Who are You?: A Data Mining Approach to Predicting Survey Non-respondents

Jaki S. McCarthy and Thomas Jacob
USDA's National Agricultural Statistics Service
Room 305, 3251 Old Lee Highway, Fairfax, VA 22030

Abstract

In almost all surveys there is some information known about sample units before data collection. Even in RDD surveys, paradata on call attempts is available and many telephone numbers can be matched to addresses which can be linked to data from external sources such as the Census Bureau (see Lavrakas, 2005). In Federal establishment surveys, such as those conducted by NASS, several classes of auxiliary variables are available. For example, frame information may include establishment or operator characteristics (such as size, type, structure, operators' race, age, gender etc). Also, many establishments will be sampled for multiple surveys, so paradata describing their reporting history (past survey response, refusals, etc.) with the survey organization is also available. Finally, characteristics of the location (county or zipcode) of the establishment are also available. All of this information can be used to model and predict likely survey non-respondents. With large datasets there are many possible ways to construct these models. This paper describes the use of a data mining approach using classification trees to predict survey non-respondents. Classification trees repeatedly divide a dataset to identify subsets of records more likely to be survey non-respondents. The results from our initial models indicate the relatively small subset of variables that were important in predicting survey non-response. This provides insight into the factors related to response propensity. This information can be used to tailor data collection strategies to reduce non-response or may be useful in making non-response adjustments to reduce non-response bias.

Key Words: Agricultural Survey, Classification Tree, Establishment Survey, Non-response, Paradata

1. Background and Introduction

Virtually all surveys suffer from some level of non-response. In order to minimize the impact of non-response on survey estimates, it is useful to model and predict those sample units most likely to be non-respondents. This knowledge can be used proactively to focus intensive field work efforts on those sample units, or can be used in post data collection processing to adjust appropriately for non-response.

Non-response modeling is somewhat limited, because information must be available for both respondents and non-respondents (who obviously do not provide any information in the survey.) Non-response models have been constructed in panel surveys using information gained in initial survey rounds to predict panel attrition in later waves. For example, several studies have been done using first wave household panel socio-

demographic variables and information about the data collection process (referred to as paradata) to predict later survey non-response (Nicoletti and Peracchi, 2005 and Lepkowski and Couper, 2002). These studies have used regression models with several classes of variables as predictors, related to both the "contactability" of the household as well as the propensity to cooperate given contact. Nicholetti and Peracchi found that children in the household, home ownership and length of residence were positively related to contactability of a household, and women in the household, college education, and being out of the labor force were related positively to response.

While less information may be available about cross-sectional survey respondents and non-respondents there are several examples of non-response models in one time surveys. Abraham, Maitland and Bianchi (2006) modeled response propensity in a large household survey conducted as a follow on to the Current Population Survey (CPS) using logistic regression. Auxiliary household and respondent characteristic data (education, income, age, race, gender, etc.) were available for cases from their CPS interviews and were used as predictors in their model. They used their model to evaluate the potential for non-response bias, finding that employment status, single marital status, and urbanicity were all related to non-response. While the survey they examined was not a panel survey, CPS non-respondents were not eligible for the sample. Therefore this study suffered from the same limitation as panel surveys in that sample units who never respond are excluded from the model.

Johansson and Klevmarken (2008) modeled non-response in a Swedish cross sectional household interview survey using a bivariate probit model with auxiliary information from administrative registers used as predictor variables. Unlike in the US, European survey organizations often have rich register data available for sampled survey units. Similar to models for panel surveys, they concluded that variables such as lower income, urbanicity, and single marital status predicted survey non-response. Logistic regression models have also been used to predict likely non-respondents in a RDD survey using available auxiliary data (Burks, Lavrakas, and Bennett, 2005). While data for non-respondents in an RDD survey is limited, auxiliary data including information about the selected telephone number, the call attempt history, interviewer's subjective ratings about the case, and census data matched to the address was available. Using their logistic regression model, they were able to correctly classify cases with respect to whether or not they would respond better than by chance. They found that variables such as interviewers' prediction of respondent cooperation, having a listed mailing address, being in a zip code with more college graduates, higher income, and more owned homes were positively correlated with response and variables such as requesting callbacks, television ownership, lower incomes or lower education levels were negatively correlated with response, although the associations between their predictor variables and non-response were small. Johnson, Cho, Campbell and Holbrook (2006) similarly found weak associations between zip code level measures of affluence and urbanicity and survey non-response.

