2007 Census of Agriculture Non-Response Methodology

Will Cecere
National Agricultural Statistics Service Research and Development Division, U.S.
Department of Agriculture, 3251 Old Lee Highway, Fairfax, VA 22030

Abstract

Every five years the USDA's National Agricultural Statistics Service (NASS) conducts a Census of Agriculture for the entire US. NASS strives to achieve the most accurate results through diligent data collection and use of the best methodology available.

For the 2007 Census of Agriculture, the non-response adjustment methodology was changed to incorporate the use of classification trees for each state in the US using a set of input variables describing several factors including size and type of farm as well as demographics of the operator. The trees split records into different groups based on which variables defined the largest difference with respect to response rate. End groups, or leaves were adjusted for non-response based on the response rate of their respective leaf. This paper will present details and results of using this methodology on the 2007 Census of Agriculture.

Key words: census of agriculture, non-response adjustment, classification trees

1. Introduction

The United States Department of Agriculture's National Agricultural Statistics Service (NASS) carries out a Census of Agriculture every five years. The most recent 2007 Census of Agriculture marked many improvements in methodology from the previous 2002 Census of Agriculture. Among them was a change in the way agricultural operations were weighted to account for non-response.

Census non-response methodology for the 2002 Census of Agriculture consisted of grouping records at the county level into one of five groups. Four of the groups were defined based upon the record's expected sales level and the fifth group was defined by whether that record had responded to any NASS surveys since 1997. If there weren't at least two records in a group for any given county, then that group was collapsed into a larger group.

Problems arose when there was a large amount of collapsing or if there were no respondents in a group. Following the 2002 Census, it was brought into question as to which covariates should be used in determining non-response weighting groups. It was not known definitively at the time which variables best predicted whether a record was a respondent or not. Also unknown was what levels of variables' values would make appropriate poststratification groupings.

1.1 Pre-Census Research

For research leading up to the 2007 Census of Agriculture, questions arose as to what variables besides expected sales should be used in grouping. A data mining approach was examined using classification trees to group records based on their likelihood of response. This is a distinct departure from the 2002 non-response methodology in that it is a data driven approach, rather than one in which the weighting cell variables and splits were predetermined. Instead of making the Census records fit into predefined groupings, the groupings are made to fit the records.

This new method was examined using Texas, California, Nebraska, Oregon, Pennsylvania, and Georgia as test states. The classification tree method was compared to the previous non-response weighting method on a number of different commodities as well as a variety of demographics. Counts by farm size were examined to note any bias associated with the expected sales class.

Research comparing methodologies on the 2002 Census of Agriculture data showed that the data mining method did not change previous commodity estimates or introduce any noticeable bias. However, it did allow NASS the opportunity to use many input variables to group records for non-response weighting. Therefore, the new methodology using classification trees was chosen to be implemented for the 2007 Census of Agriculture. In order to most effectively use the data mining methodology and while simplifying the logistics of the weighting process, the decision was made to form the non-response groupings at the state level as opposed to the county level for the 2007 Census of Agriculture.

2. Methods

Classification tree methodology is a data mining technique that is relatively recent, starting in the early sixties. A well known technique for classification trees called Chi-squared Automated Interaction Detection (CHAID) was developed by Kass (1980). This uses Chi-squared tests to partition a population into homogeneous subpopulations by the use of repeated binary splits.

In the case of the 2007 Census of Agriculture, we wish to partition records by probability of response into unique poststratification groups. These groups are referred to as leaves by the software package used, SAS Enterprise Miner 5.2. This software package has a diverse array of options when performing data mining techniques such as classification trees. Enterprise Miner is made to handle large datasets such as the 3.2 million records from the census with relative ease. It allows for many levels of control, such as which partitioning algorithm can be used, the size of the leaves, and the significance of the Chi-squared test among others.

Some advantages to using classification trees for non-response adjustment are that we can select the best combination of variable levels on which to split records based on response rate. This process allows for the use of more auxiliary variables and it follows that with additional auxiliary information, weighting models can be improved (Schouten 2003) through better explanation of response behavior. Also, the leaves are displayed in plain view from the tree, making the groups easily visible.

2.1 Input variables

For the 2007 Census, classification trees were created at the state level for all states except Alaska and Rhode Island. Candidate variables to be used as covariates were pulled from NASS frame data, information that NASS keeps about all records on the Census Mail List (CML). The reason for the use of frame data for input variables, as opposed to current survey data, is to ensure that we use information common to both respondents and non-respondents.

