

Who is Responsible for the Bias? Using Classification Trees to Identify Subgroups of Likely Nonrespondents and Assessing their Relationship to Key Survey Estimates Using Structural Equation Modeling

Morgan Earp¹ & Jaki McCarthy²

¹National Agricultural Statistics Service (NASS), 3251 Old Lee Highway Room 305, Fairfax, VA 22030

²NASS, 3251 Old Lee Highway Room 305, Fairfax, VA 22030

Abstract

This paper discusses the use of classification trees, factor analysis, and structural equation modeling (SEM) to determine the relationship between response propensity and nonresponse bias. Unlike research done by other agencies using frame or paradata, NASS possesses matching Census of Agriculture data at the record level, rather than information at the zip code or county level. Therefore, NASS is capable of making proxy comparisons to determine which estimates are most heavily influenced by various subgroup response propensities. Classification trees were used to subset ARMS sample units into subgroups with like response propensity. SEM was used to assess the relationship between subgroup response propensity scores and key survey estimates.

This research will enable NASS to flag likely influential nonrespondents and target data collection efforts to minimize bias in survey estimates.

Key Words: Nonresponse; Bias; Characteristics; Classification Trees; Exploratory Factor Analysis;

1. Introduction

As is the case in many surveys in the Federal government and elsewhere, survey response rates have been declining or have required more resources to maintain. However, the level of nonresponse is not as important as the amount of bias that the nonresponse introduces into the survey estimate. According to OMB, "...the degree of nonresponse bias is a function of not only the response rates but also how much the respondents and nonrespondents differ on the survey variables of interest" (2006, p.16). As stated by Groves, nonresponse bias is not simply a result of nonresponse: "Recent empirical findings illustrate cases when the linkage between nonresponse rates and nonresponse bias is absent. Despite this, professional standards continue to urge high response rates. Statistical expressions of nonresponse bias can be translated into causal models to guide hypotheses about when nonresponse causes bias" (2006, p.645). Currently all sample units are treated as either nonrespondents or respondents, but if not all nonrespondents share the same traits, and all do not uniformly contribute to nonresponse bias, should we not distinguish between different types of nonrespondents? Since the type of bias introduced varies along with the type of nonrespondent, it is possible that we may overlook bias introduced by subsets of nonrespondents when we only assess the overall

effect of nonrespondents as whole. Additionally, different types of nonrespondents may contribute to the bias of different estimates. According to Groves,

There are a few attributes that appear to be predictive of response propensities in a wide variety of survey settings... In at least some surveys, these influences on survey participation are correlated with the variables of interest in the survey. The practitioner must decide whether this is likely to be the case and whether, therefore, differential effort should be assigned to the groups with low base propensities. To assign more effort to subgroups with low base propensities requires identifying them. (2006, p.664)

This paper focuses on identifying characteristics of subgroups with low base propensities, while also distinguishing among the different types. If different subgroups of nonresponse can be identified, they can then be used in causal models to test and compare their effect on key survey estimates. Thus, we can identify the subsets of nonrespondents introducing bias and focus efforts on reducing or compensating for just the influential nonrespondents.

This paper examines nonresponse in the Agricultural Resource Management Survey (ARMS) conducted by the USDA's National Agricultural Statistics Service. This annual survey is one of the most complex and detailed sample survey data collections conducted by NASS and collects calendar year economic data from agricultural producers nationwide. ARMS suffers from relatively low response rates for a federal survey, and consistently falls short of the threshold set forth in OMB Guideline 3.2.9, requiring that federal surveys have a response rate of 80 percent or higher. As a result, NASS has assessed the nonresponse bias of key ARMS estimates. While NASS uses calibration weights to reduce the bias in key estimates to insignificant levels (Earp, McCarthy, Schauer, & Kott, 2008; Earp *et al.*, 2009; Earp, McCarthy, Porter, & Kott, 2010), NASS has recently focused research towards trying to preemptively address nonresponse bias by determining what the characteristics of ARMS nonrespondents are and which specific subgroups of ARMS nonrespondents influence ARMS key estimates.