Similarly, Bates, Dahlhamer and Singer (2008) used paradata available for both survey respondents and non-respondents in the National Health Interview Survey to predict survey refusals. They found that information about the concerns expressed by sampled households to interviewers in contact attempts significantly increased the predictive power of the models over those in which only information about households' location (region of the country and its urbanicity) were used. Using a logistic regression approach, Bates et al were able to identify specific types of concerns which increased the odds of a

household ultimately refusing to participate. In particular, households that stated they were "not interested" were much more likely to refuse to be interviewed.

For many large survey organizations, there may be much information known about survey sample members prior to conducting a survey. Basic information such as location, information used to identify the unit as eligible for the survey, etc. may be known in many surveys. A sample unit's location can be used to associate other external information about that location to the unit. As in the Burks et al study, Census Bureau small area demographic and socioeconomic information can be linked to sample units. In the USDA's National Agricultural Statistics Service (NASS), as in other organizations that survey establishments, there are many establishments that are selected for multiple surveys. Therefore, there may be data from previous contacts available, or other descriptive information carried on the list sampling frame for sample units in cross sectional surveys. In addition, each operation's response history with NASS, that is, whether and how often operations have been sampled in the past on other NASS surveys and whether they were respondents, refusals or non-contacts in those surveys is also known.

In addition, some establishments, unlike households, may be direct data users or have a better appreciation of the utility of the survey estimates. As suggested by Groves, Presser and Dipko (2004), respondents with more interest in the survey topic may be more likely to respond. For NASS surveys, agricultural operations who receive farm program payments from the USDA may recognize the benefits of good NASS statistics or may have a more favorable attitude to NASS. While all sampled units for our target surveys are agricultural operations, those that are larger, operate on a full time basis, derive most of their household income from the farm, or receive government program payments may be more likely to cooperate in NASS's agricultural surveys.

The variables described above can be used as potential predictors to model non-response in an individual survey. This paper presents work different from the non-response models discussed previously in two major respects. First, we examine survey non-response in an establishment survey rather than a household survey. Characteristics unique to establishments such as their size, complexity or type may impact the decision to participate in a survey, and characteristics relevant to households or individuals may be less important. Secondly, we take a different approach to modeling non-response-- the classification tree (also referred to as a decision tree) -- that has several advantages over regression models.

2. METHODS

2.1 Model Approach

For large datasets, classification trees can be used to predict a binary variable (such as survey response/non-response) from auxiliary variables. In this approach, a classification tree model is constructed by segmenting a dataset using a series of simple rules. Each rule assigns an observation to a segment 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 maximally separate the sub-segments with respect to the target variable. The 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 interested in the leaves that contain a higher proportion of records

with the target variable. We created separate models to predict survey refusals, non-contacts, and cases held out of data collection (termed office holds) as the target.

A classification tree model has several advantages over a regression model. First, cases with missing data are often dropped from regression models; in classification trees missing data is treated as valid. This is important for non-response models where ideally we would like to have data on all cases, but in practice it is often missing for subsets of records. The fact that data is missing may be an important predictor of response which would be lost if missing data were imputed or those cases excluded from analysis. Second, decision trees do not suffer from the inclusion of large numbers of predictor variables, or the inclusion of correlated variables, as they examine each predictor sequentially. Therefore, we are not forced to reduce the variables included in the model as is done in many regression models. In addition, the branches of the tree implicitly create significant variable interactions, so these need not be generated and manually included in the model as additional variables. Including interaction terms in regression models is not difficult, but with many variables and multi-way interactions, including all interactions of potential interest quickly becomes unwieldy. Finally, the subgroups of records with the highest percentage of the target of interest are explicitly defined by the resulting model and easily interpretable.

2.2 The Dataset

The first survey for this project was NASS's December 2006 Crops/Stocks Survey. The Crops/Stocks Survey provides detailed estimates of crop acreage, yields and production and quantities of grain stored on farms. It is conducted in all states with a sample of farm and ranch operations producing row crops and small grains selected by size. The data collection period for the Crops/Stocks Survey is short, approximately 2 weeks at the beginning of the reference month. Data is collected primarily by telephone, but also includes limited mail and personal interview collection. Data collection is administered by each of NASS's field offices, and results are combined to produce both National and State statistics. Each sampled operation in the survey was assigned one of the following outcome dispositions:

- Complete – respondent was contacted and data collected,
- Refusal – respondent was contacted and refused to participate,
- Non-contact – respondent was not contacted or was unavailable during the data collection period,
- Office hold – the field office held the case out of data collection (this would be the case if the operation had previously been hostile, or for some other reason should not have been contacted, this is left to the discretion of the individual field office),
- Known zero – operation was not contacted because of prior information indicating they were out of scope for the survey.

For each sampled operation (both respondents and non-respondents), a number of different auxiliary variables were available. These fell into several categories: information about the target survey, information from the 2002 Census of Agriculture, information carried as control data on the NASS list sampling frame (LSF), external county level descriptive variables and information generated from the operation's past response history with NASS.

