Evaluation of the Classification Tree Methodology Used for the Development of the 1987 Census of Agriculture Mail List

Richard L. Schmehl and Magdalena Ramos, Bureau of the Census
Washington, D.C. 20233

Keywords: Discriminant model applications, list frame development, farm/nonfarm.

Abstract. Classification tree methodology was applied to the preliminary 1987 Census of Agriculture mail list to develop a list of addresses by probability of operating a farm. Geographic area, address source list, and agricultural sales level characteristics from the 1982 mail list and mail-out to the 1987 mail list were used to develop the list. Groups were ranked in descending order according to their proportion of farms. The 1987 final census mail list was composed of the 4.1 million addresses in the highest ranking groups with 0.9 million records in the lowest ranking of these groups receiving the short census report form. A sample of addresses from the 200,000 records eliminated from the list were mailed an abbreviated survey questionnaire to determine whether the addressee operated a farm. The actual proportion of 1987 census mail list addresses operating farms within each group was determined and compared with the predicted proportion. This paper discusses the evaluation methodology and the results.

1. BACKGROUND

1.1 The Census of Agriculture Mail List Development Program

This report describes the methods used for evaluating the classification tree methodology application to the 1987 agriculture census mail list and presents results, conclusions and recommendations. The evaluation compares the results of the applications of classification tree methodology to observed responses from the 1987 Census of Agriculture and a survey of addresses excluded from census mail list operations. The methodologies are aimed at improving the performance of the classification tree methodology for the 1992 census mail list development operation. Greater detail on the design and implementation methodology is given in the paper titled "Classification Tree Methodology for Mail List Development" by Owens, Killion, Ramos, and Schmehl (1989).

The census of agriculture, taken every five years, collects data and publishes information on land in farms, operator characteristics, and agricultural production and sales in the United States. A census farm is defined as any place that, during the census year, sells or has the potential to sell $1,000 or more of agricultural products. The goal of the agriculture census is to request data from all U.S. farm operators. However, there is no comprehensive list of farm operators in the U.S. Therefore, one of the most difficult census tasks is to develop a list of addresses containing only U.S. farm operations meeting the census farm definition. The census of agriculture mail list begins as a compilation of many lists from several sources, including the previous census mail list, government or agriculture association lists, and United States Department of Agriculture lists. These lists are merged, and records from the lists are linked to eliminate duplicate operations.

Recent census operations use a pre-census Farm and Ranch Identification Survey to identify duplicate and nonfarm records. Such a farm and ranch identification survey was not approved for the 1987 census because of budget and respondent burden constraints. Hence, a method was needed to remove probable nonfarm addresses from the mail list to a predefined 4.1 million address limit. The Agriculture Division Staff selected the classification tree methodology, a type of discriminant analysis, to separate addresses belonging to the preliminary 1987 Census of Agriculture mail list into two basic categories: 1) probable farm operations and 2) probable nonfarm operations.

1.2 Application of Classification Tree Methodology to the Census Mail List Development

The classification tree methodology assigned approximately 4.1 million addresses to model groups with expected farm proportions greater than 0.1170. These addresses were designated to receive a 1987 Census of Agriculture questionnaire. Addresses belonging to model groups with expected farm proportions less than or equal to 0.1170 were designated to be removed from the mail list. The expected farm proportion value 0.1170 will be referred to as the mail list "boundary" proportion.

Census questionnaires were sent to 4,090,451 addresses belonging to model groups with expected farm proportions greater than 0.1170. Approximately 900,000 of these addresses were sent a short version of the census questionnaire. The primary intent of the short questionnaire was to reduce respondent burden while at the same time collect enough information to allow imputation of the missing data. The short questionnaires were sent to addresses in model groups with small expected farm proportions; that is, those addresses least expected to be farms. These addresses belonged to model groups with farm proportions less than or equal to 0.1170 but greater than 0.1170.

Specific modifications were made to the classification tree methodology results based on subjective judgments by Agriculture Division Staff. After implementation of the classification tree methodology but before the questionnaires were mailed, certain addresses designated to be excluded from the mail list (according to the methodology) were retained on the mail list. Agriculture Division Staff believed that these records’ likelihood of being farms was too high to exclude them from the mail list. Conversely, other records were subjectively removed. Because of these decisions, approximately 137,000 records from model groups with expected farm proportions less than 0.1170 were included in the final 1987 mail list, and approximately 39,000 records from model groups with expected farm proportions greater than 0.1170 were excluded.

