Who Makes Mistakes? Using Data Mining Techniques to Analyze Reporting Errors in Total Acres Operated

Jaki S. McCarthy and Morgan S. Earp

United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research and
Development Division
Washington DC 20250

RDD Research Report
Number RDD-09-02

April 2009

This report was prepared for limited distribution to the research community outside the U.S. Department of Agriculture (USDA). The views expressed herein are not necessarily those of NASS or USDA.

EXECUTIVE SUMMARY

Survey data collection aims to collect accurate and complete data from respondents. However, in practice, survey respondents often omit required items or provide inaccurate data, and as a result, considerable resources are spent on editing incorrect or questionable data. Although edits "correct" data, it is preferable to collect accurate data initially.

In order to improve the chances of collecting accurate data, surveys conducted by the National Agricultural Statistics Service (NASS) are always subject to pretesting such as cognitive interviewing. However, when working to reduce reporting errors, it is helpful to identify respondents most likely to report erroneous data. Unlike many of NASS surveys, the census of agriculture is self administered and unedited data are keyed and captured for most respondents. This allows an analysis of errors in respondents' initial unedited reports.

This report explores reporting errors related to total acres operated, which is calculated in two separate areas of the Census form by summing individual pieces of the operation. The first area requests total acreage operated based on ownership and the other area requests total acreage operated based on land use type. Respondents must report the designated subitems and sum them to the total acres operated. Both total acres operated figures should be equal; however, this is not always the case with reported data. Errors in summing subitems of both total acres operated and differences in the two equivalent total acres operated are seen in numerous unedited census records.

Data mining allows us to sift through large data sets, such as the census of agriculture, and identify operation characteristics related to these reporting errors. Classification trees (also known as decision trees) allow us to simply identify and describe operations who consistently exhibit reporting errors. Similar analysis could be done using logistic regression; however, logistic regression requires hypotheses regarding the characteristics related to data reporting errors, predefined interaction effects, and no missing data. Classification trees do not require any hypotheses regarding target variable relationships, they automatically identify significant interaction effects, and consider missing data as valid, making them ideal for exploring previously unexplainable relationships.

The classifications provided in this report identify characteristics of respondents and or operations that are more likely to make these errors, suggest possible reasons for the errors, identify appropriate types of respondents for inclusion in questionnaire testing, suggest content for future tests of alternative questionnaires and suggest ways to appropriately edit these items.

RECOMMENDATIONS

1. Review edits used for these items.
2. Use these classification trees as input to redesign these questions and data collection procedures.
3. Look for other areas where data mining techniques can be applied in NASS surveys and the census of agriculture.

Who Makes Mistakes?
Using Data Mining Techniques to Analyze Reporting Errors in Total Acres Operated

Jaki S. McCarthy and Morgan S. Earp[1]

Abstract

Classification trees were used to identify subgroups of respondents with higher error rates when reporting total acres operated on the 2002 Census of Agriculture. Separate trees were grown for operations exhibiting total acres summation errors, missing data, and nonequivalent sums of reported total acres. Terminal tree nodes demonstrating the greatest frequency of total acres operated errors identify characteristics of respondents and or operations that are more likely to make errors, suggest possible reasons for errors, identify content for future tests of alternative questionnaires and suggest ways to appropriately edit these items. Advantages of using classification trees over other analytic methods are also discussed.

Key Words: Classification Tree, Data Mining, Decision Tree, Reporting Errors, Item Nonresponse

1. INTRODUCTION AND BACKGROUND

In survey data collection, the objective is to collect accurate and complete data from the respondent. However, in practice, survey respondents often omit required items or provide inaccurate data. It is clear that mistakes are made when related data are inconsistent or defy logical relationships or expected data are missing. Federal statistical agencies, including the USDA's National Agricultural Statistics Service (NASS), spend considerable resources editing incorrect or questionable data.

Although avenues exist for "correcting" data, it is preferable to collect accurate data initially. In periodic surveys or those using questions common to other surveys, errors in reporting can be examined with an eye toward increasing quality of the reported data in later data collections. Traditionally, survey methodologists use cognitive interviewing and other pretesting techniques to examine reporting errors and potential solutions (i.e. changes in questionnaire design or data collection procedures). However, these small scale analyses are limited by the small *n*'s included (typically fewer than 20 respondents) and the ad hoc nature of the information gathered. In addition, pretests are often conducted with convenience samples of respondents that are readily available, not chosen

[1] Jaki McCarthy is the Chief Cognitive Research Methodologist and Morgan Earp is a Mathematical and Survey Statistician with the National Agricultural Statistics Service, Research and Development Division, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030. The authors would like to thank Thomas Jacob for help with technical advice and data preparation.

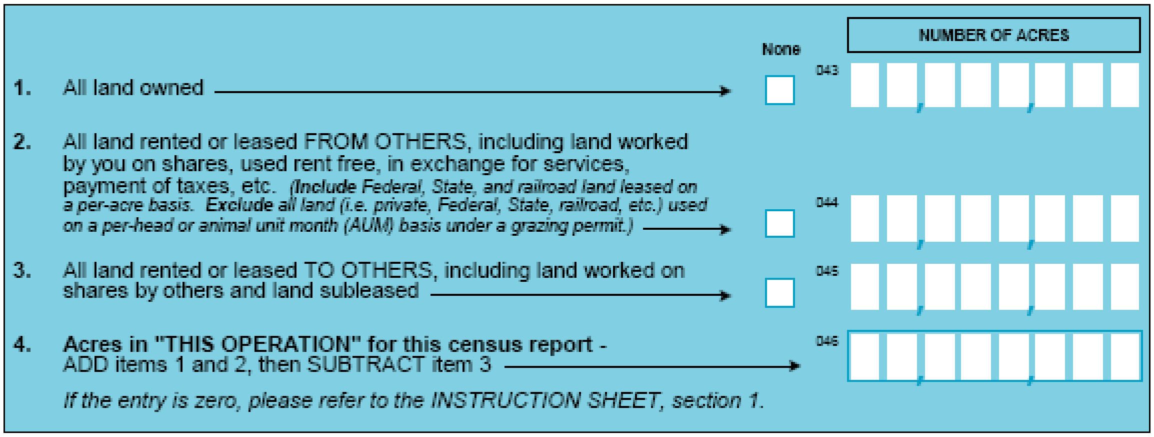
for specific characteristics potentially relevant to reporting errors. For reporting errors that do not occur in most of the population of interest, these small scale tests may not include anyone who would have reported inaccurately in the survey. In addition, they may not include appropriate questions that address information related to the inaccurate data.

In NASS, large sets of reported survey or census data are available for analysis to help improve questionnaire items. This report presents one possible way to use these large datasets for this purpose. By identifying the subsets of respondents most likely to have reporting errors, we can focus traditional techniques, such as cognitive testing, on the appropriate respondents. NASS's large datasets are ideal for the use of data mining techniques, which are used to look for patterns in large datasets. In this paper, the focus is one data mining technique, the decision or classification tree, which is used to subsegment a dataset into groups of like records. This technique is used widely in private industry and marketing for numerous purposes, including identification of specific types of customers (for example, customers likely to terminate a service, customers more likely to buy particular products) or to otherwise subsegment customers, perhaps by creditworthiness, spending patterns, etc. (Berry and Linoff, 2004). Chi-square Automatic Interaction Detection (CHAID) is one algorithm used to generate classification trees and is the basis for the approach used in this analysis. (See deVille, 2006 for a discussion of the algorithms used in SAS Enterprise Miner, the software used in this analysis.)

While large datasets are common in survey data collection, there are few documented instances of the use of data mining techniques with survey datasets. Among them, the Agency for Healthcare Research and Quality used classification trees to model the response probability and to form weighting adjustment cells for the Medical Expenditures Panel Survey (Cohen, DiGaetano and Goksel, 1999). Statistics Netherlands similarly explored the use of classification trees to form post stratification weights for respondents in the Dutch Integrated Survey on Household Living Conditions (Schouten and de Nooij, 2005). Scime and Gregg (2007) used classification trees to build models predicting likely voters based on exit poll data. NASS has used classification trees to define nonresponse weighting cells for the 2007 Census of Agriculture (Cecere, 2008) and to identify records on the initial census mail list most likely to be non-farms (Garber, 2008).

The analysis described in this paper is a novel use of classification trees to examine the relationship(s) between problems in reported data and characteristics of the records. Classification trees have not previously been used to predict census or survey reporting errors. Many problem indicators for reported data can be defined -- totals that are not summed correctly, disagreement between two equivalent numbers, etc. Data mining techniques, such as classification or decision trees (these terms are used interchangeably) allow us to examine what variables are associated with the selected targets. This paper discusses the use of decision trees to examine reporting errors in the series of questions NASS administers to define reporting units for the census of agriculture and other surveys of agricultural operations.

