CLASSIFICATION TREE METHODOLOGY FOR MAIL LIST DEVELOPMENT[1]

Dedrick Owens, Bureau of the Census, Ruth Ann Killion, Evangelical Lutheran Church in America
Magdalena Ramos, Richard Schmehl, Bureau of the Census
Dedrick Owens, Census, Agriculture Division, Washington, D.C. 20233

KEY WORDS: Discriminant Analysis; Classification Trees; Census Mail Lists; Build/Refine Trees.

1. INTRODUCTION

The census of agriculture, taken every five years, collects data and publishes information on land in farms, operator characteristics, and agricultural production and sales by farms in the United States. A *census farm* is defined as any place from which $1,000 or more of agricultural products were sold (or had the potential to be sold) during the census year. The intent of the census is to request data from operators of census farms. However, there is no comprehensive national list of farm operator names and addresses. Therefore, one of the most difficult census tasks is development of a mail list that contains all US farm operators and excludes most names or addresses that do not qualify as census farms. The mail list is recreated for each census since it is not feasible nor economical to update the previous census list on a regular basis.

Developing the mail list is a complex process that involves merging and unduplicating several source lists, including the previous agriculture census mail list, and administrative records of the Internal Revenue Service (IRS) and the US Department of Agriculture (USDA). Lists are also obtained for large or specialized operations such as nurseries, greenhouses, and specialty crop farms from government agencies and trade associations. These are referred to as special lists. Records from the merged lists are linked on name, address, social security number and employer identification number with the intention of eliminating records relating to the same operation. A detailed description of linkage operations used in development of census of agriculture lists is provided in the 1984 paper by Dea, et.al.

In recent censuses, a Farm and Ranch Identification Survey was conducted prior to the census. The purpose of this survey was to identify duplicate records and mail list records that did not represent census farms. Based on survey results, between two and three million addresses were eliminated from the mail lists prior to mailout in each census. A farm and ranch survey was not approved for the 1987 census due to budget and respondent burden constraints for the census. *(The number of persons contacted multiplied by the estimated amount of time it takes a respondent to complete a questionnaire is the "burden hours". Every federal survey or census has an assigned burden hours limit.)*

Based on previous census experience, a file of more than 5 million records was expected after merging and unduplication. Given this expectation, the census of agriculture burden constraints could not be met without

- developing a method to reduce the number of addresses on the census mail list and
- assigning some records on the final list to receive a shortened version of the census form (called the *short form* here).

The method used to meet these two needs was development of a model using a type of discriminant analysis.

The 1987 Census of Agriculture was conducted using three forms. The *regular form* has four pages of questions on general aspects of farming such as land use, crop and livestock production and sales, and use of government programs. The *sample form* is six pages long; it asks the same questions as the regular form and additional questions about production expenses, chemical use, machinery, and value of land and buildings. In order to reduce the overall burden hours of the census, a *short form* was developed for the 1987 census. This form is two pages long and asks for only the most basic farm information with intent to impute other farm information based on the basic data collected. The mailout size of the 1987 Census of Agriculture was limited to 4.1 million addresses, of which no more than 3.2 million could receive either the regular or sample forms. By using the model and process described in this paper, the final mail list contained 4,098,693 records:

- 2,084,835 regular forms
- 1,107,452 sample forms
- 906,406 short forms.

Section 2 describes the use of classification tree methodology in development of the mail list model. Section 3 describes the application of the model to the 1987 mail list. Section 4 describes the evaluation of the model.

2. THE MAIL LIST MODEL

This section outlines the methodology used for construction of the discriminant model. The approach was to recreate the 1982 census mail file prior to deleting the 1982 Farm and Ranch survey nonfarm records. This set of records corresponds to the 1987 mail list prior to application of the model. In order to be applied, the model had to be based on variables common to all records in both the 1982 and 1987 files. This limited the types of data available to geography, the estimated size and the source of the record. The estimated size is based on information contained in the source records and is meant to be an indicator of the expected total value of agricultural products sold (TVP) by each farm.

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

In order to make the process manageable, the 1982 file was divided into 29 subfiles based on groups of states. Each subfile was randomly split into two parts. Classification trees were developed (first half of subfile) and refined (second half of subfile) independently for each subfile. There were twelve possible questions (e.g., "Is this record on an IRS list?") asked in applying the discriminant process to develop classification trees. Based on the trees, groups of records with similar characteristics (model groups) were created and assigned a probability that a record in the group was a farm.

