DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 701–719, http://dx.doi.org/10.2478/JOS-2014-0044

Modeling Nonresponse in Establishment Surveys: Using an Ensemble Tree Model to Create Nonresponse Propensity Scores and Detect Potential Bias in an Agricultural Survey

Morgan Earp[1], *Melissa Mitchell*[2], *Jaki McCarthy*[3], *and Frauke Kreuter*[4]

Increasing nonresponse rates in federal surveys and potentially biased survey estimates are a growing concern, especially with regard to establishment surveys. Unlike household surveys, not all establishments contribute equally to survey estimates. With regard to agricultural surveys, if an extremely large farm fails to complete a survey, the United States Department of Agriculture (USDA) could potentially underestimate average acres operated among other things. In order to identify likely nonrespondents prior to data collection, the USDA's National Agricultural Statistics Service (NASS) began modeling nonresponse using Census of Agriculture data and prior Agricultural Resource Management Survey (ARMS) response history. Using an ensemble of classification trees, NASS has estimated nonresponse propensities for ARMS that can be used to predict nonresponse and are correlated with key ARMS estimates.

Key words: Nonresponse bias; propensity scores; classification trees; ensemble trees.

1. Introduction

In many ongoing surveys, response rates are declining or now require more resources to maintain (Curtin et al. 2005; Groves and Couper 1998; Stussman et al. 2005; Brick and Williams 2009). Reduced response rates can lead to nonresponse bias when response propensities are correlated with characteristics of interest (i.e., something we are trying to measure) or vary by subdomain (Wagner 2012). This can be a particularly serious problem for establishment surveys, because unlike household surveys, many establishment population distributions are highly skewed (Petroni et al. 2004). Thus severe nonresponse bias can occur if sample units that contribute to the estimates more than others are less likely to respond (Groves et al. 2002; Phipps and Toth 2012; Powers et al. 2006; Thompson 2009). For example, according to the 2007 Census of Agriculture, only 0.3 percent of farms had total sales of five million dollars or more, but they accounted for

[1] Bureau of Labor Statistics – Office of Survey Methods Research, PSB Suite 1950, 2 Massachusetts Avenue, NE Washington District of Columbia 20212, U.S.A. Email: earp.morgan@bls.gov
[2] USDA – National Agricultural Statistics Service, Fairfax, Virginia, U.S.A. Email: Melissa.Mitchell@nass.usda.gov
[3] USDA – National Agricultural Statistics Service, Fairfax, Virginia, U.S.A. Email: Jaki.McCarthy@nass.usda.gov
[4] University of Maryland – JPSM, 1218 Lefrak Hall, College Park, MD 20742, Maryland 20742, U.S.A. Email: fkreuter@survey.umd.edu
Acknowledgments: We would like to thank the Guest Editor and Associate Editor for the Special Issue of papers from ICES-IV, and the anonymous reviewers for their comments.

27.9 percent of total sales in the U.S. (U.S. Department of Agriculture 2007). If these operations failed to respond, it would greatly impact the estimates of total sales (and items strongly related to total sales).

Traditionally, survey methodologists use two approaches for dealing with nonresponse error. One focuses on increasing participation, for example through incentives, notification letters, or personal enumeration (Dillman 1978; Groves and Couper 1998), whereas the other tries to address potential nonresponse bias through weighting adjustment (Kalton and Flores-Cervantes 2003) or imputation (Little and Rubin 2002). However, both approaches have drawbacks. Extra efforts are costly and increased response rates could mean that only more of the same types of establishments are brought into the respondent pool, leaving nonresponse bias unchanged, or worse, increased (Groves 2006). Weighting adjustments or calibration to known population totals can be effective at reducing the nonresponse bias of the variables used in the calibration models (or for variables that are highly correlated with these covariates), but may fail to address potential nonresponse bias in other key estimates in large multipurpose surveys (Earp et al. 2010). Likewise, business programs that use imputation instead of weighting to account for unit nonresponse, for example through the use of administrative data, can induce additional bias if the imputation models are poor (Luzi et al. 2007) or if a high proportion of units in a subdomain of the imputation base have missing data (Thompson and Washington 2013).

In response to these drawbacks, survey methodologists recently started employing responsive design strategies (Groves and Heeringa 2006) that tailor fieldwork efforts to respondents with different response propensities. In order to do this, it is particularly important to identify and target the low propensity groups whose nonresponse is most likely to induce nonresponse bias and to increase their response rates. Successful examples of such approaches are the National Survey of Family Growth (Axinn et al. 2011), and several CATI surveys done by Statistics Canada (Mohl and Laflamme 2007; Laflamme and Karaganis 2010).

In this article, we present a study that assesses a method for identifying such low response propensity groups in a large-scale farm survey. Specifically, we present an application that uses an ensemble of classification trees to model establishment survey nonresponse on the Agricultural Resource Management Survey (ARMS) in relation to multiple farming characteristics collected by the 2002 Census of Agriculture. Both the Census of Agriculture and ARMS are conducted by the National Agricultural Statistics Service (NASS), making it easy to link the data and use Census data as a proxy both in modeling characteristics of nonresponse and assessing the relationship between modeled nonresponse propensity scores and nonresponse bias. To evaluate our proposed methods, we linked units from a later ARMS sample (containing missing values) to their 2007 Census of Agriculture data on numerous common agricultural characteristics using the Census data as a proxy for ARMS respondent and nonrespondent characteristics. Consequently, we can compare the relative difference of the mean between respondents and nonrespondents on several key estimates across varying response propensity groups.

In Section 2, we provide background information on the ARMS data used in the case study. In Section 3, we introduce the classification tree methodology and describe its application to the ARMS data. Section 4 presents our results. We conclude in Section 5 with a brief discussion and ideas for future research.