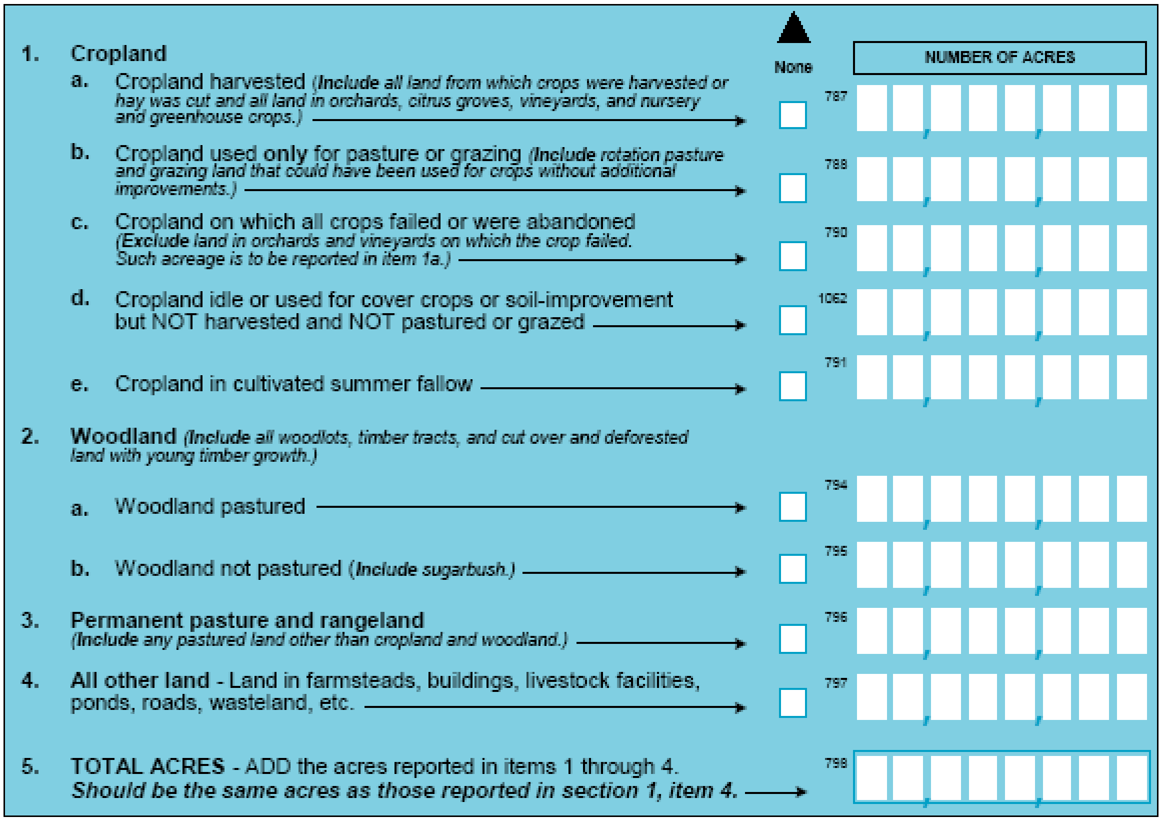
Many organizations ask standard questions to collect basic data from businesses. For NASS, questions are often asked at the beginning of surveys to define the operation (and the corresponding reporting unit). The reporting unit is defined as the "Total Acres Operated" (TAO) and can be derived by summing acres of the operation in several ways. On the Census of Agriculture and other surveys of agricultural operations, NASS asks for: 1) the number of acres owned by the operation, 2) the number of acres rented from others, and 3) the number of acres rented to others. Respondents add the number of acres owned by the operation and the number of acres rented from others, and subtract the number of acres rented to others to calculate their "total acres operated" which they then report on the survey form (Figure 1). This reported "total acres operated" is also referred to as K46.



		None		NUMBER OF ACRES
1.	All land owned →	□	043	
2.	All land rented or leased FROM OTHERS, including land worked by you on shares, used rent free, in exchange for services, payment of taxes, etc. *(Include Federal, State, and railroad land leased on a per-acre basis. Exclude all land (i.e. private, Federal, State, railroad, etc.) used on a per-head or animal unit month (AUM) basis under a grazing permit.)* →	□	044	
3.	All land rented or leased TO OTHERS, including land worked on shares by others and land subleased →	□	045	
4.	**Acres in "THIS OPERATION" for this census report -** ADD items 1 and 2, then SUBTRACT item 3 →		046	
	If the entry is zero, please refer to the INSTRUCTION SHEET, section 1.			

Figure 1. K46 Total Acres Operated items

In a subsequent section of the form, the respondent is asked for acres they operate by type (i.e. cropland, pasture, woodland, all other land) which they also have to sum and report as their "total acres" (Figure 2). This "total acres" is referred to as K798.



		None		NUMBER OF ACRES
1.	**Cropland**			
	a. Cropland harvested *(Include all land from which crops were harvested or hay was cut and all land in orchards, citrus groves, vineyards, and nursery and greenhouse crops.)* →	□	787	
	b. Cropland used only for pasture or grazing *(Include rotation pasture and grazing land that could have been used for crops without additional improvements.)* →	□	788	
	c. Cropland on which all crops failed or were abandoned *(Exclude land in orchards and vineyards on which the crop failed. Such acreage is to be reported in item 1a.)* →	□	790	
	d. Cropland idle or used for cover crops or soil-improvement but NOT harvested and NOT pastured or grazed →	□	1062	
	e. Cropland in cultivated summer fallow →	□	791	
2.	**Woodland** *(Include all woodlots, timber tracts, and cut over and deforested land with young timber growth.)*			
	a. Woodland pastured →	□	794	
	b. Woodland not pastured (*Include sugarbush.)* →	□	795	
3.	**Permanent pasture and rangeland** *(Include any pastured land other than cropland and woodland.)* →	□	796	
4.	**All other land** - Land in farmsteads, buildings, livestock facilities, ponds, roads, wasteland, etc. →	□	797	
5.	**TOTAL ACRES** - ADD the acres reported in items 1 through 4. ***Should be the same acres as those reported in section 1, item 4.*** →		798	

Figure 2. K798 Total Acres Operated Items

These two series of questions appear on facing pages in the 2002 Census of Agriculture. The total acres reported in both K46 and K798 should always be equivalent, which NASS states on the form. However, errors in these two parts of the report form result in frequent edits. Edits result when the parts do not sum to the entered totals, or when the two totals do not agree.

In many surveys, data collected via computer or by interviewers may be verified and corrected as it is reported, so errors initially reported by respondents are lost. However, for the Census of Agriculture, much of these data are collected via self-administered mail questionnaires. Respondents report both the subparts of and the summed Total Acres Operated (for both K46 and K798). Data are initially key entered directly from the forms, thus unedited data as reported by the respondents are available for many operations.

Other information about each operation is also available, including information reported on the rest of the census form, such as location, size, type, types of land operated, whether certain types of commodities are raised, if the land is grazed, is part of an Indian reservation, whether they grow organic crops, whether they operate as a partnership, participate in government agricultural programs, the age of the operator, etc. Results from data mining might indicate that problems arise more often in certain parts of the country, for respondents with specific types of land, or for operations of a certain size, type, or with other particular characteristics. This in turn helps guide further research into potential changes to the questionnaire. Data mining will not identify the reason for problems, but it can identify the subsets of operations more likely to have problems, which may in turn, lead to focused redesign of questionnaires or procedures.

Ultimately, data mining along with other questionnaire evaluation techniques can be used to improve data quality by revising questionnaires and/or data collection and processing procedures. The focus of this report is an analysis of reporting errors related to the total acres operated, specifically, those reported in the census of agriculture.

2. METHODS

The data set for this analysis consists of 496,497 unedited records from the 2002 Census of Agriculture. All records from AZ, AR, CA, CO, FL, GA, IA, MN, NY, NC, PA, SD, UT, WA, WI were selected to create a sample varying across regions and types of agriculture (Figure 3).

