**United States Department of Agriculture**

**National Agricultural Statistics Service**

Research and
Development Division
Washington DC 20250

RDD Research Report
Number RDD-10-05

November 2010

# Modeling Non-response in National Agricultural Statistics Service (NASS) Surveys Using Classification Trees

Jaki S. McCarthy, Thomas Jacob, and Amanda McCracken

This report has been prepared for limited distribution to the research community outside the U.S. Department of Agriculture (USDA). The views expressed herein are not necessarily those of NASS or USDA.

# EXECUTIVE SUMMARY

Auxiliary data from several sources were matched to all operations sampled for the March, September and December 2006 and 2007 Crops/Stocks survey samples. These data fell into several categories: information from the 2002 Census of Agriculture, information carried as control data on the NASS list sampling frame (LSF), county and zip code level descriptive variables obtained from the Census Bureau, Joint Burden Indicators (JBIs), and information generated from the operation's past response history with NASS. Using these auxiliary variables, classification tree models were built to identify those records most likely to be survey refusals or non-contacts (inaccessibles). The models indicate that the records most likely to be refusals are those that have been refusals two or more times in the past three years and have completed fewer than two surveys in the past three years. These records are more than four times as likely to be refusals as the overall sample. The records most likely to be non-contacts are those that have had more than one non-contact in the past two years, with a cumulative response rate to NASS surveys of less than 25.5 percent, and been non-contacts three or more times in the past three years. These operations were more than three times as likely to be non-contacts as the sample as a whole.

These models predict non-respondents consistently for the Crops/Stocks survey and also identify most likely non-respondents on other surveys such as cattle and labor, but are not effective in identifying non-respondents in the Agricultural Resource Management Survey (ARMS). Because no substantive variables appear in the models, this suggests that non-response is introducing little non-response bias in the Crops/Stocks surveys. However, these models can be used to rank order groups of survey samples in the future to help manage data collection.

## RECOMMENDATIONS

1. Use these tree models to score future Crops/Stocks and Cattle Survey samples and make the scores available for use by the field offices. *This recommendation is in progress, with plans in place to score the September and December 2010 Crops/Stocks samples and the January 2011 Cattle survey samples.*
2. Change data collection procedures for those records identified as most likely to be refusals or non-contacts.
3. Evaluate whether the scores were useful in altering data collection procedures by comparing predicted non-response rates for those records with actual non-response rates for the highest non-response groups.

# Modeling Non-response in National Agricultural Statistics Service (NASS) Surveys Using Classification Trees

**Jaki S. McCarthy, Thomas Jacob and Amanda McCracken**[1]

## Abstract

This paper describes the use of classification trees to predict survey refusals and inaccessibles. Data from auxiliary sources were matched to the 2006 and 2007 March, September, and December Crops/Stocks survey sample members. The data matched included variables such as establishment size (both in dollars and acres), type of commodities produced, operating arrangement, operator characteristics (such as race, age, gender, etc.) from the Census of Agriculture, paradata describing their NASS reporting history (past NASS survey response, refusals, etc.), Joint Burden Indicators, and characteristics of the location of the operation (by county and zip code) that were available from the Census Bureau. Classification trees used these data to repeatedly divide our dataset to identify subsets of records more likely to be survey non-respondents. This approach was initially applied to the NASS Crops/Stocks survey, and then applied to other NASS surveys. The results from our models indicate the relatively small subset of variables that are important in predicting survey response. The most useful variables all come from the set of NASS reporting history variables. These models work consistently for the Crops/Stocks survey and for some surveys such as Cattle, but less well for others such as ARMS. Using these models, sampled operations can be ranked based on their predicted response likelihood. These may be useful for field offices to plan alternative data collection strategies for the operations most likely to be non-respondents.

**Key Words:** Non-response Models, Classification Trees, Refusal, Non-contact

## 1. INTRODUCTION AND BACKGROUND

Virtually all surveys suffer from some level of non-response. In order to minimize the impact of non-response on survey estimates and survey costs, it is useful to model and predict those sample units most likely to be non-respondents. This knowledge can be used to proactively focus

[1] Jaki McCarthy is the Chief Cognitive Research Methodologist with the National Agricultural Statistics Service, Research and Development Division (NASS/RDD), 3251 Old Lee Highway, Room 305, Fairfax, VA 22030. Thomas Jacob is a member of NASS's Information Management Group, and Amanda McCracken contributed to this project during an internship with the Research and Development Division. The authors would like to thank Will Cecere, also of NASS/RDD, for his assistance on this research project.

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 typically limited, because information must be available for both respondents and non-respondents (who obviously do not provide any information in the survey). One approach has been to construct non-response models in panel surveys using information gained in initial survey rounds to model 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 the contactability of a household, and women in the household, college education, and being out of the labor force were related positively to response. The obvious drawback to these studies of non-response is that they do not include non-respondents from the initial panel wave.

Although not in a panel survey, Abraham, Maitland, and Bianchi (2006) similarly 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 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 were excluded from the model.

Johansson and Klevmarken (2008) modeled nonresponse 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. Burks, Lavrakas, and Bennett (2005) used logistic regression models to predict likely non-respondents in a random digit dial survey. While data for non-respondents in an RDD survey were 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 were 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.

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 demographic and socioeconomic information at the county or zip code level can be linked to sample units. In counties or zip codes with higher percentages of people speaking foreign languages, it may be more difficult to elicit survey cooperation. In addition, cooperation rates may be lower in areas with lower education rates.

In NASS, as in other organizations that survey establishments, there are many establishments that are selected for multiple surveys. Therefore, there may also 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 almost all sampled units for our 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 have more interest (and thus be more cooperative) 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 respect to the target variable. We created separate models to predict survey refusals and non-contacts 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 are treated as valid. This is important for non-response models where ideally we would like to have data on all cases, but in practice they are often missing for subsets of records. The fact that data are missing may be an important predictor of response. 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 them makes a regression model unwieldy and hard to interpret. Finally, the subgroups of records with the highest percentage of the target of interest are explicitly defined by the resulting model and are easily interpretable.

### 2.2 The Dataset

The non-response models we built were based on the NASS Crops/Stocks survey, which provides detailed estimates of crop acreage, yields, and production and quantities of grain stored on farms. It is conducted quarterly 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 two weeks at the beginning of the reference month. Data are collected primarily by telephone, but also include limited mail and personal interview collection. Response is voluntary, and non-response rates typically are between 20 and 30 percent. 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,
- Known zero – operation was not contacted because of prior information indicating they were out of scope for the survey,
- 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, as determined by the individual field office).

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), county and zip code level descriptive variables from sources outside NASS, and information generated from the operation's past response history with NASS.

Information from the 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 variables describing both the agricultural operation (such as the commodities raised) and the individual 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, and whether or not they have received 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. Variables on the list frame include the operations' current business status (operating/out of business), the expected farm type, the age of the operator, the number of years they have been on the list frame, and their location. The operation address was used to attach location-specific information to the records for analysis, such as descriptive information from sources outside NASS about the county or zip code 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 used urbanicity indicators 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 addition, information from the US Census Bureau regarding the population characteristics of the zip code or county was attached to each operation. These variables included the total population and the population density of the county, as well as the percent of population below the poverty line, percent foreign born, percent speaking a foreign language, percent of high school graduates, and the total population in the zip code.