2. Background on the ARMS

The ARMS collects calendar year economic data from agricultural producers nationwide that describe the financial performance and operational characteristics of farm households. These data are used to inform the U.S. Farm Bill and are used extensively for analysis by the United States Department of Agriculture's Economic Research Service (ERS) to understand the financial performance and household characteristics of farms. The ARMS is conducted in three phases. Phase I screens for potential samples for Phases II and III using a mail questionnaire. Phase II collects data on cropping practices and agricultural chemical usage. Phase III (also referred to as ARMS III) collects detailed economic information about the agricultural operation as well as the operator's household. ARMS III data (referred to as ARMS from here on) are primarily collected through personal interviews and mail questionnaires. There are multiple versions of the ARMS questionnaire. Some versions are administered by mail and personal interview and some by personal interview only. In addition, there are several commodity-specific versions of the questionnaire that vary by year: Examples include organic agriculture, apples, and poultry.

The ARMS is a probability sample, drawn from both a list and an area frame. The list frame is stratified by farm total sales, farm type, and farm region; the area frame is stratified by land use (U.S. Department of Agriculture 2012). Units are selected from the stratified frame using a Sequential Interval Poisson (SIP) design. By utilizing SIP, NASS is able to decrease the probability of sample selection for operations previously sampled for ARMS and other NASS surveys and thus reduce respondent burden (Miller et al. 2010). Note that sample design weights were not used in the creation of the tree models discussed in Subsection 3.2.1, since the purpose was to model the expected response rates specifically for ARMS using the ARMS sample design, and not to model the expected response rates for the entire originating population of farms (Phipps and Toth 2012).

ARMS response rates (see Table 1) have been fairly stable over the years but consistently fall below the target of 80 percent; studies below 80 percent are required to complete a nonresponse bias analysis according to the standards issued by the U.S. Office of Management and Budget in 2006 (United States Executive Office of the President 2006).

Table 1. ARMS response rates 2000–2008

Year	Sample size	Response rate (%)
2000	17,903	63
2001	13,313	64
2002	18,219	74
2003	33,861	63
2004	33,908	68
2005	34,937	71
2006	34,203	68
2007	31,924	70
2008	36,388	66

3. Methodology

3.1. Classification and Ensemble Trees

Often, logistic regression models are used to relate covariates to nonresponse and to compare response rates across subgroups (Axinn et al. 2011; Johansson and Klevmarken 2008; Johnson et al. 2006; Abraham et al. 2006; Little and Vartivarian 2005; Nicoletti and Peracchi 2005; Lepkowsi and Couper 2002; Little 1986; Rosenbaum and Rubin 1983). In many applications, however, classification trees are considered easier to specify and interpret, specifically with regard to interaction effects (Phipps and Toth 2012; Schouten 2007; Schouten and de Nooij 2005). Moreover, classification tree models also have the added benefits of

1. automatically detecting significant relationships and interaction effects without prespecification, thus reducing the risk of selecting the wrong variables or other model specification errors;
2. identifying the variables that are correlated with the target variable, along with the optimal breakpoints within these variables for maximizing their correlation;
3. identifying hierarchical interaction effects across numerous variables and summarizing them using a series of simple rules;
4. incorporating missing data into the model and assessing whether missingness on a given variable is related to the target; and
5. creating a series of simple rules that are easy to interpret and use for identifying and describing subgroups with higher propensities.

While using classification trees provides some advantages over logistic regression, the results from a single tree are also considered to be potentially unstable. Therefore, it is recommended that trees be modeled and validated using independent data. As is typical in classification tree modeling, the dataset used in the creation of our trees was randomly split into three independent sets. An initial training subset of the data is used to grow the tree, and an independent subset is subsequently used to validate the results of the initial model. Finally, a third subset of the data can be used to further test the reliability of the model. This guards against overfitting the model.

A classification tree considers all input variables (independent variables) and grows branches using input variables that demonstrate significant relationships with the target (dependent variable), while also considering interaction effects among the various inputs. Classification tree models work by segmenting the data using a series of simple rules. 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 (cf., interaction effects in a logistic regression model). The rules are chosen to subdivide cases into segments that have the largest difference with respect to the target variable, which in this case is nonresponse. Thus the rule selects both the variable and the best breakpoint to separate the resulting subgroups maximally. The breakpoints also take into consideration whether data are missing for an item and either uses a surrogate item (something closely related) or classifies missing data into whichever group is most similar in terms of the target. If the observations that have missing data are distinctly different from those not missing data, then the tree will break the item on the

missing classification. These rules make it easier to describe the likely nonrespondents specifically and to identify what characteristics contribute to nonresponse. By itself, a propensity score helps predict the likely nonrespondents and identifies which inputs in the model are positively or negatively related to nonresponse. However, the propensity scores do not provide a specific description of who the nonrespondents are, whereas this is explicit in the classification tree model (Phipps and Toth 2012).

The break points of variables 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. In the application described in Subsection 3.2.1, interval variables were assessed using 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 (SAS 2009). Bonferroni adjustments are applied to the *p*-value before selecting the split to ". . .mitigate the bias towards inputs with many values" (Neville 1999, 18). Ordinal variables were assessed using entropy, which measures the reduction in variance. The same variables may appear multiple times throughout a tree to introduce further segmentation.

After the initial split, the resulting leaves are considered for splitting using a recursive process that ends when no leaves can be split further (SAS 2009). A leaf 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.

Using a single classification tree approach, the best initial splitting variable is chosen and significant subsequent splits are selected based on the initial split. However, if the initial splitting variable is chosen based on the significance level using only the training data, it may not actually be the ideal initial splitting variable given all the data; furthermore, it is important to recognize that the effect of subsequent splits is not considered when choosing the optimal initial split. The initial split selected directly affects the optimality of variables considered for subsequent splits. Although one split may be optimal for maximizing the dichotomy at a given level of the tree, there is no guarantee that given subsequent splits, a tree selecting the optimal initial split will correctly identify the greatest number of observations with the target.

To mitigate this, ensemble tree models are used instead. Ensemble trees grow multiple trees each with varying initial splits. With varying initial splits, each of the trees within the ensemble is capable of identifying different (but possibly overlapping) subgroups with high occurrences of the target. Using the ensemble of classification trees results in a more accurate, stable, powerful, and generalizable model than using a single classification tree (Breiman 1998; Dietterch 2000; Matignon 2008). An ensemble tree can either use voting or the average of the propensity scores across all the trees to identify those likely to exhibit the target (SAS 2009). We utilized the average propensity score across all of the trees, since we were interested in the overall propensity to respond as opposed to nonresponse classification.

