishing between production versus demographic type items, we found that the direct method resulted in the least amount of bias on average for both of the tree methods. Single tree weights created directly performed slightly better in terms of average bias than the ensemble tree weights; however, ensemble tree weights provided a better estimate for more variables than the single tree method. If classes are used, ensemble tree weights provide better estimates than single tree or logistic regression models. All of the weighting methods performed essentially the same in terms of demographic type items. Calibration did the best job of adjusting for nonresponse bias overall, but trees performed slightly better for production type items. Weights created using tree models could provide a good alternative to calibration when external benchmarks are not available, but rich auxiliary data are.

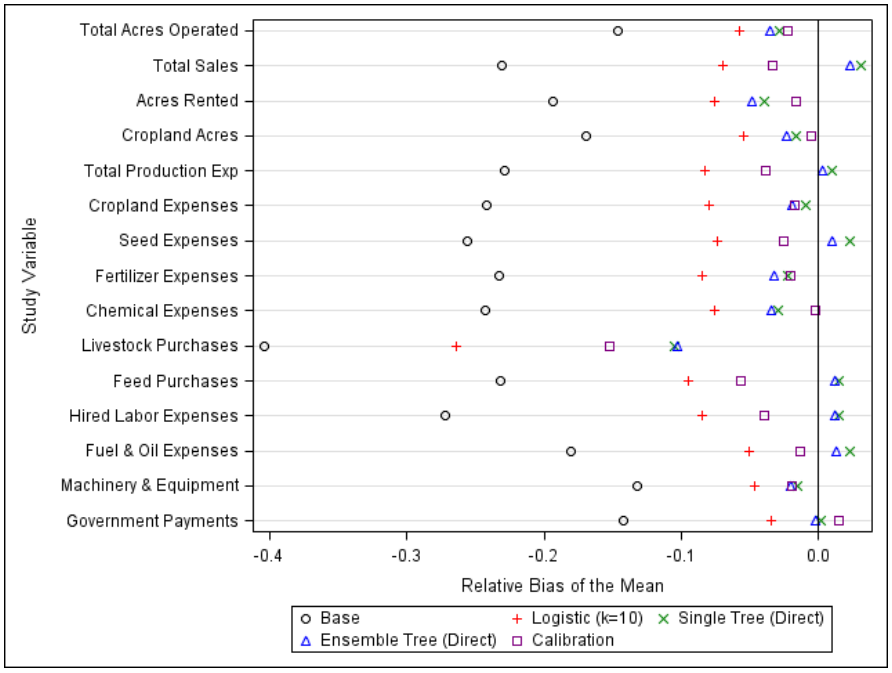

Figure 1: Comparison of Best Weighting Methods Using Production Items

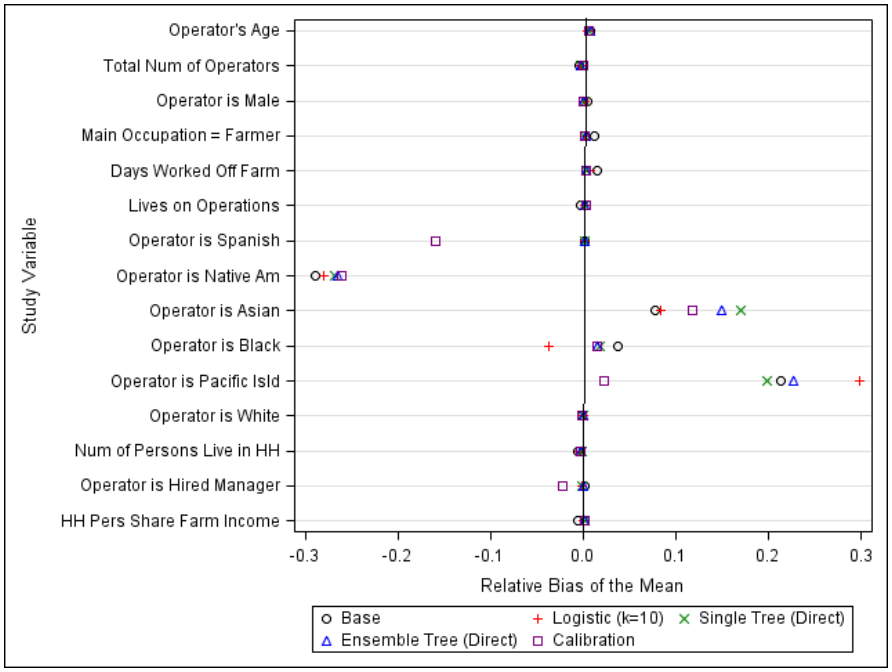

Figure 2: Comparison of Best Weighting Methods Using Demographic Items

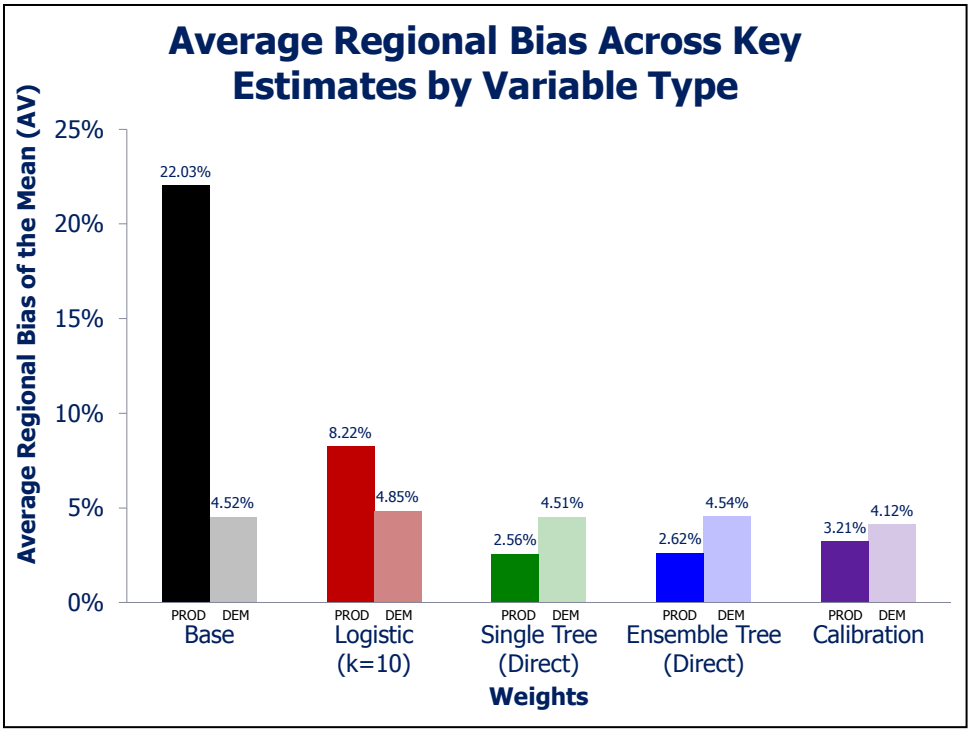

Figure 3: Comparison of Average Nonresponse Bias (Absolute Values) Across Production and Demographic Type Items Using Best Weighting Methods

