Identifying Person Duplicates of Short Geographic Distance by Computer Matching

Thomas Mule, U.S. Census Bureau, Washington, D.C.

Abstract

The Census Bureau conducted evaluations of person duplication in Census 2000. Duplicates of short geographic distances were identified by both clerical and computer matching. The evaluations showed that for these short distance duplicates that the computer matching algorithms were not able to find all of the duplicates identified by the clerks. However, the computer matching algorithms in the previous evaluations were primarily developed to identify duplicates of longer distances. This report analyzes the potential of computer matching when the focus is on short distance duplicates. I used the Bureau's record linkage software to do the computer matching. Using SAS®, I was able to compare the computer matching results to the clerical results. First, I attempted to identify groups of links with high concentrations of true duplicates. I used Enterprise Miner® to generate decision trees for several approaches and compared their results. Second, I analyzed clerical duplicates that were not identified by the computer matching to try to identify any patterns in these cases.

Introduction

In several evaluations, the Census Bureau examined person duplication in the census. One was part of the evaluations for the Executive Steering Committee on A.C.E. Policy (ESCAP II) decision made in October 2001 to not use adjusted estimates from the Accuracy and Coverage Evaluation (A.C.E.) survey. This analysis focused on identifying person duplication outside of the search area for A.C.E. The search area for A.C.E. was typically a block cluster, which is a contiguous group of census blocks with approximately 30 housing units. As a benchmark, Mule (2001) showed that this matching algorithm was able to identify only 37.8 percent of the duplicates found by A.C.E. clerical matching.

A second was the Further Study of Person Duplication (FSPD). See Chapter 5 of Kostanich 2003, Mule 2002 and Fay 2002 for more information. We again identified duplicates to help produce the coverage estimates of the census for the A.C.E. Revision II. The FSPD made more use of computer matching methods and developed a new methodology to assign probability of duplication for exact matches. Again, the analysis focused on identifying duplication outside of the search area. As a benchmark, Mule (2002) compared the duplicates identified within the search area to those identified by the A.C.E. clerical matching. His overall analysis showed that the FSPD was able to identify 65 percent of the duplicates identified by A.C.E. clerks. This analysis showed that the FSPD methodology was more efficient at identifying whole household and partial household duplication and was less efficient at identifying one person being duplicated between different housing units.

This research examines using computer matching to identify duplicates within the A.C.E. cluster area. The Census Bureau has not focused on how we can use computer matching techniques to identify duplicates of a short distance. I hypothesize that we can make more use of computer matching when identifying duplicates of a short geographic distance. I believe we can do this because there are fewer coincidental agreements in a shorter geographic area. I will compare our computer matching results to the duplicates identified by the A.C.E. clerical staff.

This analysis focuses on detecting duplicates to *other* housing units. It does not focus on duplication to group quarters or detecting duplicates in the same housing unit. I would want to investigate duplicates to the same housing unit separately because of twins and the likelihood of the last names being the same. The A.C.E. clerical staff did not search for duplicates between the housing unit population and the group quarters population.

Background and Methods

Duplicate Search by A.C.E. Clerical Staff

In order to produce population estimates using dual system estimation, the A.C.E. selected a probability sample of enumerations from the census. The A.C.E. attempted to determine for each sample enumeration if it was a correct or erroneous enumeration for April 1, 2000. As part of this determination, the A.C.E. clerical staff examined the enumerations in the search area to determine if any of the sample cases were enumerated more than once. This analysis focused on duplicates within the cluster and does not include any duplicates to the one ring of surrounding blocks.

The clerical staff identified 6,234 person duplicates to an enumeration in another housing unit within the A.C.E. clusters. This analysis shows how well computer matching can do at isolating these cases.

Computer Matching

I matched the Enumeration sample records in the A.C.E. clusters against all of the enumerations within each cluster using the Census Bureau's BigMatch software (Yancey 2002) that did one-to-many matching. Six characteristics common to both files, called matching variables, were used to link the records. Matching parameters were associated with each matching variable that measure the degree to which the matching variables agree between the two records, ranging from Full Agreement to Full Disagreement. The measurement of the degree to which each matching variable agreed was called the variable match score. If a field was blank on one or both records, then a score of 0 is returned. The overall match score for the linked records was the sum of the variable match scores.

The matching variables were first name, last name, middle initial, month of birth, day of birth, and computed age. Census 2000 was the first census to have the name fields captured. Computed age was determined by the responses to the age and year of birth fields on

the form. The matching variables and parameters are given in Table 1. The agreement weight and the disagreement weight are the matching parameters of each variable. I used standard matching parameters. The relationship of the agreement and disagreement parameters translated into the match score for each variable. I used the Jaro-Winkler string comparator to compare the first name and last name fields. This allowed me to quantify the partial agreement in these two fields. The ages were compared using an algorithm that allows slight variations to receive the full agreement score. All other variables were compared exactly. For example, the full agreement value for first name was 2.1972; whereas, the full disagreement match score was -2.1972. "Steve" compared with "Steven" generates a partial agreement score of 2.04. The sum of the variable match scores was the total match score. When the match score was 9.4006, this indicated full agreement of all variables. A match score of -9.4006, on the other hand, indicated full disagreement.