1.3 Description of the Evaluation Data

We created data files for the evaluation from information supplied from the following three sources: 1) results from the classification tree methodology described above, 2) observed responses from the 1987 Census of Agriculture, and 3) observed responses from the survey of addresses omitted from the final mail list.

From the first source, we recorded an expected farm proportion for each model group. From the second source, we recorded for each model group the observed number of farm and nonfarm respondents, and the total number of classified operations (farm and nonfarm). Note that an observed farm status is defined as the final classification (farm/nonfarm) assigned to a respondent based on whether their reported data meets the census farm definition. We defined the observed and expected farm frequencies for each model group as follows:

- An observed farm frequency (OFF) was defined as the number of 1987 observed farm respondents in a model group that were classified as farms.
- An expected farm frequency (EFF) was defined as the expected number of farm respondents in a model group. This was determined by multiplying the 1987 observed number of respondents (classified operations) in a model group by the group’s corresponding expected farm proportion (EFF).

This paper reports the general results of research by the U.S. Bureau of the Census. The views expressed are attributable to the authors and do not necessarily reflect that of the Census Bureau.
The 1987 census file from which the observed data was taken contained 4,090,451 records. Of these, 3,355,349 respondents were farm or nonfarm operations. The rest of the cases were nonrespondents or post master returns. For processing reasons, 9,171 respondents were classified as either farm or nonfarm and hence were excluded from the evaluation. Also excluded from some analyses were 110,615 records with EFPs < 0.1170. These are nonrespondents from the 127,000 records that were not selected for inclusion based on model group assignment but were added subjectively to the mail list by the Agriculture Division Staff. Since these 110,615 cases were only portions of model groups (those records subjectively expected to be farms), we were able to show that the model group observed farm proportion (OFP) for these records was not representative of the total model group farm proportion. These records were then effectively removed from the mail list after the application of the classification tree methodology to the mail list. Further note that stratum records were subjectively removed from the mail list by Agriculture Division staff even though the classification tree methodology applied to those records was not representative of the stratum. A total of 2,463 survey cases (49.5 percent) responded but only 2,475 (46.4 percent) of them could be classified as either farm or nonfarm. There were 2,528 nonrespondents and 189 post master returns.

2. EVALUATION OF THE CLASSIFICATION TREE METHODOLOGY USING MEASURES OF ASSOCIATION

2.1 Analysis Methodology

The data were partitioned by two criteria of classification: ranges of EFPs and observed farm/nonfarm status. Using this classification, the data were arranged into a 2x2 contingency table. Only mail list respondents were used to construct the table. The row classification criterion was the 1987 census observed farm status and the column classification criterion was based on EFP. The column criterion partitioned model groups with EFPs < 0.4322 (short farm boundary proportion) into class 1 and model groups with EFPs > 0.4322 into class 2.

We computed a Pearson chi-square statistic to test if the observed farm status (the row classification) was independent of the EFPs (the column classification). The odds ratio which measures the odds of an address being a farm given that it is a member of one of the two column classification criteria was also computed. We also applied a goodness-of-fit procedure to a 2x9 contingency table to test for departures of the OFPs from an expected proportion ordering. The classes were formed by partitioning records into nine classes based on EFP. The OFPs were compared to hypothesized farm proportions: the midpoints of the expected ranges for each class.