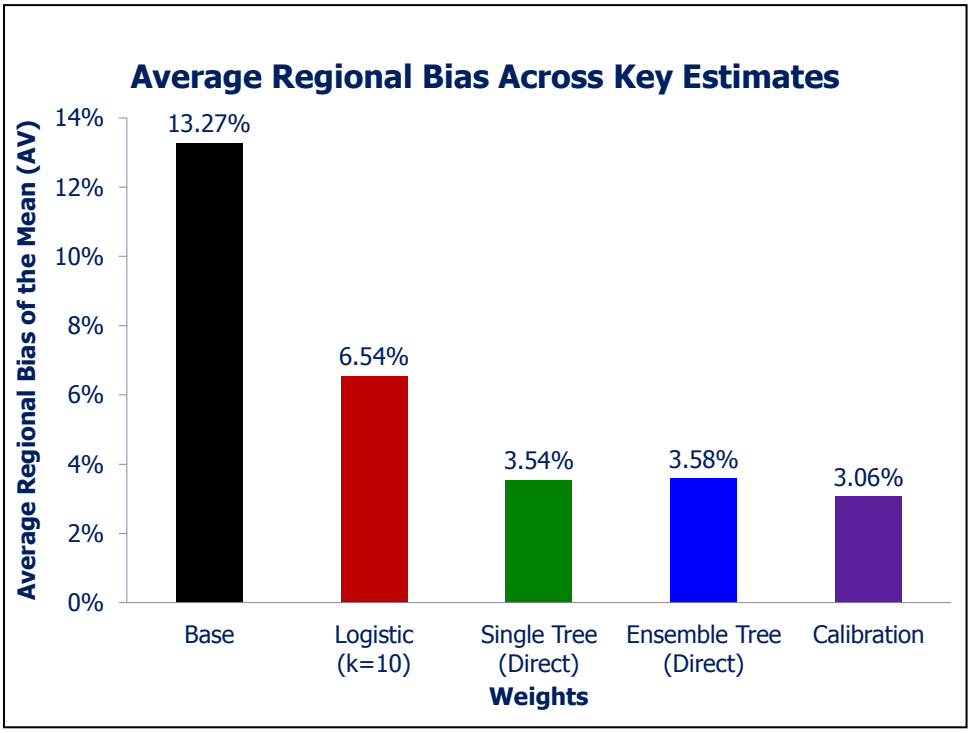

Figure 4: Comparison of Average Nonresponse Bias (Absolute Values) Across All Items Using Best Weighting Methods

All of the models had difficulty adjusting for race of the operator. All of the weighting methods underestimated the number of Native American operators and over estimated the number of Asian and Pacific Islander operators. The logistic regression weights underestimated the number of Black operators, where the other weighting methods overestimated the number of Black operators. Calibration underestimated the number of Spanish operators, where the other weighting methods resulted in relatively zero bias. Calibration did a considerably better job than the other weighting methods at adjusting for the number of operators that are Pacific Islander.

Variance estimation was outside the scope of this analysis. The ARMS III presently uses a replication approach to variance estimation, which seems well suited for handling nonresponse adjustment for tree methods.

The results of our study were limited in that our logistic regression model only looked at main effects, where the trees looked at interaction effects. The logistic regression model would likely perform better if: 1) indicators of item missingness were developed for all 69 variables; and 2) all six way interaction effects were explored across the 69 variables and 69 indicators of item missingness; however, the logistic regression model would still not be capable of identifying optimal breakpoints to distinguish between nonrespondents and respondents. The results would only show that more or less of something was indicative of nonresponse, not specifically how much or how little. Another limitation is that for purposes of this study, we had rich auxiliary data; therefore, we are unsure whether trees would perform as well as calibration using only limited frame data. On the other hand, when we do have rich auxiliary data, a tree is capable of including a number of variables in the model; however, the same is not true for calibration. The more variables we include in the calibration process, the more difficult it can become to meet all of the specified targets and thus converge at a solution. The fact that the trees were able to account for so many other characteristics that the calibration weights did not, may in part explain why they performed slightly better when adjusting for the production type items.

4. Discussion

In the case of the ARMS survey, NASS has good external estimates to use as calibration targets. This analysis shows that this weighting scheme considerably reduces the bias that would be introduced into the selected survey estimates using only the survey's base weights. Indeed, the objective of calibration is not just to meet the calibration benchmarks, but to improve all of the statistics produced by the survey. The correlation between the calibration variables and survey estimates of economic activity is likely high. For example, an operation's "corn acres" (the calibration benchmark) is likely correlated with its "cropland acres" and "seed expenses" (the survey variables of interest discussed in this analysis). Although, for other variables of interest, the correlation is likely lower (for example with variables such as "acres rented" or "operator's age"). The analysis also shows that a single-tree or ensemble tree weighting scheme is more effective at reducing nonresponse bias of the mean than calibration for selected production items, but not for demographic type items.

While these results are limited to the 30 variables assessed in relation to nonresponse for the ARMS III 2008 sample, this research suggests that trees work better than logistic regression and are comparable to calibration, which is not always an option. The results also indicate that while on average using a single-tree approach results in less bias across variables, an ensemble-tree approach provides a better estimate for more variables than a single-tree approach. If a survey administrator is more concerned about the average bias across estimates, then a single tree appears to work best; however, if they are more concerned with how frequently they produce the best estimate, they may want to consider creating an ensemble of classification trees. While calibration works well for ARMS III, calibration is not a viable option in surveys when good external benchmarks for calibration are not available. Our analysis provides evidence that tree methods may provide a comparable alternative to calibration when rich auxiliary data is available.