In an attempt to monitor and reduce survey burden NASS also computes several indicators termed the Joint Burden Indicators (JBIs). Separate variables are computed for individual operations each year reflecting the projected number of surveys for which each has been sampled, 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 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 inaccessible, refused or was held in the field office, 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, and the number and percentage of these was also computed. The full list of variables is shown in Appendix A.

### 2.3 The Target Surveys

Separate models were built for survey refusals, survey non-contacts, and survey office held cases. Models were built using a combined dataset with data from the March, September, and December 2006 and 2007 Crops/Stocks surveys.

The sample size and response outcomes for the combined survey dataset follow:

**Table 1. Sample disposition in model dataset**

| Sample Outcome | N | Percent |
|---|---:|---:|
| Complete | 330,536 | 69.05 |
| Refusal | 60,773 | 12.70 |
| Inaccessible | 59,383 | 12.41 |
| Known zero | 17,314 | 3.62 |
| Office hold | 10,695 | 2.23 |
| TOTAL | 478,701 | 100.00 |

Since survey office held cases and operations known to be out of scope for the survey (known zero operations) are not non-respondents, these cases were removed from the data sets used for analysis.

### 2.4 Building Classification Tree Models

Classification trees describe subsets of data and are constructed without any theoretical guidance beyond the inclusion of the variables in the dataset. 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, and 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 value for an individual variable, and also selects the amount of that variable 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.

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.

Typically, in this type of analysis, the dataset 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. For simplicity, only the training data are shown.

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

## 3. RESULTS AND DISCUSSION

### 3.1 Univariate Comparisons of Respondents and Non-respondents

Simple univariate comparisons of completes, refusals, and non-contacts show that virtually all of the variables we included in our dataset are significantly different among these groups. There were several notable differences between respondents, refusals and non-contacts. Refusals are more rural on average (by several measures) than cooperators, they are more likely to be full time farmers, and more likely to be grain or oilseed farm types. A sample of the variables is shown in the Table 2.

Given the size of the dataset, it is not surprising that most variables show significant statistical differences. However, the effect sizes are small and do not provide practical ways to identify specific operations likely to be respondents or non-respondents. For example, 70.45 percent of refusals received some government payments, versus only 62.36 percent of respondents. This difference is significant, and even though it is relatively large, knowing whether or not an individual operation received government payments is not particularly helpful in identifying likely refusals. Respondents and refusals are both more likely than not to have received government payments. Similarly, even though the average level of county urbanicity is less for refusals, targeting just the most rural counties will not identify all or even most of the refusals. Interestingly, the JBI's, while significantly different, show the opposite pattern than what might be expected, with larger JBI's for completions than for either refusals or non-contacts.

**Table 2. Selected univariate comparisons**

| Measure | Complete | Refusal | Non-contact |
|---|---|---|---|
| Expected Farm Type: Grains, Oilseeds, Dry Beans and Dry Peas | 37% | 51% | 39% |
| Expected farm type: Cattle and Calves | 24% | 18% | 20% |
| CRP reported | 18% | 23% | 18% |
| Soybeans reported | 73% | 85% | 76% |
| Hired Manager reported | 5% | 4% | 5% |
| Hours worked at an off farm job= 0 | 65% | 72% | 64% |
| Operator lives on operation | 82% | 84% | 79% |
| Major occupation is farming | 86% | 91% | 86% |
| Average % of the R's county in farmland | 63.00 | 73.73 | 64.97 |
| Average % of R's county population that is foreign born | 3.5 | 3.07 | 3.86 |
| Average number of people per square mile in the R's county | 95.09 | 68.12 | 94.08 |
| 2003 Rural Urban continuum Code = 8 or 9[2] | 19% | 26% | 18% |
| Total Minutes joint burden index (JBI) for current year | 112.83 | 112.69 | 111.04 |
| Total Contacts JBI for current year | 6.25 | 5.86 | 5.80 |
| Total Surveys JBI for current year | 6.59 | 6.22 | 6.19 |
| Individual operation response rate (past 3 years) | 77.02% | 37.78% | 48.76% |
| Average number of completions (past 3 years) | 4.66 | 2.26 | 3.01 |
| Average number of refusals (past 3 years) | .46 | .80 | 2.24 |
| Average number of non-contacts (past 3 years) | .49 | 1.44 | .67 |

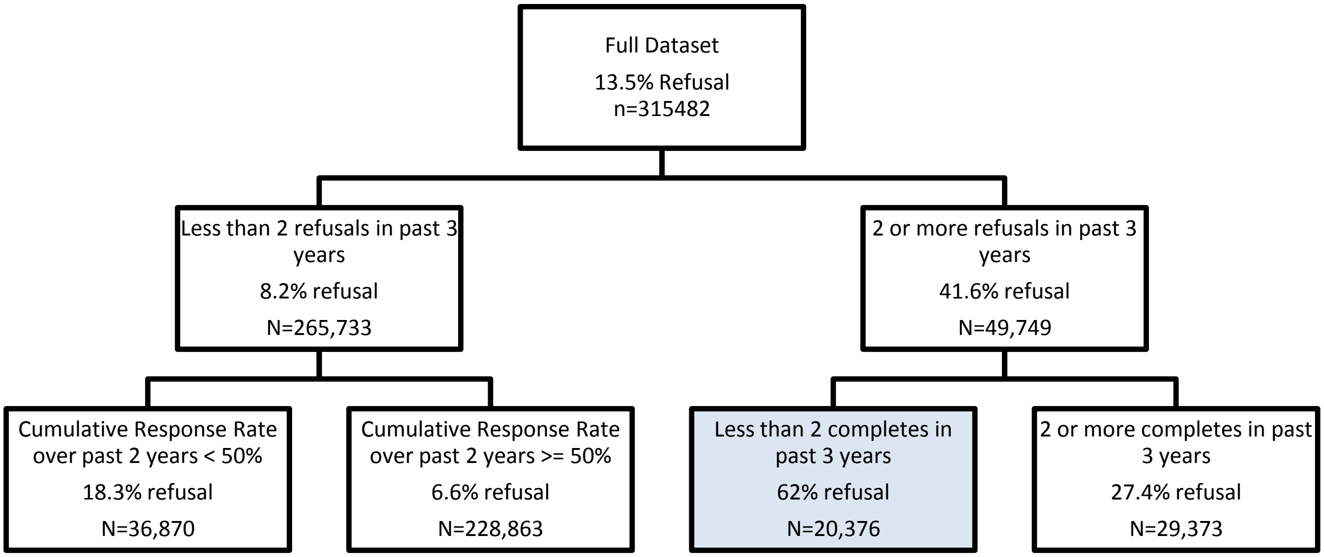
*all variables were significantly different (p < .01)

[2] Rural Urban Continuum code 8 = Completely rural or less than 2,500 urban population, adjacent to a metro area
Rural Urban Continuum code 9 = Completely rural or less than 2,500 urban population, not adjacent to a metro area

### 3.2 Classification Tree Models

More useful than simple univariate comparisons are the classification trees constructed to identify non-respondents  (Again, for clarity, only the training dataset results are shown.  Results were similar for the subsample used for the verification and testing datasets.).  The tree for refusals using the combined dataset is shown below:

**Refusal Classification Tree**



Full Dataset
13.5% Refusal
n=315482
Less than 2 refusals in past 3 years
8.2% refusal
N=265,733
2 or more refusals in past 3 years
41.6% refusal
N=49,749
Cumulative Response Rate over past 2 years < 50%
18.3% refusal
N=36,870
Cumulative Response Rate over past 2 years >= 50%
6.6% refusal
N=228,863
Less than 2 completes in past 3 years
62% refusal
N=20,376
2 or more completes in past 3 years
27.4% refusal
N=29,373

In this model, the first split uses number of refusals over the prior three years, with those operations having two or more over three times as likely to be refusals as the entire dataset. Continuing down this branch, within that group, of those operations who also had fewer than two completes in the past three years, 62 percent 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 those with few complete reports in other surveys.  The exact splits defining this group can be used to identify this group.