Trees are constructed by creating a vector for each record, where the data are responses (yes/no) to the twelve questions. Using an optimization rule, the question which best splits the vectors into groups of farms vs. nonfarms is determined. An iterative process of determining optimal splits is used until a stopping rule is satisfied, ending the construction stage. The refinement procedure is applied to yield the tree with the minimum misclassification rate. Combined results from all 29 classification trees were used to define model groups for the mail list model.

The following subsections describe the data used, the optimization rule, construction of a classification tree, the refinement method used to determine the optimal tree, and the model specification from the groups defined by the 29 classification trees.

2.1 *Data for Classification Tree Development*

The classification trees leading to the mail list model were developed using 1982 census mail files. These files contained data which could be used to ask classification questions and, most importantly, an indication of farm status for each record as of the 1982 agriculture census. The 1982 census file was supplemented by the 1982 Farm and Ranch survey records to provide a good approximation to the set of records that existed in the preliminary 1987 mail file.

The file of 1982 data used for tree development contained about 5.3 million records. It was necessary to split the file into several smaller files to minimize computer costs. The file was ultimately divided into the nine census geographical divisions and into categories A and B. Category A consisted of records from

- two or more reliable sources (IRS, USDA, or 1978 Census farm) or
- nonfarm sources only (1978 Farm and Ranch survey nonfarm or 1978 Census nonfarm).

The remaining records were placed into Category B.

Some geographic division files were still too large and were further divided by state or groups of states. The total number of subfiles created was 29. *(See Table 1.)*

2.2 *Classification Questions*

The questions to be asked were limited to those that could be answered for each record by data existing on both the 1982 and 1987 files. Geography was used in defining the subfiles, so this restriction limited questions to two major topics: source of the record and expected total value of products.

The twelve questions used for model development were:

- 1978 census nonfarm?
- On a 1982 IRS list?
- 1978 census farm?
- 1978 census nonrespondent?
- 1978 Farm and Ranch survey nonfarm?
- On any 1982 special list?
- On a 1982 USDA list?
- 1982 expected TVP unknown?
- 1982 expected TVP <$2,500 or unknown?
- 1982 expected TVP <$5,000 or unknown?
- 1982 expected TVP <$60,000 or unknown?
- 1982 expected TVP >$60,000 or is this record a "multiunit" or "abnormal" farm? *(multiunit farms conduct operations in more than one location; abnormal farms have some atypical characteristic, such as being on an Indian reservation, university, grazing association, or prison grounds)*

2.3 *Data Format*

Each subfile was divided randomly into two sets of records, X1 for tree construction and X2 for tree refinement.

For each record i in the set a vector, $q_i = [q_1, q_2, \ldots q_{13}]$, was created. The first twelve vector elements are the answers to the twelve questions used for classification and are valued 0(no) or 1(yes). The last element is an indication of nonfarm(0) or farm(1) status in 1982.

TABLE 1
Files for Tree Development

DIVISION	CATEGORY	STATES
1	A,B	ME,NH,VT,MA,RI,CT
2	A,B	NY,NJ,PA
3	A1,B1	OH,IN
	A2,B2	IL
	A3,B3	MI,WI
4	A1,B1	MN,ND,SD
	A2,B2	MO,KS
	A3,B3	IA,NE
5	A	DE,MD,DC,VA,WV,NC,SC,GA,FL
	B1	DE,MD,DC,VA,WV,NC
	B2	SC,GA,FL
6	A	KY,TN,AL,MS
	B1	KY,TN
	B2	AL,MS
7	A	AR,LA,OK,TX
	B1	AR,LA,OK
	B2	TX
8	A,B	MT,ID,WY,CO,NM,AZ,UT,NV
9	A,B	WA,OR,CA,AK,HI

2.4 *Developing a Classification Tree*

The process for developing a classification tree is an iterative one which selects the best question to separate farm and nonfarm records at each step of the process. This selection employs the use of an optimization criterion. The initial set is all the records in a particular subfile X1 (tree building half). The end result is many sets that have common characteristics within each set but differing characteristics across sets. Since a stopping rule is used, not all questions are asked of each set (i.e., the q_i vector for each record in a set will not necessarily be identical, but the elements related to the set-defining questions will be the same). The end result sets are defined by a model group vector, $M_b = [m_{b1}, m_{b2}, \ldots, m_{b12}]$, where the elements are the answers to the twelve questions: 0(no), 1(yes) or 2 (not asked for this model group). The range on the subscript b is from 1 to B, the number of branches in the tree.

