Evaluation of Classification and Regression Tree Generated Model Groups for the 1992 Census of Agriculture

Stephen Ash, Melinda Kraus and Anne Peterson, Bureau of the Census
Stephen Ash, AGFS, Bureau of the Census, Washington D.C. 20233[1]

Keywords: Classification and Regression Tree (CART), Mail List Development

Census of Agriculture Overview

A census of agriculture, taken every five years, collects data and publishes information on agricultural production and sales, land in farms, and operator characteristics. A census farm is any agricultural operation that sells or has the potential to sell $1,000 or more of agricultural products during the census year. The census of agriculture mail list is a multiple list frame. It is not constantly maintained, but is recreated prior to each census. Sources of the mail list include Internal Revenue Service (IRS), National Agricultural Statistics Service (NASS), the previous census mail list, and several special lists.

The initial mail list contains more records than the census is able to mail, due to budget considerations. For the 1992 Census of Agriculture, the initial mail list needed to be reduced from 3.78 to 3.55 million records. A systematic procedure of selecting records for mailing and nonmailing was developed using CART Methodology. This methodology classified records into groups of probable farm and nonfarm operations. Groups of records least likely to be farm operations were dropped from the list until the 3.55 million record cutoff was reached.

Given the present climate of cost-cutting within the federal government, including the census bureau, the accuracy of determining which records to drop from the mail list is becoming an increasingly important issue as mailout sizes may be further reduced. This paper presents the evaluation of the 1992 Census of Agriculture CART Methodology, the first step in improving this procedure for the 1997 Census of Agriculture.

Constructing the CART Model

The model was constructed using CART software purchased from the California Statistical Software, Inc. CART software constructs binary trees from the independent input variables. At each step of the binary tree construction, or the creation of two branches, CART selects the independent variable that maximizes the homogeneity of the dependent variable within each of the branches. This continues until a "large" tree is constructed. The "large" tree was then pruned back to a sub-tree which most efficiently uses the independent variables. The endpoints of the binary tree branches or the terminal nodes are identified as the final "model groups" in the census application.

Data from the 1987 Census of Agriculture was used to construct the CART model. Input variables were common to both the 1987 and 1992 census mail lists. In total, fifteen input variables were used. Fourteen variables were related to the source of the record and were formatted as a yes or no question. (Example: Was the record from a NASS Source, and was it a NASS Farm?) One variable indicated the expected size for the record, as derived from the record's source(s). This variable had a range of 17 possible values. Each of the 15 variables had previously shown to be related to the farm and nonfarm status of the record.

Since the dependent variable was the record's farm and nonfarm status, only farm and nonfarm records from the 1987 Census of Agriculture could be used to construct the model. Nonrespondents, undeliverable as addressed, non-classified records, and records dropped from the 1987 mail list could not be considered. This implicitly assumed that the classes of records not considered had the same properties as those considered; i.e., non-classified records had the same proportion of farms.

Each state was run separately, creating 50 different CART models. This was done since previous census research showed that each state's characteristics were different with respect to farm and nonfarm status. In total, 757 model groups were identified, each group containing records from one state.

Application to the 1992 Mail List

The 1992 Census of Agriculture Mail List was divided into the same model groups identified above. An expected farm proportion equal to the 1987 census farm proportion was associated with each group. For the 1992 Census of Agriculture, records were divided into a total of 734 model groups (23 model groups had no records from the 1992 Census of Agriculture Mail List).

The model groups were sorted in descending order by the expected farm proportion in each group. Starting with the model groups with the highest expected farm proportion, the model group whose cumulative record count was closest to the designated census record count was selected as the mail list cutoff. All records in model groups in and above the cutoff were kept on the mail list and all other records were dropped. For the 1992 Census of Agriculture, all model groups with an expected farm proportion less than or equal to 0.188 (a total of 276,078 records in 62 model groups) were dropped from the mail list. These model groups were dropped regardless of state.

Subject Analyst Review

Census of agriculture subject matter analysts reviewed all records dropped from the mail list by CART. Analysts changed the nonmail to mail status for specific model groups or subsets of model groups in an attempt to maximize the mail list coverage across states and to include groups of records which they considered historically significant (for example, those records with large expected sales and those records which were farms in the previous census). Approximately the same number of records were also switched from mail to nonmail status to offset the initial adjustments. These records were also selected by the analysts. Revisions of mail and nonmail status were made to all or some of the records within 142 model groups.