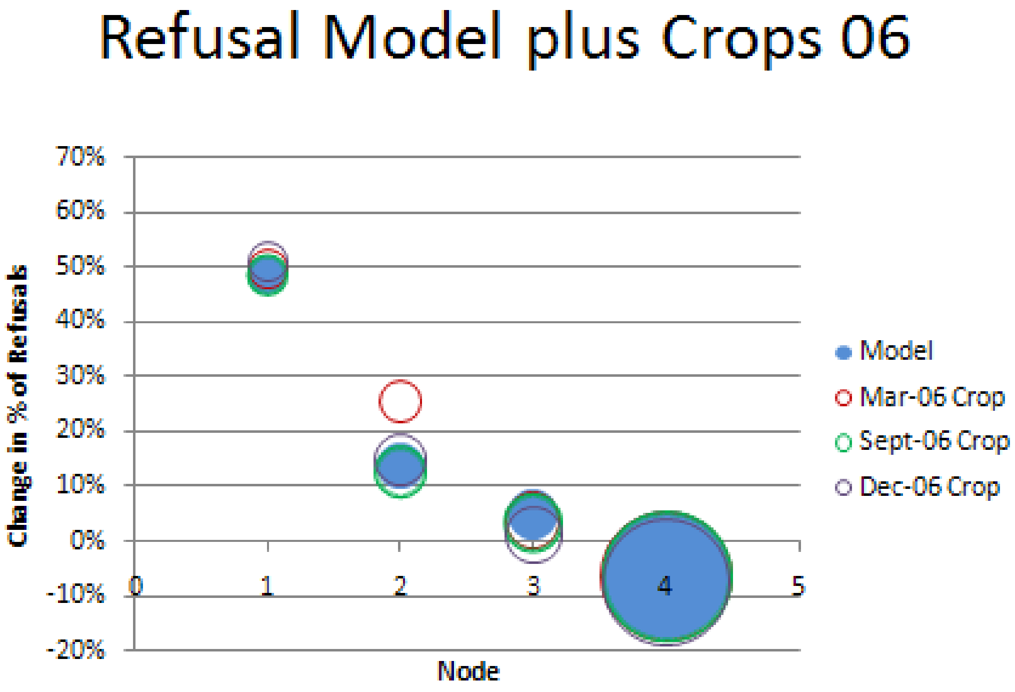
One way to evaluate the usefulness of this tree is to consider the gain in the percent of the target in the group with the most non-respondents, along with the size of that group.  In the case of the refusal model, we have increased the percent of refusals from 13.5 percent in the overall sample to 62 percent.  However, the number of records in this group is 20,376.  Therefore, this node has correctly identified only 12,633 refusals (or approximately 30 percent of all refusals). The remaining nodes can be rank ordered by the percent of non-respondents identified in each. This is shown in the first column of Table 3.

We applied this model individually to each of the quarters used to construct the model, and the results showed that it performed similarly for each one. Individual quarters for 2006 are shown in the additional columns in Table 1. Each of the quarters showed similar levels of non-respondents in the high non-response node (node 1 in the table) and included a similar proportion of the dataset.

**Table 3. Records included in each node**

| | **Model Tree** | | **Mar 2006 Crop** | | **Sept 2006 Crop** | | **Dec 2006 Crop** | |
|---|---|---|---|---|---|---|---|---|
| **Node** | **Node N** | **Percent Refusals** | **Node N** | **Percent Refusals** | **Node N** | **Percent Refusals** | **Node N** | **Percent Refusals** |
| 1 | 20376 | 62.0 | 5456 | 63.2 | 3926 | 60.9 | 5137 | 66.2 |
| 2 | 29373 | 27.4 | 6337 | 25.5 | 6491 | 24.9 | 9218 | 30.1 |
| 3 | 36870 | 18.3 | 10494 | 17.4 | 8443 | 15.4 | 10989 | 16.6 |
| 4 | 228863 | 6.6 | 59139 | 7.1 | 44421 | 5.6 | 55614 | 7.9 |
| Total Sample<br>(training data only) | 315482 | **13.5** | 81426 | **13.6** | 63281 | **12.3** | 80958 | **15.3** |

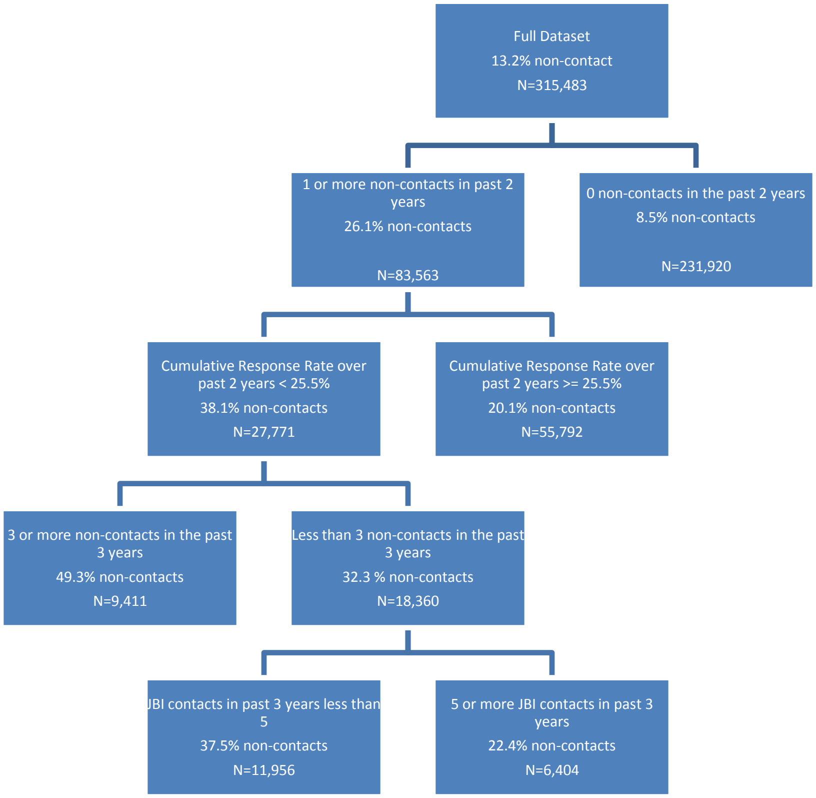
This is also shown graphically in the figure below. In this chart, each node is plotted depicting the change in the percent of refusals in the node from the refusal rate in the overall dataset. The size of the circle depicts the relative number of records in that node. Thus, you can see from this chart that the model consistently identifies a group of records with a substantially higher percentage of refusals (48.5 percentage points above the overall sample), but this group is relatively small. Node 4, the largest node, contains the most records and the percent of refusals in this group is 6.9 percentage points less than the overall sample.