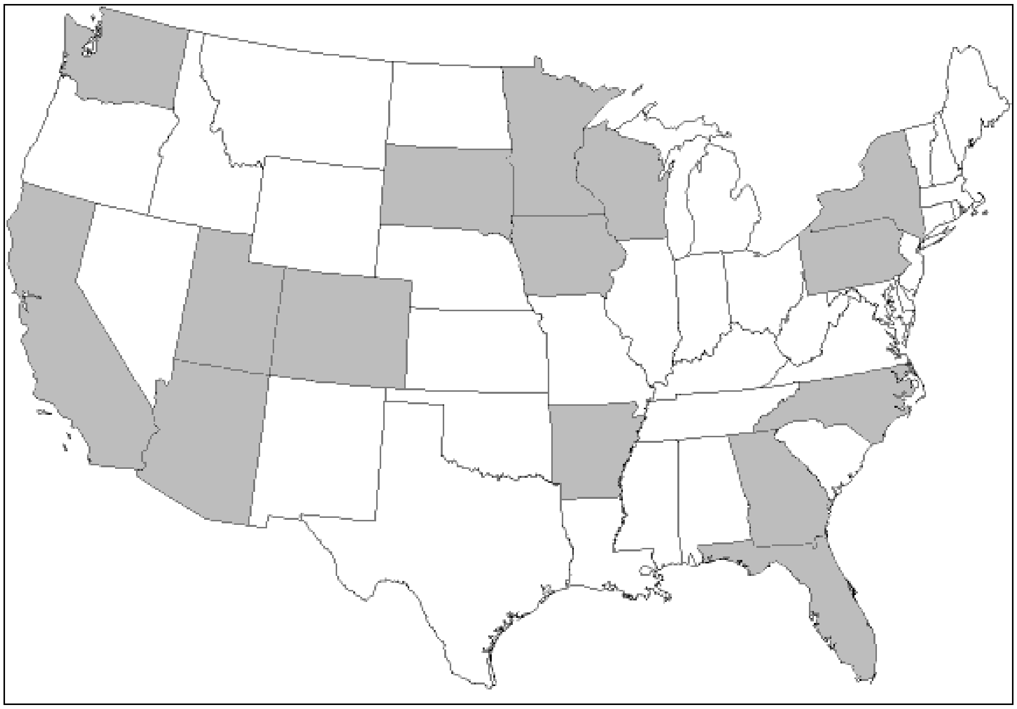


Figure 3. States included in analysis (shaded)

Records included only self administered forms returned by mail and key entered by the National Processing Center (NPC) in Jeffersonville, IN. From the reported data, all of the variables used to derive the two separate total acres operated figures were included. Records for which the values of either total acres operated variables were not keyable (illegible entries, etc.) were excluded from the dataset. In addition, we included a number of variables from the remaining reported data on the form. Some of these variables were collapsed to create presence/absence indicator variables (e.g. any cattle reported, any aquaculture reported, berries, etc.) and others were used as reported. The full list of variables included in the analysis is included in Appendix A.

Defining Reporting Errors

The first step in this analysis was to define reporting errors. Three separate types of errors were examined:

- Total Acres Operated not equal to reported subparts;
- Total Acres Operated missing;
- Equivalent Total Acres Operated figures not equal.

Respondents are required to enter *both* the subitems used to calculate their Total Acres Operated (TAO), *and* the totals. In many cases one or both of the TAO figures was missing and not entered on the form. Also examined was the "error" of item nonresponse where either the TAO (K46 or K798) was missing. In addition, the error of individual TAO figures that were present but not summed correctly using the reported subparts was examined. Finally, both TAOs should be the same (i.e. K46 = K798). However, in many cases, different numbers were entered for the two individual TAOs, logically creating a inconsistency or error. To eliminate small rounding errors, only errors of 10 or more acres were examined. Records were coded using binary variables indicating presence of the above errors. These variables were used as the targets for the classification trees.

Using Classification Trees

The primary data mining technique used in these analyses was the classification or decision tree. In this type of analysis, the full data set is first randomly partitioned into three subsets. These subsets are termed the training, validation, and test sets. In this analysis, the data were apportioned across the three subsets as follows: 40% training, 30% validation, and 30% test. The training dataset is used to construct the initial tree model. This model is then applied to the validation dataset in order to prevent generating a model for the training data that does not fit other data or is unreliable (i.e. overfitting). The validation data are used when pruning the initial tree to generate the final model. Finally, the test set is used to evaluate the model's performance on independent data not used in creating or pruning the model.

A classification tree model is constructed by segmenting the data through the application of a series of simple rules. Each rule assigns an observation to a segment based on the value of one input variable. For example, the segmenting rule may be to divide the dataset into groups, one with records reporting a certain commodity, and one with records that do not report the commodity. One rule is applied after another, resulting in a hierarchy of segments within segments. The rules are chosen to maximally separate the subsegments with respect to the target variable. Thus, the rule selects both the variable and the best breakpoint to maximally separate the resulting subgroups. In other words, the segmenting rule divides records into groups with more and less of the target based on their reports of a commodity, and also selects the amount of that commodity that maximally separates the groups. For categorical variables, the rule will select the groups of categories that maximally separate the groups. The categorical groupings and continuous variable breakpoints are not defined by the researcher but are dictated by the data.

The resulting hierarchy is called a tree, and each segment is called a node. The original segment contains the entire data set and is called the root node of the tree. A node with all its successors is termed a branch of the node that created it. The final nodes are called leaves. Each record in the dataset will appear in one of the tree leaves, and the leaves will collectively contain all records in the dataset. In our analysis, the leaves of interest were those containing a higher proportion of records with the target/error.

Classification trees describe subsets of data and are constructed without any theoretical guidance. Variables are chosen to maximally separate the subsegments, so only one or a few of similar correlated variables (which individually might be related to the target) may appear in the tree. There are several alternative methods for constructing decision trees; our trees were grown using the chi-square approach available in SAS Enterprise Miner, which is similar to the chi-square automatic interaction detection (CHAID) algorithm. There are multiple stopping criteria used to decide how large to grow a decision tree. Generally, these trees were pruned so leaves represented at least 500 records or when adding additional leaves did not markedly improve the overall misclassification rate of the tree as a whole. All trees had similar misclassification rates for the training and validation datasets used to grow the trees and for the test data used to verify reliability of the trees after construction. Since the objective of this analysis was using the trees for error classification, rather than for error prediction, the trees grown were relatively small. By growing larger trees with longer branches, the predictive power of the trees could have been incrementally increased, but the leaves would then have become harder for users to interpret and would have contained fewer and fewer records.

For most of our analyses, the data were oversampled so the target occurred 25% of the time in the root node, if it was less frequent in the original dataset. This is standard practice in tree modeling to allow for enough cases for analysis after multiple branch splits. Because of this oversampling of the records with errors for our analyses, we will report results generally and with respect to the oversampled dataset in the root nodes.

3. RESULTS

3.1 Errors in Summation of Total Acres Operated (K46)

The first type of error examined was that when the parts of a TAO did not sum correctly from its reported subparts. K46 is the sum of: Land Owned (K43), plus land rented from others (K44), minus land rented to others (K45) (Figure 1). These subparts did not sum correctly 8.7% of the time for our dataset; and 5.7% of the time they differed by more than 10 acres.

Using summation errors of more than 10 acres as the target, a decision tree was generated (Figure A-1). In Figure A-1 the full dataset is shown in the box at the top of the tree. The number of records in each of the training and validation sets is shown and the percent of those records with the target of interest (K46 summation errors) is also shown. The training and validation sample sizes are the result of splitting the full dataset into training, validation and test subsets, followed by oversampling the data to bring the percent of the target up (from 5.7%) to 25% of the dataset. Thus, this tree was generated using a training dataset of 45,090 records, 25% of which had summation errors. In each subsequent box in the tree, the number of records in the training and validation datasets and the percent of those with the target after the specified branch split are shown.

In this tree, the first split selected was on land rented to others. The classification tree algorithm systematically examines each variable to determine which maximizes the difference in the number of records with the target when it is selected to split the set of records in the parent box. The software will select the variable (using a chi-square criteria), and the optimum breakpoint in that variable to generate the split for that box. In this case, the first split is on "acres rented to others" with respondents with greater than zero acres rented twice as likely as the full dataset in the root node to have a K46 summation error. Thus the software selected both the variable "acres rented to others" and the optimum value of this variable (greater than 0) to split the records in the root node. The breakpoints do not have to be specified in advance, all possible breakpoints are evaluated and the best one is selected.

Continuing down the right hand side of the tree, of those records renting land to others, those with 48 or more acres of cropland harvested were even more likely to have summation errors. In this case, the variable maximally separating these subsegments was "cropland harvested" and the break point maximizing the split for this variable was 48 acres.

The group of respondents renting no land to others, with K798 summation errors, and reporting more than 673 acres of land owned were also more than twice as likely to have errors as the root dataset.

To summarize, the two resulting branches of interest in this tree end in the following nodes:

Branch 1 (highlighted in blue):

- 1 or more acres of land rented to others (K45) and;
- 48 or more acres of cropland harvested[2].

Branch 2 (highlighted in red):

- 0 acres of land rented to others (K45);
- Parts of the K798 that did not sum to the reported K798 and;
- Land owned (K43) over 673 acres.

[2] One further split is shown in the tree in this branch ("total sales not under production contracts less than $1201"). While the difference is significant, these leaves do not have substantially more records with errors. Due to the small number of cases in these leaves and the small difference between them we did not include these in the discussion.

For the group of records in Branch 1, one possibility is that respondents do not subtract the land rented to others and incorrectly include these acres in their total acres operated. Indeed, of the 28,752 records with summation errors in the dataset, 6537 (22.7%) of them reported some land rented to others (K45) and had a total acres operated number (K46) that was equal to their calculated total acres operated plus their land rented to others.

We also looked at the decision tree made after forcing the initial split on whether there had been a summation error in the K798 to examine errors separately for these two groups. The resulting tree was similar, with land rented to others still the first selected splitting variable. In the subtree with K798 summation errors, those records with 15 or more acres of land rented to others were over twice as likely to have K46 summation errors. For those without K798 summation errors, similar to the previous tree, those with acres of land rented to others (in this case, more than 10 acres) and 50 or more acres of cropland harvested were almost twice as likely to have K46 summation errors.