Model Drop Survey

After subject analyst review, a total of 229,180 records were finally dropped from the mail list, 107,467 by CART and 121,713 by analyst adjustment. Only 10 entire model groups were dropped, the remaining model groups contained partial drops. A survey of the records dropped from the census was conducted to estimate the farm status of the dropped records and make inferences about the records dropped by CART versus those dropped by analyst adjustments. The model drop survey sample was selected to produce national level estimates with a 5% coefficient of variation. A systematic sample of 7,897 records was selected at a rate of 1 in 29. This was based on an expected proportion of farms equal to 0.10 and inflating for an expected response rate of 54%. The final response rate was 82%.

In this evaluation, the model drop survey results are compared to information from the 1992 Census of Agriculture. Since the techniques for gathering and processing the data from these two sources were not identical, a probable processing or interviewing mode bias exists. The model drop survey primarily used computer assisted telephone interviewing (CATI) for data collection. Survey records unresolved after the CATI operation were mailed one census form. The 1992 Census of Agriculture used several mailings to gather census information but limited CATI use to low-response counties, large and abnormal farms.

Evaluation Overview

The primary goals of this evaluation are to:

- Examine the associations between the expected and observed farm proportions for all model groups on the mail list;
- Evaluate the differences between the expected and observed farm proportions. Determine possible variables which explain these differences, such as state, expected farm proportion and the number of records in a model group;
- Examine measures for comparing the 1992 and 1987 models;
- Evaluate differences between original CART drops and analyst adjustments and compare to what was expected;
- Examine the overall performance of the model including measures for correct classification of records;
- Provide recommendations for the 1997 Census of Agriculture.

Statistical Methods

Nonparametric comparisons were used since the differences between the expected and observed proportions of farms were shown to be nonnormal using the Shapiro-Wilk statistic. One sample paired replicate tests were used to analyze the expected and observed farm proportions since each proportion can be viewed as a measurement for a specific model group. All hypothesis tests were conducted at a $\alpha = 0.05$ level of significance.

Split model groups containing both mailed and dropped records were not used when comparisons between observed and expected farm proportions between model groups were made. However, the expected farm proportions were regenerated for each half of the split groups in order to evaluate final records dropped from the mail list.

Multi-units, "abnormal" farms, special list cases, and records that were part of the census evaluation of mail list coverage were excluded from the CART model, since these records were included in the census with certainty. These groups were not included in this evaluation.

The records on the preliminary mail list can be classified into four categories based on analyst adjustments and CART drops. The CART discriminant model divided the original mail list into two categories, records kept on the mail list (record types A and C) and records dropped (record types B and D). Analysts further divided the list into two more categories, records originally dropped by the model but later restored to the list (record type B), and records originally kept on the mail list but later dropped from the list by analysts (record type C).

Record Type	Description
A	Model selected records for mailout, analyst kept mail status.
B	Model dropped records, analyst changed to mailout.
C	Model selected records for mailout, analyst changed to nonmail.
D	Model dropped records, analyst kept mail status.

Examination of Expected and Observed Farm Proportions

First, we determined whether there was a dependence between the expected and observed paired observations. The expected and observed farm proportions were compared using a distribution-free test for independence (Kendall). This test examined the null hypothesis, H_o: $P(X \leq x \text{ and } Y \leq y) = P(X \leq x) \cdot P(Y \leq y)$ for all x and y, of the paired proportions in a large sample approximation. The null hypothesis was rejected, and therefore we concluded that the expected and observed proportions were dependent.

To describe whether the expected and observed farm proportions had a positive or negative association, Kendall's τ was also estimated. This measure of association ranges from -1 to 1, where -1 represents a negative association, 1 a positive association and zero, no association. Kendall's τ was estimated as 0.661, where the 95% confidence interval ranged from 0.625 to 0.697. This demonstrated that the expected and observed farm proportions had a strong positive association.

Since dependence was established, we completed a statistical test of the "interchangeability" or "exchangeability" of the expected and observed farm proportions. This distribution-free test for bivariate symmetry (Hollander) examined the null hypothesis H_o: $P(X \leq x \text{ and } Y \leq y) = P(X \leq y \text{ and } Y \leq x)$, for all x and y, in a large sample approximation. The null hypothesis was not rejected, and therefore we concluded that there was not enough information to determine whether the expected and observed proportions are not interchangeable.