The approach described in this paper differs from similar research comparing respondents and nonrespondents based on auxiliary data in some important respects. First, the modeling approach we used could be classified as a data mining approach, rather than a traditional hypothesis testing approach. Rather than restrict the variables used in our models, the classification tree allows us to include all the available auxiliary variables. This allows identification of a restricted set of records of interest in our dataset. Other nonresponse models have been developed using auxiliary data, but most begin with hypotheses about a small set of relevant variables and have generated response propensity scores based on regression or similar models (Johansson and Klevmarcken, 2008, Johnson, Cho, Campbell, and Holbrook, 2006, Abraham, Maitland, and Bianchi, 2006, Nicoletti and Peracchi, 2005, Lepkowski and Couper, 2002). These types of models may increase accurate prediction of nonrespondents, but they do not typically include large sets of auxiliary variables. We included 70 variables (many of them correlated) as well as all possible interaction effects across those variables, something impossible using logistic regression. In addition, with a large dataset such as ours, many relationships may be statistically significant, but not practically useful. Our models identify some of these relationships, but the objective is only to identify subsets of records of interest. The use of the classification tree also allows us to include auxiliary variables with missing data as

possible “characteristics” of a sample unit, something potentially important in analysis of nonresponse. Finally, while our models can be used to generate propensity scores, more importantly, they provide an exact description of the characteristics of each identified nonrespondent group. These groups were numerous, so we followed the development of the classification trees with exploratory factor analysis (EFA) to combine the groups into nonresponse “factors.”

Since the survey we examined collects economic data from agricultural operations, the specific variables we examined are not likely to be important indicators of nonresponse for surveys of other populations. However, the modeling approach we used is unique and could be applied in any survey where auxiliary matching data is available for sample units.

This analysis does not speak directly to nonresponse bias, but identifies likely nonrespondents. Future work will ultimately assess the effect of these various types of nonrespondents on the bias of key survey estimates.

2. Method

In order to identify characteristics of nonrespondents and compare their effects on ARMS survey estimates, we identified multiple subsets of the sample with high nonresponse propensities, based on auxiliary data describing the sample units from another source (the Census of Agriculture). This was done using classification trees. These subsets are neither mutually exclusive nor independent, so factor analysis was used to combine them into factors comprising similar groups. Finally, the relationship between the factors and the key survey estimates were evaluated using structural equation modeling. This paper describes the first two parts of this work -- the identification of sample units with high nonresponse propensities and the construction of nonresponse factors from these subgroups. This approach to characterizing survey nonresponse is unique and well suited to the situation for our survey, where rich auxiliary data are available and our analyses involve large data sets.

Using classification trees, we assessed the relationship between 71 variables (many of them correlated) and ARMS nonresponse. Seventy of the 71 candidate variables from the 2002 Census of Agriculture were significantly related to the target variable of survey nonresponse ($p < .20$). These 70 variables were selected and used to explore respondent characteristics in the ARMS. The variables included descriptive information about the operation such as its size, acreage, the type of commodities produced, expenses, its location, etc. as well as information about the principal operator, such as the operator's race, gender, number of days worked off the farm, etc. The full list of variables used is shown in Table 1.

Data from the 71 census of agriculture variables were matched to both respondents and nonrespondents in the ARMS III 2000-2008 samples. Matching census of agriculture data was available for 78 percent of the records (199,042/254,632). In order to ensure reliability of results, data were partitioned into three groups: training, validation, and test. Associated characteristics of nonrespondents were identified using 40 percent of the data ($n = 79,616$). We validated nonrespondent characteristics using 30 percent of the data ($n = 59,713$). We tested the reliability of the validated nonrespondent characteristics using 30 percent of the data ($n = 59,713$).

The matched variables from the census of agriculture were used to identify subsets of the ARMS 2000-2008 sample that exhibited nonresponse rates of 70 percent or greater. Classification trees model relationships with a categorical outcome (e.g., respondent or nonrespondent) using a tree-like structure.

Table 1: 2002 Census Operational Characteristic Variables (Ordered by Strength of Correlation to Nonresponse Propensity)

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
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
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 ($p > .20$)

1

For the purposes of this study, the target was ARMS III nonresponse. Operations responding to the ARMS III were marked with a "0" and those not responding with a "1" in a new survey nonresponse target variable. A classification tree considers all input variables (independent variables) and grows branches using input variables that demonstrate significant relationships with the target, while also considering interaction effects between the various inputs. The classification trees described in this study explored the relationship between operation characteristics and survey response.