Causes of non-contacts are likely different from survey refusals, so a separate model was built to predict survey non-contacts (Again, for simplicity, only the training data are shown.).

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

## Non-contact Classification Tree

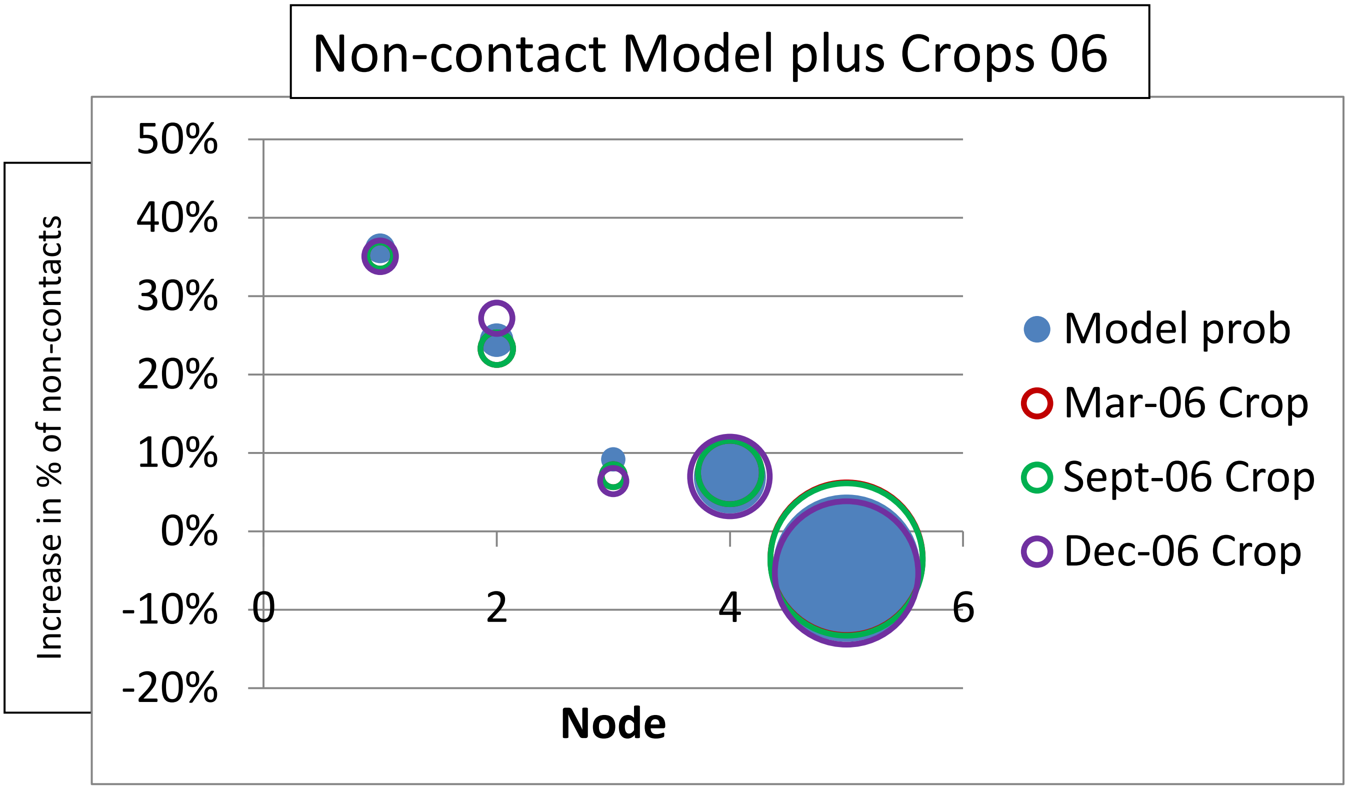

Full Dataset
13.2% non-contact
N=315,483
1 or more non-contacts in past 2 years
26.1% non-contacts
N=83,563
0 non-contacts in the past 2 years
8.5% non-contacts
N=231,920
Cumulative Response Rate over past 2 years < 25.5%
38.1% non-contacts
N=27,771
Cumulative Response Rate over past 2 years >= 25.5%
20.1% non-contacts
N=55,792
3 or more non-contacts in the past 3 years
49.3% non-contacts
N=9,411
Less than 3 non-contacts in the past 3 years
32.3 % non-contacts
N=18,360
JBI contacts in past 3 years less than 5
37.5% non-contacts
N=11,956
5 or more JBI contacts in past 3 years
22.4% non-contacts
N=6,404

In this tree, the leaf of interest is:

- those operations who have had one or more non-contacts in a NASS survey in the past two years,
- had a less than 25.5 percent response rate over the last two years, and
- three or more non-contacts in the past three years,

Forty-nine point three percent of the operations in this group were non-contacts in the target survey, compared to 13.2 percent in the original dataset. This model did not perform as well as the refusal model, producing a more modest gain in the percent of non-contacts in node 1. Again, no descriptive variables from the census of agriculture, list frame or external variables were useful in the model, with the exception of the final split on the contact JBI. 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 that had not provided data often in the past. The actual tree splits indicate the optimum threshold for splitting on these variables.

This model also showed consistent performance on the individual quarters used to construct the model, indicating that the predictions of non-contacts are similar throughout the survey year.



An additional model was built for those records which were non-respondents because they were held out of data collection by the field office. This tree is shown in Appendix B. This model is not particularly insightful, as it merely shows that cases held in the office tend to be those which have been held in the office before, and the number of cases held varies by state. Since the decision to hold the case is determined by the field office and NOT by the respondent, this tree has little utility for data collection management.

### 3.3 Building Models without Using Response History

The models built using all of the variables available to us allow us to identify the groups most likely to be non-respondents. However, if we were to use these models to alter data collection strategies and convert these non-respondents into good respondents, ultimately we would change the nature of the relationship between an operation's response history and its likelihood of responding. Thus, it is unclear whether or not these models would remain useful in the future. Therefore, we built new tree models with our dataset after excluding all of the response history variables. This model is shown in Appendix C. Unfortunately, this model is quite weak, producing only marginal gains over using no model at all. While the model for refusals increased the percent of refusals from 13.4 percent to 55 percent, it included less than 4 percent of the total dataset, identifying only 575 out of the 36,236 refusals in the complete dataset. In addition, as shown in the model, it also includes only a subset of states, thus is not useful for all field offices. This model demonstrates that our dataset does not contain good predictors of non-response other than the response history variables.

### 3.4 Extending the Models to Other NASS Surveys

While the Crops/Stocks models identify the groups most likely to be non-respondents in the Crops/Stocks survey, it was unclear whether these models would work on other surveys. In order to evaluate this, we applied the Crops/Stocks refusal model and the non-contact model to the quarters of the Crops/Stocks survey that were not in the model's dataset (2008), the Cattle survey, the Agricultural Labor survey, and the Agricultural Resource Management Survey (ARMS).