3.2 Errors in Summation of Total Acres Operated (K798)

The next type of error examined was errors summing the reported subparts of the second TAO (K798). K798 is a sum of the following subparts: cropland harvested (K787), cropland with failed crops or abandoned (K790), cropland in summer fallow (K791), cropland idle (K1062), permanent pasture (K796), woodland pasture (K794), cropland pasture (K788), woodland (K795), and all other land (K797). These nine reported subparts did not sum correctly 17.2% of the time for our dataset; 11.2% of the time, they were different by more than 10 acres.

Again, errors in summation of 10 or more acres were used as the target. The first split in the decision tree generated for this target (Figure A-2) divided the records into groups based on acres of cropland pasture, with the optimum breakpoint at 16 acres. The two branches of interest end in the following nodes:

Branch 1 (highlighted in blue):

- Cropland pasture less than 16 and;
- Cropland in summer fallow greater than 3 acres.

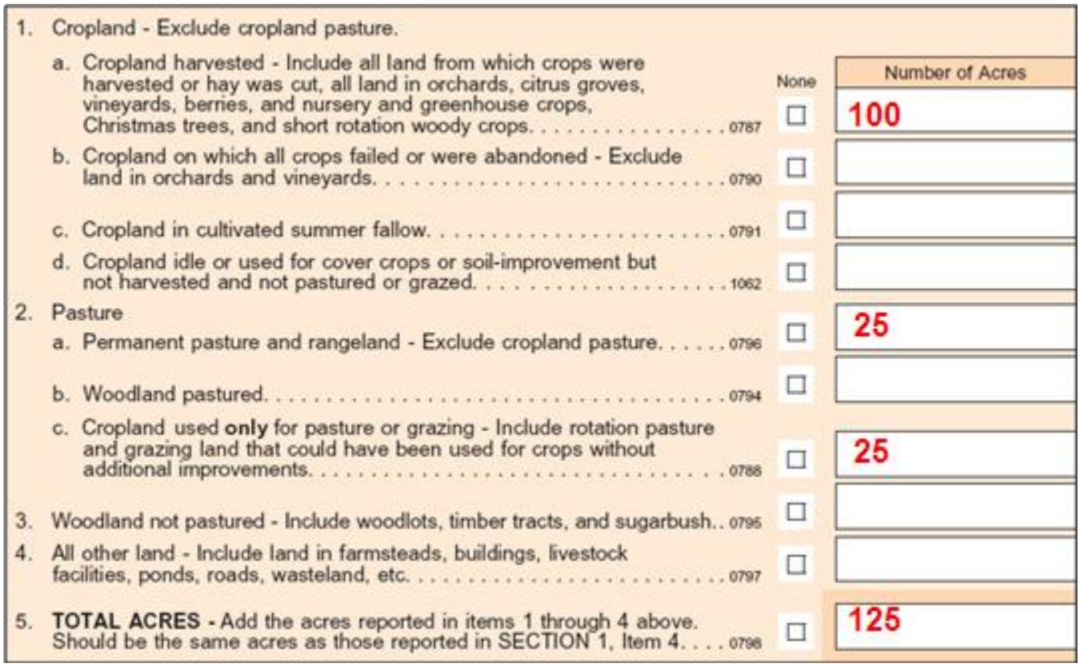
Branch 2 (highlighted in red):

- Cropland pasture greater than or equal to 16 acres and;
- Permanent pasture greater than or equal to 9 acres.

Those in Branch 1 were twice as likely as the root dataset to have summation errors, while those in Branch 2 were almost 3 times as likely to have errors. Branch 2 suggests that respondents may be incorrectly reporting the same acres in both cropland pasture and permanent pasture.

In Branch 1, it is possible that cropland in summer fallow is being erroneously reported twice, but the overreporting is not limited to another single land use category.

In order to look further at possible causes of repeated erroneous acreage reporting in this section, new indicator variables were created. If the errors reported in any of the types of land that are subparts of the K798 total acres equaled the K798 summation error, this was flagged as a double counting error. For example, looking at *Figure 4*, if an operation reported 100 acres of cropland harvested, 25 acres of permanent pasture, 25 acres of cropland pasture and 125 acres in K798, this was a double counting error.



	None	Number of Acres
1. Cropland - Exclude cropland pasture.		
a. Cropland harvested - Include all land from which crops were harvested or hay was cut, all land in orchards, citrus groves, vineyards, berries, and nursery and greenhouse crops, Christmas trees, and short rotation woody crops. 0787	□	100
b. Cropland on which all crops failed or were abandoned - Exclude land in orchards and vineyards. 0790	□	
c. Cropland in cultivated summer fallow. 0791	□	
d. Cropland idle or used for cover crops or soil-improvement but not harvested and not pastured or grazed. 1062	□	
2. Pasture		
a. Permanent pasture and rangeland - Exclude cropland pasture. 0796	□	25
b. Woodland pastured. 0794	□	
c. Cropland used **only** for pasture or grazing - Include rotation pasture and grazing land that could have been used for crops without additional improvements. 0788	□	25
3. Woodland not pastured - Include woodlots, timber tracts, and sugarbush. . 0795	□	
4. All other land - Include land in farmsteads, buildings, livestock facilities, ponds, roads, wasteland, etc. 0797	□	
5. **TOTAL ACRES -** Add the acres reported in items 1 through 4 above. Should be the same acres as those reported in SECTION 1, Item 4. 0798	□	125

Figure 4. Example of K798 Error Due to Double Counting Acres

This is a conservative measure since any acres that were reported in more than one item, but did not *exactly* match the reporting error (e.g. partial acres double counted, double counted acres split between items, etc.) were not classified as being double counted. In the group of records with a K798 summation error (over 10 acres), nearly one third of these records (31.5%) had a double counting error. Obviously, this is a major source of reporting errors in this section.

To examine this within a classification tree, an initial split was forced using the double counting error variable. The resulting tree is shown in Figure A-3. The branch of the tree without double counting errors included records reporting 17 or more acres of cropland pasture, 3 or more acres of permanent pasture, and less than 70 acres of cropland harvested. Thus even excluding the double counting errors, errors in summing K798 are still related to a few of the types of land being reported in this section.

3.3 Missing "Total Acres Operated" (K46)

K46 is the total acres operated derived by adding owned land, plus land rented from others, minus land rented to others. Respondents are supposed to report both the parts and the total on the form. K46 was missing for 11.2% of the records in the complete dataset and this was the target for this analysis. The most immediately striking finding was that many of the splits in the tree were not the result of splitting based on a value of another variable, but instead split between records with and without item non-response for a variable.

Using the decision tree technique, we generated the tree, shown in Figure A-4. The two branches of interest for this tree ended in the following nodes:

Branch 1 (highlighted in blue):

- Records with a reported value in K798
- Acres rented from others missing/unreported;
- Acres of land owned zero or unreported.

Branch 2 (highlighted in red):

- Records with K798 missing;
- Acres of land rented from others missing/unreported.

Respondents in both branches were over 3 times as likely to have an error as the root dataset. This likely indicates that those respondents not reporting K46 do not know or are unwilling to report their total acres. They do not (or cannot) report their total acres rented, or owned and rented, therefore, they do not report their total acres. It does not appear to be related to any specific type of land use, since none of the land use variables appear in the tree.

3.4 Missing "Total Acres Operated" (K798)

K798 is the total acres operated derived by summing acres operated by individual use (i.e. cropland, pastureland, other land). In 11.9% of the records, this number was not entered on the form. For this analysis, we set our target as records with K798 missing and generated the tree in Figure A-5.

The branches of interest for this tree ended in the following nodes:

Branch 1 (highlighted in red):

- Records with K46 missing, and
- Acres of All other land (K797) missing.

Branch 2 (highlighted in blue):

- Records with K46 missing, and
- Acres of all other land (K797) zero,

Branch 3 (highlighted in yellow):

- Records with K46 reported,
- Acres of all other land (K797) zero or missing,
- Acres of cropland harvested (K787) 0 or missing, and
- Acres of cropland idle (K1062) missing.

Respondents in Branches 2 and 3 had over twice as many errors as the full dataset in the root node, while Branch 1 had more than 3 times the number of errors.

This may indicate that respondents not reporting K798 are NOT those that did not want to sum their acreage, but instead were not able (or were unwilling) to report the subfigures of "other land" or "cropland harvested" and therefore did not sum this to obtain the TAO (either the K798 or the K46). This has implications for how editing may be done for the TAO. If the TAO is not reported by the respondent, simply summing the reported subtotals likely omits figures that the respondent did not report (but possibly should have) and may lead to an underestimate of TAO. This is true both for K798 and K46.

Indeed, there were 39,862 records in this dataset with K798 missing and either 0 or missing cropland harvested (K787). Of these, 11,763 (29.5%) had 1 or more acres of crops reported elsewhere on the report form.

3.5 Total Acres K46 and K798 Unequal

Both of the TAO figures (K46 and K798) reported on the census should be exactly the same. In the 2002 Census, these items appeared on facing pages of the questionnaire, and it is clearly stated on the form that they should be the same acres. In all cases where both were reported and they are not equal, there is a reporting error of some kind. In our dataset, 13.4% of records had nonzero differences between the two TAO, and 8.8% of the records had absolute differences between the 2 TAO of more than 10 acres.