In a typical classification tree approach, the best initial splitting variable would be chosen and a single model built; however, the initial splitting variable is chosen based on the significance level using only the training data, and therefore, may not actually be the ideal initial splitting variable given all the data. Furthermore, the effect of subsequent splits is not considered when choosing the initial split, but the initial split directly affects

¹ See the *PRISM II Code Book* (United States Department of Agriculture, 2008) for variable descriptions.

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 using the initial optimal split will correctly identify the greatest number of observations with the target. By varying the initial splitting variables, we can grow multiple trees using a single data set each of which are capable of identifying different (but possibly overlapping) subgroups with high occurrences of the target.

In this type of analysis, the full data were comprised of the 2002 Census of Agriculture data for the 2000-2008 ARMS III sample. A classification tree model is constructed by segmenting the data through the application of a series of simple rules. Each rule assigns an observation to a subsegment based on the value of one input variable. One rule is applied after another, resulting in a hierarchy of segments within segments. The rules are chosen to dichotomize maximally the subsegments with respect to the target variable, in this case, nonresponse. Thus, the rule selects both the variable and the best breakpoint to separate maximally the resulting subgroups. Variables may appear multiple times throughout the tree for further segmentation. The resulting hierarchy is called a tree, and each segment is called a node. The original segment contains the entire data set 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 records with the target (nonresponse).

The data were randomly broken into subsets to be used as the training, validation, and test sets, with 40%, 30%, and 30% in each, respectively. The training dataset was used to construct each initial tree model that identified subsets of records that responded at lower rates than the overall sample. This model was applied to the validation dataset in order to prevent generating a model for the training data that would not fit other data or that would be unreliable (i.e. overfitted). The validation data were used when pruning the initial tree to generate the final model. Finally, the test data were used to evaluate the model's performance on independent data not used in the creation of the model.

Like other data mining techniques, classification trees describe subsets of data and are constructed without any theoretical guidance. Variables are chosen to separate maximally the subsegments, 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 several alternative methods for constructing classification trees. For these models, trees were grown using the chi-square approach available in SAS Enterprise Miner, which is similar to the chi-square automatic interaction detection (CHAID) algorithm (deVille, 2006).

There are multiple stopping criteria used to decide how large to grow a decision tree. After the initial split, the resulting nodes are considered for splitting using a recursive process that ends when a node can no longer be split (SAS, 2009). A node can no longer be split when the number of specified observations is too low, the specified maximum depth (hierarchy of the tree) is too deep, or no significant split can be identified. For purposes of our research the minimum number of observations was set to five, the maximum depth was set to six, and the significance level was set to .20.

For this study, we explored the trees growing from all possible initial splits. Since our dataset contained 71 variables, there were 71 possible variables on which to conduct the initial split, 70 of which provided significant initial splits ($p < .20$). All seventy

significant initial splits were explored, which resulted in growing 70 different classification trees. Typically, a tree will be grown using the best split at each level, including the initial split. However, we can dictate which variable is used for the initial split. Each tree was forced to split initially on one of the 70 available variables. Forcing each of the 70 different variables to serve as the initial split ensured that each variable was considered when assessing the characteristics of nonrespondents. After the initial split, all variables were available for subsequent splits, which were determined automatically by the splitting algorithm.

All leaves with a 70 percent nonresponse rate or higher in both the training and validation data were selected from each tree. The logic rules leading to each of these nodes were used to create node membership indicators. Operations meeting the criteria for membership in a given node were coded “1,” and those not meeting the criteria were coded “0.” Each node membership indicator was considered a unique indicator of nonresponse. Records coded “1” by a given indicator were considered likely nonrespondents. Each indicator of node membership was named to indicate which tree and node it belonged to. For example, if an indicator was created from the 70th node in the 15th tree, it was named “tr15_070” and was coded “1” if the criteria for being a member of that node were met; furthermore, if the criteria for membership within that node were met, the operation would be considered likely to be a nonrespondent. These indicators of node membership are referred to as indicators of nonresponse. An example of the logic rules leading to a single leaf and thus a corresponding indicator of nonresponse using a classification tree is shown in Figure 1.