Next, we determined whether a "treatment effect" or a systematic difference between the expected and observed farm proportions was present. A distribution-free test associated with Friedman rank sums examined the null hypothesis H_o: $\tau_{expected} = \tau_{observed}$, given the model $X_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$, where i denotes the model group and j the expected or observed status. The null hypothesis was rejected, and we concluded that a treatment effect was present. This treatment effect is illustrated in Table 1.

Table 1: Observed Farm Proportions by Selected Expected Model Group Ranges. [2]

Expected Farm Proportion (p)	1987 Farm Proportion	1992 Farm Proportion
$0.2 < p \leq 0.3$	0.316	0.334
$0.3 < p \leq 0.43$	0.321	0.390
$0.43 < p \leq 0.5$	0.422	0.483
$0.5 < p \leq 0.6$	0.464	0.601
$0.6 < p \leq 0.7$	0.532	0.674
$0.7 < p \leq 0.8$	0.661	0.794
$0.8 < p \leq 0.9$	0.818	0.816
$0.9 < p \leq 1.0$	0.88	(N/O)
Overall	0.539	0.584

Note: (N/O) identifies that no observations were present.

In the 1992 Census of Agriculture, the observed farm proportions were consistently greater than the expected farm proportions for each model group (the opposite of what occurred in 1987). The reason for this is an overall shift in the proportion of farms on the mail list from census to census. When the farm proportion decreases between censuses, the expected farm proportion will be, on average, greater than the observed, as from 1982 to 1987 (the overall farm proportions for the 1982 and 1987 censuses were 0.5836 and 0.5427, respectively). Likewise, if the farm proportion increases between censuses, the expected farm proportion will be less than the observed, as from 1987 to 1992. There is no method present to reconcile these differences, since we do not have this information prior to conducting the census.

Examination of Differences Between Expected and Observed Farm Proportions

The average absolute difference between the expected and observed farm proportions was 0.06. Only 35.9% of the model groups had a smaller farm proportion than we expected. Approximately 75% of the absolute differences were smaller than 0.09 and 95% of the absolute differences were smaller than 0.22. The condoles for both the difference and absolute differences are shown in Table 2.

Table 2: Condoles of Differences of Expected and Observed Farm Proportions.

Quantiles	Absolute Differences	Differences
25%	0.02	0.02
50%	0.04	0.02
75%	0.09	0.06
95%	0.22	0.20

Since a different CART model was created for each state, we tested whether any state relationship contributed to the difference of the expected and observed farm proportions. A distribution-free test for independent samples (Kruskal-Wallis) compared H_o: $\tau_1 = \tau_2 = ...\tau_{50}$, in the model $X_{ij} = \mu + \tau_i + \epsilon_{ij}$, where τ_i denotes the i^{th} state effect. The null hypothesis was not rejected, and therefore we concluded that there was not enough information to determine that any one of the state's differences between the expected and observed proportion of farms was greater than any other state's differences.

We also looked at the size of the model group to see if this factor contributed to the differences. Although the differences were not affected by the size of the model group, there was considerably more variation in model groups containing a small number of records. A statistical test also determined whether the dispersion of the differences for model groups with 150 records and less against greater than 150 records was significantly different. A distribution-free test for dispersion with unknown or unequal medians (Moses) compared H_o: $\gamma^2 = 1$, where $\gamma = \sigma_2 / \sigma_1$, in the model $X_i = \mu_1 + e_i \sigma_1$ and $Y_j = \mu_2 + e_j \sigma_2$ for I = 1,..m and j = m+1,...2m. The null hypothesis was rejected, and therefore we concluded that the dispersion of the differences between the expected and observed farm proportions for model groups with 150 and less records was not equivalent to the dispersion for model groups with greater than 150 records.

The expected proportion of farms was also examined to determine whether it could explain any of the variation in the differences. We tested this by forming 50 model group categories with similar farm proportions. We used the distribution-free test for independent samples (Kruskal-Wallis), which compares H_o: $\tau_1 = \tau_2 = ... \tau_{50}$, in the model $X_{ij} = \mu + \tau_i + \epsilon_{ij}$, where τ_i denotes the i^{th} model groups category effect. The null hypothesis was not rejected, and therefore we concluded that there was not enough information to determine that the differences in any one farm proportion category were different than any other.