3.2. ARMS Application

3.2.1. Building the Initial Model

To evaluate the classification tree procedure described in Subsection 3.1, 2002 Census of Agriculture data were matched to all available ARMS 2000–2008 sample units. These data were then used to construct classification trees predicting ARMS non-

respondents and to estimate their nonresponse propensities. 78 percent ($n = 185{,}767$) of all ARMS cases sampled for data collection between 2000 and 2008 had 2002 Census of Agriculture data available.

The dependent (target) variable for our model was ARMS nonresponse. Operations responding to the ARMS were coded "0" and those not responding were coded "1". The classification trees described in this study explored the relationship between key agriculture characteristics collected on the 2002 Census for the ARMS 2000–2008 samples and nonresponse.

All of the classification trees were created using a randomly selected subset of the data (66,876 of 167,190 farms), which is referred to as the training data. The same training data were used for all trees in the ensemble tree. The results were evaluated and tested using the remainder of the data (93,314 farms). The average squared error of the model applied to the training, validation, and test data was equivalent (average squared error = .34), indicating that the model performed equally well on the training dataset used to create the model and on the two independent validation and test datasets.

Using an ensemble tree approach, we grew multiple trees, forcing each one to initially split on one of the 70 of 71 variables significantly related to nonresponse ($p < .20$); Table A1 in the Appendix provides the list of studied variables. We set the minimum number of observations for a leaf to five, the maximum depth of the tree to six, and the significance level to 0.20. A liberal criterion is used to assess the significance of main effects, since classification trees are primarily interested in the subsequent interaction effects and use independent sources of data to evaluate the results. According to Uther and Veloso (1998, 4), "In the tree based learning literature, it is well known that stopping criteria often have to be weak to find good splits hidden low in the tree."

A popular form of an ensemble tree model called random foresting randomly selects subsets of variables to split on, since in most cases it is not possible to grow all possible trees (Banfield et al. 2007). In our case, we did not grow all possible trees, but we explored all initial splitting variables. We forced each of the 70 variables to be used as an initial splitting variable, so that we could ensure that each of these variables was considered at least once in the overall model. This was important for us in being able to assure operational and field staff that each of the variables in Table A1 were tested in relation to nonresponse. While some of these variables may not be as strongly related to nonresponse as total sales or total acres operated, they are still important to NASS in terms of nonresponse bias. For example, by forcing Tree 66 to split on acres of certified organic farming, we were able to model the relationship between number of certified organic acres and nonresponse. Only significant initial and subsequent splits were retained in our model. After the initial split, all significant subsequent splits were detected automatically using the splitting algorithm described above. Out of 71 initial forced splits shown in Table A1, only one was considered nonsignificant – whether the farm operator was Native Hawaiian or Pacific Islander.

Each tree identified unique subsets of likely nonrespondents based on varying initial splits, and therefore provided unique indicators and thus probabilities of nonresponse. This resulted in a richer and more inclusive model that included not only the characteristics we knew were related to nonresponse, such as total sales and total acres operated, but also the gender of the operator and the number of female operators, which we previously did not know were related to nonresponse. The overall ensemble tree propensity score for each

sample unit was estimated by taking the average of all 70 individual tree nonresponse propensities for that unit. The average propensity score from multiple trees with varying significant splits is considered to be more accurate and generalizable than those taken from an individual tree (Bauer and Kohavi 1999; Breiman 1998). The segmentation rules for all 70 trees were saved into a score code that could be used to estimate nonresponse propensities of future ARMS samples using their 2002 census data.

3.2.2. Evaluating the Model for Nonresponse Predictive Power

Once the ensemble model was created, we evaluated the model using the 2009 ARMS sample, a completely independent ARMS sample which had not been used in creating any of the trees. By pulling the 2002 Census data for the 2009 ARMS sample, we were able to apply the model specification rules to the 2009 sample and evaluate the predictive power of the ensemble tree nonresponse propensity scores using a logistic regression model. The logistic regression model specified the ARMS 2009 nonresponse as the dependent variable and the ARMS ensemble tree nonresponse propensity score as the independent variable, controlling for Census 2007 total sales and total acres operated. By controlling for total sales and total acres operated, we could determine whether the ensemble tree propensity scores were significantly correlated with future ARMS nonresponse beyond just farm size and total sales. The logistic regression analysis was run using the 21,969 of the 34,429 operations for which we had both 2002 and 2007 Census data; 2002 data were needed to generate the nonresponse propensities and 2007 data were necessary as the proxy data for the 2009 sample. Census 2002 data were available for 24,264 (70%) of the ARMS 2009 sampled operations, and Census 2007 data were available for 27,830 (81%) of the ARMS 2009 sampled operations; both Census 2002 and 2007 data were available for 64 percent of all operations sampled for the 2009 ARMS.

3.2.3. Evaluating the Model for Nonresponse Bias Predictive Power

If the nonresponse propensities are correlated with 2009 ARMS nonresponse beyond just measures of farm size, they can be used to classify the 2009 sample into nonresponse subgroups with similar response propensities. According to Eltinge and Yansaneh (1997), nonresponse propensity score deciles are considered to be more stable than the individual propensity scores, and therefore can be used to distinguish less likely respondents from more likely respondents. Using deciles, we classified the ARMS 2009 sample into ten groups using their nonresponse propensity scores. We then compared the ten nonresponse propensity groups on key estimates (by using their Census 2007 data as a proxy of key ARMS estimates for this sample) (see Table 3). Finally, we plotted the relative difference of the mean (and median) as shown in Equation 1 for all ten deciles in order to determine if the groups least likely to respond might contribute substantively more to nonresponse bias on the studied characteristics than those more likely to respond.

$$\text{Relative Difference of the Mean} = \frac{\bar{y}_c - \bar{y}_o}{\bar{y}_o} \tag{1}$$

where,

$\bar{y}_c =$ Class Mean
$\bar{y}_o =$ Overall Mean of Full Sample Results.

4. Results

Figure 1 demonstrates a weak positive relationship between the ensemble tree nonresponse propensities and probability of ARMS 2009 nonresponse given an operation's modeled nonresponse propensity score, value of Census 2007 total sales, and Census 2007 total acres operated.