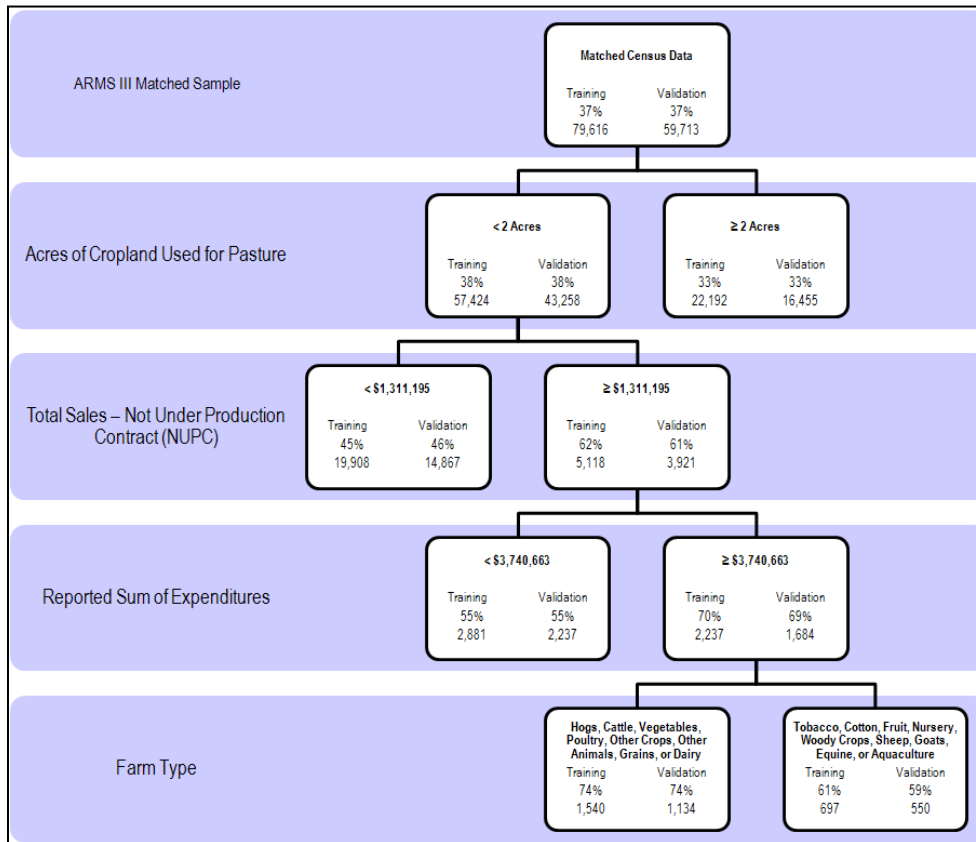


Figure 1. Example Tree - Acres of Cropland Use for Pasture

In interpreting the tree depicted in this figure, note that the logic rules used to specify the leaf at the bottom left are as follows:

- Acres of Cropland Used For Pasture < 2
- Total Sales – Not Under Production Contract (NUPC) $\geq \$1,311,195$
- Reported Sum of Expenditures $\geq \$3,740,663$
- Farm Type = Hogs, Cattle, Vegetables, Poultry, Other Crops, Other Animals, Grains, or Dairy

Both the number of operations meeting these criteria in the training and validation data (1,540 and 1,134) and the percentage of these who were nonrespondents (74%) are shown.

An indicator of node membership was created using the above criteria. Operations meeting all of the above criteria were coded “1,” indicating that they were members of this node, while operations not meeting all of the above criteria were coded “0,” indicating that they were not members of this node. Operations coded “1” are considered likely nonrespondents, and thus the indicator of node membership is in turn an indicator of nonresponse. Each node demonstrating greater than a 70 percent nonresponse rate in both the training and validation data was assigned a unique indicator of nonresponse.

Each tree identified unique subsets of nonrespondents based on varying initial splits, and therefore provided unique indicators of nonresponse. By creating several complementary trees, we created more indicators of nonresponse and thus identified more nonrespondents than we could have using a single tree.

The significance of potential splitting variables was assessed using the LogWorth statistic, which measures how well a given input variable measures the target using only the training data. All 70 classification trees were explored for two reasons:

- 1) The LogWorth of initial split variables were calculated using only the training data. Therefore, although it may be highly significant in the training phase, it may prove unreliable using the validation data, or the test data. Therefore, competing models may be just as likely to identify consistent nonrespondents. Although the 70 variables varied in significance, all 70 variables provided significant initial splits.
- 2) The characteristics identified in a given tree vary given the variable used in the initial split; therefore, each tree is capable of identifying unique subsets of respondents.