As expected, the refusal model works quite well on additional quarters of the Crops/Stocks survey, with the relative gains in percent of refusals identified similar to the original model and the sizes of the groups identified also similar. The models also appear to work well for the Cattle survey, with the July Cattle survey performing a bit better than January. Charts depicting the refusal models relative to the original model are shown in Appendix D, and those for the non-contact models are shown in Appendix E.

Results from the Agricultural Labor survey are a bit different and are likely influenced by the survey's sample year, which begins in July. Eleven replicates are marked in the survey sample, and only six of these are used in any given quarter. Each of the eleven replicates is included in at least two of the quarters of the year (See the Survey Administration Manual for the Agricultural Labor Survey for more details.). Therefore, beginning in October and continuing through April, at least part of the replicates for each quarter have been included in the sample for a previous quarter. Thus, the operations included in the Agricultural Labor survey sample may not have enough response history in the first quarters of the sample year for the models to be effective. As the survey year progresses, there is at least a response history for the current year's survey available for some of the sample and the models improve over that time. In the case of the non-contact model, it appears to identify higher levels of non-response for both nodes 1 and 2.

In contrast to these surveys, results show that these models do not predict survey non-response for the ARMS very well. The gain in identifying refusals is lower than the model, and importantly, the size of the group included in this node is very small. This is likely the result of the concerted efforts made to minimize the overlap between the ARMS samples and other surveys (see the ARMS Survey Administration Manual for more details on the Sequential Interval Poisson sampling used). In this case, the amount of recent response history for the ARMS sample is expected to be much less than for other surveys. In addition, unlike the other surveys examined, the ARMS is conducted primarily by in-person interviews, which independently would be expected to influence survey response.

Efforts are underway to use classification trees to identify ARMS nonrespondents (Earp and McCarthy, 2009, 2010). A more expansive list of variables from the census of agriculture and

ARMS survey outcomes are being used to predict ARMS survey non-response, as compared to the set used for the Crops/Stocks survey. However, it appears that these models, while useful, still identify a much smaller set of nonrespondents than was identified for the Crops/Stocks survey.

## 4. CONCLUSION

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. While many of these variables are significantly different for respondents and non-respondents, they are not practically useful in identifying likely non-respondents. Even variables which we hypothesized might indicate greater interest in USDA surveys, such as the receipt of payments from USDA, 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 prove useful in predicting non-respondents. The only variables included in our models for refusals and non-contacts were the response history variables that described how cooperative sample units had been in the past.

The approach we have used in modeling non-response has several advantages over other methods such as logistic regression. For example, classification trees can include many variables, including those that are correlated. Classification trees also identify the variables most related to your target, but importantly, identify the optimum break point for those variables. The branches of the tree are the result of variable interactions, which do not have to be pre-specified as they would in a regression model. In addition, classification trees consider missing data as valid, which may be particularly important for modeling non-response. One of the most useful features of the classification tree is that each node is clearly described by the rules used to create it. This makes the group easily interpretable.

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). 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 respondents and non-respondents is good news. While we weren't able to easily characterize our non-respondents, this also suggests that the non-response is not introducing bias into the Crops/Stocks survey estimates (upon which the models were based). Non-response is not concentrated in certain sizes or types of operations. The classification tree models built without our response history variables are quite weak and do not provide any increase in non-response classification accuracy. Classification tree approaches have been used elsewhere to create non-response weighting groups, but it does not appear that classification trees using the variables in our dataset will provide any substantial benefits in this area.

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 can be 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 face to face enumeration, etc.

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. The next step in this project will be to test methods to incorporate these models into ongoing data collection planning and operations.

## 5. RECOMMENDATIONS

1. Use these tree models to score future Crops/Stocks and Cattle survey samples and make the scores available for use by the field offices. *This recommendation is in progress, with plans in place to score the September and December 2010 Crops/Stocks samples and the January 2011 Cattle survey sample.*
2. Change data collection procedures for those records identified as most likely to be refusals or non-contacts.
3. Evaluate whether the scores were useful in altering data collection procedures by comparing predicted non-response rates for those records with actual non-response rates for the highest non-response groups.

## 6. 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.

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.

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

Earp, M. and J. McCarthy. (2009). Using Respondent Prediction Models to Improve Efficiency of Incentive Allocation. In *JSM Proceedings*. Fort Lauderdale, FL: American Association of Public Opinion Research.

Earp, M. and J. McCarthy. (2010). 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. In *JSM Proceedings*. Fort Lauderdale, FL: American Association of Public Opinion

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.

Lepkowski, J.M. and Couper, M.P. (2002). Nonresponse in the Second Wave of Longitudinal Household Surveys. In Survey Nonresponse, 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 Nonresponse: If We Bother Them More, Are They Less Cooperative? Journal of Official Statistics, 22(1), 97-112.

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

Appendix A. Variables included in the model

| VARIABLE | DESCRIPTION |
|---|---|
| *Target Survey Variables:* | |
| Response | Response outcome (complete, refusal, inaccessible, other) |
| *Census of Agriculture variables:* | |
| Bees | Bee indicator: Yes/No |
| Cattle | Cattle indicator |
| CRP | Conservation Reserve Program indicator (based on acres) |
| Government Payments | Operation received 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 |
| Horses | Horse indicator |
| K46 | Total Acres Operated, computed based on ownership |
| 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 |
| Census response | Census response outcome (complete, refusal, non-contact) |

*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 farmtype |
| NAICS | North American Industrial Classification code |
| 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 1993 |
| 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 |