Table 1: Parameters for Computer Matching

Matching Variables	Type of Comparison	Matching Parameters		Match Score	
		Agreement Weight (m)	Disagreement Weight (u)	Agreement ln(m/u)	Disagreement ln(1-m/1-u)
First Name	String	0.9	0.1	2.1972	-2.1972
Last Name	String	0.9	0.1	2.1972	-2.1972
Middle Initial	Exact	0.7	0.3	0.8473	-0.8473
Month of Birth	Exact	0.8	0.2	1.3863	-1.3863
Day of Birth	Exact	0.8	0.2	1.3863	-1.3863
Computed Age	Age	0.8	0.2	1.3863	-1.3863
Total				9.4006	-9.4006

The search for duplicate links was limited to those pairs that agree on certain identifiers or blocking criteria. Blocking criteria were sort keys and were used to increase the computer processing efficiency by searching for links where they were most likely to be found. Using multiple sets allowed us to identify more duplicates.

I used 2 sets of blocking parameters in this matching:

1. Cluster Number, First Initial of First Name and First Initial of Last Name,
2. Cluster Number, Month of Birth and Day of Birth

Analysis

For this analysis, I required that the first and last name fields on both records be filled. The A.C.E. required a census enumeration to have sufficient information to be eligible for the clerical duplicate search. Part of the sufficient information was a first and last name. Since I used the A.C.E. clerical data as a benchmark, I analyzed links generated by BigMatch where both the first and last names were filled. I started by examining the links with an overall match score greater than 0. There were 17,047 person links that met this criteria.

Of the 17,047 links, the A.C.E. clerks identified 5,506 (32 percent) of them. These 5,506 links give me a detection rate of 88 percent (5,506 of the 6,234) of the duplicate links identified by the A.C.E. clerks. Figure 1 shows two lines. The first is the *density*, the percent of links for an *overall match score* that were identified by the clerks. The second is the *cumulative density*, percent of links for an *overall match score or higher* that the clerks identified. For example, of those links with a score greater than or equal to 4.5, the cumulative density is approximately 90 percent. Approximately 90 percent of these links were designated to be duplicates by the clerical staff. The other line shows that links with a score of 4.5, the density of duplicates was just fewer than 80 percent.

The cumulative density appeared to have 3 break points. The first midpoint was at 4.4, the second was at 1.4 and the third was at 0. I decided to focus on the links greater than the second break midpoint, 1.4 for our data mining analysis. Table 2 shows the trade off of making this decision. While the cumulative density has increased from 32 percent to 69 percent, the price is that the detection rate has slipped from 88 percent to 83 percent (5,151 of 6,234).

Table 2: Cumulative Density and Detection

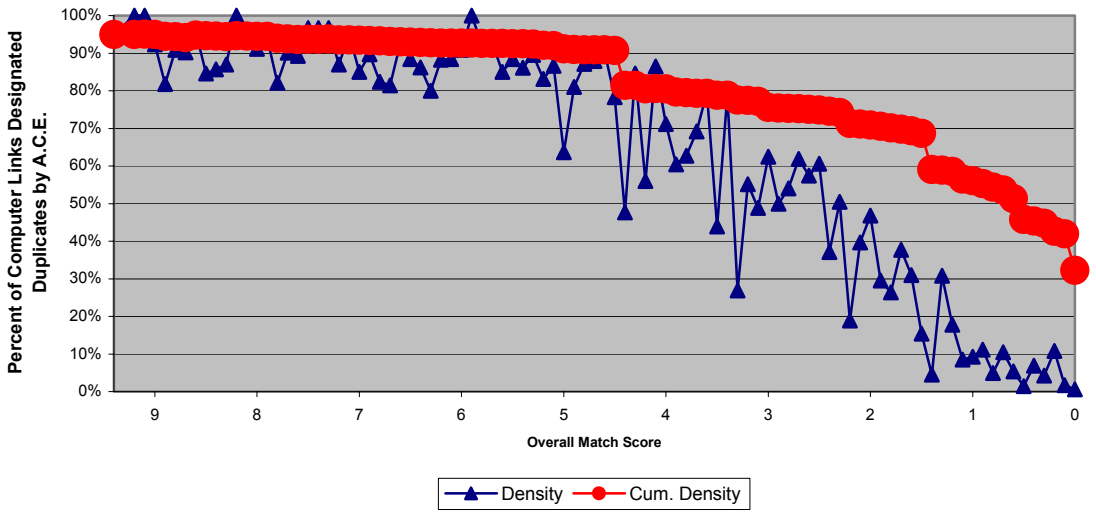
	Overall Match Score ≥ 0	Overall Match Score ≥1.45
Total Computer Links	17047	7500
Computer Links Identified as Duplicates by Clerks	5506	5151
Cumulative Density	32%	69%
Percent of the 6,234 A.C.E. Duplicates Detected by Computer Matching	88%	83%

Using Enterprise Miner® to Identify Efficient Groupings of Duplicates

Using a cutoff of 1.45, I have isolated links where 69 percent are duplicates. I would like to identify groups of these links that have high densities. If a telephone call or field visit is required to resolve the potential duplication, I would like to identify groups which represent real duplicates so we are maximizing our resources.