Due to the large number of indicators identified and the fact that the indicators of nonresponse were not mutually exclusive, we used exploratory factor analysis (EFA) to assess the communality across indicators and thus identify the main factors of nonresponse. EFA is used to explore the underlying structure of items by assessing their communality. In our case EFA was used to explore the underlying structure of the indicators of nonresponse by assessing their communality. EFA allowed us to identify and distinguish between different factors (types) of nonresponse.

3. Results

Seventy trees were grown by forcing an initial split using each of the 70 variables. This process ensured that all 70 variables were considered at least once when assessing characteristics of nonrespondents. Sixty-nine of the 70 trees grown identified and validated unique nonrespondent groups with nonresponse rates of 70 percent or greater. Interestingly enough, the only tree that did not identify and validate a single nonresponse group with a nonresponse rate of 70 percent or greater was the second tree grown, using the second most optimal initial split; thus supporting our rationale for exploring trees grown from splits beyond the optimal splits initially selected by LogWorth. The 69 trees identified and validated 226 nodes with nonresponse rates of 70 percent or greater, which resulted in the creation of 226 corresponding indicators of nonresponse. A test of the full data set demonstrated an average classification accuracy rate of 80.71 percent ($s = 7.76\%$, $k = 226$, $n = 199,042$). The number of likely nonrespondents classified by each indicator ranged from 7 to 5,989 with an average membership of 747 operations.

To gain a broader understanding of nonresponse, EFA was used to determine the main factors of nonresponse. EFA allowed us to identify the main factors of nonresponse as opposed to describing all 226 possible indicators of nonresponse in detail. EFA was initially run using all 226 groups; however, this approach resulted in model convergence issues, since the number of likely nonrespondents classified by some groups was so low. EFA was rerun including groups classifying 100 or more operations ($k = 122$), 500 or more operations ($k = 56$), and 1,000 or more operations ($k = 45$). Only the model using nonrespondent indicators classifying 1,000 or more operations converged, resulting in our including only 45 of the original 226 indicators of nonresponse in the EFA portion of this research; Therefore, we could not classify the remaining 181 indicators. While this may appear to limit the power of our study, it is important to point out that due to the overlap of nonrespondents identified by indicators, we were still able to accurately identify 83.56 percent of the nonrespondent identified by using all 226 indicators of nonresponse (9,828/11,762). Of the 17,355 operations predicted to be nonrespondents using all 226 indicators, 11,762 were actually nonrespondents, resulting in an overall classification accuracy rate of 67.77 percent. Using all 226 indicators, we correctly identified 16.08 percent (11,762/73,126) of the nonrespondents in the ARMS III 2000-2008 samples. Of the 14,625 operations predicted to be nonrespondents using only the 45 indicators, 9,828 were actually nonrespondents, resulting in an overall classification accuracy rate of 67.20 percent -- a 0.57 percent reduction in classification accuracy. Using just the 45 indicators, we correctly identified 13.44 percent (9,828/73,126) of the nonrespondents in the ARMS III 2000-2008 samples, which reduced our predictive power by 2.64 percent.

Examination of a scree plot demonstrated that the four factor solution best fit the data (Figure 2).

