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 group addressees by probability of operating a farm. Geographic area, address source list, and agricultural sales level characteristics from the 1982 mail list and farm status from the 1982 census were used to define the groups. 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 addressees 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-out. Recommendations are aimed at improving the performance of the classification tree methodology for the 1992 census mail list development operation. Greater detail on the classification tree 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 censuses conducted a precensus 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 questionnaires. 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 0.4322 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 127,000 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 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 (EFP).

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 classified as either farm or nonfarm operations. The rest of the cases were nonrespondents or post master returns. For processing reasons, 9,171 respondent records were not assigned model group numbers and hence were excluded from the evaluation. Also excluded from some analyses were 110,615 records with EFPs < 0.1170. These are the respondents 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 expected to have a higher proportion of farms than predicted by the model and therefore, were excluded from some of the analyses. So, the observed census results presented in this paper are based on the 3,235,563 remaining classified records.

From the third source, the observed responses were obtained from a sample survey conducted during 1988, of approximately 5,300 addresses that were removed from the final 1987 Census of Agriculture mail list. These responses provided additional insight into the classification status (i.e., farm or nonfarm) of the sampled addresses. The survey sample size was inflated for an expected 40 percent nonresponse rate. The sampling frame consisted of five strata defined below by specific source combinations and size characteristics (not by inclusion in specific model groups). Note that records from strata A through D were records removed from the mail list after applying the classification tree methodology to the mail list. Further note that stratum E records were subjectively removed from the mail list by Agriculture Division staff even though the classification tree methodology included them in the list. A total of 2,643 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 168 post master returns.

Stratum	Description
A	Records with any (or combinations of) IRS source, a 1982 Census nonfarm or 1982 Farm and Ranch Identification Survey nonfarm source, and expected TVP[1] less than $1,000 (excluding combinations with 1982 Census or USDA farm source or special list sources). A sample of 1,248 records was selected from a total of 86,987 records in this stratum.
B	Records with any (or combinations of) 1982 census nonrespondent source or USDA nonfarm source, with TVP between $20,000 and $99,999. A sample of 1,202 records was selected from a total of 15,155 records in this stratum.
C	Records with any (or combinations of) 1982 census nonrespondent source, or USDA nonfarm source, with TVP between $2,500 and $19,999. A sample of 1,229 records was selected from a total of 33,243 records in this stratum.
D	Records excluded from the census mail list by the classification tree methodology that were not included in strata A through C. All 424 records in this stratum were included in the sample.
E	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. These records have the same characteristics as those from stratum C, except that they had model group farm proportions greater than 0.1170 due to the classification tree methodology. A sample of 1,236 records were selected from a total of 39,052 records in this stratum.

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 form 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, 1,510,376 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 1. 2x2 Contingency Table

	Frequency Percent Row Pct Col Pct	EXPECTED FARM PROPORTION CLASSIFICATION Class 1 (.1170 < EFP ≤ .4322)	 Class 2 (EFP > .4322)	Total
OBSERVED RESPONSE	Farm	285470 8.82 15.90 31.06	1510376 46.68 84.10 65.21	1795846 55.50
	Nonfarm	633752 19.59 44.02 68.94	805965 24.91 55.98 34.79	1439717 44.50
	Total	919222 28.41	2316341 71.59	3235563 100.00

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.

TABLE 2. Goodness-Of-Fit Test Results

EFP (X)	Total Classified Respondents	Observed Farm Proportion	Hypothesized Class Mid-point Farm Proportions
.1170 < X ≤ .2000	123,099	0.2563	0.1585
.2000 < X ≤ .3000	368,768	0.3162	0.2500
.3000 < X ≤ .4322	427,355	0.3213	0.3661
.4322 < X ≤ .5000	317,545	0.4224	0.4661
.5000 < X ≤ .6000	329,051	0.4635	0.5500
.6000 < X ≤ .7000	318,417	0.5321	0.6500
.7000 < X ≤ .8000	392,317	0.6610	0.7500
.8000 < X ≤ .9000	785,304	0.8177	0.8500
X > .9000	173,707	0.8798	0.9500

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 orderings 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 EFF to its corresponding OFF. Likewise, since the expected model group frequencies were derived from the EFPs, the methodology would have done well predicting the farm 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 used to create the sorted frequency distributions. (A total of 551 model groups (345+206) were excluded.) Index numbers 1 through 1,633 were assigned independently in descending frequency order to the model groups of the expected and observed distributions.