Information from the recent census of agriculture is available for most of the sampled operations in any NASS survey, since the census includes all known and potential agricultural operations and response is required by law. There are both variables describing the agricultural operation (such as the commodities raised) and the operator (such as age, race and gender) on the census. Additional variables which may be associated with interest in the survey topic such as size, whether they are run by full time farmers, the percent of their household income derived from the farm, or receipt of government agricultural program payments, are also available.

Information carried on the list frame comes from many sources, including pre-census and survey screening, previous NASS survey data collections, administrative records, other external lists, etc. The location of the operation is one variable on the NASS list frame. This was used to attach descriptive information from sources outside NASS about the county to each operation. This included how much of the county was in farmland and how urban the county was. Other studies of non-response have found that urbanicity is negatively correlated with response and the dataset used in this analysis includes the Urban Influence Codes which are based on metro status as classified by the Office of Management and Budget. Because of the nature of our population, i.e. farms and ranches, a simple urban/rural location indicator may not sufficiently capture the geographic differences among our survey sample locations. Therefore, an urbanicity measure that considered finer degrees of rural classification which might be relevant to agricultural operations was also added. This measure was the Rural-Urban Continuum Codes developed by the USDA's Economic Research Service (available at http://www.ers.usda.gov/Data/RuralUrbanContinuumCodes/) which distinguishes metropolitan counties by size and nonmetropolitan counties by degree of urbanization and proximity to metro areas. A comparison of the two coding schemes can be found in Ghelfi and Parker (1997). We included the most recent classifications (2003) as well as the previous classifications based on 1993 information.

In an attempt to monitor and reduce survey burden NASS also computes several indicators termed the Joint Burden Indicators (JBIs). Separate indicators are computed each year for the projected number of NASS surveys each operation has been sampled for, the total number of survey contacts (since data for several surveys can be collected in a combined contact), and the total number of estimated minutes for the contacts. JBIs for the previous 3 years were included as predictors.

While the JBIs estimate the maximum NASS burden that would be imposed on each operation per year, they do not measure how responsive operations have been. Therefore, an individual NASS response rate was computed for each operation for the previous 1, 2 and 3 year periods. Also computed was the number and percentage of time each operation was a non-contact, refused or was an office hold, i.e. no contact was attempted. Operations known to be out of scope for a particular survey are termed "known zeros" and are also excluded from data collection, this was also computed.

The full list of predictor variables is shown below:

VARIABLE	DESCRIPTION
Target Survey Variables:	
Response	Response outcome (complete, refusal, noncontact, office hold, other)

Census of Agriculture variables:	
Bees	Bee indicator: Yes/No
Cattle	Cattle indicator
CRP	Conservation Reserve Program indicator (based on acres)
Government Program	Receipt of Government Program payments
Current_Status Code	Census of Agriculture response outcome (complete, refusal, inaccessible, non-contact)
Exp_Farmtype	Expected Farm Type used in Census (e.g. grain, tobacco, hog, nursery, cattle, poultry, aquaculture, etc.)
Fruits	Fruit indicator
Hay	Hay indicator
Hogs	Hog indicator
Horse	Horse indicator
K46	Total Acres Operated, (i.e. land owned, minus land rented out, plus land rented in)
K787	Acres of Cropland Harvested
Nursery	Nursery indicator
Organic	Organic indicator
Poultry	Poultry indicator
Sheep	Sheep indicator
Tenure	Farm tenure (1=full owner, 2=part owner, 3=tenant)
Vegetables	Vegetable indicator
Po_Box_Flag	Mailing address was a PO box
Off_farm_job	Principal operator works at off farm job
Operator_living	Principal operator lives on the farm
Hired Manager	Operator is a hired manager
Occupation	Principal operator's primary occupation is farming
%_HHincome	% of Household income produced by the operation
Yr_begin_operation	Year the operator began operating this operation
Raceid	Race of principal operator
Sex	Sex of principal operator
Spanishoriginid	Spanish ethnicity indicator
TVP	Total Value of Production
Activestatusid	Census active status code
List Frame Variables	
District_Code	Agricultural Statistics District
Elmo_In_Census_Flag	Operation was on Census Mail List
Elmo_age	Age of Operator
Elmo_dtActiveStatus	Current Operating Status (in business)
Elmo_dtAdded	Years on the NASS list frame
Expected_Sales_Group	Expected Sales Group
Farmtype	NASS farm type
Nass_State_Fips	State identifier