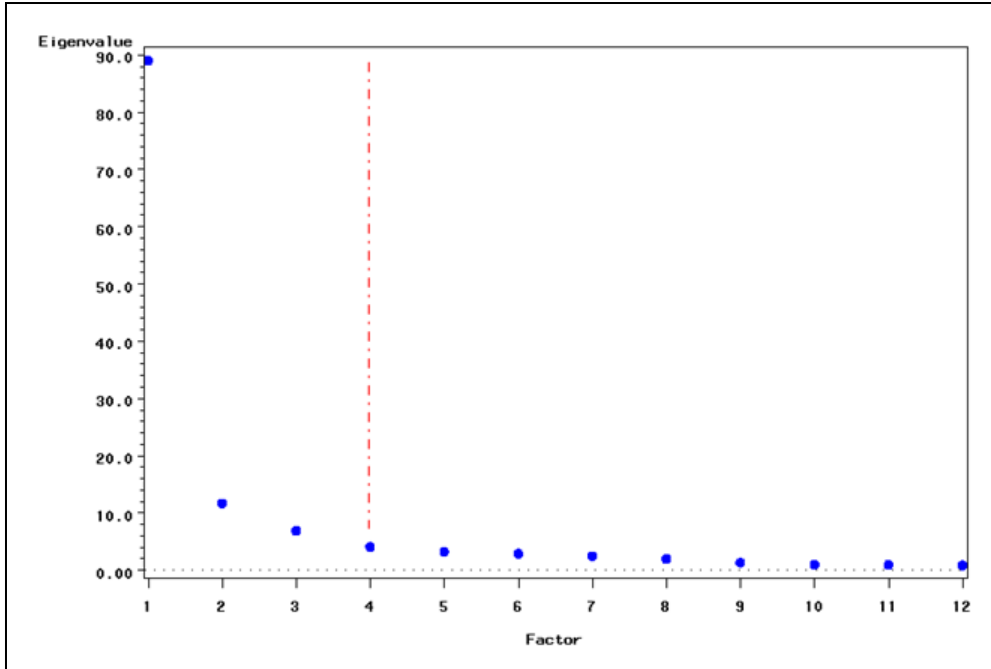


Figure 2. Scree Plot



A scree plot indicates the importance of the factors using eigenvalues. In factor analysis, eigenvalues indicate the amount of variance that is explained by a factor. The elbow of the scree plot indicates when additional factors cease to explain additional variance. Of the 45 indicators tested, 22 loaded under one of the four factors with a loading greater than .50 (Table 2). The remaining 23 indicators were not assessed since they loaded either under multiple factors or under factors with two or fewer indicators, thus creating a weak factor.

Table 2: Rotated Pattern Matrix

Variables	Factor 1	Factor 2	Factor 3	Factor 4
tr67_085	0.93759	0.18461	0.23278	0.08356
tr57_103	0.92625	0.18070	0.22349	0.08110
tr14_107	0.79992	0.09283	0.09259	0.10109
tr23_092	0.79906	0.32947	0.14179	0.08008
tr26_093	0.79104	0.16076	0.25146	0.10927
tr06_092	0.78981	0.16321	0.06826	0.10772
tr01_115	0.78325	0.43464	0.10325	0.09009
tr10_117	0.71191	0.40487	0.05341	0.09175
tr11_094	0.63138	0.12691	0.31005	0.18279
tr15_106	0.62941	0.18979	0.28767	0.02129
tr29_120	0.62022	0.45746	0.09585	0.11056
tr36_076	0.31826	0.74364	0.22951	0.17517
tr37_071	0.24827	0.71130	0.22842	0.16893
tr38_096	0.24829	0.70946	0.30176	0.10896
tr12_086	0.33390	0.59485	0.24446	0.21130
tr28_108	0.21006	0.24080	0.84460	0.05184
tr19_090	0.25298	0.21051	0.73425	0.09723
tr41_109	0.26395	0.36327	0.69351	0.07883
tr24_075	0.08330	0.07346	0.04863	0.74148
tr09_090	0.15836	0.09789	0.04162	0.73337
tr22_075	0.12019	0.13073	-0.00896	0.61205
tr15_070	-0.01317	0.07124	0.09138	0.51796

Using EFA, we identified four main factors of nonresponse. The following six key variables were used to distinguish between different types of nonrespondents: Reported Sum of Expenditures, Total Sales (NUPC), Farm Type, State, Cropland Harvested, and Percentage of Male Operators. Given a standard logistic regression, we could have determined that these variables were all significantly related to survey nonresponse; however, we would not have been able to distinguish between the different levels in relation to other characteristics and far more variables than would be practically useful would have been identified, given the size of our dataset. Classification trees allowed us to identify the optimal break points for maximizing the dichotomy between respondents and nonrespondents using a combination of multiple characteristics. According to the EFA, Reported Sum of Expenditures was an important variable for all four factors; however, the level at which this characteristic influences propensity to respond varies depending on other characteristics (Table 3). According to factor one, a cut off of eight million dollars or more in Reported Sum of Expenditures for operations with greater than 700 thousand dollars in Total Sales (NUPC) that are predominantly male operated is indicative of survey nonresponse (Table 3). For factor two, a cut off of 1.9 million dollars or more in Reported Sum of Expenditures for operations with greater than 800 thousand dollars in Total Sales (NUPC) that produce/raise grain, oilseeds, dry beans, dry peas, vegetables, or hogs in California, Nevada, Arizona, New Mexico, Colorado, South Dakota, Kansas, Oklahoma, Iowa, Indiana, Florida, New York, and Connecticut is indicative of survey nonresponse (Table 3). For factor three, a cut off of three million dollars or more in Reported Sum of Expenditures for operations with greater than 1.3 million dollars in Total Sales (NUPC) that produce/raise vegetables, “other” crops, hogs, pigs, milk, dairy products, cattle, poultry, eggs, or “other” animals and their products is indicative of survey nonresponse. Lastly, according to factor four, a cut off of 1.8 million dollars or more in Reported Sum of Expenditures for operations with greater than 2,400 acres of Cropland Harvested in California, Washington, Arizona, Wyoming, Colorado, New Mexico, South Dakota, Nebraska, Kansas, Oklahoma, Minnesota, Iowa, Missouri, Michigan, Indiana, Florida, and Vermont is indicative of survey nonresponse (Table 3).