Table 2 shows that even though the relationship between the ensemble tree nonresponse propensity score and 2009 ARMS nonresponse appeared weak, it was still a significant predictor of 2009 ARMS nonresponse, even after controlling for the operation's 2007 total sales and total acres operated; indicating that ARMS nonresponse is not completely explained by farm value and size.

Having evaluated our classification tree nonresponse propensities on an "independent" dataset, we then grouped the nonresponse propensity scores for the ARMS 2009 sample into deciles so that we could distinguish between operations expected to be more likely versus less likely to respond. Lower classes were expected to have lower rates of nonresponse and higher classes were expected to have higher rates of nonresponse. Figure 2 shows that the percent of nonrespondents within each class generally increases from Class 1 (C1) (the group most likely to respond) through Class 10 (C10) (the group least likely to respond), although counter to expectation the group with the highest predicted nonresponse propensities did not have the highest nonresponse rate.

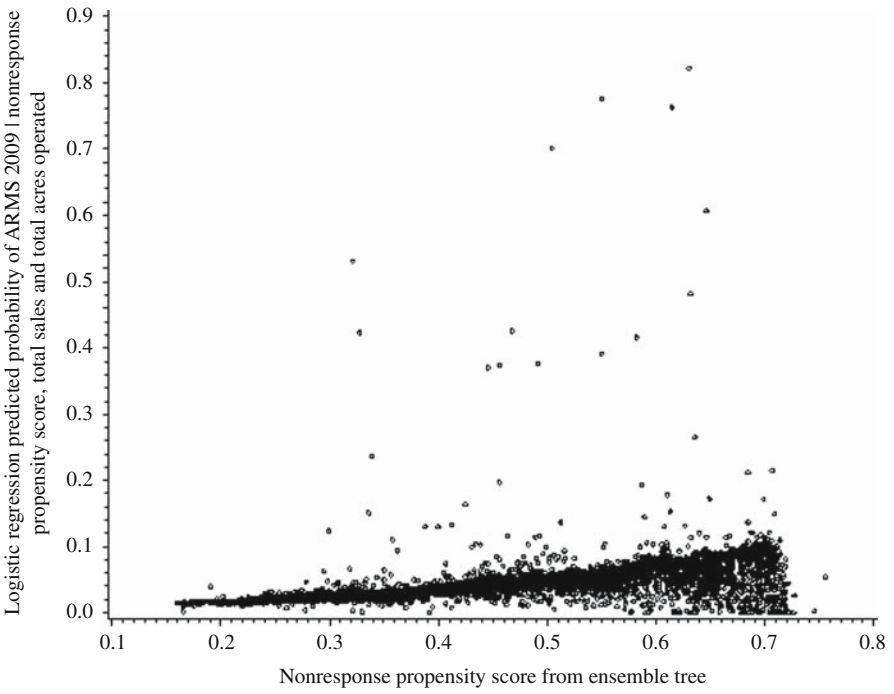


Fig. 1. Plot of the logistic regression predicted probability of 2009 ARMS nonresponse given the ensemble tree nonresponse propensity score, 2007 total sales, and 2007 total acres operated, by the ensemble tree nonresponse propensity score

Table 2. Logistic regression model fit statistics

Analysis of maximum likelihood estimates					
Predictor	β	SE β	Wald's χ^2 ($df = 1$)	p	e^{β} Odds Ratio
Constant	−4.77	0.14	1191.55	<.0001	
Propensity score	3.76	.34	118.99	<.0001	42.93
Total sales	−9.02−08	2.11E-08	18.35	<.0001	1.00
Total acres operated	2.0E-05	3.19E-06	40.67	<.0001	1.00

Finally, we compared the 14 key agricultural estimates (again, using their 2007 Census data as a proxy) across the ten nonresponse propensity classes to see whether these estimates varied by class. Table 3 provides the mean value of the 14 key estimates by the ten nonresponse propensity classes, Class 11 (C11) identifies the ARMS 2009 sampled operations that were missing Census 2002 data and therefore have missing nonresponse propensity estimates, but were not missing Census 2007 data. This allowed us to assess how operations missing nonresponse propensity scores compared to those not missing nonresponse propensity scores. Using the overall mean and the class means shown in Table 3, we calculated the relative difference of the mean for each class shown in Figure 3 (Särndal 2011).

Given the significant correlations between the modeled nonresponse propensity scores, ARMS 2009 nonresponse, and the key estimates shown in Table A2, we expected to see a relationship between nonresponse propensity classes and relative difference of the mean.