Comparison to the 1987 Model

In an attempt to create a single measure for comparing the 1992 and 1987 models, a simple regression model was applied to the expected and observed farm proportions. This model used the expected farm proportion and the mailed state values to predict the observed farm proportion. The mailed state value was included since there was a state effect in the 1987 model. For comparability, the state effect was also included in the regression model for the 1992 data, although it was already known that there was no state effect in the 1992 model.

The 1992 model had an R^2 value of 0.71 and the 1987 model had an R^2 value of 0.36. We interpret this to say that the 1992 model explained approximately twice as much of the variation in the observed farm proportion as the 1987 model.

Model Drop Survey Analysis

Results of the model drop survey and the data from the 1992 Census of Agriculture are combined to create the farm proportions for all four record categories previously defined. Table 3 shows the effect of analyst adjustments on the final mail list as well as the differences between the expected and observed farm proportions in each category. The expected farm proportion was estimated from the 1987 Census mailout whereas the observed farm proportion was calculated using both the 1992 Census mailout and model drop survey.

Table 3: Summary of Model Drop Survey Results. [3]

	CART Status					
Ana-lyst Status	Mail		Nonmail		Total	
	Obs'd (%)	Exp'd (%)	Obs'd (%)	Exp'd (%)	Obs'd (%)	Exp'd (%)
Mail	59.17 A	56.44	40.77 B	16.02	58.39 A&B	55.39
Nonmail	25.26 C	19.17	26.66 D	16.08	25.91 C&D	17.28
Total	57.99 A&C	55.98	34.94 B&D	16.05	56.42	54.22

Comparisons between the observed and expected farm proportions showed that the observed farm proportions of all four record categories in the table were larger than their respective expected farm proportions. The largest differences were found in the original CART drops (categories B, D, B & D). The difference between the observed and expected original CART drops (category B & D) was much higher than the average differences found on the mail list. The difference in final drops (category C & D) was only slightly larger than the average differences.

Comparisons between CART and analyst dropped records showed that the analyst adjustments were beneficial to the accuracy of the mail list. The 40.77% farm proportion of records added back to the mail list (category B) is larger than the proportion for both the original CART dropped records (category D) and the analyst dropped records (category C). In contrast, 25.26% of records dropped by analysts were farms much smaller than the percentage added back. The farm proportion on the mail list increased from 57.99% to 58.39% after analyst adjustments. In addition, the farm proportion on the drop list decreased from 34.94% to 25.91%.

Evaluation of Records with a NASS Source

Approximately 64% of the records dropped from the mail list were records received from NASS. Over half of these NASS dropped records have a NASS farm source only. The remaining records have a combination of NASS farm source and a weak census source such as census nonrespondent, previous census nonfarm and NASS nonfarm. The observed farm proportions for NASS source records are shown in Table 4.

Table 4: Farm Proportions for NASS Source Records.

	Observed Farm Proportion	
Drop Type	NASS source only records	NASS source and weak census sources
CART	26.67%	13.41%
Analyst	30.19%	14.21%
Total	28.66%	13.87%

Dropping NASS records combined with weak census sources was the best decision made by both the model and analysts. Of the source combinations and mail size codes dropped by the model and analysts, the NASS farm source combined with a weak census source resulted in the smallest farm proportion (13.87%). This was much lower than the 25.91% overall proportion of the model drop survey. However, dropping NASS records not matched to any census source was not significantly different than the overall survey farm proportion.

Evaluation of the Mail List Cutoff

The percent of records which were correctly assigned a mail or nonmail status by the model was also of interest to us. This data is shown in Table 5. Census data and model drop survey data were used to calculate these percents.

Table 5: Summary of Mail List Cutoff Performance.[4]

Number of Records Percent of All Records Column Percent		Observed Farm Proportion		
		≥ 0.188	< 0.188	All
Expected Farm Proportion	≥ 0.188	3,125,547 88.0% 90.3%	66,459 1.9% 73.6%	3,192,006 89.9%
	< 0.188	335,673 9.4% 9.7%	23,772 0.7% 26.4%	359,445 10.1%
	All	3,461,220 97.4%	90,231 2.6%	3,551,451 100.0%

Table 5 shows that 88.7% of all records were correctly assigned a mail or nonmail status. Of the records with an observed farm proportion 0.188 or greater, 90.3% were on the mail list. However only 26.4% of the records with an observed farm proportion less than or equal to 0.188 were correctly dropped from the mail list.