2.2 Measures of Association Results

Table 1 provides the class definitions and results of the 2x2 contingency table classification. Each cell of the contingency table lists four values: 1) frequency (number of respondents), 2) percent of the overall total, 3) percent of the row total, and 4) percent of the column total. For example, 1,210,376 census farm respondents belonged to model groups with an EFP > 0.4322. These farm respondents are 46.68 percent of the total classified mail list respondents (3,235,563), 84.10 percent of the row (observed farm respondents) total (1,795,846), and 65.21 percent of the column (class 2) total (2,316,341). The margins contain the row or column totals and percentages.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Records with any (or combinations of) IRS source, a 1982 Census nonfarm or 1982 Farm and Ranch Identification Survey nonfarm source, and expected TVP &lt; 0.250 and &lt; $7,000.</td>
</tr>
<tr>
<td>B</td>
<td>Records with any (or combinations of) 1982 Census nonrespondent or USDA nonfarm source, with TVP between $5,000 and $19,999.</td>
</tr>
<tr>
<td>C</td>
<td>Records with any (or combinations of) 1982 Census nonrespondent source or USDA nonfarm source, with TVP between $2,500 and $19,999.</td>
</tr>
<tr>
<td>D</td>
<td>Records excluded from the census mail list by the classification tree methodology that were not included in strata A through C.</td>
</tr>
<tr>
<td>E</td>
<td>Records designated by the classification tree methodology to be included on the mail list that were subjectively removed from the 1987 final mail list by Agriculture Division Staff.</td>
</tr>
</tbody>
</table>

The chi-square test statistic was computed to be 310,741 with one degree of freedom which is significant beyond the 0.0001 level. Therefore, the null hypothesis is rejected, indicating that the observed farm status and the EFPs are dependent.

The odds ratio was determined to be 0.240 with 0.239 and 0.242 being the 95 percent confidence bounds. This indicates that, under the given classification, the odds that a farm respondent belongs to a model group with farm proportion < 0.4322 are 0.240 times those of a nonfarm respondent belonging to such a model group. This result indicates that the classification tree methodology was able to distinguish farm from nonfarm addresses.

Table 2 provides the nine classes for the goodness-of-fit test, the total farm respondents and OFPs for each class, plus each class hypothesized midpoint for the chi-square goodness-of-fit test.
Since the model groups were arranged into classes based on EFP values, we hypothesized that the proportion of farm respondents in these classes should increase as do the EFPs. The chi-square test statistic was computed to be 0.1380 with eight degrees of freedom. This was not significant and therefore we were unable to conclude that the proportions of the expected class midpoints and the observed proportions are different.

3. EVALUATION OF THE CLASSIFICATION TREE METHODOLOGY USING FARM FREQUENCIES

3.1 Analysis Methodology

We evaluated the EFP values and the OFP values using the expected and observed farm frequencies from 1,839 model groups. Respondents were not observed for 345 model groups out of the 2,184 model groups created by the classification tree methodology. We compared the two sets of frequency distributions: the expected frequency distribution and the observed frequency distribution, to determine differences in distributional behavior.

To understand our motivation for this analysis, first consider one model group at a time. An indication that the classification tree methodology did well predicting the number of farms for a model group would occur if the EFP for that model group closely matched the OFP for the same model group. An equivalent measure would match a model group's EFP to its corresponding OFP. Likewise, since the expected model group frequencies were derived from the EFPs, the methodology would have done well predicting the proportions of all model groups if the expected frequency distribution behaved similar to the observed frequency distribution. Hence, our approach was to compare the two frequency distributions with plots and data analyses.

First we plotted the expected and observed frequency distributions, respectively, for the 1,839 model groups with 1987 observed classified respondents. The distributions were sorted by ascending model group number. The distributions appeared to be very similar.

Next we plotted both sets of distributions (observed and expected) sorted by descending farm frequency. Of the 1,839 model groups that had 1987 classified respondents, 206 of them did not have observed or expected classified farm records when the two data sets were paired by frequency ranking. Exclusion of these 206 model groups left 1,633 model groups that were observed with an expected frequency greater than their corresponding observed frequency. In other words, the classification tree methodology assigned too many farms to over half of the model groups.

We also applied a sign test to compare the model group arrangement between the expected and observed sorted frequency distributions. We used this test to measure the differences between the model group arrangement of the two frequency distributions. (A sign test was defined as if a model group's number from the observed frequency distribution was greater than its complement model group's number from the expected frequency distribution. A perfect match would result in a zero sign statistic. Tests were conducted using all 1,633 model group numbers and five breakouts of these numbers.)