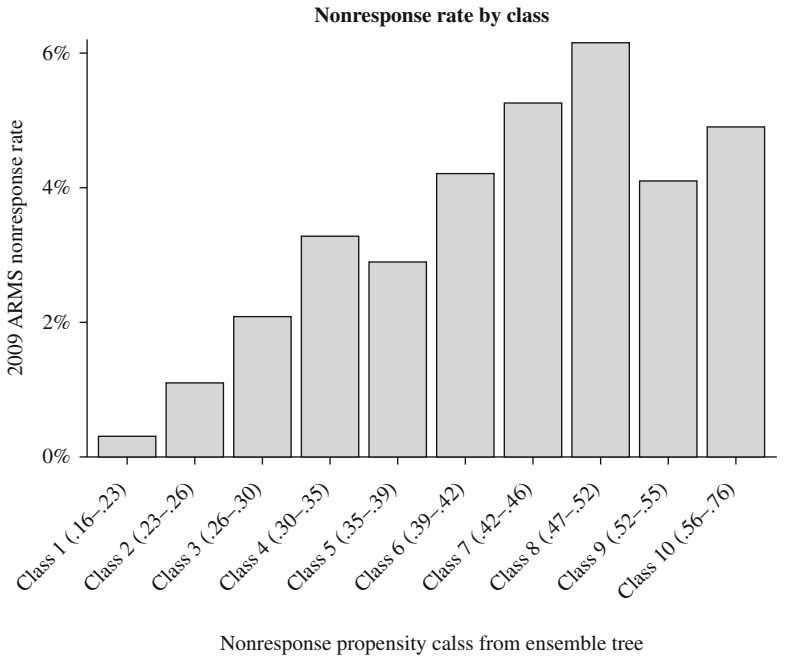
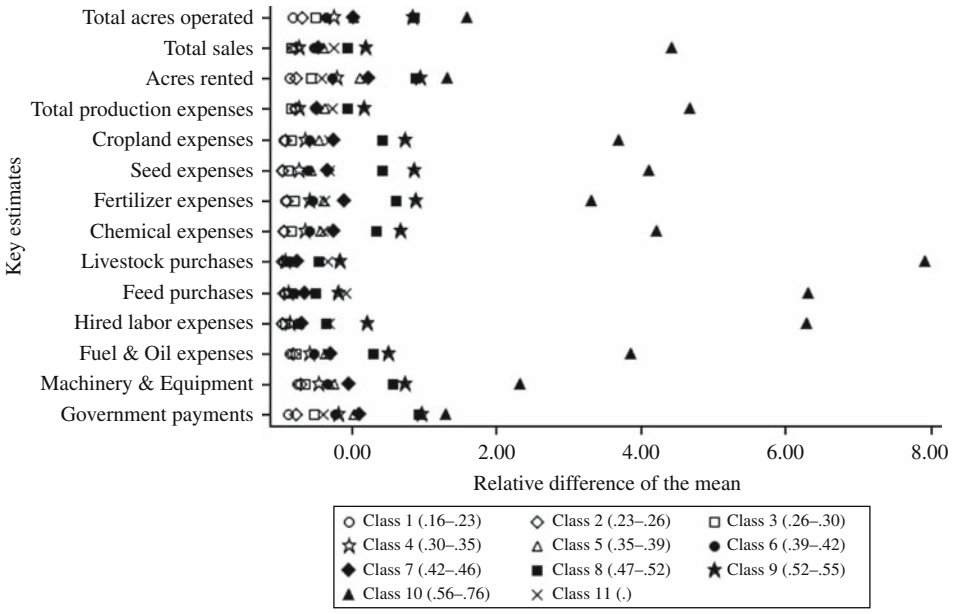


Fig. 2. ARMS 2009 nonresponse rate by ensemble tree nonresponse propensity class

Table 3. Key estimate means by nonresponse propensity class. (Note that the deciles were created for all ARMS 2009 sampled operations with 2002 Census data (n = 24,264) and since not all sampled operations also had 2007 Census data, the ns for the deciles are not all equal, but are very close.)

		Nonresponse propensity classes										
Key estimates	Overall	C1 (.16–.23)	C2 (.23–.26)	C3 (.26–.30)	C4 (.30–.35)	C5 (.35–.39)	C6 (.39–.42)	C7 (.42–.46)	C8 (.47–.52)	C9 (.52–.55)	C10 (.56–.76)	C11 (.)
Total acres operated	1,536	311	489	764	1,166	1,586	998	1,565	2,855	2,844	3,992	952
Acres of land rented	772	116	179	341	618	854	564	950	1,458	1,502	1,791	465
Seed expenses	37,359	1,730	1,999	5,111	10,519	16,873	14,962	25,366	53,272	69,670	190,396	25,701
Fertilizer expenses	50,597	4,765	4,929	10,562	21,291	31,046	23,763	45,683	81,720	95,061	217,912	32,519
Chemical expenses	32,877	2,214	2,128	5,498	12,162	18,333	14,059	24,546	44,349	54,783	171,250	20,491
Feed expenses	110,157	8,830	6,961	12,887	12,681	22,160	23,647	39,969	56,465	88,685	804,299	101,551
Hired labor expenses	96,288	5,131	5,021	7,517	14,474	26,752	24,781	31,127	63,732	116,288	701,271	67,082
Fuel & oil expenses	32,942	5,015	6,656	7,568	14,048	20,874	15,982	23,386	42,801	49,942	159,622	22,544
Machinery & equipment value	319,088	79,207	92,313	119,093	172,158	239,160	215,385	307,331	501,652	554,382	1,066,464	229,141
Total government payments	15,577	2,137	3,568	7,663	12,806	16,064	11,991	17,452	29,977	30,646	35,830	9,690
Crop expenses	127,985	9,477	9,939	22,394	45,701	69,254	54,059	97,000	181,969	222,969	598,429	88,920
Livestock expenses	92,153	5,977	3,921	9,395	7,976	15,734	12,964	23,101	49,761	77,755	822,388	61,539
Total sales	1,049,540	209,400	255,876	183,701	282,365	637,756	512,065	566,066	996,348	1,254,680	5,702,015	788,303
Total production expenses	797,476	165,106	188,632	140,530	213,950	508,571	400,236	422,024	754,359	933,910	4,520,589	586,415
n	27,830	2,449	2,445	2,438	2,432	2,426	2,419	2,411	2,414	2,410	2,419	3,566



Relative difference of the mean = [(class mean – overall mean)/overall mean]

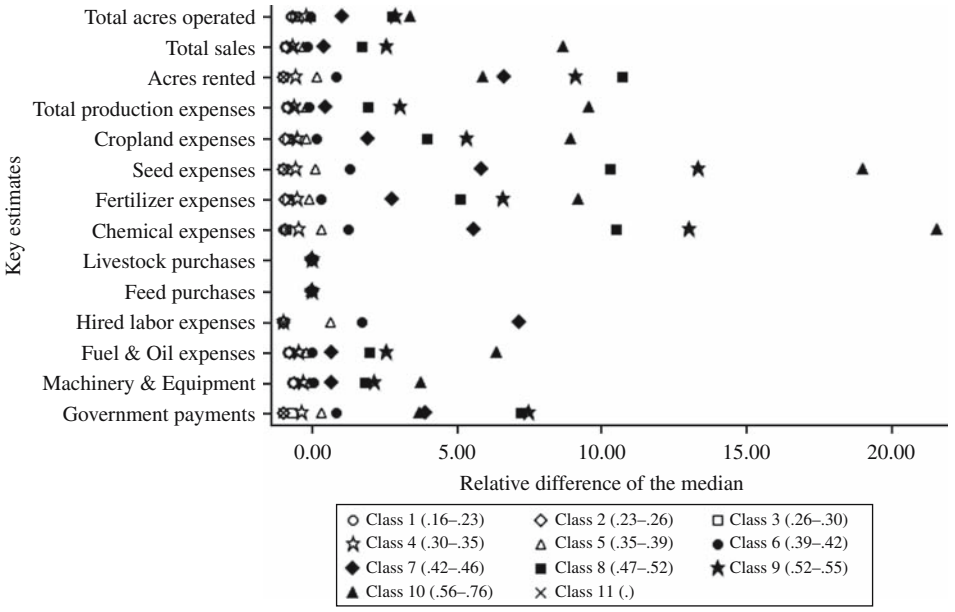
Fig. 3. Relative difference of the mean for key estimates by nonresponse propensity class