2.4.1 *Optimization Criterion*

The *optimization criterion* provides a standard rule for judging the merit of the sets created by asking each particular question. It is a measure assessing the quality of the data split. A split is of high quality if it does a good job of splitting the original set of records according to farm/nonfarm status.

The measure of the distribution of a set by farm/nonfarm status is called the *set impurity*. The set impurity is zero when all records in that set have the same status and increases to 0.693 when the records are equally divided according to status.

The impurity of data set t is given by the expression:

$$I(t) = - \sum_{j} p_{jt} * \ln[p_{jt}] \qquad (1)$$

where j = the status (farm or nonfarm),
$p_{jt} = n_{jt}/n_t$, the probability of status j in set t,
n_{jt} = the number of status j records in set t,
n_t = the number of records in set t.

The impurity is computed for the set t -- I(t) -- and the two sets that result from asking a question: records answering "no" -- $I(t_N)$ -- and records answering "yes" -- $I(t_Y)$. A measure of *impurity reduction* caused by asking question s is given by:

$$R(s,t) = I(t) - [P_N * I(t_N)] - [P_Y * I(t_Y)] \qquad (2)$$

where $P_N + P_Y = 1$ are the proportion of "no" and "yes" records in set t, respectively.

The question, s*, that maximizes R(s,t) is the one that yields the best split of set t.

2.4.2 *The Iterative Process*

Beginning with the complete set of records from the tree building half of a subfile, X1, each of the twelve questions is used to split the data set into two parts, those records with a 0(no) and those with a 1(yes). The best split (i.e., the question that does the best job of dividing the total set of records into farms and nonfarms) is determined by maximizing R(s,t).

Once the entire set has been optimally split, there are two "branches" represented by sets t_1 and t_2. For each of these sets, the eleven remaining questions are asked to determine the one that maximizes $R(s,t_1)$ and the one that maximizes $R(s,t_2)$. *The questions that maximize $R(s,t_1)$ and $R(s,t_2)$ are likely to be different questions.* The process continues with each newly formed set being asked the remaining questions. The process ends for a particular branch of the tree when the stopping rule is reached.

2.4.3 *The Stopping Rule*

The final definition of a branch of the tree, M_b, is determined by the questions which have been asked to that point. Every branch keeps splitting until:

- the set of records at the end of the branch all have the same farm status, or
- there are five or fewer records in the set, or
- the records in the set have the same vector values for the unasked questions, or
- all twelve questions have been asked.

These conditions define the *stopping rule*. Once the set at the end of a branch meets the stopping rule it is called a *terminal node*.

2.5 *A Representation*

Figure 1 is a representation of a completely built tree, assuming four (rather than twelve) questions are possible. In this tree, the original set, X1, is split into t_1 and t_2 by asking question 3. Each of these sets is then split by asking another question: t_1 yields t_3 and t_4 by asking question 1 and t_2 yields t_5