The data was plotted in its discrete form using histograms. Observations of the expected and observed histograms revealed that both distributions had similar shapes. These shapes appeared to approximate the distribution of an exponentially distributed random variable.

Next, we generated plots using continuous versions of the sorted frequency distributions which were defined to be linear between their values. These will be referred to as continuous plots. Several continuous plots were constructed to focus on specific segments of the horizontal axis.

The continuous plots of the observed frequencies were superimposed on the expected frequencies. These plots illustrate the similarities between observed and expected frequency distributions and their likeness to the distribution of an exponentially distributed random variable. The most obvious difference between the two distributions was the larger expected frequencies over most of the domain, except for the right tail of the distributions where the two frequencies converged.

3.2 Model Group Comparisons

We compared the values of the frequencies of the two distributions, pairing them by model group number, to determine the number of model groups that had greater expected than observed frequencies. This comparison was conducted on the 1,839 model groups with classified respondents. We determined that 902 model groups had greater expected than observed frequencies, 814 model groups had greater observed than expected frequency, and 123 model groups had occurrences of equal frequencies.

If the classification tree methodology correctly predicted model group farm proportions, we would expect approximately half of the model groups to have greater expected than observed frequencies, and vice versa. So we compared the number of model groups with the larger expected frequency to the number of model groups with the larger observed frequency using a two-tailed binomial proportion test of hypothesis. Of the 1,716 model groups 52.6 percent had greater expected than observed frequencies. We conducted this test to determine if these 902 model groups represented more than half of the total (1,716), that is, we compared 0.526 to 0.5.

The test statistic was computed to be 2.13 with standard error 0.0121 which was significant at the .05 level. Hence, we conclude that, even though 52.6 percent is relatively close to 50 percent, more than half of the model groups 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. The sign test would measure the magnitude by which the paired model group numbers (sorted by frequency) are mismatched. A positive difference occurring 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 test statistic. Tests were conducted using all 1,633 model group numbers and five breakouts of these numbers.

TABLE 3. Sign Test Results

Test Number	Index Number Range	Sample Size	Test Statistic	90% Critical Region Bounds Lower	Upper
1	1 - 1,633	1,609	842**	772	837
2	10 - 35	16	10	4	11
3	50 - 400	348	177	159	189
4	400 - 750	347	180	158	189
5	750 - 1100	350	184	160	190
6	1100 - 1,633	567	285	247	285

** Significant at the 90% level

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 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 $P(x) = Pr[X=x_j]$ where X is a random variable and x_j, $j = 1, 2, \cdots, 1633$ are the decreasing frequency values. If we let y_j be the farm frequency values less than or equal to x_j, the cdfs of the expected and observed frequency distributions are obtained by

$$F(x) = \sum_{y \le x} P(y) \text{ where } P(y) > 0.$$

Both cdfs, $F_e(x)$ and $F_o(x)$, for the expected and observed frequency distribution functions, respectively, are continuous and non-decreasing functions of x where $F(x) = Pr[X \le x]$, $0 < x < \infty$.

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 $F(Q(u)) = u$, $0 < u < 1$, when $F(\cdot)$ is a continuous function. The quantile function is given by

$$Q(u) = F^{-1}(x)$$

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.