Email	Operation has email address on file
County Level Variables	
1993_Rural_Urban_Continuum_Code	ERS Rural/Urban indicator as classified in 1993
_1993_Urban_Influence_Code	OMB Urban influence code as classified in 1993
_2003_Rural_Urban_Continuum_Code	ERS Rural/Urban indicator as classified in 2003
_2003_Urban_Influence_Code	OMB Urban influence code as classified in 2003
Percent_farmland	Percent of the total land in the county in farmland
2000_persons_persq_mile	Population density of county in 2000
_2000_population	Total population of the county in 2000
Response History Variables	
I51_cur_yr	Joint Burden Index (JBI) projected number of surveys for the current year
I51_prior1_yr	JBI projected number of surveys for prior year
I51_prior2_yr	JBI projected number of surveys for 2 years ago
I52_ cur_yr	JBI projected number of contacts for current year
I52_ prior1_yr	JBI projected number of contacts for prior year
I52_ prior2_yr	JBI projected number of contacts for 2 years ago
I53_ cur_yr	JBI projected number of OMB minutes for current year
I53_ prior1_yr	JBI projected number of OMB minutes for prior year
I53_ prior2_yr	JBI projected number of OMB minutes for 2 years ago
Knownzero_cur_yr	Number of surveys for which operation was sampled but classified as out of scope in current year
Knownzero_ prior1_yr	… in prior year
Knownzero_ prior2_yr	… in 2 years ago
Cum_knownzero_2yr	Cumulative number of surveys for which operation was sampled but classified as out of scope in past 2 years
Cum_knownzero_3yr	Cumulative number of surveys for which operation was sampled but classified as out of scope in past 3 years
Complete_ cur_yr	number of contacts operation responded in current year
Complete_ prior1_yr	… in prior year
Complete_ prior2_yr	… in 2 years ago
Cum_complete_2yr	Cumulative number of surveys for which operation responded in past 2 years
Cum_complete_3yr	Cumulative number of surveys for which operation

	responded in past 3 years
Inaccessible_ cur_yr	number of contacts operation was non-contact in current year
Inaccessible_prior1_yr	… in prior year
Inaccessible_ prior2_yr	… in 2 years ago
Cum_ Inaccessible_2yr	Cumulative number of contacts operation was non-contact in past 2 years
Cum_ Inaccessible_3yr	Cumulative number of contacts operation was non-contact in past 3 years
Officehold_ cur_yr	number of contacts operation was held in the office (no contact attempt made) in current year
Officehold_ prior1_yr	… in prior year
Officehold_ prior2_yr	… in 2 years ago
Cum_officehold_2yr	Cumulative number of contacts operation was held in office in past 2 years
Cum_officehold_3yr	Cumulative number of contacts operation was held in office in past 3 years
Refusal_ cur_yr	number of contacts operation refused in current year
Refusal_ prior1_yr	… in prior year
Refusal_ prior2_yr	… in 2 years ago
Cum_refusal_2yr	Cumulative number of contacts operations refused in past 2 years
Cum_refusal_3yr	Cumulative number of contacts operations refused in past 3 years
Responserate_1year	% of contacts with positive response in prior year
Responserate_2year	% of contacts with positive response in prior 2 years
Responserate_3year	% of contacts with positive response in prior 3 years

2.3 The Target Survey

Several models were developed for individual surveys. This paper will describe the models built for the NASS December 2006 Crops/Stocks survey. Separate models were built for survey refusals, survey non-contacts, and survey office held cases.

The sample size and response outcomes for the survey are shown below:

December 2006	N	Percent
Complete	56,594	65.96
Refusal	12,367	14.41
Non-contact	11,997	13.98
Known zero	3,312	3.86
Office hold	1,528	1.78
TOTAL	85,798	100%

2.4 Building Classification Tree Models

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 segment based on the value of one input variable. For example, the segmenting rule may be to divide the dataset into groups, one with records reporting a certain commodity, and one with records that do not report the commodity. One rule is applied after another, resulting in a hierarchy of segments within segments. The rules are chosen to maximally separate the subsegments with respect to the target variable. Thus, the rule selects both the variable and the best breakpoint to maximally separate the resulting subgroups. In other words, the segmenting rule divides records into groups with more and less of the target based on their reports of a commodity, and also selects the amount of that commodity that maximally separates the groups. For categorical variables, the rule will select the groups of categories that maximally separate the subgroups. The categorical groupings and continuous variable breakpoints are not defined by the researcher but are dictated by the data.

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. Each record in the dataset will appear in one of the tree leaves, and the leaves will collectively contain all records in the dataset. In our analysis, the leaves of interest were those containing a higher proportion of records with the target.