Table 1 is a list of candidate variables used for the classification trees. Expected sales group is a variable that gives an indication of size through one of 17 ordinal categories. Several demographics are included such as binary variables for the indication of whether or not the primary operator is of Hispanic origin, American Indian race, and female gender. Type of operation covers things such as if the operation is a partnership, co-op, etc. The status of the operation is an indicator as to whether it is out of business or a duplicate among others. A geographic variable is also included which identifies the agriculture district within a state.

Table 1: Covariates or input variables used for classification trees

Expected sales group for the operation
Race identifier variable
Farm type
Binary indicator variable for American Indian primary operator
Type of operation
Status of operation
Gender of the primary operator
Hispanic origin
Agriculture district identifier
Indicator variable for presence of telephone number

2.2 Classification Tree Rules

In order to make any classification tree, rules must be set to guide the program in deciding when to split a group and when to stop.

To split groups we use the CHAID algorithm as described previously. This sorts through all possible combinations of each input variable until it finds a partition that produces the maximum logworth, where the logworth = $-log_{10}$(P-value), of the Chi-squared tests as long as it meets the significance criteria. Once the best partition is determined for each input variable, the partitions are compared across variables and assessed on maximum logworth once again. Variables with more input levels tend to have a larger maximum logworth resulting from more degrees of freedom.

The tree will continue making binary partitions until it meets a preset stopping criterion. A consequent stopping rule is when no splits meet the Chi-squared test significance criterion. In our application, to insure that group sizes didn't get small, a minimum of 100 records were required in each partition. This aided in limiting the non-response weight of any one group. The final stopping criterion was to limit the number of levels that a tree could split so the leaves could be interpreted in a reasonable fashion.

Once the tree was finished we had formed G mutually exclusive groups based on each record's likelihood of being a respondent. For each group g, we then calculated the non-response weight as

$$W_g = \frac{N_g}{R_g}$$

where N_g is the total number of records in group g and R_g is the total number of respondents in group g.

2.3 Oversampling

When modeling rare or somewhat rare events using classification trees, it is common to take a sample of the population, stratified by the dependent variable in order to allow the algorithm a better chance to detect differences. For instance, without any sampling, if NASS had a state with a 95% response rate the classification tree algorithm might say that the state does not need to be partitioned, resulting in a model with only one group that would predict 95% of the data accurately.

In order to make a better model we use a technique called oversampling, in which we take a sample of respondents to come closer to an even ratio of respondents to non-respondents. This allows the algorithm to more easily disseminate new groupings. In order not to lose too much data, oversampling was done to target a 60/40 ratio of respondents to non-respondents.

The oversampling rates were programmed to roughly follow the guidelines of Table 2.

Table 2: Oversampling rates

State response rate	Oversampling rate
<%70	No oversampling
%70-74.9	0.6
%75-79.9	0.48
%80-84.9	0.40
%85-89.5	0.27
>%90	0.21

2.4 Follow-on Survey

One assumption made for Census non-response weighting is that non-respondents have the same in-scope rate as respondents. NASS uses the term in-scope to refer to operations qualifying as farms. Not all of the records on the CML are in-scope.

One problem category that we ran into with the 2007 Census was with operations that didn't respond to our pre-census screening questionnaire. These were referred to as screener non-respondents. Little was known about these screener non-respondents and their associated records were considered by many to be junk records for the most part. However, they accounted for over 400,000 of the 3.2 million records on the CML. These operations were identified early on as the group with the lowest response, with response rates typically lower than 40% depending on the state.

From the early results and due to the nature of the screener non-respondents, the question naturally arose as to whether or not our assumption of equal in-scope rate was valid for these records. Largely as a result of this concern, NASS performed a non-response follow-on survey by enumeration to determine the in-scope rate of screener non-respondents who also didn't respond to the Census for each of the 48 states involved in the non-response weighting. The survey was completed in one month between the preliminary and final Census weighting runs.

As suspected, NASS found from the follow-on survey that the in-scope rate for screener non-respondents was lower for Census non-respondents than for respondents. This indicated that the previous estimate of the number of screener non-respondents was inflated. To account for this in the final run of the Census non-response weighting, a composite estimator (smoothing) was used in order to correct for the different in-scope rate for screener non-respondents.

2.5 Categorization of records

After the results of the follow-on survey, there were three categories of records for the final run of Census non-response weighting.

Category 1:	Records to receive a non-response weight of one
Category 2:	Screener non-respondents
Category 3:	Records weighted based on their state's classification tree grouping

The records falling into category 1 were excluded from any non-response weighting adjustment for various reasons. Prior to the census forms being received, there were records that were flagged for having values of sales and farm acres falling above state defined thresholds. A second type of excluded records were those operations that exceed commodity thresholds based on their information from the Census form. Other types of exclusions included records newly added after the Census forms were mailed and imputed records.