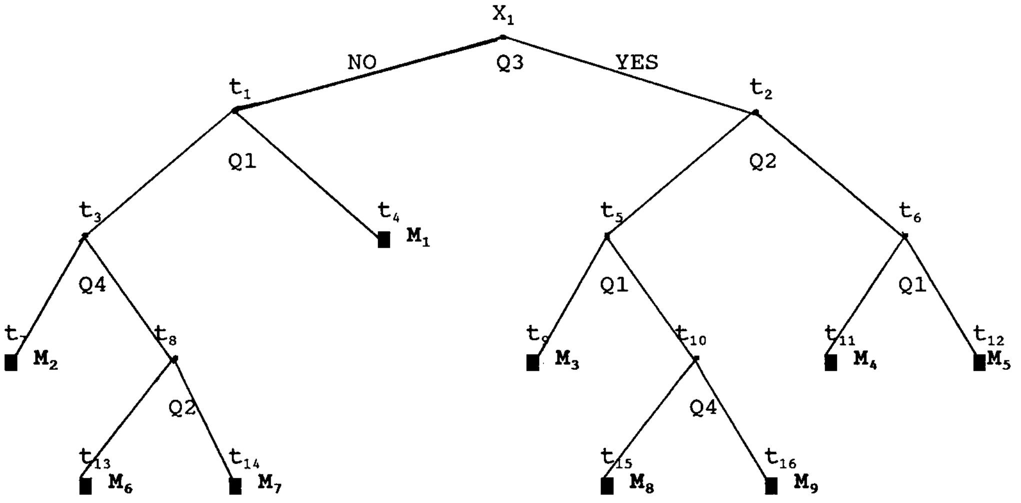

Figure 1

and t_6 by asking question 2. This completely built tree has nine branches; numbering the terminal nodes from top to bottom and left to right, t_4 is the first terminal node, M_1. The last terminal node in this tree, M_9, is set t_{16}.

Assuming that the branch to the left answers the question "no" and the branch to the right answers the question "yes", the model group vectors for these nine terminal nodes are:

- M_1 = [1,2,0,2]
- M_2 = [0,2,0,0]
- M_3 = [0,0,1,2]
- M_4 = [0,1,1,2]
- M_5 = [1,1,1,2]
- M_6 = [0,0,0,1]
- M_7 = [0,1,0,1]
- M_8 = [1,0,1,0]
- M_9 = [1,0,1,1]

The appendix provides data to illustrate the building and refinement of a classification tree.

2.6 *Tree Refinement*

The process of tree building determines, for each set, whether the records in that set can be split by asking an additional question. As is the case with all stepwise procedures, the constructed tree may not be the "best" tree that can be attained. In order to determine the best tree, the concept of *minimum tree misclassification rate* is used.

Each terminal node, M_1 to M_B, is assigned a status based on the farm/nonfarm status that is predominant in the node. The *within node misclassification rate* is the proportion of records in the terminal node that have a status different from the one assigned to the node. Thus, in the representation in Figure 1, if M_6 is assigned the status "farm" and three of nine records in set t_{13} are nonfarm, the misclassification for that node is 0.33. The tree misclassification rate is the sum of the within node rates weighted by the proportion of all records in the node. The object of tree refinement is to determine the tree with the minimum tree misclassification rate.

Breiman, et.al. (1984) show that the tree misclassification rate varies according to the number of terminal nodes (B). The tree misclassification rate has a minimum value at some tree, usually between the largest and the smallest tree. The method is to grow the largest tree possible by going to the stopping rule for all branches and then prune to obtain a series of nested trees, one of which has the minimum tree misclassification rate. This is called *minimum cost complexity pruning*.

The measure of complexity used, C(T,α), is based on the tree misclassification rate, C(T) and a parameter α representing the cost per node. The measure for tree T is:

$$C(T,\alpha) = C(T) + \alpha B \qquad (3)$$

where $C(T) = \sum_b C(b) = \sum_b c(b)p(b)$,

$c(b) = 1 - [\max (n_{jb}/n_b)]$ -- within node misclassification rate for node b
$p(b) = n_b/n$ -- probability that a record is in node b
$b = 1,2,\ldots,B$
B = the number of terminal nodes in tree T.

Breiman et.al. show that as α increases from zero, a series of trees $T_1,T_2,\ldots$, each a subtree of the one before it, minimizes C(T,α). Tree T_i minimizes C(T,α) until α reaches a breakpoint, when tree T_{i+1} minimizes C(T,α). The process continues until the root node, X1, is reached.

The series of trees is defined by pruning *branches* at nonterminal nodes. The branch, K_y, related to a nonterminal node, y, consists of all nodes that descend from y. In Figure 1, the branch from the nonterminal node $y = t_5$ consists of nodes M_3, t_{10}, M_8 and M_9.

Starting with the tree, T, that exists at the end of the tree building process, for each nonterminal node y and branch K_y, set $C(y,\alpha) = C(K_y,\alpha)$ and solve for α. The result is:

$$\alpha_y = [c(y) - C(K_y)]/[B_y - 1]. \qquad (4)$$

Determine the pair (y,K_y) that satisfies

$$\alpha_1 = \min_{y\varepsilon T} [p(y)\cdot(\alpha_y)] \qquad (5)$$

The branch K_y that satisfies (5) is the weakest branch and is pruned. Complexity parameter α_1 is the breakpoint at which subtree $T_1 = T - K_y$ becomes the minimizing subtree. The same process is used to find the weakest branch of T_1, determining α_2 in the process. Pruning of branches and defining subtrees continues until the root node, X1, is reached. This yields a set of nested subtrees, T, T_1, $T_2,\ldots$,X1.