As discussed above in the analysis of either of these figures being missing, in many cases, it appears that this may be the result of the respondent underreporting their acreage for one or more types of land. Therefore, for the following comparison of reported TAOs, we will include only those cases where both numbers were reported by the respondent. Records with one or both missing TAO, were excluded from this tree.[3]

Records where K46 did not equal K798 were more likely than the population as whole to have either (or both) K46 and K798 summation errors. K798 summation errors were more common than K46 errors. This is not surprising as many more sub-items are included in the K798 summation. Therefore, it is no surprise that a variable indicating summation errors for TAO is selected early in the decision tree splits. As shown in the decision tree generated (Figure A-6.), those respondents with summation errors (of 10 or more acres) in K46 are more than twice as likely as the entire population to have differences between their K46 and K798 reported values.

The decision tree generated for this error is shown in Figure A-6 and included the following branches of interest:

[3] Initially, we calculated a new TAO for any records which had some of the subparts of the TAO reported, but not the total. However, as discussed previously, many of these totals were missing because some of the required subparts to the TAO were also missing. Therefore, calculating a TAO from the parts generated many more errors and likely often underestimated the true TAO.

Branch 1 (highlighted in red):

- Errors in summation of K46 > 10 acres

Branch 2 (highlighted in blue):

- Error in summation of K46 ≤ 10 acres (or no error)
- Acres of Land Rented > 10 acres
- Acres of Cropland Harvested > 10

Another tree (not shown) was generated excluding the two summation error variables, yielding similar but slightly different results. Similar to the later splits in the tree above, the first split grouped respondents with more than nine acres of land rented to others reported, which had about twice as many respondents with TAO not in agreement as the original dataset. Continuing down this branch of the tree (within this subgroup), records with more than 10 acres of cropland harvested and less than $6012 in total sales, or less than 10 acres of cropland harvested and more than 26 acres of cropland used for pasture both had more than twice the percentage of records with TAO unequal.

It is not surprising that records with summation errors in one of the TAO figures would also report inconsistently, as in Branch 1 of this tree. This tree also includes a group with land rented to others and cropland harvested, similar to the tree for the K46 summation errors.

4. DISCUSSION

The use of decision trees to examine the two series of total acres operated questions (K46 and K798) provides insights into errors that might not have been discovered using traditional statistical techniques. A similar analysis for this data could have been done using logistic regression. However, such an analysis would require that all independent variables as well as all possible interaction effects be specified, programmed, and summarized. The resulting regression equation would have made it much more difficult to see the key characteristics of the subsets of records with the highest rates of errors than eyeing a simple classification tree. Using classification trees is a clear cut method for identifying key characteristics related to a given target. Unlike a regression analysis, the trees also provide the optimum break points for each variable by collapsing nominal or ordinal variables into fewer categories or by identifying cutoffs within continuous variables that result in the greatest dichotomies. For example, when a variable expressed in acres is selected as a splitting variable, the tree will also select the specific cut off (e.g. 47 acres, 674 acres) which will provide the greatest separation between the subsequent leaves. Importantly, in many cases it is also the missing values in other variables which were highly predictive of errors – decision trees easily incorporate missing data as valid values. In a regression model, most of these data would not have been used.

The rationale for using decision trees, like many other “data mining” techniques, was to help narrow the scope of our examination of a large dataset. The subgroups identified in the trees often suggested additional analysis of the data and prevented fruitless exploration into differences between respondents where there were none.

One important thing that is not apparent from the decision trees alone are the variables included in the analysis that never appear in the trees. For example, there were no instances where characteristics of the operator (such as age or gender), the types of commodities raised or the location (i.e. state) were important in finding subgroups of records more likely to make the errors studied. Therefore, it is assumed that errors are evenly distributed across these dimensions. This is a positive finding as concentration of these errors in specific states or in operations with certain types of commodities might increase the possibility of systematic biases being introduced by these errors. It also suggests that respondents chosen for testing of modifications to these questions can be chosen with less concern for their location or other characteristics that did not appear in our trees.

Since the interest was in using the trees for classification rather than prediction of errors, the trees we grew were relatively small. By growing the trees larger with longer branches, models incrementally increase in predictive power, but the leaves become harder to interpret and contain fewer and fewer records.

5. RECOMMENDATIONS

1. Review edits used for these items.

The results of this analysis can be used for several follow on activities. First, the edits that involve these items should be reviewed. For example, actions taken to impute missing total acres operated figures may need to take into account the idea that missing figures generated by summing reported TAO subparts may understate the true TAO. If other reported figures on the census report form (i.e. individual crop acres) sum to a greater total than a calculated TAO, it may be a reasonable action to impute this larger figure for the TAO rather than performing the simple summation of the reported TAO subparts.

Another example of an edit suggested by this analysis is for errors in summation of the K46 total acres operated. If a respondent reported land rented to others, and their K46 is overstated by this figure, then the respondent likely did not subtract out the land they rented to others. The figure most likely in need of editing is the K46 TAO, and not the relevant subparts of the TAO.

2. Use these classification trees as input to redesign these questions and data collection procedures.

The classification trees presented in this report should be reviewed by the next Census Content Test Team. The results of this analysis should be used to suggest potential areas for changes in the questionnaire, the types of respondents to include in testing, and questions they should be asked in tests of any changes. This would most likely increase the information gained in small scale cognitive or pre-testing over the traditional guidelines to simply include a wide variety of types of respondents. For example, respondents reporting land rented to others and with over 47 acres of cropland harvested were much more likely to sum K46 incorrectly. These respondents should be included in testing and specifically asked about reporting for rented land. Testing alternative formats of these questions with changes to the rented land question (perhaps in wording or format) should also be considered.

Similarly, those respondents with summation errors in the K798 total acres operated often appear to erroneously report the same acres in multiple items in this section. Perhaps changes to the format or instructions in this section could be considered to clarify that acres should be reported only once. And certainly, respondents with multiple types of land (particularly pasture) must be included in any tests of alternative formats.

In addition, it may be helpful for enumerators to be aware of those respondents most likely to misreport. Enumerator training can include an emphasis on the proper verification of data from these respondents, how to provide appropriate assistance for these respondents, and what notes may be helpful for analysts editing and summarizing these data.

3. Look for other areas where data mining techniques can be applied in NASS surveys and the census of agriculture

The use of classification trees is a general technique that can be useful in many areas in survey data collection where large data sets, such as the census of agriculture, are available. Certainly, the same approach we took in these analyses could be repeated for other reporting errors or targeting items with a high incidence of edits. NASS is currently using this technique in other areas such as identifying records most likely to be non-farms (Garber, 2008), and creating non-response weighting groups (Cecere, 2008). However, many other applications of this technique can be explored. Another potential use could include identifying characteristics of records more likely to be survey non-respondents either to evaluate (or remediate) potential non-response bias in our estimates or to alter data collection techniques for these operations. Work on developing these models is currently underway. In addition, classification trees are just one of many “data mining” techniques that are useful in analysis of large data sets. Where large survey datasets are available, other data mining techniques should be explored for their potential to improve NASS’s survey processes and data.

6. REFERENCES

Cecere, W. (in preparation) “2007 Census of Agriculture Non-response Methodology.” US Department of Agriculture, National Agricultural Statistics Service. Research and Development Division Report.

Cohen, S.B., DiGaetano, R., and Goksel, H. (1999). "Estimation Procedures in the 1996 Medical Expenditure Panel Survey Household Component," Agency for Health Care Policy and Research, MEPS Methodology Report No. 5, AHCPR Publication No. 99-0027, Rockville, MD.

deVille, B. (2006) “Decision Trees for Business Intelligence and Data Mining Using SAS Enterprise Miner.” Cary, NC:SAS Institute, Inc.

Garber, S.C. (in preparation) “Census Mail List (CML) Trimming using SAS Data Mining.” US Department of Agriculture, National Agricultural Research Service, Research and Development Division Report.

Schouten, B. and de Nooij, G. (2005) “Nonresponse adjustment using classification trees,” Discussion Paper 05001, Voorburg/Heerlen: Statistics Netherlands.