*Zipcode Level Variables*

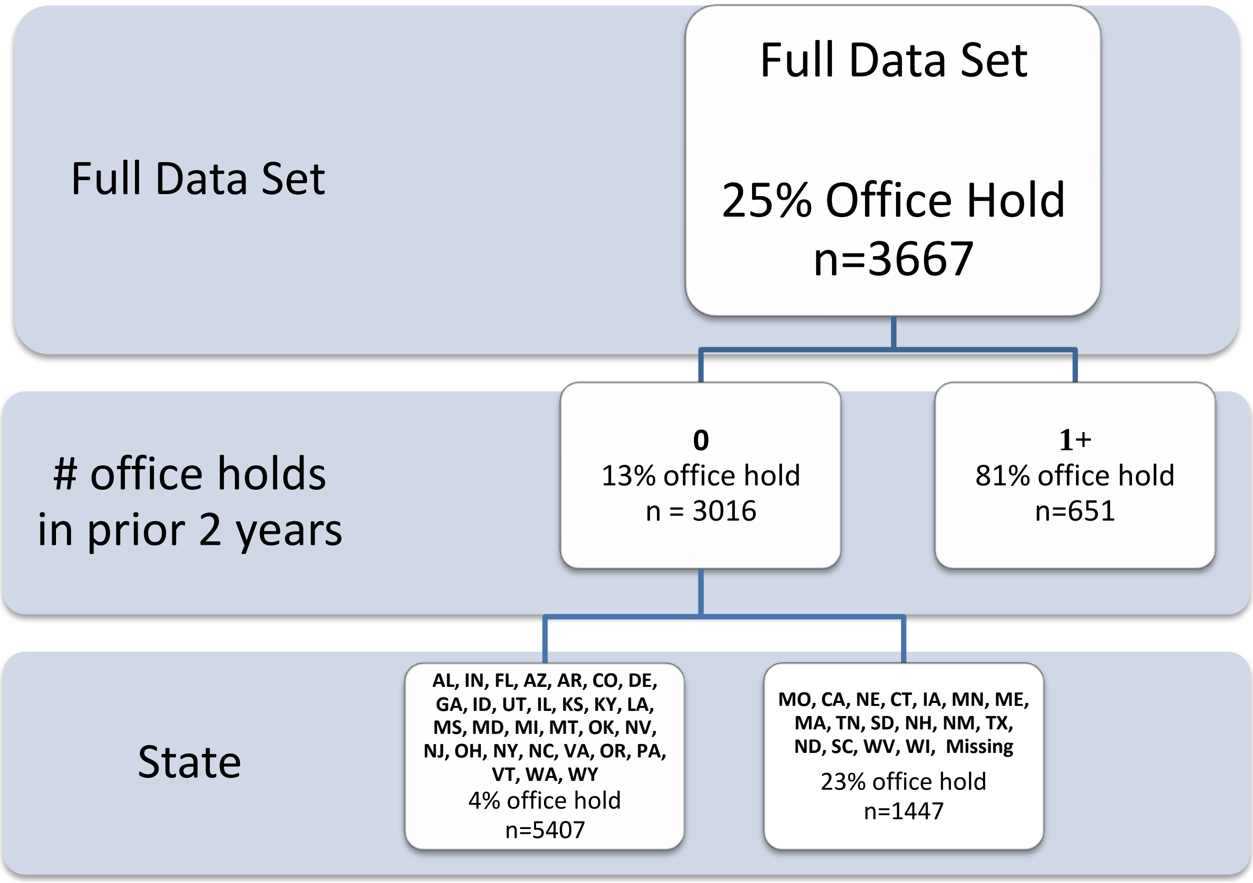
| | |
|---|---|
| %_poverty | % of population with income below the poverty level in 1999 |
| %_foreignborn | % of population foreign born |
| %_foreign lang | % of population over age 5 who speak a foreign language |
| %_highschool | % of population over 25 who graduated from high school |
| Total pop | Total population in the zipcode in 2000 |

*Response History Variables*

| | |
|---|---|
| I51_2005 | Joint Burden Index (JBI) projected number of surveys for 2005 |
| I51_2006 | JBI projected number of surveys for 2006 |
| I51_2007 | JBI projected number of surveys for 2007 |
| I52_2005 | JBI projected number of contacts for 2005 |
| I52_2006 | JBI projected number of contacts for 2006 |
| I52_2007 | JBI projected number of contacts for 2007 |
| I53_2005 | JBI projected number of OMB minutes for 2005 |
| I53_2006 | JBI projected number of OMB minutes for 2006 |
| I53_2007 | JBI projected number of OMB minutes for 2007 |

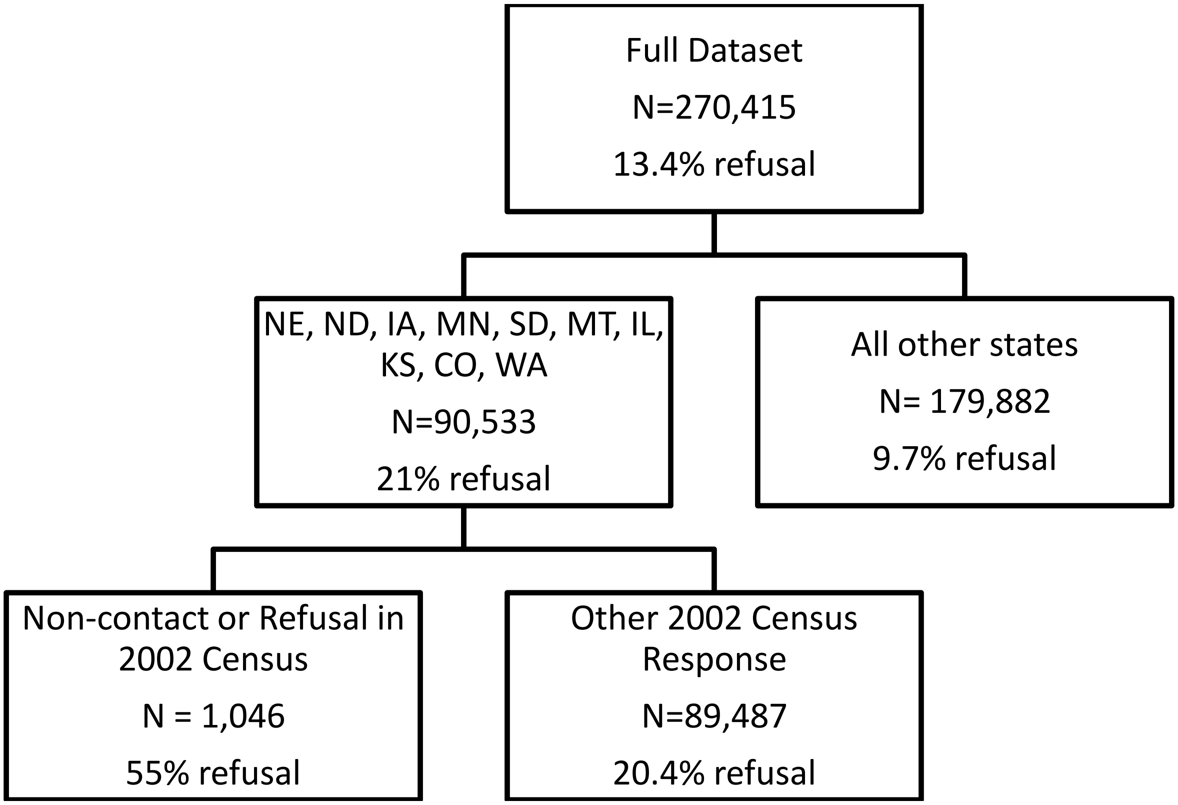
| | |
|---|---|
| Knownzero_2005 | Number of surveys for which operation was sampled but classified as out of scope in 2005 |
| Knownzero_2006 | … in 2006 |
| Knownzero_2007 | … in 2007 |
| Complete_2005 | number of contacts operation responded in 2005 |
| Complete_2006 | … in 2006 |
| Complete_2007 | … in 2007 |
| Inaccessible_2005 | number of contacts operation was non-contact in 2005 |
| Inaccessible_2006 | … in 2006 |
| Inaccessible_2007 | … in 2007 |
| Officehold_2005 | number of contacts operation was held in the office (no contact attempt made) in 2005 |
| Officehold_2006 | … in 2006 |
| Officehold_2007 | … in 2007 |
| Refusal_2005 | number of contacts operation refused in 2005 |
| Refusal_2006 | … in 2006 |
| Refusal_2007 | … in 2007 |
| 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 |

Appendix B.  Classification Tree for Office Hold Cases


Full Data Set
Full Data Set
25% Office Hold
n=3667
# office holds
in prior 2 years
0
13% office hold
n = 3016
1+
81% office hold
n=651
State
AL, IN, FL, AZ, AR, CO, DE, GA, ID, UT, IL, KS, KY, LA, MS, MD, MI, MT, OK, NV, NJ, OH, NY, NC, VA, OR, PA, VT, WA, WY
4% office hold
n=5407
MO, CA, NE, CT, IA, MN, ME, MA, TN, SD, NH, NM, TX, ND, SC, WV, WI,  Missing
23% office hold
n=1447