2.7 *Determining the Optimal Tree*

The subfile X2 is used to determine the optimal tree. This process provides an independent verification of the tree construction. (Recall that X1 and X2 are randomly selected halves of the entire data in one of the 29 subfiles.)

The X2 records are assigned to terminal nodes by matching their vectors to the M_b vectors associated with the terminal nodes for each of the trees $T,T_1,\ldots$,X1. A record is misclassified if its known farm status does not agree with the status assigned to the node it is in. The *optimal tree* is the one with the fewest records misclassified, based on the status assigned to each node during the tree construction phase.

For each terminal node, M_b, of the optimal tree, the set X2 is used to determine the probability that a record in that group is a farm:

$$P_b(f) = (n_{fb}/n_b),$$

where n_{fb} is the number of farms in node M_b and
n_b is the total number of addresses in M_b.

The final result of the tree building and refinement process is B vectors, $M_1,M_2,\ldots,M_B$, each with twelve elements (0,1,or 2), and a farm probability $P_b(f)$ assigned to it.

2.8 *Specification of the Mail List Model*

The development of an optimal tree occurred independently in each of the 29 subfiles. These 29 trees had 2,284 mutually exclusive terminal nodes. An example of a terminal node description is: 1978 census farm, expected TVP less than $60,000; located in Arkansas, Louisiana, Oklahoma or Texas. The 2,284 vectors defining the terminal nodes in the 29 optimal trees, along with their associated probability of being a farm, P(f), define the model groups applied to the 1987 mail file.

The model group vectors were sorted in descending P(f) order and sequentially assigned model group numbers from 1 to 2,284. Thus, groups with small model group numbers represent groups with high probability of being farms; those with high model group numbers have low probability.

3. *APPLICATION TO THE MAIL LIST*

The mail list development operations for the 1987 Census of Agriculture began with approximately 13.5 million name and address records. After several phases of matching records and linking duplicates a preliminary mail file of approximately 6 million records remained. Most addresses not responding to the 1982 census and those with indications of nonfarm status from the census, previous surveys or special lists were dropped from the mail list, reducing the preliminary list to approximately 4.3 million addresses. Prior to record linkage this preliminary list was expected to have as many as 5 million addresses. The model was used to further trim the mail list and to identify those records which would receive the short form. This procedure resulted in cost and respondent burden in accordance with the general parameters set for the size and composition of the census mail list.

The main purpose of applying the model was to determine which addresses fell in model groups with lowest farm probability and exclude those records from the mail list. The assumption in developing the model is that record groups with low farm probability in 1982 will have similar low farm probability in 1987.

In order to apply the model to the preliminary mail list, the twelve questions were redefined to ask about the 1987 census rather than 1982, and 1982 Farm and Ranch survey instead of 1978. The questions used to determine the twelve element vector for each record in the 1987 preliminary file were :

- 1982 census nonfarm?
- On a 1987 IRS list?
- 1982 census farm?
- 1982 census nonrespondent?
- 1982 Farm and Ranch survey nonfarm?
- On any 1987 special list?
- On a 1987 USDA list?
- 1987 expected TVP unknown?
- 1987 expected TVP <$2,500 or unknown?
- 1987 expected TVP <$5,000 or unknown?
- 1987 expected TVP <$60,000 or unknown?
- 1987 expected TVP >$60,000 or multiunit or abnormal?

Model groups were assigned to the 4.3 million records on the preliminary mail file using these questions. Records were then sorted by model group number. Extensive review of the group definitions with the lowest probability of being farms was conducted by staff of the Agriculture Division. Based on that review, some records were assigned probability 1.0 and forced into the mail file. A total of 174,834 records in model groups with P(f) < .117 were identified and removed from the final mail file. This file of records is called the *model drop file.*

The 1,107,452 records which would receive the sample form were selected using normal census sample identification procedures. The remaining records in the file were sorted in descending model group order. Starting with the highest model group numbers (lowest farm probability), groups of records were assigned to receive the short form until the group that had the 900,000th record was reached. All remaining records received the regular census form.