Summary of Results - How do the expected and observed farm proportions differ?

We conclude that the expected farm proportions are, in general, good indicators of the observed farm proportions. The expected and observed proportion of farms are dependent, positively associated and possibly "interchangeable", however there is an effect or shift due to the different overall farm proportions from 1987 to 1992. Due to the differences in the mail list formation and processing, we cannot anticipate the magnitude or direction of this shift.

The two conclusions, the farm proportions are possibly interchangeable and there was a treatment effect, may initially seem contradictory. If farm proportions are interchangeable, one would assume that there would be no treatment effect. It should be noted that the interchangeability test is sensitive to differences of treatment, dispersion and "more general deviations of H_0 ", and therefore less sensitive to specific differences than specialized tests.

What possible sources for the differences between the expected and observed farm proportions were significant?

There was no state effect. It is suggested that no state effect existed since individual state CART models were created. We have also shown that the differences of the expected and observed farm proportions have a greater dispersion as the number of records within a model group decreases.

How did the 1992 and 1987 CART models compare?

Given the increase of R^2 from 0.36 to 0.71 and the removal of the state effect, we conclude that the 1992 model was an improvement over the 1987 model.

What farm proportions were dropped compared with the expected ?

An estimated 25.91% of the 229,180 records dropped from the mail list were farms. There was no significant difference between the analyst drops and model drops that were left by analysts. However, these figures were all much higher than the expected 17.28%. Probable reasons for this include the data collection methodology and the differences in the mail lists. The expected figure was estimated from the 1987 mailout in 1992 whereas the Model Drop Survey was conducted primarily using CATI.

Did the analyst adjustments to the records dropped from the list decrease the proportion of farm records dropped?

The model and analysts were able to identify farms at similar rates but the analysts were better at identifying nonfarms. This could occur because the analysts used information not represented in the model. For example, the model dropped records regardless of state leaving the analysts to adjust the coverage rates of the smaller states. Dropping NASS records combined with weak census sources was the best decision made by both CART and analysts.

How did the mail list cutoff perform?

Approximately 88.7% of all the records on the mail list were assigned a mailout status correctly. Of the records with an observed farm proportion of 0.188 or greater, 90.3% were on the mail list. However only 26.4% of the records with an observed farm proportion less than or equal to 0.188 were correctly dropped from the mail list.

Future Recommendations

- Given the ideal input values, the CART software would be able to construct a perfect model. Therefore, we recommend that further research should examine improving CART input.
- Increase the minimum model group size during CART's model group generation. A minimum size should reduce the dispersion of the difference of the expected and observed proportion of farms.
- Develop additional steps in the CART methodology to reduce the number of analyst adjustments, thereby reducing time spent in analyst review and the number of revisions required. Recommendations include:
  - Include 100% of all states with "small" expected farm counts on the mail list, i.e. New England, Alaska and Hawaii.
  - Identify groups that must be included or excluded prior to model application, for example, records with large expected mail sizes.
  - Drop model groups by state to insure approximately equal mail list farm coverage across all states under consideration.
- If a survey of nonmailed records is conducted again, select a larger sample to allow for testing at the state or model group level. Also, select the sample prior to the census and include the cases in the census mailout. This would remove any possible processing or interviewing mode bias.
- Investigate other possible methodologies for determining expected farm status of mail list records, i.e. logistic regression.

Footnotes

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

[2] Table 1 excludes model groups 1 through 4.

[3] Table 3 excludes model groups 1 through 4. The expected farm proportions do not include three states where CART results could not be reproduced.

[4] Some model groups from the model drop survey were not included, because of sparse response. We also used split model groups with a recalculated expected farm proportion from the 1987 mail list with the same splits.

References

Conover, W.J. (1980). Practical Nonparametric Statistics. New York, NY.: John Wiley & Sons, Inc.

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

Hollander, Myles and Wolfe, Douglas A. (1973). Nonparametric Statistical Methods. New York, NY.: John Wiley & Sons, Inc.

Schmehl, Richard L. and Ramos, Magdalena (1990). "Evaluation of the Classification Tree Methodology Used for the Development of the 1987 Census of Agriculture Mail List." American Statistical Association 1990 Proceedings of the Section on Survey Research Methods, 308-313.