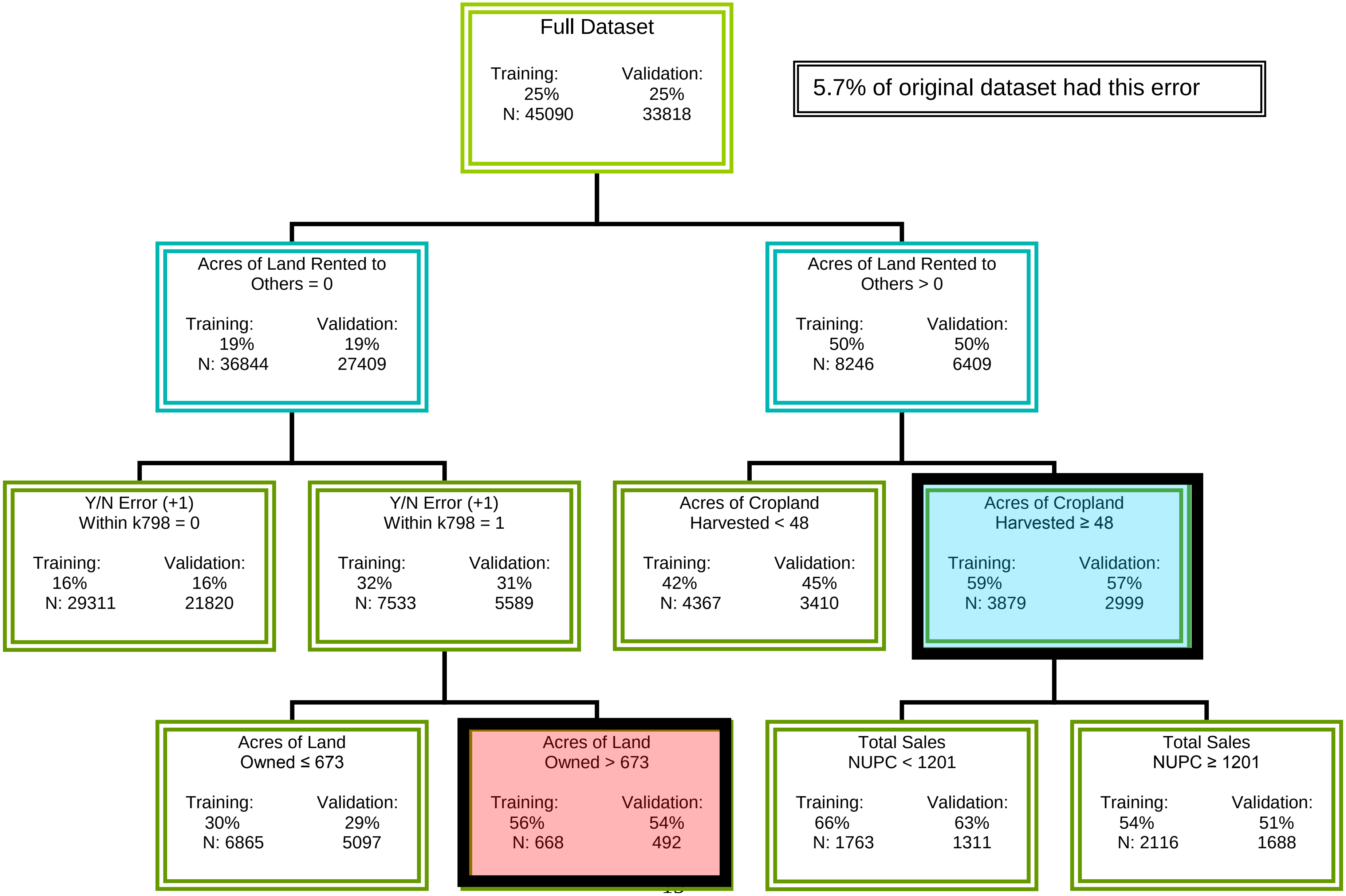
Scime, Anthony and Murray, Gregg R. (2007). “Vote Prediction by Iterative Domain Knowledge and Attribute Elimination,” International Journal of Business Intelligence and Data Mining, 2(2), 160-176.

Appendix A. Input Variables Used in Analysis

Census Variable Name	= Variable Description
K43	= Acres of Land Owned
K44	= Acres of Land Rented from Others
K45	= Acres of Land Rented to Others
K46	= Total Acres Operated - Reported
K787	= Acres of Cropland Harvested
K788	= Acres of Cropland Used for Pasture
K790	= Acres of Cropland for Which All Acres Failed
K1062	= Acres of Cropland Idle or Used for Cover Crops
K791	= Acres of Cropland in Summer Fallow
K794	= Acres of Woodland Pasture
K795	= Acres of Woodland Not in Pasture
K796	= Acres of Permanent Pasture & Rangeland
K797	= Acres of All Other Land
K798	= Total Acres - Reported
K943	= Machinery and Equipment Value
K923	= Principal Operator—residence on Place
K925	= Principal Operator—age
K926	= Principal Operator—sex
K927	= Principal Operator—Spanish Origin
K928	= Principal Operator—Principal Occupation
K929	= Principal Operator Days Worked off Farm
K930	= Principal Operator—year Began Operation
K1701	= Principal Operator –Race, White
K1702	= Principal Operator –Race, Black
K1703	= Principal Operator –Race, American Indian
K1704	= Principal Operator –Race, Native Hawaiian or Pacific Islander
K1705	= Principal Operator –Race, Asian
K1567	= Partnership Registered under State Law? Y/N
K1568	= Any Fertilizer or Chemicals--Y/N
K1569	= Acres on Which Manure Was Applied
K1573	= Any Migrant Workers--Y/N
K1574	= Number of Women Operators
K1575	= Number of Operators
K1576	= Any Hired Manager? Y/N
K1577	= Principal Operator—# of Persons Living in Household
K1578	= Principal Operator—% of Income from the Operation
K1602	= Computer used Y/N
K1603	= Internet access Y/N
K1608	= Households sharing in net farm income
K1578	= % of Household income from operation
K1080	= Possible duplicate Y/N
K1086	= Have other farm Y/N
K1050	= Ag on Indian Reservations Y/N
K60	= State
K1021	= Acres of all Hay and Forage
K1022	= Acres of all Irrigated Hay and Forage Harvested

K803	= Total Cattle and Calf Inventory
K815	= Total Hog and Pig Inventory
K1347	= Total Sales – Not Under Production Contracts (NUPC)
K1069	= Acres of Certified Organic Farming
K941	= Hired Workers Working Less Than 150 days
K942	= Hired Workers Working Greater Than or Equal to 150 days
K1314	= Total $ - Under Production Contract
TCL	= Cropland Acres
TFPE	= Total Production Expenses
TVPG	= Total Value of Products Sold + Government Payments
VEGA	= Sum Acres of Vegetables
RCROP	= Sum of Cropland Harvested
RBERA	= Sum of All Reported Berry Acres
RPLTINV	= Sum of Poultry Inventory Data
RSUMFA	= Sum of All Reported Fruit Acres
SHEP	= Sheep and Lamb Indicator
TCTA	= Total Citrus Acres
TOTOTLVK	= Other Livestock Animals
REXP	= Reported Sum of Expenditures
NGFS	= Nursery Indicator
OAQ	= Aquaculture Indicator
CTWDYCP	= Xmas Trees and Short Rotation Woody Crops
K684	= Total Government Payments
FARMTYPE	= Farm Type
TENURE	= Operation Farm Tenure(1=full owner, 2=part owner, 3=tenant)
	= Summation in K46 error over 10 acres
	= Summation in K798 error over 10 acres
	= Difference between K46 and K798 over 10 acres
	= K46 Missing Y/N
	= K798 Missing Y/N
	= Double Counted Acres in K798