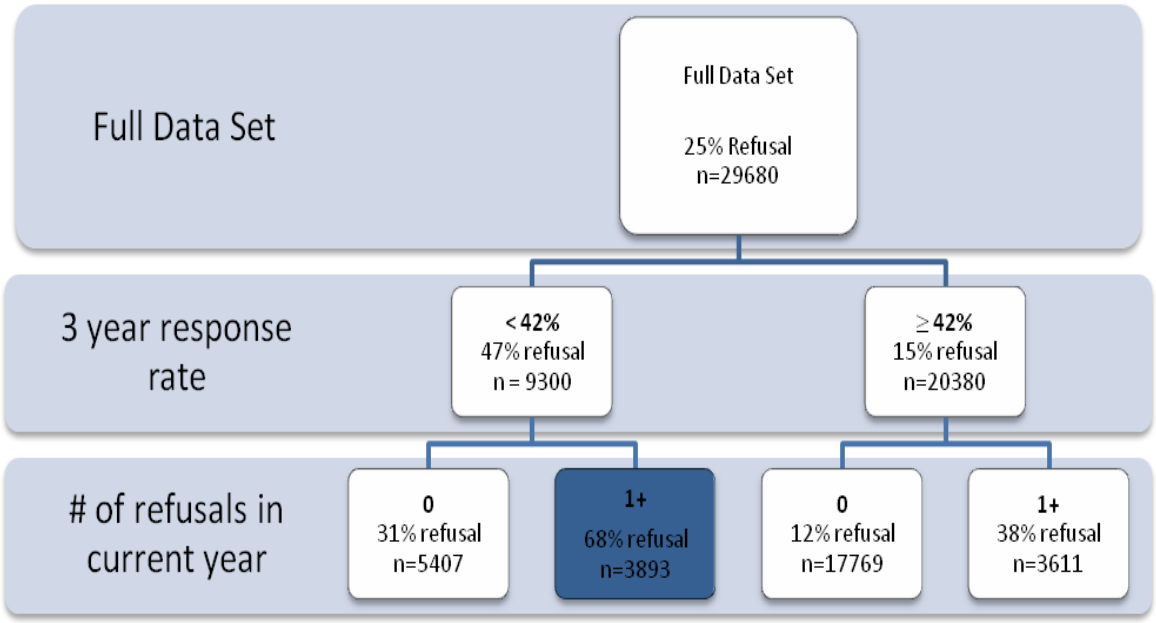
Decision trees describe subsets of data and are constructed without any theoretical guidance. Variables are chosen that maximally separate the sub-segments, so only one or a few similar correlated variables (which individually might be related to the target) may appear in the tree. There are several alternative methods for growing decision trees; our trees were grown using the chi-square approach available in SAS Enterprise Miner 5.2, which is similar to the chi-square automatic interaction detection (CHAID) algorithm. (See deVille, 2006 for a discussion of the algorithms used in SAS Enterprise Miner, the software used in this analysis.) There are multiple stopping criteria that can be used to decide how large to grow a decision tree. Generally, we pruned the trees when there were no appreciable gains in the misclassification rates (or mean squared error rates) of the trees.

In this type of analysis, the dataset is first oversampled to increase the percent of the data with the target of interest. This is to allow for enough data for analysis, once the dataset is split multiple times. As is standard practice in data mining applications, we retained all records with the model target and randomly sampled the remaining records so that the target appeared 25% of the time in each of our initial datasets. Then the full data set is randomly partitioned into 3 subsets. These subsets are termed the training, validation and test sets. For our analysis, 60%, 30% and 10% of the data were apportioned into these subsets, respectively. The training dataset is used to construct the initial tree model. This model is then applied to the validation dataset in order to prevent generating a model for the training data that does not fit other data (i.e.overfitting). Finally, the test set is used to evaluate the model's performance on independent data not used in the creation of the model. All trees had similar misclassification rates for the training and validation datasets used to grow the trees and for the test data used after the trees were constructed. Similar results were obtained for training, validation and test datasets, therefore, for simplicity, only the training data are shown.

Decision tree models were generated separately for refusals, non-contacts and office hold cases, since these likely have different causes.

3. RESULTS AND DISCUSSION

The first tree for refusals is shown below:

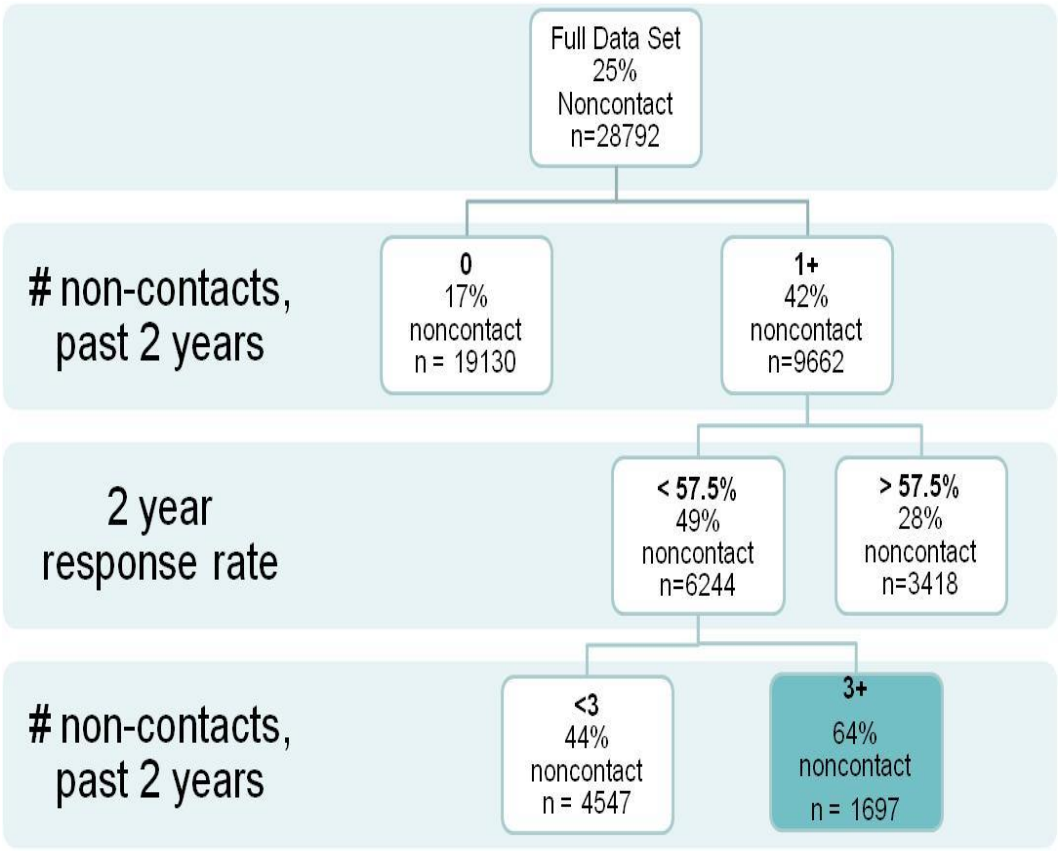


In this model, the first split uses the response rate calculated over the prior 3 years, with those operations having a 42% or lower response rate being almost twice as likely to be refusals as the entire dataset. Continuing down this branch, within that group, of those operations who also had one or more refusals in the current year 68% were refusals. This is shown in the highlighted node. It is interesting to note that none of the census of agriculture, list frame, or external variables appears in the tree. This tree illustrates that operations most likely to be refusals in the current survey are those that have not cooperated in the past and with recent refusals in other surveys.

One way to evaluate a classification tree is by the misclassification rate in the prediction of the target. In using this tree model, the rule generated for each terminal node would be used to classify individuals in that node. Individuals falling in a node with less than 50% of the target would be classified as "not target" and those in nodes with 50% or greater would be classified as "target." Since none of the nodes in our trees produces a group with 100% (or 0%) of the target, using these rules will always misclassify some individuals. Since there are 25% of the cases that are the target in the full dataset, if we used that node as the model, we would classify all records as not having the target, producing a misclassification rate of 25%. Instead, using each of the terminal nodes to classify records in the full tree, all records except in the blue highlighted box (those in nodes with less than 50% refusals) would be classified as "target" (refusals). Using these

rules, the misclassification rate on the test dataset was 19%. Thus, our model reduced the misclassification rate from 25 to 19%.

The tree generated for the non-contact cases is shown here:

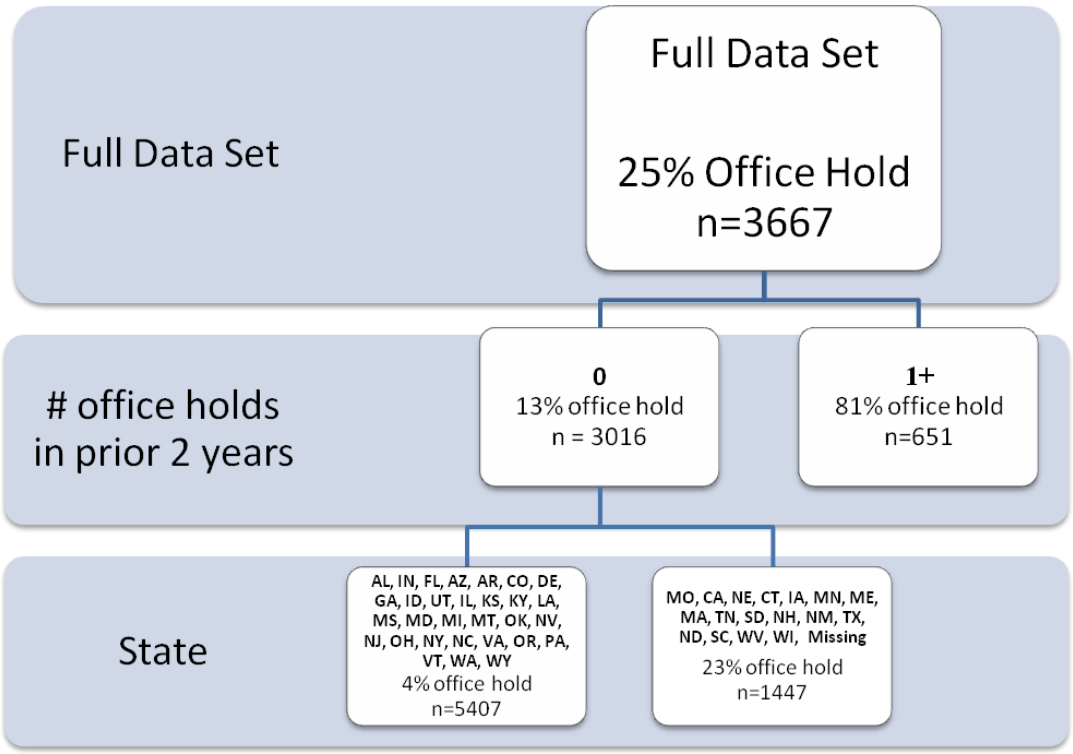


In this tree, the branch and leaf of interest is:

- those operations who have had 1 or more non-contacts in a NASS survey in the past 2 years,
- had a less than 57.5% response rate over the last 2 years, and
- 3 or more non-contacts in the past 2 years.

Sixty four percent of the operations in this node were non-contacts in the target survey, compared to 25% in the original dataset. This model did not perform as well as the refusal model producing a more modest gain to only a 23% misclassification rate. Again, no descriptive variables from the census of agriculture, list frame or external variables were useful in the model. Similar to refusals, those operations most likely to be non-contacts in the current survey were those that had been non-contacts in the past and had not provided data often in the past.

A final model was built for office held cases. The resulting tree is shown below:



It appears that field offices are more likely to hold cases that have been held in the office for another survey in the prior 2 years. This appears to be more often the case for some states than others, likely based on the particular management in those offices. The interpretation for this tree is a bit different from the prior two, since the decision to hold a case in the office is made by NASS staff, not by the potential respondent. Therefore, this model merely identifies the criterion used by the field offices to keep a case out of data collection. For that reason, it is not surprising that this model performs better than either of the preceding two, with a reduction in the misclassification rate from 25% to 13%. The factors impacting the decision to hold cases in the office are likely not captured in any of the variables we were able to include in the model (e.g. an operation has threatened an enumerator or is considered dangerous).

4. GENERAL DISCUSSION

The research discussed in this paper presents a novel approach to modeling survey non-response. Using classification trees to identify the most likely non-respondents offers several advantages over alternative methods, such as regression models. For example, we included a large set of variables, including many that were highly correlated. The models identify those which are the most important predictors and including other variables does not dilute the results. Similar to regression models, these trees can be used to score records. These models also have the advantage that the subgroups of interest are clearly defined and thus easily interpretable. Although it was not critical for our non-response models, classification trees also have the advantage of including missing data. In previous research (McCarthy and Earp, 2009) we have found that missing data can be an important characteristics of sample records (and one that would not have been identified had these records been either imputed or excluded from analysis.) Finally, classification trees have the advantage that they do not require the researcher to

hypothesize interaction effects in advance. Interactions are inherent in the structure of the tree, so relationships need not be linear.

Similar to research by others, we did not find that characteristics of the sampled units such as race, gender, urbanicity, or other descriptive variables were strong useful predictors of survey non-response. Even variables which we hypothesized might indicate greater interest in USDA surveys, such as participation in USDA agricultural programs, having farming as their primary occupation, or the size of the operation (which increases the probability that they directly use the statistics produced by the survey), did not greatly distinguish respondents from non-respondents. The only variables included in our models for refusals and non-contacts were those that described how cooperative sample units had been in the past.

Interestingly, the variables which measured the prior burden NASS has placed on these operations were also not helpful in predicting response. Previous research conducted by NASS (McCarthy, Beckler and Qualey, 2006) supports the idea that burden, by any traditional definition (i.e. number, frequency or length of contacts) does not predict non-response. The variables that are necessary to accurately predict survey refusals appear to be other than any that are typically available to survey researchers. The strongest correlates of survey non-response we have seen have been measures of the knowledge and attitudes our respondents have about NASS (McCarthy, Johnson and Ott, 1999). These may be particularly relevant in establishment surveys where the link between an establishment and potential uses of the survey statistics may be clearer. Of course, information such as the respondents' attitudes toward the survey sponsor, belief in the lack of utility of the outcome of the survey, etc. are not the type of data typically available about survey respondents.