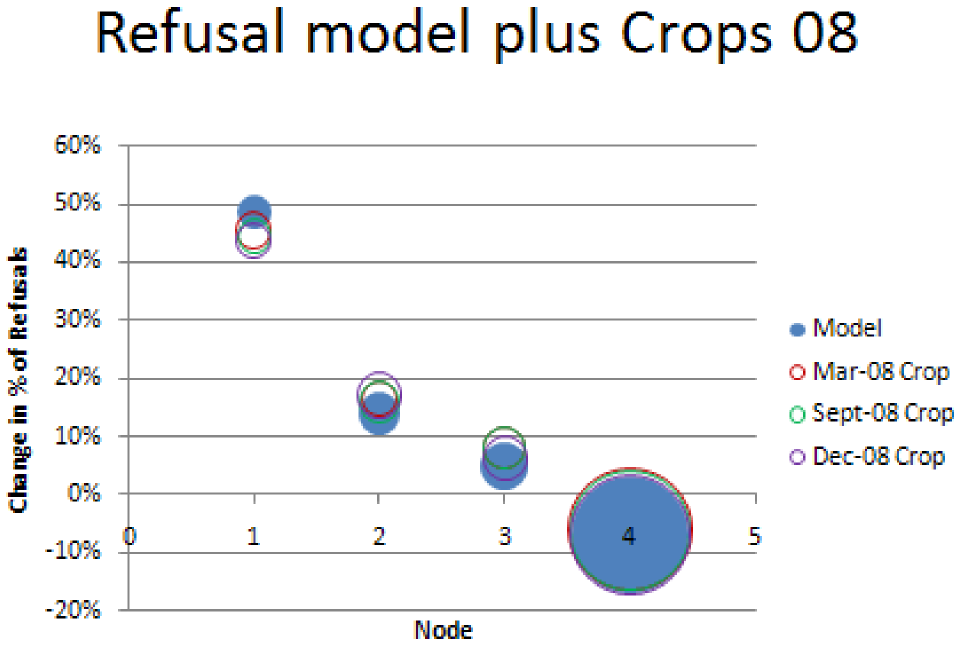
Because the number of office hold cases was small, for this tree the data was sampled to increase the percent of office hold cases in the sample to 25%.  The biggest predictor of whether a case was held in the office was whether or not it had been held in the office before for another survey in the prior two 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.  The factors impacting the decision to hold cases in the office is 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).

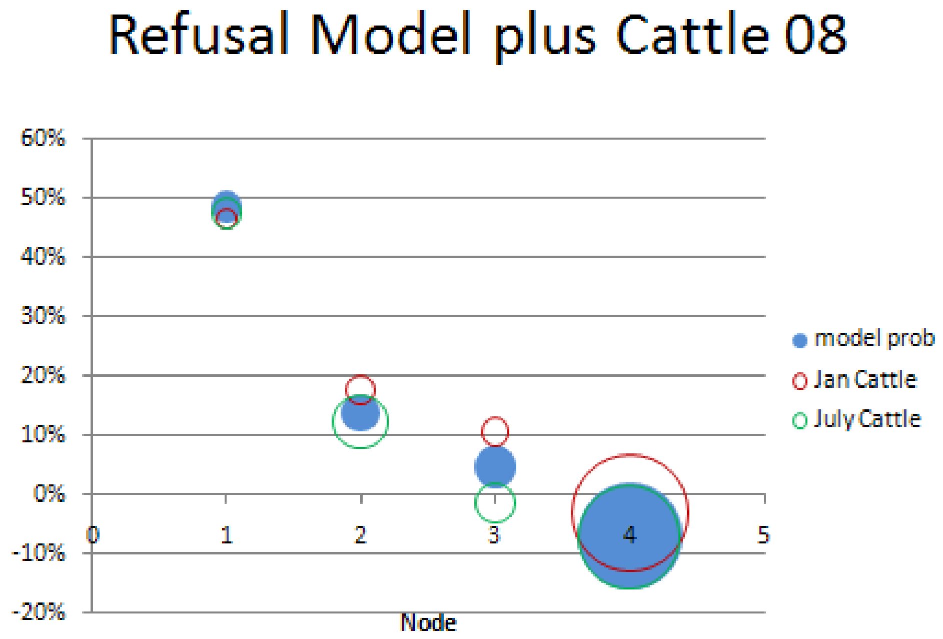
Appendix C. Classification Tree for Refusals, All Quarters Crops/Stocks with Response History Variables Excluded.



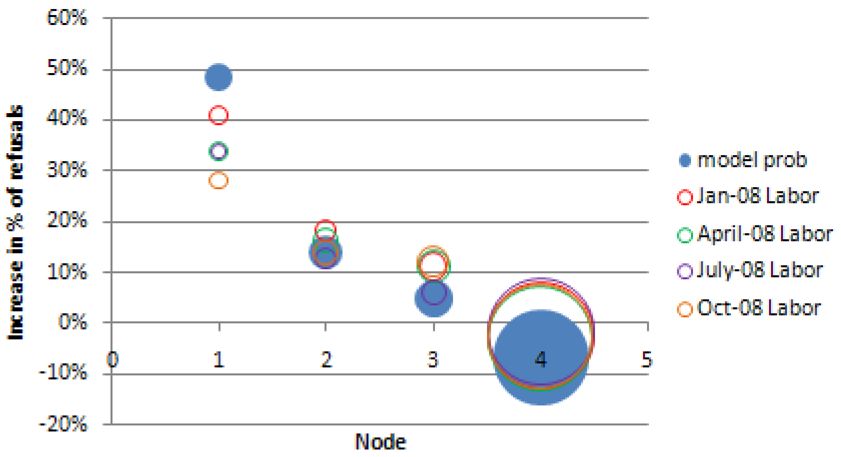
As is clear in the diagram, the model is only useful for a handful of states, and ultimately only identifies 1,046 operations from the initial set of nearly 270,415 as most likely to be refusals. Even within those operations, only slightly more than half (55%) are refusals. Thus, the model includes less than 1 percent of all refusals in the dataset in the node of likeliest refusals.