The relative difference of the mean as plotted in Figure 3 indicates that the group least likely to respond in terms of their propensity score (Class 10) also poses the greatest threat in terms of relative difference of the mean. Without response from this group, all 14 key estimates would be underestimated. In order to mitigate the potential impact of a few extreme values in any class, we also show the same comparison for the medians by class (see Figure 4). This may be a problem particularly for highly skewed establishment populations. These results are comparable to the estimate means.

Note that the relative difference of the class median for hired labor expenses was extremely high for Classes 8 ($62.86), 9 ($112.68) and 10 ($476.47) and has been omitted from the chart for clarity. Furthermore, the overall median and almost all of the class medians were zero for both livestock and feed expenses, indicating zero relative difference; however, in the few instances where the class median was not zero (livestock: Class 10 ($3,125); feed: Class 1 ($1,000), Class 2 ($89), and Class 3 ($23)) we were unable to estimate the relative difference of the median since the overall median was zero and would have resulted in dividing the class medians by zero.

While Class 11 operations were missing Census 2002 data and thus propensity scores, the relative difference for Class 11 did not stand out in comparison to the other classes shown in Figure 3 or 4.

These results demonstrate that by using an ensemble of classification trees with Census of Agriculture data, we created nonresponse propensity scores that were significantly correlated with future ARMS nonresponse and with all 14 key agricultural estimates from the ARMS (see Table A2). The farms classified into the lowest expected response propensity had the greatest relative difference of the mean and therefore posed the greatest potential threat in terms of both nonresponse and nonresponse bias.



Relative difference of the median = [(class median – overall median)/overall median]

Fig. 4. Relative difference of the median for key estimates by nonresponse propensity class

5. Discussion

This article presents a procedure that uses an ensemble of classification trees to produce robust nonresponse propensity estimates. By examining the individual trees used to create the average nonresponse propensity, we can easily identify the characteristics of various types of nonrespondents. These models not only considered the most obvious and significant predictors of nonresponse in the studied program, but they also identified the rare and yet important variables that are also related to nonresponse. The resulting average nonresponse propensity scores from all the trees may not be greatly influenced by these less predictive or important variables, but they are at least considered given the forced initial-split method we used, which is important to operational and field office staff.

While the logistic regression model's pseudo R^2 (McFadden 1974) was low ($= 0.03$), this may be partly due to the fact that the nonresponse rate was much lower for those operations that had both Census 2002 data and Census 2007 data than for the overall ARMS 2009 sample. Operations with both 2002 and 2007 Census data had a nonresponse rate of 0.03 ($n = 21{,}969$) compared to 0.32 for the overall ARMS 2009 sample ($n = 34{,}429$). Had we been able to estimate propensity scores and had proxy data for the entire 2009 ARMS sample, the propensity scores might have been more strongly related to future ARMS nonresponse.

The results of the study might have differed had we included sample design weights. Using sample design weights and or calibration weights in the models would allow development of prediction models that identify nonrespondent characteristics and estimate nonresponse propensities for the entire population, instead of being restricted to the

selected sample (Phipps and Toth 2012). We may consider including sample or calibration weights in a future model to gain a more general understanding. Given that Class 10 had the greatest relative difference of the mean for all 14 key estimates, NASS may consider using adaptive design efforts to increase the likelihood of response for operations that fall into this class, and potentially Classes 9 and 8 as well depending on funding.

While we did not evaluate whether using varying forced initial splits works as well as random forests, this method did provide us with a level of control that made our methods and results easy to explain to operational and field staff. This was important given that this was only the third operational use of classification trees at NASS, and the first in relation to survey nonresponse. In a future article, we would like to explore the performance of this method in comparison to random foresting. We would be specifically interested in assessing the relative difference of specialty crops and rare operator characteristics such as being female, since we believe this may be a potential strength of using initial forced splits.

The ensemble tree method of modeling survey nonresponse introduced in this article can be helpful in identifying and describing characteristics of influential nonrespondents in other surveys. It provides a tool that allows the researcher to assess the impact of multiple establishment characteristics and interaction effects on nonresponse. Classification trees provide a series of simple rules that can be used to describe specific characteristics of likely nonrespondent subgroups to operational and field staff. The modeled nonresponse propensities can then be used to create nonresponse subgroups. These subgroups can then be used to evaluate the potential impact on survey estimates, or as inputs to adaptive design strategies targeting different data collection strategies to different subgroups of a sample.

Appendix

Table A1. Census of agriculture operational characteristic variables in ranking order of initial split significance

Rank	Variable name
1	Total sales not under production contract (NUPC)
2	Total value of products sold + government payments
3	Total production expenses
4	The number of hired workers employed more than 150 days
5	Machinery and equipment value in Dollars
6	Acres of cropland harvested
7	Cropland acres
8	Total reported acres of crops harvested
9	Acres of land owned
10	State
11	Total acres operated
12	The number of hired workers employed less than 150 days
13	Any migrant workers Y/N
14	Total cattle and calf inventory
15	Total expenditures
16	Farm type code
17	Type of organization
18	Percent of principle operator's income from the farm operation
19	Computer used for the farm business Y/N