Only the test of all 1,633 model group numbers showed a significant difference between the model group number arrangements along the horizontal axes of the two frequency distributions. (This significant result for test number 1 was caused by the removal from this analysis of 551 model groups lacking any (farm and/or nonfarm) respondents [see Section 3.1].) Our method of computing the expected frequencies biased their values and was observed from this comparison. The computed expected frequency for model groups with small EFPs and small numbers of classified respondents were assigned farm frequencies of less than one-half. We assigned a zero frequency to those

<table>
<thead>
<tr>
<th>Test Number</th>
<th>Index Number Range</th>
<th>Sample Size</th>
<th>Test Statistic</th>
<th>90% Critical Region Bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 - 1,633</td>
<td>1,609</td>
<td>842*</td>
<td>772 817</td>
</tr>
<tr>
<td>2</td>
<td>10 - 35</td>
<td>16</td>
<td>4</td>
<td>4 11</td>
</tr>
<tr>
<td>3</td>
<td>50 - 400</td>
<td>348</td>
<td>177</td>
<td>159 189</td>
</tr>
<tr>
<td>4</td>
<td>400 - 750</td>
<td>347</td>
<td>168</td>
<td>15 298</td>
</tr>
<tr>
<td>5</td>
<td>750 - 1,100</td>
<td>350</td>
<td>144</td>
<td>160 190</td>
</tr>
<tr>
<td>6</td>
<td>1,100 - 1,633</td>
<td>567</td>
<td>265</td>
<td>247 285</td>
</tr>
</tbody>
</table>

* Significant at the 90% level
model groups. This rounding procedure resulted in more expected than observed frequencies with zero values and a sign statistic that was significant on the high end of the critical region. Hence, the absence of these 551 model groups accounted for the significant sign test statistic. The tests numbered 2 through 6 gave a better indication of the model group arrangement differences since we restricted the domain. The results of tests numbered 2 through 6 indicate that the differences in model group arrangements between the two sorted frequency domains were random for those segments of the horizontal axes that were tested.

3.3 Sorted Frequency Distribution Comparisons

We compared the behavior of the expected and observed sorted frequency distributions. To make this comparison, the cumulative distribution function (cdf) of both the expected and observed frequency distributions were determined.

The expected and observed sorted frequency distributions are monotonically decreasing in frequency and both distributions can be given by

\[ P(x) = \frac{k}{y} \]

where \( x \) is a random variable and \( k \) is an integer.

The two distributions were compared using quantile function methodology (Parzen 1979, 1980). The quantile function \( Q(u) \) is defined as the inverse of the cdf \( F(x) \) in the sense that

\[ Q(u) = F^{-1}(u) \]

The quantile function is given by

\[ Q(u) = \sum_{j=1}^{n} P(y_j) \]

where \( x = Q(u) \) and \( u = F(x) \).

We computed quantile functions for samples taken from the expected and observed frequency distributions. Using a systematic sampling scheme, five samples, each of size fifty, were drawn from both the expected and observed frequency distributions. Ten quantile functions were computed from the samples; five each from the expected and observed frequency distributions. These will be referred to as the expected and observed quantile functions.

Tables of location (means), scale (twice the inter-quartile range), and tail measures were determined from each of the ten quantile functions. Measures of tail behavior are values of the standardized quantile function evaluated at 0 and 1 and indicate how the distribution function behaves as \( x \) approaches zero and positive infinity. The standardized quantile function evaluated at 0 measures the behavior of the left tail and evaluated at 1 measures the behavior of the right tail. The expected and observed measures of location, scale, and tail behavior for the five samples are given below on Tables 4 and 5.

### TABLE 4. Expected Quantile Function Measures

<table>
<thead>
<tr>
<th>Sample Number</th>
<th>Mean</th>
<th>Twice the Inter-Quartile Range</th>
<th>Tail Measures</th>
<th>Left</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1089.34</td>
<td>2018.00</td>
<td>-0.039</td>
<td>5.352</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1276.58</td>
<td>2127.50</td>
<td>-0.070</td>
<td>14.670</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2302.10</td>
<td>1292.50</td>
<td>-0.136</td>
<td>23.995</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2422.20</td>
<td>4376.01</td>
<td>-0.062</td>
<td>6.798</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2550.60</td>
<td>1568.02</td>
<td>-0.092</td>
<td>18.253</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 5. Observed Quantile Function Measures