Appendix D.  Crops/Stocks Refusal Models Applied to Other Surveys (model is shown blue )

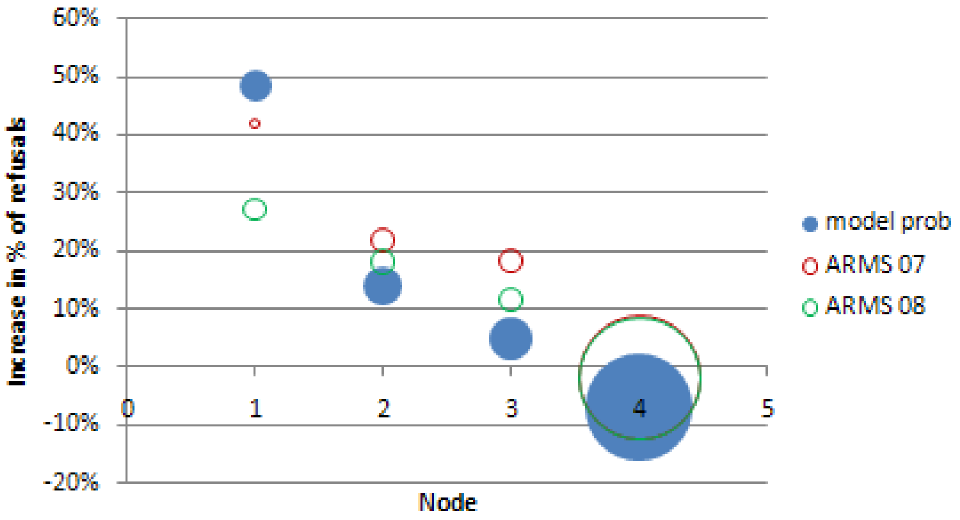




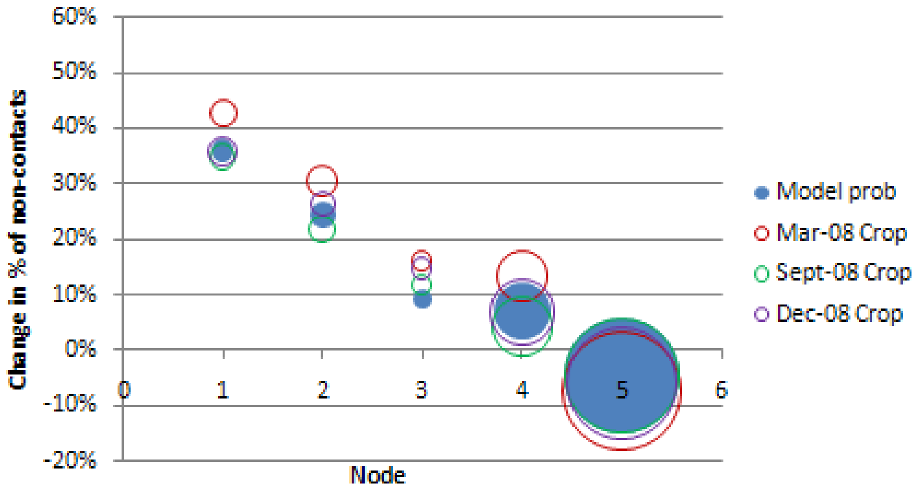
## Refusal Model plus Labor 08

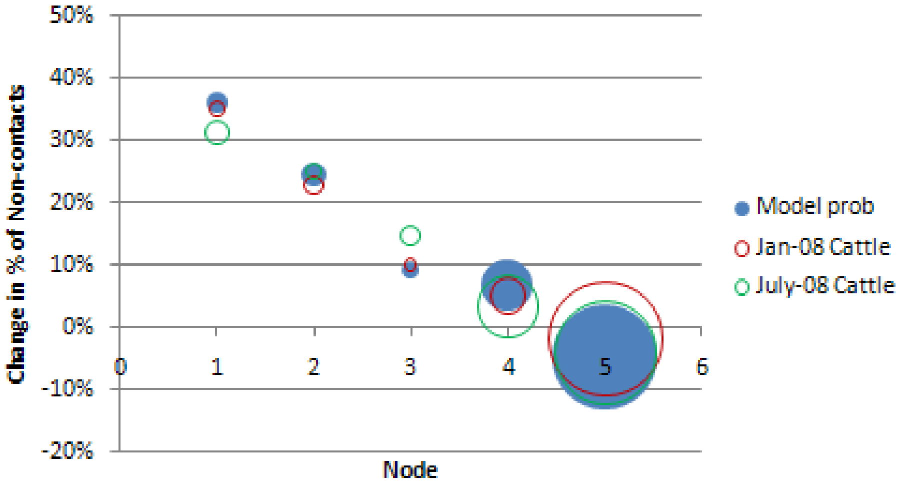


## Refusal Model plus ARMS

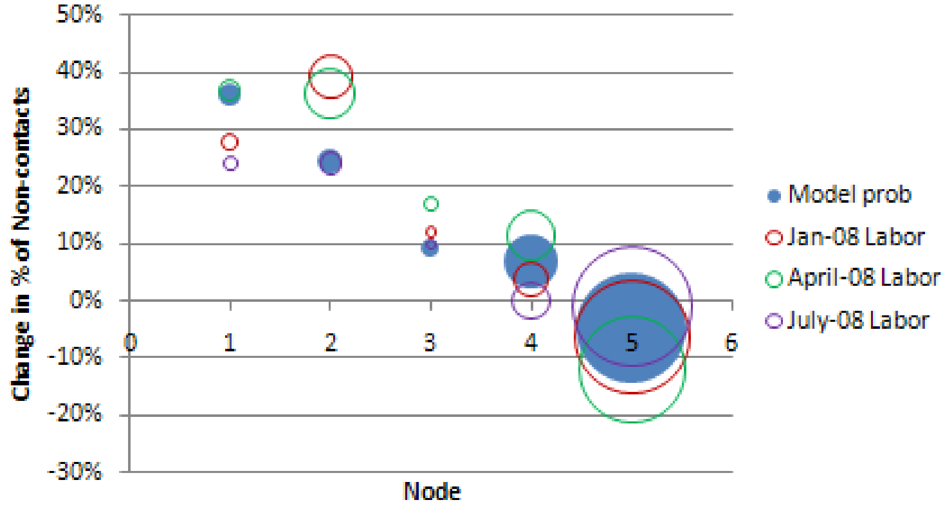


Appendix E.  Crops/Stocks Non-contact Models Applied to Other Surveys

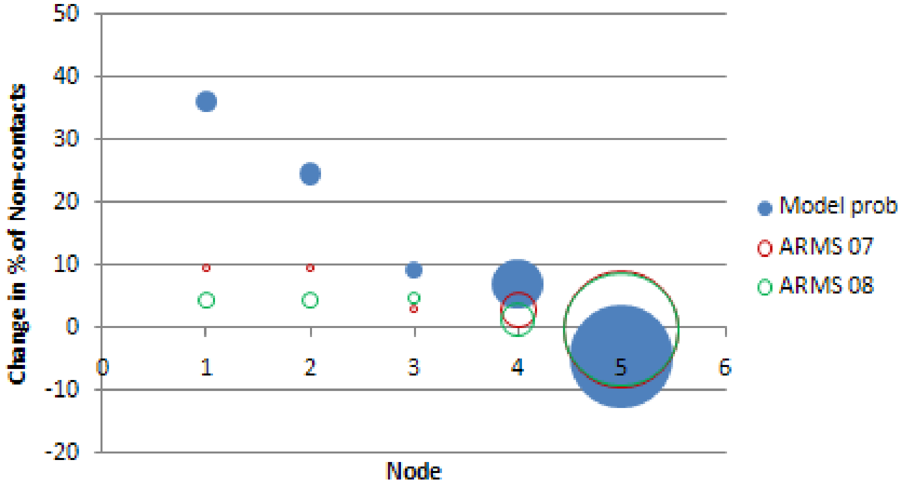

Non-contact Model plus Crops 08


Non-contact Model plus Cattle 08

## Non-contact Model plus Labor 08



## Non-contact Model plus ARMS 08


Change in % of Non-contacts
Node
● Model prob
○ ARMS 07
○ ARMS 08