Category 2 records account for one group per state based on each state's smoothed estimator. The weighting methods for category 2 and 3 records were explained previously.

3. Results

The final performance of the census non-response weighting showed a diverse use of auxiliary variables for each state. Many of the states in similar regions showed similar types of groupings. The smallest amount of non-response groupings or leafs for records in classification tree was three, exhibited by several New England states. The largest number of groupings was shown by Texas, with 24 poststratification groups. Group sizes varied depending on how many records a state had as well as the amount of variation of response rate across the auxiliary variables.

Figure 1 shows the classification tree for Nebraska. It should be noted that the total number of records listed in the parent node as 16,395 is the number of records after oversampling. In this example we have reached our target of a 60/40 ratio of respondents to non-respondents. With a total of 4 weighting groups, this is a small number of groupings relative to the state's size. This may be attributed to low variance of response rate within the auxiliary variables.

In the Nebraska tree, we see that the first variable selected for partitioning was Farm Type. Since Farm Type is a nominal variable, the tree shows a list of Farm Type levels for each group. The group on the right of the first partition, with a 72% response rate has met its stopping criteria, making it a leaf or a weighting group. The opposing group is split again on Expected Sales, with left group being split again on Hispanic Origin. A Hispanic Origin value of 3 indicates no Hispanic Origin while the group labeled Others accounts for missing values as well as those with Hispanic Origin.

To illustrate how a group would be weighted, assume that the first leaf to end splitting, or the right most group in the Nebraska tree, is in fact the true non-response weighting group. This group would then receive a non-response weight of $w_g = N_g/R_g$ = 5201/(5201*0.72) = 1.39.

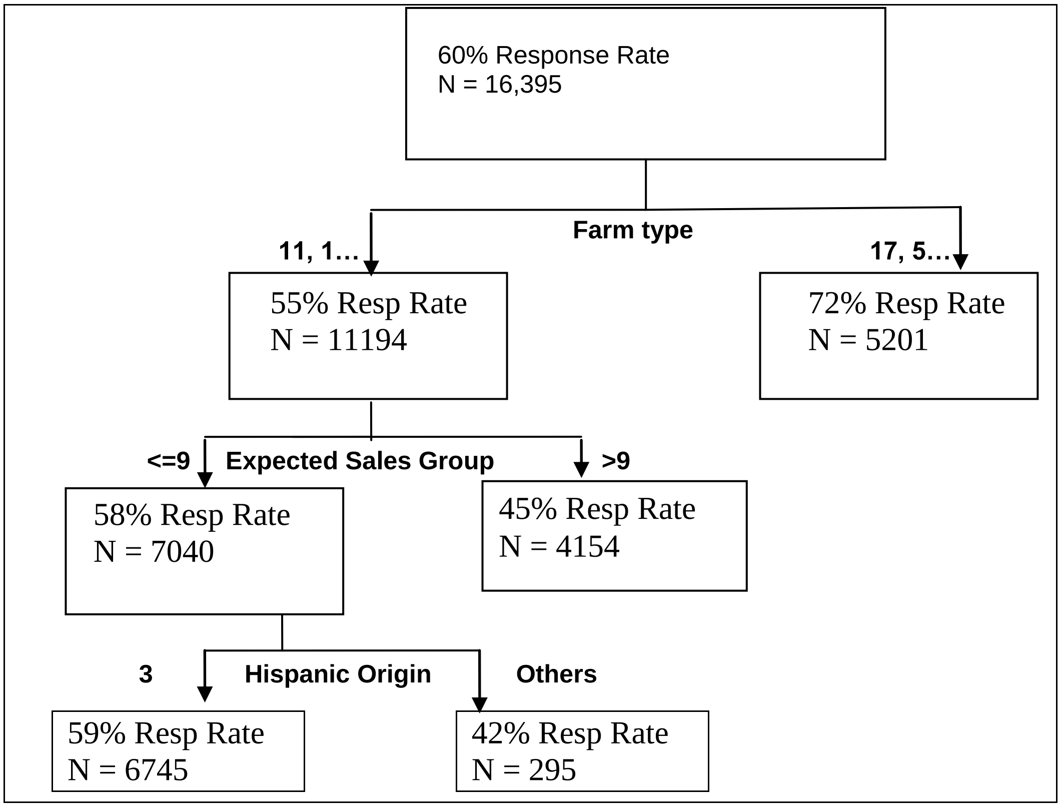

Figure 1: The classification tree for Nebraska

To show a picture of the first input variables chosen for partitioning, Figure 2 is a choropleth map of all US states. This was used to examine regional similarities among first split variables as well as which variables if any were predominant. Hispanic Origin and Active Status appear to be the most common variables for primary partitions. Some variables appear to be regional such as Farm Type and Race Code.