<table>
<thead>
<tr>
<th>Sample Number</th>
<th>Mean</th>
<th>Twice the Inter-Quartile Range</th>
<th>Tail Measures</th>
<th>Left</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>305.90</td>
<td>1599.00</td>
<td>-0.042</td>
<td>4.742</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1229.16</td>
<td>1735.00</td>
<td>-0.070</td>
<td>16.080</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2697.60</td>
<td>2128.01</td>
<td>-0.133</td>
<td>23.142</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2146.38</td>
<td>2810.01</td>
<td>-0.070</td>
<td>9.422</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2409.46</td>
<td>1392.01</td>
<td>-1.106</td>
<td>16.902</td>
<td></td>
</tr>
</tbody>
</table>

We applied the Wilcoxon sign rank test to test the hypothesis that all ten samples were drawn from the same population. The means of the ten samples were combined and ranked. The test statistic, the sum of the ranks associated to the means of the expected frequency samples, was computed to be 33. We accepted the hypothesis that a difference does not exist between the sample means computed from the ten frequency distribution samples based on the upper tail probabilities for the null distribution of Wilcoxon's rank sum statistic (Hollander and Wolfe, 1973). Observations of the tail measures on tables 4 and 5 reveal that the tail behavior of the two distributions are close in value. The results indicate that the two frequency distributions are very similar since no difference among the means was detected and both tails behave similarly.

4. EVALUATION OF THE SURVEY DATA

The data obtained from the model drop survey were evaluated using binomial proportion tests of hypothesis. Our objective was to determine if the observed survey farm proportions were significantly different from the mail list boundary proportion value, 0.1170. The survey of addresses belonging to strata A through D were assigned to three categories based on expected model group farm proportion. Addresses belonging to model groups with EFP < 0.05 were assigned to category 1, model groups with EFP ≥ 0.05 but < 0.1 were assigned to category 2 and model groups with EFP ≥ 0.1 were assigned to category 3. We created these additional categories to determine if a model group categorization behaved differently from the strata. Stratum E was not included in this categorization because, as a result of the classification tree methodology, the addresses would have been included in the census and hence, all model groups belonging to stratum E had EFP > 0.1170.

OCFs were determined for each individual stratum, strata A through D combined, and categories 1 through 3. Using two-tailed, large-sample (normal approximation) tests of hypothesis for proportions, we compared the individual farm proportions for each classification to 0.1170. Finally, using normal probabilities, we computed the probability of getting the observed farm proportions.

A summary of the survey respondents by stratum is given below. Note that since the response rates were very low for some strata, the estimates of proportion of farms may be upwardly biased. Farm operators are more likely to respond to an agriculture survey than nonfarm addresses. The low response rate for strata B, C, and E may be a reflection of the lower expected farm proportions for the survey cases.

### TABLE 6. Survey Counts By Stratum

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Sample Size</th>
<th>Total Response</th>
<th>Classified Farm</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>80,947</td>
<td>1,248</td>
<td>1,049</td>
</tr>
<tr>
<td>B</td>
<td>15,135</td>
<td>1,024</td>
<td>958</td>
</tr>
<tr>
<td>C</td>
<td>33,243</td>
<td>1,229</td>
<td>1,172</td>
</tr>
<tr>
<td>D</td>
<td>42,424</td>
<td>1,224</td>
<td>1,176</td>
</tr>
<tr>
<td>E</td>
<td>50,222</td>
<td>1,225</td>
<td>1,178</td>
</tr>
</tbody>
</table>

We compared the results from the sample survey of addresses removed from the mail list to the boundary proportion value for mail list development. If the null hypothesis is true, an observed farm proportion is distributed approximately normally with mean 0.1170. We expected their farm proportions to be less than the mail list boundary proportion. Similarly, since the sample
drawn from stratum E represents those records included on
the final mail list according to the classification tree
methodology, we expected its farm proportion to be greater
than the mail list boundary proportion. In other words,
we tested if the classification tree methodology correctly
predicted farm proportions for the sampled model groups
relative to the mail list boundary proportion. The re-
results indicate that strata A, C, A through D combined,
and all three model group categories are significantly above
the mail list boundary proportion. This is contrary to
what was expected. Based on the test results for strata
B, D, and E, we were unable to conclude that the observed
proportions were different from the boundary proportion.
Tables 7 and 8 provide the analysis results.