Measures of location (means), scale (twice the interquartile range), and tail behavior 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

Sample Number	Mean	Twice the Inter-Quartile Range	Tail Measures Left	Tail Measures Right
1	1089.34	2018.00	-0.039	5.352
2	1276.58	2127.50	-0.070	14.670
3	2304.10	1192.50	-0.136	23.995
4	2422.20	4374.01	-0.062	6.788
5	2355.80	1568.02	-0.092	18.253

TABLE 5. Observed Quantile Function Measures

Sample Number	Mean	Twice the Inter-Quartile Range	Tail Measures Left	Tail Measures Right
1	905.90	1698.00	-0.042	4.742
2	1226.16	1566.01	-0.061	20.022
3	2057.60	1218.01	-0.133	23.142
4	2146.98	2810.01	-0.070	9.422
5	2188.46	1392.01	-0.106	16.902

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 assigned to the means from 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 measures 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 but ≤ 0.1170 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, these addresses would have been included in the census and hence, all model groups belonging to stratum E had EFP > 0.1170.

The OFPs 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 on Table 6. 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

Stratum	Stratum Size	Sample Size	Total Number of Respondents	Response Rate	Classified Respondents	Farm Respondents
A	86,987	1,248	1,087	0.87	1,058	183
B	15,155	1,202	376	0.31	336	45
C	33,243	1,229	438	0.36	412	63
D	424	424	292	0.69	256	39
E	39,052	1,236	450	0.36	409	59
Totals	174,861	5,339	2,643	0.50	2,475	386

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 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

Stratum	Observed Farm Proportion	Standard Error of Estimate	Test Statistic
A	0.1730	0.0116	4.81*
B	0.1355	0.0186	0.91
C	0.1514	0.0176	1.96*
D	0.1523	0.0225	1.57
E	0.1443	0.0174	1.57
A - D	0.1597	0.0081	5.30*

* Significant at the 95% level

TABLE 8. Survey Results By Model Group Category

Category	Observed Farm Proportion	Standard Error of Estimate	Test Statistic
1	0.1561	0.0161	2.425*
2	0.1636	0.0118	3.967*
3	0.1541	0.0153	2.429*

* 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 between the expected and observed frequencies; particularly, which questions were most associated with records classified 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" response, respectively, where an "unknown" response indicates 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 observed 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 tallies 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 differences 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 contributing to the classification process. Question 12 contributed 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 addresses from these classes had an "unknown" response to question 12. Further examination of the effect of these differences in counts on the classification tree methodology 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

Number	Question
3	Is this record a 1982 Census farm?
5	Is this record a 1982 Farm and Ranch Survey nonfarm or on any special list?
8	Is the 1987 expected total value of products unknown or less than $1,000?
12	Is the 1987 expected total value of products greater than $60,000 or is this address a multi-unit or an abnormal farm*?* (Multi-unit farms conduct sizeable operations in more than one location. Abnormal farms have some atypical characteristic, such as being an Indian reservation, university, grazing association, or prison ground.)

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. Conversely, 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 E in Section 4.

We compared the OFPs of the 127,000 addresses designated to be excluded from the mail list but that were retained on the final mail list to OFPs of addresses included 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 0.0106 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

Category	Size	Estimated Farm Proportion	Standard Error of Estimate
1	20,739	0.2282	0.0031
2	1,944	0.1590	0.0083

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 dependence 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.

[1]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

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984), Classification and Regression Trees, California: Wadsworth International Group

Hollander, M. and Wolfe, D.A. (1973), Nonparametric Statistical Methods, New York: John Wiley and Sons.

Killion, R.A. (1987), "Model Evaluation and Specification," unpublished document, Bureau of the Census, Washington, D.C.

Owens, D. (1989), "Classification Model Documentation," unpublished document, Bureau of the Census, Washington, D.C.

Owens, D., Killion, R.A., Ramos, M., and Schmehl, R.L. (1989), "Classification Tree Methodology for Mail List Development," American Statistical Society Association 1989 Proceedings of the Section on Survey Research Methods.

Parzen, E. (1979), "Nonparametric Statistical Data Modeling," Journal of the American Statistician, 74, 105-131.

Parzen, E. (1980), "Quantile Functions, Convergence in Quantile, and Extreme Value Distribution Theory," Technical Report No. B-3, Texas A&M Research Foundation Project No. 4226, Texas A&M University, College Station, Texas.

ACKNOWLEDGEMENTS

We would like to acknowledge the efforts of Agriculture Division Staff for their contributions. Cynthia Clark, Ann Vacca, and Tommy Gaulden 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.