Table 3: Factor Characteristic Summary Table

Variables	Factor One	Factor Two	Factor Three	Factor Four
Reported Sum of Expenditures	> \$8,000,000	> \$1,900,000	> \$3,000,000	> \$1,800,000
Total Sales (NUPC)	> \$700,000	> \$800,000	> \$1,300,000	
Farm Types		Grains, Oilseeds, Dry Beans, Dry Peas Vegetables Hogs & Pigs	Vegetables Other Crops Hogs & Pigs Milk & Dairy Products Cattle & Calves Poultry & Eggs Other Animals & Their Products	
States				
Cropland Harvested				> 2,400 Acres
Percentage of Male Operators	> 50 Percent			

Factors one through four classified 12,027 operations as nonrespondents, 8,277 of which were actually nonrespondents, resulting in an overall classification accuracy rate of 68.82 percent. Factors one through four correctly identified 11.32 percent of all nonrespondents (8,277 of 73,126).

Although there was overlap across factors, meaning that some operations were identified as likely nonrespondents by multiple factors, almost half of the operations classified as likely nonrespondents, were identified by a single factor (Table 4). This further underscores the idea that there are distinct and separate groups of nonrespondents. Operations classified as nonrespondents by all four factors had the highest classification accuracy, but they only accounted for 4.92 percent of all operations classified as nonrespondents and were only capable of identifying 0.66 percent of all nonrespondents.

Table 4: Operation Nonrespondent Classification Accuracy by Number of Factors

Number of Factors	Number of Operations Classified as Nonrespondents		Number of Operations Correctly Classified as Nonrespondents		Percent of Nonresponding Operations Identified <i>(n = 73,126)</i>
	Count	Percent	Count	Percent	
One	5,605	46.60	3,557	63.46	4.86
Two	3,242	29.96	2,263	69.80	3.09
Three	2,588	21.52	1,976	76.35	2.70
Four	592	4.92	481	81.25	0.66
Total	12,027		8,277		11.32

Of those operations classified as likely nonrespondents by a single factor, over half were identified by factor one (Table 5). Of those operations classified by multiple factors, factor four appeared to have the smallest amount of classification overlap (Table 5).

Table 5: Operation Nonrespondent Classification across Factors by Number of Factors

Factor	Operations Classified as Nonrespondents By One Factor (<i>n</i> = 5,605)		Operations Classified as Nonrespondents By Two Factors (<i>n</i> = 3,242)		Operations Classified as Nonrespondents By Three Factors (<i>n</i> = 2,588)		Operations Classified as Nonrespondents By Four Factors (<i>n</i> = 592)	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
One	2,807	50.08	2,153	66.41	2,368	91.50	592	100
Two	1,389	24.78	2,230	68.78	2,531	97.80	592	100
Three	213	3.80	1,173	36.18	2,045	79.02	592	100
Four	1,196	21.34	928	28.62	820	31.68	592	100
Total	5,605		N/A		N/A		592	