4. *MODEL EVALUATION*

The validity of using a discriminant model for the future can be determined by an evaluation of the model predictive power. Two methods will be used for the evaluation. A model drop survey of approximately 5,300 addresses that were assigned to the model drop file was conducted during 1988. Tabulation of the data is underway to verify the decision to delete them from the 1987 census mail list.

In addition, the proportions of farms by model groups will be calculated for the entire 1987 mail list. The final mail list file will be sorted by model groups and the observed 1987 farm proportions calculated for each group. These observed 1987 model group farm proportions will be compared to the model group farm probability which is based on 1982 data.

Recommendations for further applications of this discriminant analysis methodology to the development of the census mail list will be based on results of the model evaluation and the census coverage evaluation to be completed during 1990.

REFERENCES

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

Clark,Cynthia Z.F. (1989) "Mail Data Collection Methodology and Research for the U.S. Census of Agriculture," Proceedings of the 47th Session of the International Statistical Institute.

Dea,J.,Gaulden,T.,Prochaska,D. (1984) "Record Linkage for the 1982 Census of Agriculture Mail List Development using Multiple Sources", Proceedings of the American Statistical Association Survey Research Section, p.576ff.

APPENDIX

Table 1A provides data for use in an example of the construction and refinement of a classification tree. The questions used are:

Q1: A previous census farm?
Q2: On a current IRS list?
Q3: Expected TVP > $60,000, a multiunit or abnormal?

TABLE 1A

Set X1:		
q_1=[1,1,0,1] (1)[1]	q_2=[1,1,1,1] (1)	q_3=[0,1,0,0] (3)
q_4=[1,0,1,1] (2)	q_5=[1,0,1,0] (1)	q_6=[1,0,0,0] (2)
q_7=[1,0,0,1] (1)	q_8=[0,1,1,1] (2)	q_9=[0,1,1,0] (1)
q_{10}=[0,0,1,1] (1)	q_{11}=[0,0,1,0] (3)	q_{12}=[0,1,0,1] (1)
q_{13}=[0,0,0,0] (1)		
Set X2:		
q_1=[0,0,0,0] (1)	q_2=[0,1,1,0] (1)	q_3=[0,1,1,1] (4)
q_4=[0,0,1,0] (2)	q_5=[1,0,0,1] (1)	q_6=[1,0,0,0] (2)
q_7=[1,1,1,1] (2)	q_8=[1,0,1,1] (2)	q_9=[1,0,1,0] (1)
q_{10}=[0,1,0,1] (1)	q_{11}=[0,1,0,0] (2)	q_{12}=[1,1,0,1] (1)

[1] *The first three elements are answers to the three questions (0=no,1=yes); the fourth element is indication of farm status (0=nonfarm,1=farm). The number in parentheses indicates the number of cases with this vector.*

Find the best split by computing the impurity reduction caused by asking each question, using equations 1 and 2.

$$I(t) = - \Sigma p_{jt} {*} ln[p_{jt}] \qquad (1)$$

$$R(s,t) = I(t) - [P_N {*} I(t_N)] - [P_Y {*} I(t_Y)] \qquad (2)$$

$I(X1) = - [(9/20)(ln9/20) + (11/20)(ln11/20)] = 0.688$
Q1: $I(X1_Y) = - [(5/8)(ln5/8) + (3/8)(ln3/8)] = 0.662$
$I(X1_N) = - [(4/12)(ln4/12) + (8/12)(ln8/12) = 0.637$
$R(1,X1)= 0.688 - [(8/20)(.662) + (12/20)(.637)] = 0.041$
Q2: $I(X1_Y) = 0.687$; $I(X1_N) = 0.656$; $R(2,X1) = 0.018$
Q3: $I(X1_Y) = 0.689$; $I(X1_N) = 0.637$; $R(3,X1) = 0.022$.

Since Question 1 has the largest reduction in impurity (0.041), it is the optimal split of X1. Continue to the stopping rule for all branches, obtaining tree T of Figure 1A.

The weakest branch, K_y, of T is determined by minimizing $p(y){*}\alpha_y$ over all nonterminal nodes -- X1, t_1, t_2, t_3, t_4, t_5. Equations 4 and 5 are used in this step:

$$\alpha_y = [c(y) - C(K_y)]/[B_y - 1] \qquad (4)$$

$$\alpha_1 = \min_{y \epsilon T} [p(y){*}(\alpha_y)]. \qquad (5)$$

Table 2A provides the c(y) values for the nodes.