**TABLE 7. Survey Results By Stratum**

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Farm Proportion</th>
<th>Standard Error</th>
<th>Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1730</td>
<td>0.0116</td>
<td>4.81*</td>
</tr>
<tr>
<td>B</td>
<td>0.1355</td>
<td>0.0186</td>
<td>0.91</td>
</tr>
<tr>
<td>C</td>
<td>0.1514</td>
<td>0.0176</td>
<td>1.96</td>
</tr>
<tr>
<td>D</td>
<td>0.1523</td>
<td>0.0225</td>
<td>1.57</td>
</tr>
<tr>
<td>E</td>
<td>0.1343</td>
<td>0.0174</td>
<td>1.57</td>
</tr>
<tr>
<td>A - D</td>
<td>0.1597</td>
<td>0.0081</td>
<td>5.30*</td>
</tr>
</tbody>
</table>

* Significant at the 95% level

**TABLE 8. Survey Results By Model Group Category**

<table>
<thead>
<tr>
<th>Category</th>
<th>Farm Proportion</th>
<th>Standard Error</th>
<th>Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1561</td>
<td>0.0161</td>
<td>2.425*</td>
</tr>
<tr>
<td>2</td>
<td>0.1636</td>
<td>0.0138</td>
<td>3.867*</td>
</tr>
<tr>
<td>3</td>
<td>0.1541</td>
<td>0.0153</td>
<td>2.429*</td>
</tr>
</tbody>
</table>

* Significant at the 95% level

5. EVALUATION OF THE CLASSIFICATION VARIABLES

We evaluated the responses to the twelve questions derived
from the classification variables that were used for the
methodology. Evaluation of the responses helped associate
which questions may have contributed to differences be-
tween the expected and observed frequencies; particularly,
which questions were most associated with records classi-
fied in error. Each model group is defined by a twelve
element vector. Each element corresponds to a response to
one of the 12 questions and is represented by a value of
zero, one, or two. The zero, one, and two values indicate
an "unknown" response, a "yes" response, and a "no" re-
response, respectively, where an "unknown" response indi-
cates that the question was not asked. Described below
are two specific classes of model groups that were used
for this evaluation.

The first class are the 902 model groups that had an
expected frequency greater than their corresponding ob-
served frequency. Class 2 are the 814 model groups that
had an expected frequency less than their corresponding
observed frequency.

We performed the evaluation by first tallying the
responses to the twelve questions for both classes. The
tables for the 814 model groups were weighted upwards by
a factor of 1.108 (= 902/814). Comparison of the question
response counts between classes 1 and 2 revealed differ-
ences greater than ten percent between the two classes for
questions 3, 5, and 8. We further examined the question
counts to determine if all twelve questions were contrib-
uting to the classification process. Question 12 contrib-
uted the least. Only 1.2 percent of the addresses from
class 1 and 2 combined responded with either a "yes" or
"no" to question 12. That is, 98.8 percent of the ad-
resses from these classes had an "unknown" response to
question 12. This suggests that there is no special exam-
ination of the effect of these
differences in counts on the classification tree method-
ology will be required. Table 9 lists questions 3, 5, 8,
and 12.

**TABLE 9. Questions 3, 5, 8, and 12 Used By the Methodology
for the 1987 Addresses**

<table>
<thead>
<tr>
<th>Question</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>

6. EVALUATION OF MODIFICATIONS MADE TO THE CLASSIFICATION TREE METHODOLOGY

We evaluated the results of the modifications made by
Agriculture Division Staff to the classification tree
methodology results before the questionnaires were mailed.
Approximately 127,000 addresses designated to be excluded
from the mail list according to the classification tree
methodology were retained on the final mail list. Con-
versely, approximately 39,000 addresses designated to
receive a questionnaire according to the methodology were
excluded from the final mail list. The evaluation of the
approximate 39,000 addresses designated to receive a
questionnaire by the methodology but that were excluded
were presented in the survey results for Stratum 2 in
Section 4.