5. References

Abraham, K.G., Mailand, A., and Bianchi, S.M. (2006). Nonresponse in the American Time Use Survey. Who is Missing from the Data and How Much Does it Matter? *Public Opinion Quarterly, 70* (5), 676-703.

Bauer, E. a. (1999). An Emperical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning, 36* 105-132.

Brieman, L. (1998). Arcing Classifiers (with discussion). *Annals of Statistics 26*(3), 801-849.

deVille, B. (2006). *Decision Tress for Business Intelligence and Data Mining using SAS Enterprise Miner*. Carey, NC: SAS Institute, Inc.

Earp, M., McCarthy, J., Porter, E., and Kott, P. (2010). *Assessing the Effect of Calibration on Nonresponse Bias in the 2008 ARMS Phase III Sample Using Census 2007 Data.* In JSM Proceedings, Government Statistics Section. Vacouver, CA: American Statistical.

Earp, M., Mitchell, M., and McCarthy, J. (2011). *Who is Responsible for the Bias? Using Proxy Data and Tree Modeling to Identify Influential Nonrespondents & Reduce Bias*. Proceedings of the Fourth International Conference of Establishment Surveys, June 11-14, 2012, Montréal, Canada [CD-ROM]: American Statistical Association.

Groves, R. M. (2006). Nonresponse Rates and the Nonresponse Bias in Household Surveys. *Public Opinion Quarterly, 70* (5), 646-675.

Grobes, R. M. and Couper, M. (1998). *Nonresponse in Household Interview Surveys.* New York: John Wiley.

Johansson, F. and Klevmarken, A. (2008). Explaining the Size and Nature of Response in a Survey on Health Status and Economic Standard. *Journal of Official Statistics, 24* (3), 431-449.

Johnson, T.P., Cho, I.K., Campbell, R.T., and Holbrook, A.L (2006). Using Community-Level Correlates to Evaluate Nonresponse Effects in a Telephone Survey. *Public Opinion Quarterly, 70* (5), 704-719.

Kott, P. And Chang, T. (2010). Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse. *Journal of the American Statistical Association, 4*, 501-514.

Lepkowsi, J.M. and Couper, M.P. (2002). Nonresponse in the Second Wave of Longitudinal Household Surveys. In R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (Eds.), *Survey Nonresponse*. New York: Wiley and Sons.

Little, R. (1986). Survey Nonresponse Adjustments for Estimates of Means. *Journal of the American Statistical Association, 77*, 237-250.

Little, R. And Varivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology, 31*, 161-168.

Neville, P. (1999). Decision Trees for Predictive Modeling. Carey, NC: SAS Institute, Inc.

Nicoletti, C. and Peracchi, F. (2005). Survey Response and Survey Characteristics: Microlevel Evidence from the European Community Household Panel. *Journal of the Royal Statistical Society, A, 168* (4), 763-781.

Phipps, P. and Toth, D. (2012). Analyzing Establishment Nonresponse Using and Interpretable Regression Tree Model with Linked Administrative Data. *Annals of Applied Statistics, 6* (2), page numbers are forthcoming.

Potts, W. J. (2006). *Decision Tree Modeling Course Notes.* Cary: SAS Institude Inc.
SAS. (2009). *SAS Enterprise Miner 6.1.* Cary, NC: SAS Institude Inc.

Rosenbaum, P. and Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70*, 41-55.

SAS Institute Inc., Enterprise Miner 6.2 Help and Documentation, Cary, NC: SAS Institute Inc., 2009

Seastrom, M. K., Kaufman, S., Lee, R. (2002). *2002 National Center for Education Statistics Statistical Standards, Appendix A.* Washington, DC: National Center for Education Statistics: http://nces.ed.gov/statprog/2002/appendixa.asp

Singer, E. (2006). Nonresponse Bias in Household Surveys. *Public Opinion Quarterly, 70* (5), 637-645.

United States. Department of Agriculture. (2007). *2007 Census of Agriculture.* Washington, DC: U.S. Department of Agriculture: http://www.agcensus.usda.gov/Publications/2007/Full_Report/

United States. Executive Office of the President. (2006). *Office of Management and Budget Standards and Guidelines for Statistical Surveys.* Washington, DC: U.S. Executive Office of the President.