4. Discussion

Using multiple classification tree models, we are able to identify numerous groups of likely nonrespondents based on auxiliary data available for both survey respondents and nonrespondents. However, these nonrespondent groups are too numerous and overlapping to use effectively in nonresponse bias studies. Our trees specified 226 subgroups or indicators of nonresponse that ultimately identified 16.08 percent of the total nonrespondents in our samples. Overall, it may appear that the predictive power of our indicators is limited; however, considering the breadth of characteristics covered by the 70 Census of Agriculture variables selected and the use of multiple classification trees which considered all of these variables, the limited predictive power suggests that the majority of our nonrespondents (83.92 percent) are missing at random with respect to the auxiliary variables. In this case, many of our auxiliary variables are proxies for or related to the key estimates of the ARMS survey. The groups of nonrespondents we identified are not missing at random and thus have the potential to contribute to bias in our survey estimates, and ultimately these are the ones of most interest. Our four factors, including the final 22 indicators, correctly identified 11.32 percent of the total nonrespondents in our sample, accounting for 70.40 percent of the operations identified by the 226 original indicators that are clearly not missing at random (Table 6).

Table 6: Operation Nonrespondent Classification Accuracy by Number of Indicators

Number of Indicators	Number of Operations Classified as Nonrespondents (<i>n</i> = 199,42)		Number of Operations Correctly Classified as Nonrespondents		Percent of Nonresponding Operations Identified (<i>n</i> = 73,126)
	Count	Percent	Count	Percent	
226	17,355	8.72	11,762	67.77	16.08
45	14,625	7.35	9,828	67.20	13.44
22	8,277	6.04	8,277	68.82	11.32

Using EFA, we are able to distil these groups into a small number of nonrespondent factors. The above results indicate that it is possible to accurately identify and distinguish between considerable numbers of nonrespondents using this smaller number of factors. Building on this work, we will proceed to determine which of the nonrespondent factors contribute to the bias of individual estimates, using structural equation modeling (SEM). The propensity of some nonrespondents may be related to the bias of certain estimates, while the propensity of others may be related to bias in different estimates. Each nonresponse factor and estimate can be examined individually. Furthermore, it is possible that some nonrespondents are completely unrelated to the survey estimate of interest, and thus do not contribute to the bias of that estimate. The number of nonrespondents ultimately included in the factors was only a limited percentage of all likely survey nonrespondents, due to factor loading and identification constraints; therefore, it may also be useful to assess the direct effects of indicators not included in the factors using SEM. It is possible that some indicators identify a type of nonrespondent that is not represented by any other indicators and thus failed to load into a factor, or that the indicator taps into traits specified by all the factors and thus failed to load under a single factor.

While the results we have obtained are specific to the ARMS survey, the approach we have taken here to identify likely nonrespondents can be applied to other surveys where auxiliary data can be matched to sample units. When compared to other approaches such as logistic regression models (Nicoletti & Peracchi, 2005; Lepkowski & Couper, 2002) classification trees have several distinct advantages. For example, they can 1) automatically detect significant relationships and interaction effects without pre-specification, reducing the risk of variable selection or model specification bias; 2) identify not only variables significantly correlated with the target, but also the optimal breakpoints within these variables for maximizing the propensity of the target; 3) identify hierarchical interaction effects across numerous variables, and summarize them using a series of simple rules; 4) treat missing data as valid, and assess whether variable missingness is related to the target; and 5) create a series of simple rules that are easy to interpret and use for identifying subgroups with higher propensities of the target.

Using a data driven approach as this allows us to utilize more of the information in our data set. However, because the initial results with 226 indicators were still unwieldy, we

followed it with EFA. This focused our analysis with four factors that can then be analyzed with respect to their impact on bias in key estimates of the ARMS survey. We are following up work reported in this paper with that analysis. Ultimately this can be used to propose changes in data collection, sampling, or nonresponse adjustment.

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.
- Comrey, A. L., and Lee, H. B. (1992). *A First Course in Factor Analysis* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum Associates.
- deVille, B. (2006). *Decision Tress for Business Intelligence and Data Mining using SAS Enterprise Miner*. Carey, NC: SAS Institute, Inc.
- Groves, R.M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70 (5), 646-675.
- Hair, J.F., Anderson, R.E., Tatham, R.L., and Black, W.C. (1995). *Multivariate Data Analysis*, Upper Saddle River, NJ: Prentice-Hall.
- Johansson, F. and Klevmarcken, 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.
- Lepkowski, 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.
- 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.
- United States. Department of Agriculture (2008). *PRISM II Code Book*. Washington, DC: U.S. Department of Agriculture.
- 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.