We compared the OFPs of the 127,000 addresses designat-
ed to be excluded from the mail list but that were re-
tained on the final mail list to OFPs of addresses includ-
ed in the model drop survey using the two-tailed binomial
proportion test of hypothesis. We will refer to these
addresses from the final mail list as category 1 addresses
and addresses from the survey as category 2 addresses.
Table 10 provides the number of addresses, their estimated
farm proportions, and standard error of estimates for
these two categories. The test statistic was computed to
be 6.52 with standard error 0.0.016 which is significant
beyond the 95 percent level. We expected this result
since the addresses from category 1 were selected to be on
the final mail list based on their high potential for farm
classification.

**TABLE 10. Estimated Farm Proportions for Common Mail List
and Survey Records**

<table>
<thead>
<tr>
<th>Category</th>
<th>Size</th>
<th>Estimated Farm Proportion</th>
<th>Standard Error of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20,739</td>
<td>0.2282</td>
<td>0.0031</td>
</tr>
<tr>
<td>2</td>
<td>1,944</td>
<td>0.1590</td>
<td>0.0083</td>
</tr>
</tbody>
</table>

7. CONCLUSIONS AND RECOMMENDATIONS

In summary, we conclude that the classification tree
methodology was successful in selecting which addresses
were to be on the final 1987 Census of Agriculture mail
list. This conclusion is based on the following:

1. The measures of associations all gave indications that
the classification tree methodology performed well. The
chi-square tests of independence indicated a strong depen-
dence between the expected model group proportions and the
observed responses. The chi-square goodness-of-fit test
showed that the ranking used by the methodology performed
well. The odds ratio also showed that the classification
tree methodology identified addresses most likely to be
farms. Therefore, based on the measures of association,
the classification tree methodology accurately determined
the proportion of farms for each model group.

2. The classification tree methodology successfully
predicted the number of farm respondents for the 1,633
model groups that had classified respondents. This was
evident by the similarities of the expected and observed frequency distributions. We concluded that the means from these two distributions were not different and the tails of both distributions behaved similarly.

3. The sign test indicated that the model group arrangement along the horizontal axes of the two distributions were very similar. Even though the number of model groups that had a higher expected frequency was significantly different from 0.5, the observed difference was close to 0.5. This agrees with the sign test statistic that was significant when all index numbers were tested. The sign test statistic extended only slightly into the critical region.

The plots, sign tests, and sample survey results all indicate that the accuracy of the classification tree methodology can be improved. We base this inference on the following:

1. The survey expected farm proportions were found to be significantly lower than the observed proportions for five out of the eight strata and categories.

2. The results from the evaluation of the survey of non-mailed addresses indicate that the model groups excluded from the final mail list had a farm proportion greater than 0.1170 even when the subjectively expected more likely farm records were added to the mail list. Even though two of the strata did not show significant results, their farm proportions were observed to be higher than 0.1170. This indicates a need to review the assignment of expected farm proportions by the classification tree methodology.

3. The survey's stratum A had a significantly greater farm proportion (0.1730) than the boundary proportion (0.1170), which suggests the need for an in-depth examination of its characteristics. Among other things, all these addresses seem to have an IRS source and this might be an indication not to delete records with IRS sources from the mail list.

4. Visual inspection of the continuous plots show that the expected frequency is greater than the observed frequency over most of their domain.

5. Results of the evaluation of the subjective modifications indicate that Agriculture Division Staff made good decisions when selecting addresses to be retained on the final list. Examination of the criteria used to select these records is needed so that they can be incorporated in the classification tree methodology.

A problem encountered with the current classification tree methodology was that the controls on the number of cases that composed the final model groups was not appropriate. In some instances, farm proportions were estimated from too few cases. Further research into this area is recommended.

TVP is an indicator of expected total value of agriculture products sold by each farm operation. It is derived from the estimated size of farm information contained in the source records.

REFERENCES


ACKNOWLEDGEMENTS

We would like to acknowledge the efforts of Agriculture Division Staff for their contributions. Cynthia Clark, Ann Vacca, and Tommy Gauleen reviewed this report and gave helpful comments; and Melody Atkinson and Bruce Hughes produced some of the graphs; Dedrick Owens conducted the model drop survey. Further, we would like to thank the Census Bureau reviewers: Easley Hoy and William Winkler.