Table 2A

c(X1,0)=9/20	c(t_1,0)=4/12	c(t_2,1)=3/8	c(t_3,0)=1/5
c(t_4,0)=3/7	c(t_5,0)=3/6	c(t_6,1)=0/2	c(t_7,0)=0/1
c(t_8,0)=1/4	c(t_9,0)=1/4	c(t_{10},1)=1/3	
c(t_{11},0)=1/3	c(t_{12},1)=1/3		

The second entry in parentheses is the farm status for the node.

$p(X1){*}\alpha_{X1} = \underline{1/30}$ $\qquad p(t_1){*}\alpha_{t1} = 12/20{*}1/36 = \underline{1/60}$
$p(t_2){*}\alpha_{t2} = 8/20{*}1/16 = \underline{1/40}$ $\quad p(t_3){*}\alpha_{t3} = 5/20{*}0 = \underline{0}$
$p(t_4){*}\alpha_{t4} = 7/20{*}1/7 = \underline{1/20}$ $\quad p(t_5){*}\alpha_{t5} = 6/20{*}1/6 = \underline{1/20}$

Since the branch from t_3 satisfies Equation 5, it is trimmed, creating tree T_1. Performing the calculations of Equation 4 for tree T_1 shows both t_1 and t_2 yield the minimum value. The tree is randomly trimmed at t_2, yielding tree T_2, which is pruned at t_1 yielding tree T_3. X1 is the final tree.

The tree refinement set X2 (Table 1A) is used to determine the optimal tree. The q_i vectors are matched to the terminal node vectors for each tree. A record is misclassified if its known status (farm or nonfarm) does not agree with the status of its terminal node. The optimal tree is the one with the fewest misclassified records. Trees T, T_1, T_2, T_3 and X1 misclassify 4, 4, 5, 8 and 9 of the 20 records, respectively. Since T and T_1 have the same number misclassified, the smaller tree is selected. **T_1 is the optimal tree.**

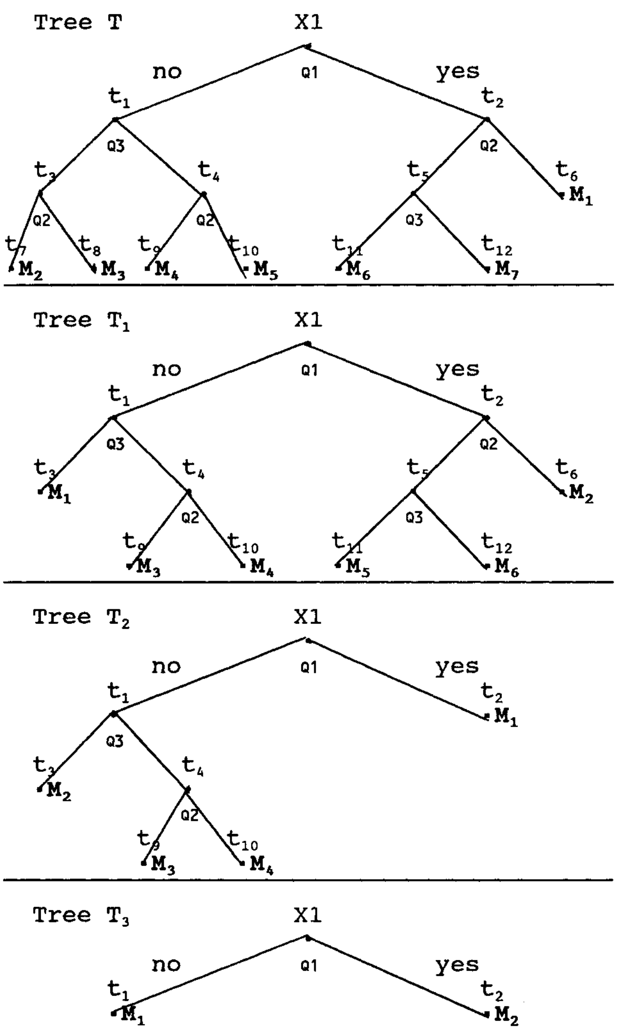

Figure 1A