In some respects, the fact that we didn't find large differences between our respondent and non-respondents is good news. Since non-respondents do not appear to fall into specific sizes or types of operations, this suggests that the non-response is not introducing bias into our survey estimates. We did build classification tree models without our response history variables, and the models are quite weak and do not provide any increase in non-response classification accuracy (McCarthy, Jacob and McCracken, in preparation). Classification tree approaches have been used to create non-response weighting groups (Cohen, DiGaetano and Goksel, 1999 and Cecere, 2008), but it does not appear that this method will produce any substantial gains for this survey.

While these models do not provide much insight into the causes or correlates of non-response, they can be used to modify data collection techniques. The groups identified as most likely to be non-respondents can be identified (or the terminal nodes can be used to rank order subgroups of the sample) before data collection and the likeliest non-respondents targeted with alternative data collection strategies. For example, these can be assigned to more experienced enumerators, can be contacted earlier in the data collection period, assigned to personal enumeration, etc. Of course, if we are successful in converting potential non-respondents into good respondents, this will affect their response histories in future surveys. Therefore, it is unclear how useful these models will continue to be if we are able to successfully use them. Our models should be rebuilt in the future to determine whether they remain predictive.

The trees shown in this paper are the first step in ongoing efforts to predict survey non-respondents and ultimately use that information to increase future response rates. In this

paper, we examined only one quarter (in one year) of the quarterly Crops/Stocks Survey. Future work will include building similar models using datasets for other NASS survey periods, and other NASS surveys (see McCarthy, Jacob and McCracken, in preparation). This will enable us to see whether the variables important in these initial models are the same for other surveys. From there, we will work on methods to incorporate these models into ongoing data collection planning and operations.

References

Abraham, K.G., Mailand, A. and Bianchi, S.M. (2006). Non-response 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.

Bates, N., Dahlhamer, J. and Singer, E. (2008). Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Survey Non-response. Journal of Official Statistics, 24(2), 591-612.

Burks, A.T., Lavrakas, P.J., and Bennett, M. (2005). Predicting Sampled Respondents' Likelihood to Cooperate: Stage III Research. Presented at the Annual Conference of the American Association for Public Opinion Research.

Cecere, W. (2008) 2007 Census of Agriculture Non-response Methodology. US Department of Agriculture, National Agricultural Statistics Service. Research and Development Division Report.

Cohen, S.B., DiGaetano, R., and Goksel, H. (1999). "Estimation Procedures in the 1996 Medical Expenditure Panel Survey Household Component," Agency for Health Care Policy and Research, MEPS Methodology Report No. 5, AHCPR Publication No. 99-0027, Rockville, MD.

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

Ghelfi, L.M. and Parker, T. S. (1997). A County-Level Measure of Urban Influence. Rural Development Perspectives, 12(2), 32-41.

Groves, R.M., Presser, S. and Dipko, S. (2004). The Role of Topic Interest in Survey Participation Decisions. Public Opinion Quarterly, 68(1), 2-31.

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, Y.I., Campbell, R.T., and Holbrook, A.L. (2006). Using Community-Level Correlates to Evaluate Non-response Effects in a Telephone Survey. Public Opinion Quarterly, 70(5), 704-719.

Lepkowski, J.M. and Couper, M.P. (2002). Non-response in the Second Wave of Longitudinal Household Surveys. In *Survey Non-response*, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds.). 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.

McCarthy, J., Beckler, D. and Qualey, S. (2006). An Analysis of the Relationship Between Survey Burden and Non-response: If We Bother Them More, Are They Less Cooperative? Journal of Official Statistics, 22(1), 97-112.

McCarthy, J. and Earp, M. (2009). Who Makes Mistakes? Using Data Mining Techniques to Analyze Reporting Errors in Total Acres Operated. US Department of Agriculture, National Agricultural Statistics Service. Research and Development Division Report RDD09-02.

McCarthy, J., Jacob, T. and McCracken, A. (in preparation). Modeling NASS Survey Non-response Using Classification Trees. US Department of Agriculture, National Agricultural Statistics Service. Research and Development Division Report.

McCarthy, J. Johnson, J and Ott, K. (1999). Exploring the Relationship Between Survey Participation and Survey Sponsorship: What do Respondents and Non-respondents Think of Us? Presented at the International Conference on Survey Non-response, Portland, Oregon.