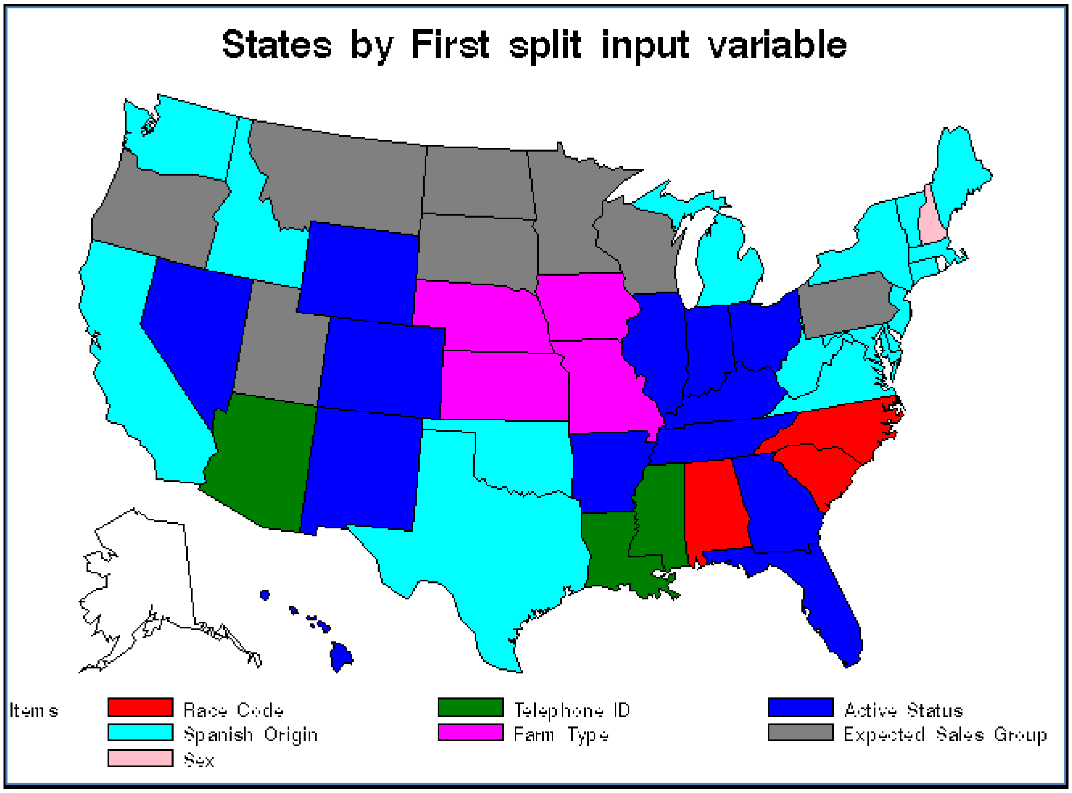

Figure 2: Choropleth map of first split variables by region

4. Discussion

Overall, the use of classification trees proved to be effective in forming useful non-response weighting groups in a timely fashion. The construction of trees was automated with the ability to produce well justified models using an unlimited number of auxiliary variables. This was especially valuable given the short time frame that is given to perform the non-response weighting.

One disadvantage of using classification trees in a short time frame is that there is little time to compare different tree models. With the construction of 48 classification trees, there is relatively no time to evaluate each individual model. However, this problem can be accounted for by checking for consistency with multiple preparation tests, which were performed prior to the final run of the census weighting.

Future census research can be done to examine a number of different options for non-response weighting. Since there is no penalty for additional auxiliary variables in classification tree modeling, new variables can be examined to determine their relevance in predicting non-response. Identifying specific groups with high non-response can aid in data collection for the next census. Another item for evaluation is the use of state level weighting groups versus county level weighting groups. A technique for using state level classification trees at the county level has been proposed and can be evaluated for future censuses.

References

Kass, G.V. (1980), An explanatory technique for investigating large quantities of categorical data, *Journal of the Royal Statistical Society C*, *Applied Statistics* 29, 119-127

Potts, W.J.E. (2006), Decision Tree Modeling, SAS Institute Inc. Course Notes

Schouten, J.G (2004), A selection strategy for weighting variables under a Not-Missing-at-Random assumption, paper submitted to journal.

Schouten, B., de Nooij, G. (2005), Nonresponse adjustment using classification trees, *Statistics Netherlands*