Figure A-1. Decision Tree: Errors in Summation over 10 Acres for K46



5.7% of original dataset had this error

Figure A-2. Decision Tree: Errors in Summation over 10 Acres for K798

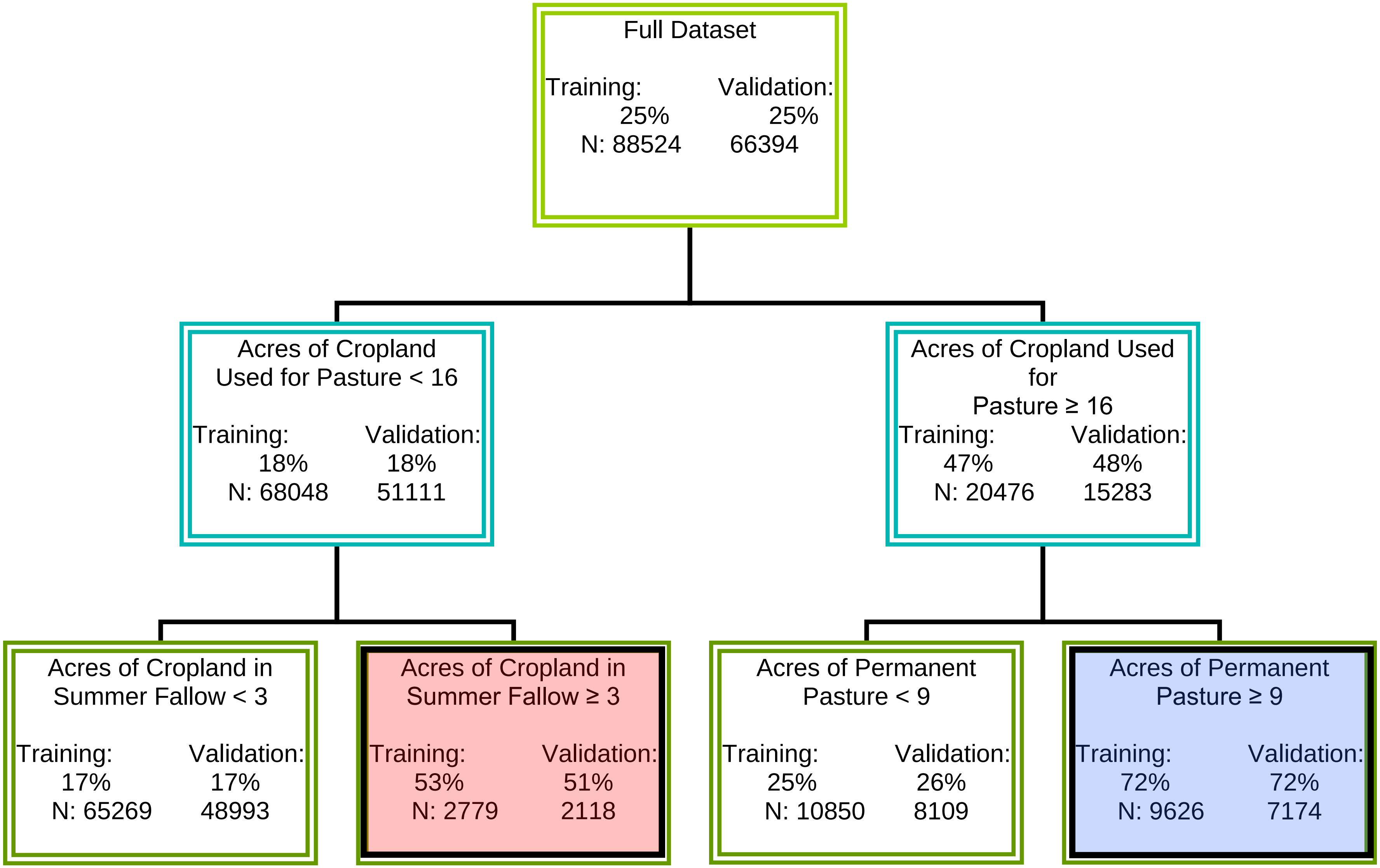


Figure A-3. Decision Tree: Errors in Summation over 10 Acres for K798, initial split on Exact Double Counting Error