Table A1. Continued

Rank	Variable name
20	Acres of all other land
21	Principal occupation of principle operator is farming Y/N
22	Total government payments
23	ARMS III production region (Atlantic, South, Midwest, Plains, or West)
24	Acres of land rented from others
25	Any hired manager Y/N
26	Operation had internet access Y/N
27	Number of households sharing in net farm income
28	Acres of all irrigated hay and forage harvested
29	Number of days principle operator worked off farm
30	Total fruit acres
31	Total acres of vegetables
32	Acres of woodland pasture
33	Principal operator's age
34	Acres of woodland not in pasture
35	Number of operators
36	Acres on which manure was applied
37	Acres of permanent pasture & rangeland
38	Acres of all hay and forage harvested
39	Total poultry inventory
40	Partnership registered under state law Y/N
41	Acres of cropland used for pasture
42	Total hog and pig inventory
43	Principal operator lives on operation Y/N
44	Percent of operators that are women
45	Acres of cropland for which all crops failed
46	Acres of cropland in summer fallow
47	ARMS III questionnaire version
48	Total sales under production contract (UPC)
49	Total citrus acres
50	Nursery indicator Y/N
51	Principal operator's sex
52	Principal operator – race, black
53	Acres of land rented to others
54	Operation farm tenure (1 = full owner, 2 = part owner, or 3 = tenant)
55	Number of persons living in principle operator's household
56	Acres of cropland idle or used for cover crops
57	Have other farm Y/N
58	Principal operator – race, white
59	Sheep and lamb indicator Y/N
60	Year principal operator began this operation
61	Number of women operators
62	Other livestock animals
63	Agriculture on indian reservations Y/N
64	Principal operator – race, american indian
65	Acres of Christmas trees and Short rotation woody crops
66	Acres of certified organic farming
67	Possible duplicate Y/N

Table A1. Continued

Rank	Variable name
68	Principal operator is of Spanish origin Y/N
69	Principal operator – race, Asian
70	Aquaculture indicator Y/N
71	Principal operator – race, native Hawaiian, or Pacific Islander[5]

[5] Not significant at the 0.20 level.

Table A2. Pearson & Point biserial correlation matrix of nonresponse propensity score, indicator of 2009 ARMS response, and key estimates

		Nonresponse propensity score	2009 ARMS nonrespondent
Nonresponse propensity score		1.00	0.08
	p		<.0001
	n	24,264	24,264
2009 respondent		0.08	1.00
	p	<.0001	
	n	24,264	34,429
Total acres operated		0.20	−0.03
	p	<.0001	<.0001
	n	21,969	27,830
Acres of land rented from others		0.15	−0.02
	p	<.0001	0.00
	n	21,969	27,830
Seed expenses		0.23	−0.02
	p	<.0001	0.00
	n	21,969	27,830
Fertilizer expenses		0.35	−0.03
	p	<.0001	<.0001
	n	21,969	27,830
Chemical expenses		0.30	−0.03
	p	<.0001	<.0001
	n	21,969	27,830
Feed expenses		0.20	−0.01
	p	<.0001	0.18
	n	21,969	27,830
Labor expenses		0.29	−0.03
	p	<.0001	<.0001
	n	21,969	27,830
Fuel & oil expenses		0.37	−0.04
	p	<.0001	<.0001
	n	21,969	27,830
Machinery & equipment value		0.44	−0.04
	p	<.0001	<.0001
	n	21,969	27,830
Total government payments		0.26	−0.04
	p	<.0001	<.0001
	n	21,969	27,830

Table A2. Continued

		Nonresponse propensity score	2009 ARMS nonrespondent
Total sales		0.34	−0.03
	p	<.0001	0.00
	n	21,969	27,830
Livestock		0.12	−0.01
	p	<.0001	0.16
	n	21,969	27,830
Crop expenses		0.35	−0.02
	p	<.0001	<.0001
	n	21,969	27,830
Total Production expenses		0.30	−0.03
	p	<.0001	<.0001
	n	21,969	27,830

6. References

Abraham, K.G., A. Mailand, and S.M. Bianchi. 2006. "Nonresponse in the American Time Use Survey. Who is Missing from the Data and How Much Does it Matter?" *Public Opinion Quarterly* 70: 676–703. DOI: http://dx.doi.org/10.1093/poq/nfl037.

Axinn, W., C. Link, and R. Groves. 2011. "Responsive Survey Design, Demographic Data Collection, and Models of Demographic Behavior." *Demography* 48: 1127–1149. DOI: http://dx.doi.org/10.1007/s13524-011-0044-1.

Banfield, E., L.O. Hall, K.W. Bowyer, and W.P. Kegelmeyer. 2007. "A Comparison of Decision Tree Ensemble Creation Techniques." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29: 173–180. DOI: http://dx.doi.org/10.1109/TPAMI.2007.250609.

Bauer, E. and R. Kohavi. 1999. "An Emperical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants." *Machine Learning* 36: 105–132. DOI: http://dx.doi.org/10.1023/A:1007515423169.

Breiman, L. 1998. "Arcing Classifiers (with discussion)." *Annals of Statistics* 26: 801–849.

Brick, J.M. and D. Williams. 2009. "Reasons for Increasing Nonresponse in U.S. Household Surveys." Paper presented at the Workshop of the Committee on National Statistics, Washington, DC, December 14.

Curtin, R., S. Presser, and E. Singer. 2005. "Changes in Telephone Survey Nonresponse over the Last Quarter Century." *Public Opinion Quarterly* 69: 87–98. DOI: http://dx.doi.org/10.1093/poq/nfi002.

Dietterich, T.G. 2000. "Ensemble Methods in Machine Learning." In Proceedings of the Multiple Classifier Systems: First International Workshop, MCS 2000, June 21–23, Cagliari, Italy. Available at: http://www.eecs.wsu.edu/~ holder/courses/CptS570/fall07/papers/Dietterich00.pdf (accessed August 2014).

Dillman, D. 1978. *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley & Sons.

Earp, M., J. McCarthy, E. Porter, and P. Kott. 2010. "Assessing the Effect of Calibration on Nonresponse Bias in the 2008 ARMS Phase III Sample Using Census 2007 Data." In Proceedings of the Joint Statistical Meetings: American Statistical Association. Alexandria, VA: American Statistical Association. Available at: http://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/conferences/JSM-2010/earp-2010_jsm_paper_arms_calibration.pdf (accessed August 2014).