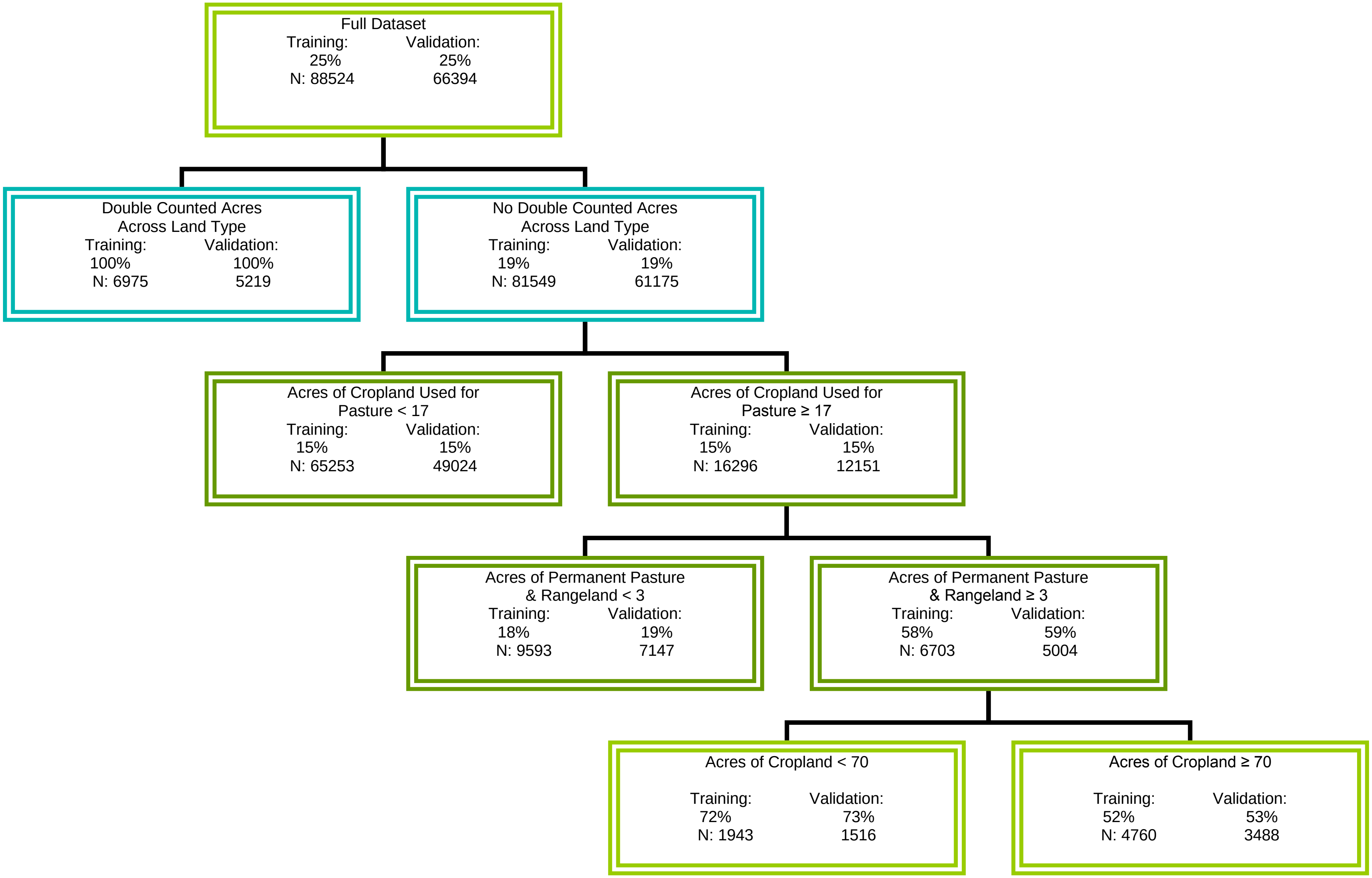


Figure A-4. Decision Tree: K46 Missing

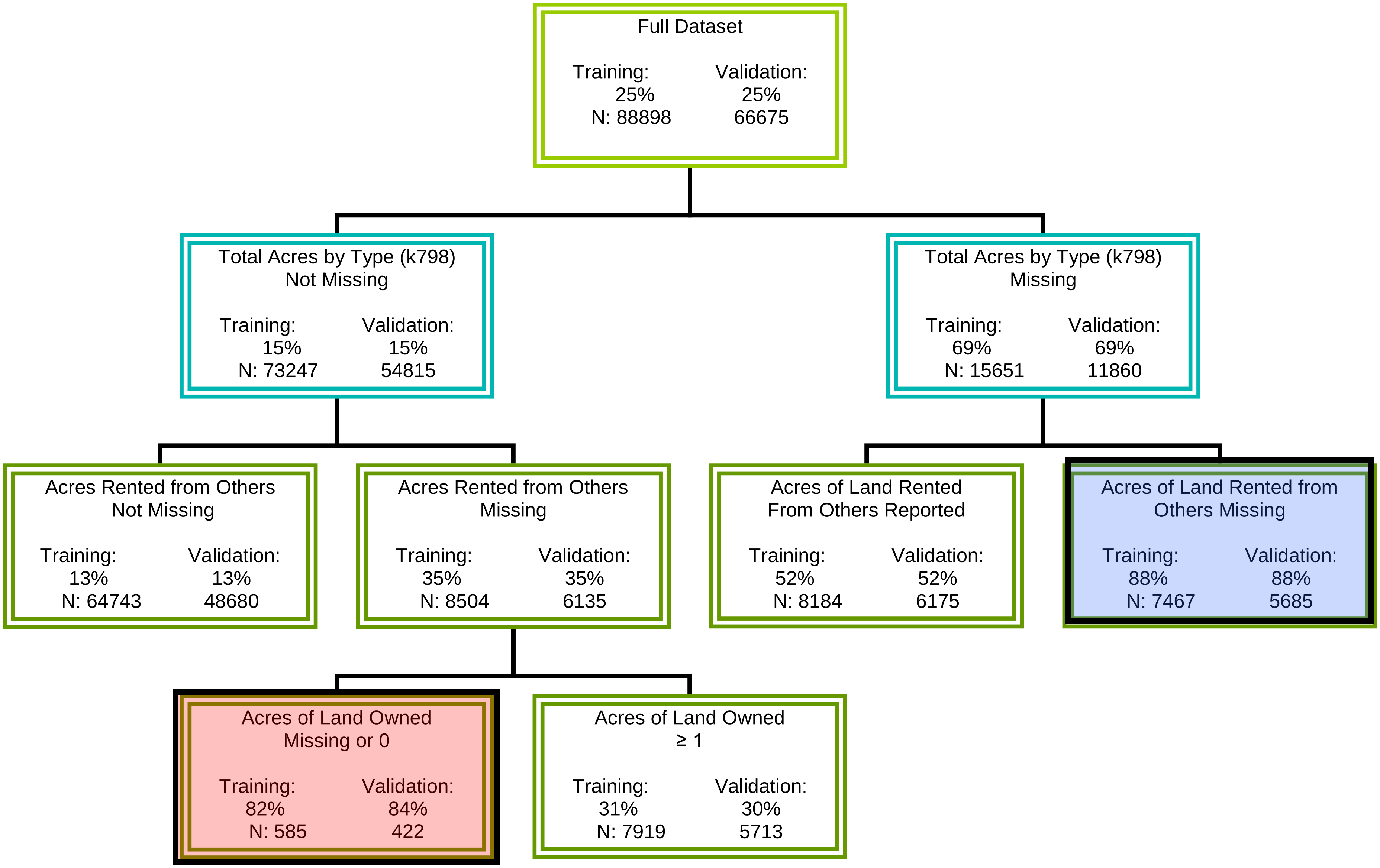
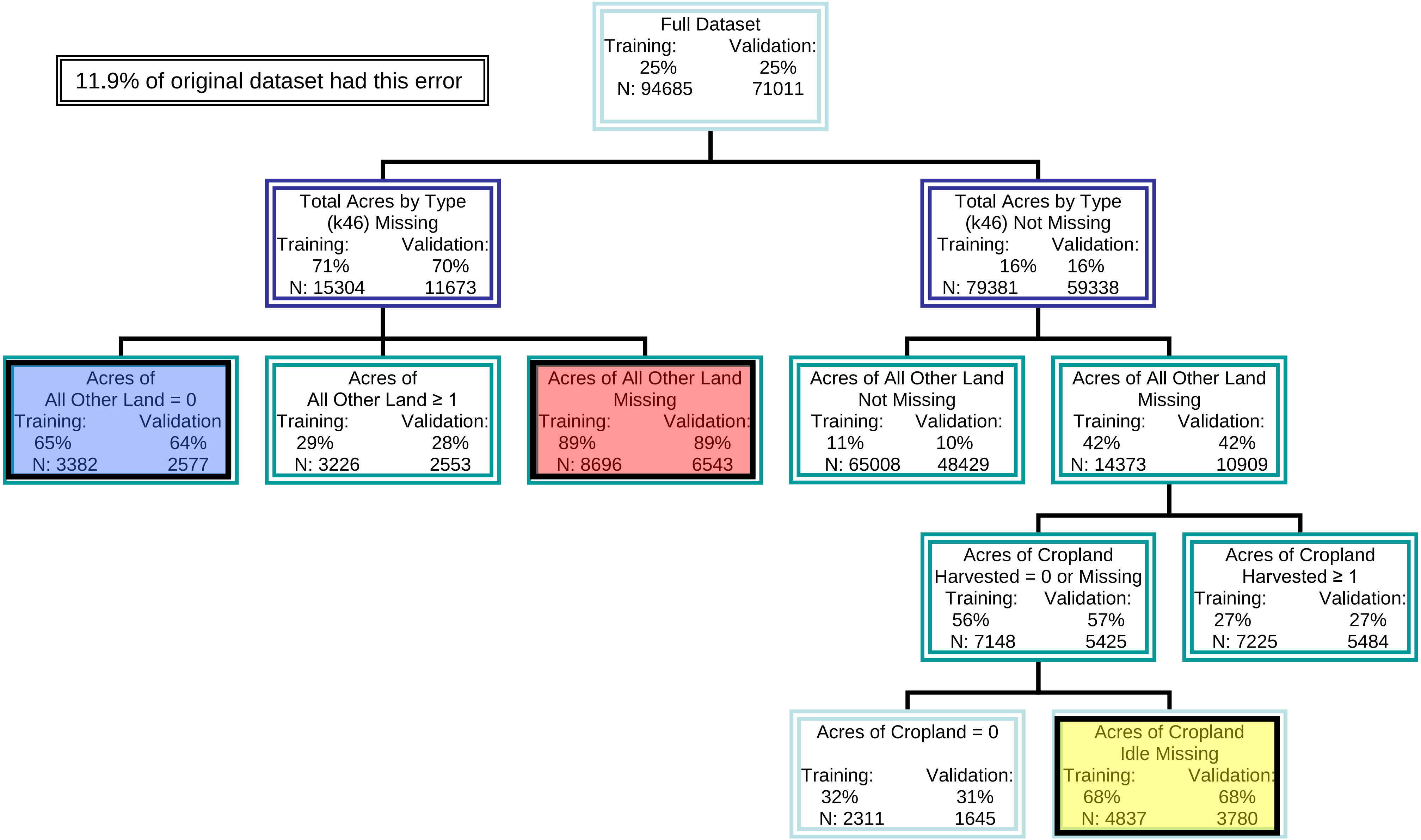
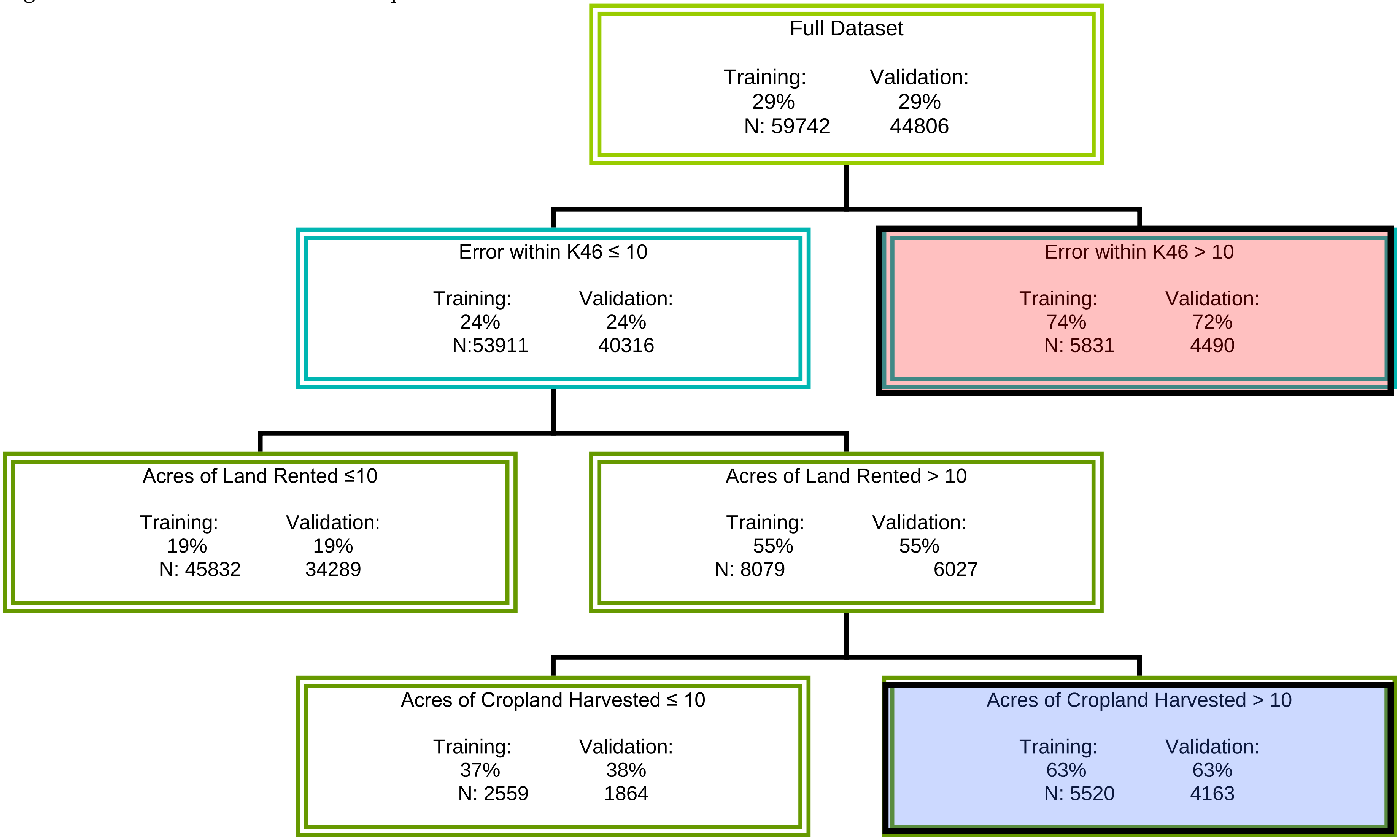


Figure A-5. Decision Tree: K798 Missing



11.9% of original dataset had this error

Figure A-6. Decision Tree: K46 Not Equal to K798


Full Dataset
Training: 29% N: 59742
Validation: 29% 44806
Error within K46 ≤ 10
Training: 24% N:53911
Validation: 24% 40316
Error within K46 > 10
Training: 74% N: 5831
Validation: 72% 4490
Acres of Land Rented ≤10
Training: 19% N: 45832
Validation: 19% 34289
Acres of Land Rented > 10
Training: 55% N: 8079
Validation: 55% 6027
Acres of Cropland Harvested ≤ 10
Training: 37% N: 2559
Validation: 38% 1864
Acres of Cropland Harvested > 10
Training: 63% N: 5520
Validation: 63% 4163