Eltinge, J.L. and I.S. Yansaneh. 1997. "Diagnostics for Formation of Nonresponse Adjustment Cells, with an Application to Income Nonresponse in the US Consumer Expenditure Survey." *Survey Methodology* 23: 33–40.

Groves, R. 2006. "Nonresponse Rates and the Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70: 646–675. DOI: http://dx.doi.org/10.1093/poq/nfl033.

Groves, R. and M. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R.M., D. Dillman, J.L. Eltinge, and R.J. Little. 2002. *Survey Nonresponse*. New York: Wiley.

Groves, R. and S. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society Series A: Statistics in Society* 169: 439–457. DOI: http://dx.doi.org/10.1111/j.1467-985X.2006.00423.x.

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

Johnson, T.P., I.K. Cho, R.T. Campbell, and A.L. Holbrook. 2006. "Using Community-Level Correlates to Evaluate Nonresponse Effects in a Telephone Survey." *Public Opinion Quarterly* 70: 704–719. DOI: http://dx.doi.org/10.1093/poq/nfl032.

Kalton, G. and I. Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19: 81–97.

Laflamme, F. and M. Karaganis. 2010. "Development and Implementation of Responsive Design for CATI Surveys at Statistics Canada." In Proceedings of the European Quality Conference: Helsinki, Finland.

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

Little, J. and D. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *Journal of the American Statistical Association* 77: 237–250.

Little, R. and S. Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31: 161–168.

Luzi, O., T. De Waal, B. Hulliger, M. Di Zio, J. Pannekoek, D. Kilchmann, and C. Tempelman. 2007. *Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys*. Italian Statistical Institute ISTAT.

Matignon, R. 2008. *Data Mining Using SAS Enterprise Miner*. Cary, NC: SAS Institute Inc.

McFadden, D. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by P. Zarembka. New York: Academic Press.

Miller, D., M. Robbins, and J. Habiger. 2010. "Examining the Challenges of Missing Data Analysis in Phase Three of the Agricultural Resource Management Survey." In Proceedings of the Joint Statistical Meetings: American Statistical Association. Alexandria, VA: American Statistical Association. Available at: https://www.amstat.org/sections/srms/proceedings/y2010/Files/306438_56491.pdf (accessed August 2014).

Mohl, C. and F. Laflamme. 2007. "Research and Responsive Design Options for Survey Data Collection at Statistics Canada." In Proceedings of the Joint Statistical Meetings: American Statistical Association. Alexandria, VA: American Statistical Association. Available at: https://www.amstat.org/sections/srms/proceedings/y2007/Files/JSM2007-000421.pdf (accessed August 2014).

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

Nicoletti, C. and F. Peracchi. 2005. "Survey Response and Survey Characteristics: Microlevel Evidence from the European Community Household Panel." *Journal of the Royal Statistical Society* Series A: 168: 763–781. DOI: http://dx.doi.org/10.1111/j.1467-985X.2005.00369.x.

Petroni, R., R. Sigman, D. Willimack, S. Cohen, and C. Tucker. 2004. "Response Rates and Nonresponse in Establishment Surveys – BLS and Census Bureau." *Federal Economic Statistics Advisory Committee*, 1–50.

Phipps, P. and D. Toth. 2012. "Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data." *Annals of Applied Statistics* 6: 772–794. DOI: http://dx.doi.org/10.1214/11-AOAS521.

Powers, R., J. Eltinge, and M. Cho. 2006. "Evaluation of the Detectability and Inferential Impact of Nonresponse Bias in Establishment Surveys." In Proceedings of the Joint Statistical Meetings: American Statistical Association. Alexandria, VA: American Statistical Association. Available at: http://www.bls.gov/ore/pdf/st060130.pdf (accessed August 2014).

Rosenbaum, P. and D. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55. DOI: http://dx.doi.org/10.1093/biomet/70.1.41.

Särndal, C.-E. 2011. "The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation." *Journal of Official Statistics* 27: 1–21.

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

Schouten, B. 2007. "A Selection Strategy for Weighting Variables Under a Not-Missing-at-Random Assumption." *Journal of Official Statistics* 23: 51–68.

Schouten, B. and G. de Nooij. 2005. *Nonresponse Adjustment Using Classification Trees*. CBS, Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/1245916E-80D5-40EB-B047-CC45E728B2A3/0/200501x10pub.pdf (accessed August 2014).

Stussman, B., J. Dahlhamer, and C. Simile. 2005. "The Effect of Interviewer Strategies on Contact and Cooperation Rates in the National Health Interview Survey." Federal Committee on Statistical Methodology, Washington, DC

Thompson, K.J. 2009. "Conducting Nonresponse Bias Analysis for Two Business Surveys at the US Census Bureau: Methods and (Some) Results." In Proceedings of the Section on Survey Research Methods: American Statistical Association Alexandria, VA: American Statistical Association. Available at: http://www.scs.gmu.edu/~wss/wss100922linebackpaper.pdf (accessed August 2014).

Thompson, K.J. and K.T. Washington. 2013. "Challenges in the Treatment of Unit Nonresponse for Selected Business Surveys: A Case Study." *Survey Methods: Insights from the Field*. Retrieved from http://surveyinsights.org/?p=2991.

United States Department of Agriculture. 2012. *2012 Agricultural Resource Management Survey – Phase III Cost and Returns Report Survey Administration Manual*. Washington, DC: US Department of Agriculture.

United States Department of Agriculture. 2007. *2007 Census of Agriculture*. Washington, DC: US Department of Agriculture. Available at: http://www.agCensus.usda.gov/Publications/2007/Full_Report/ (accessed August 2014).

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. Available at: http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf (accessed August 2014).

Uther, W.T.B. and M.M. Veloso. 1998. "Tree Based Discretization for Continuous State Space Reinforcement Learning." In Proceedings of AAAI-98, the Fifteenth National Conference on Artificial Intelligence: 769–774. Available at: http://www.cs.cmu.edu/~mmv/papers/will-aaai98.pdf (accessed August 2014).

Wagner, J. 2012. "A Comparison of Alternative Indicators for the Risk of Nonresponse Bias." *Public Opinion Quarterly* 76: 555–575. DOI: http://dx.doi.org/10.1093/poq/nfs032.

Received December 2012
Revised August 2014
Accepted September 2014