Table 3 shows four models explored to make these groupings. I decided to compare using just the overall score to using the six results as individual predictors to see if they can better identify groupings. Also, Mule (2002) showed the computer matching was more

Figure 1: Density and Cumulative Density of A.C.E. Duplicates by Overall Match Scores



efficient when there was multiple links between the units. I decided to use the number of links as another variable to see if that improved the results.

Table3: Models for Grouping Duplicates

Model	Matching Output	Number of Links Between Units Used in Model
1	Overall Match Score	No
2	Results of 6 Matching Variables	No
3	Overall Match Score	Yes
4	Results of 6 Matching Variables	Yes

Enterprise Miner® has many tools for data mining analysis including decision trees, neural networks and regression analysis. I decided to use decision trees because of their explicability. Each analysis generated a tree showing how the splits were made and how the final groupings were formed.

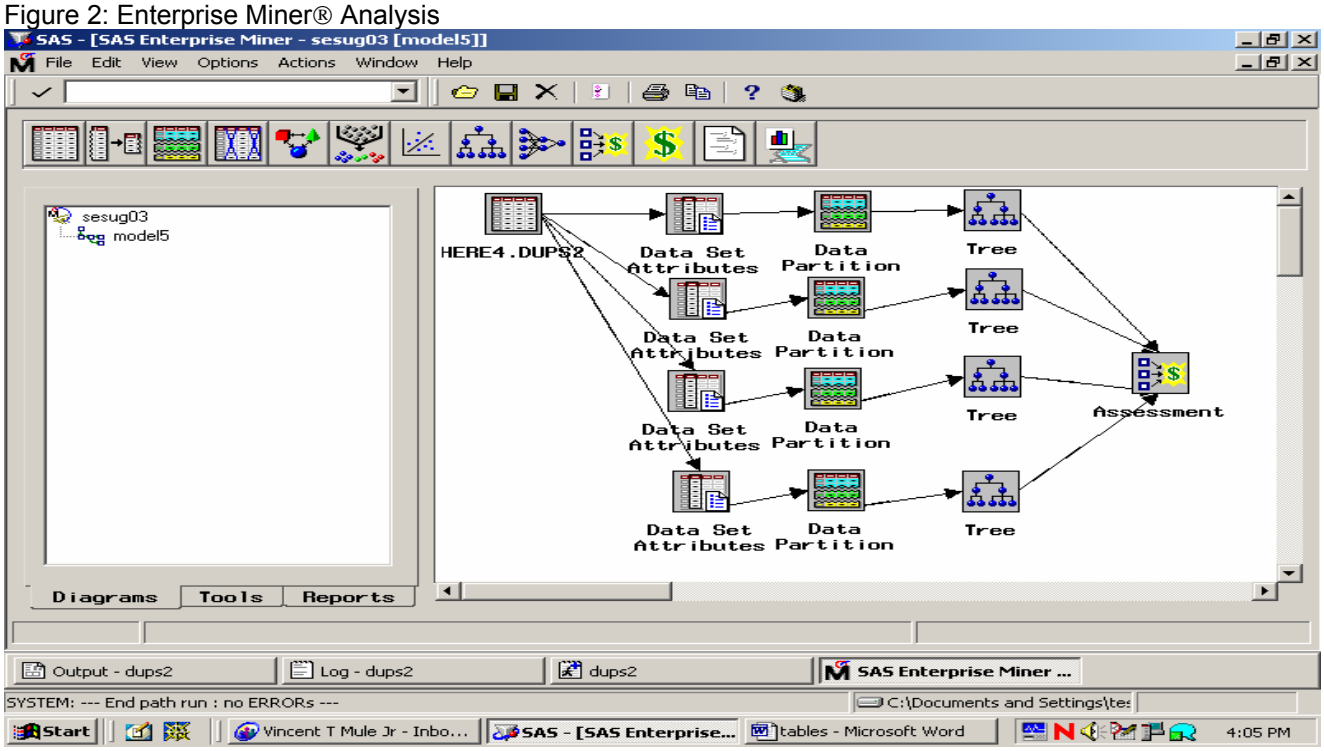
The Decision Tree node has three splitting criteria available: Chi-Square Test, Entropy Reduction and Gini Reduction. I chose Gini Reduction since it is recommended for a binary outcome variable as in this case. Table 4 shows the parameters used in the analysis. For Models 2, 3 and 4, I chose to do binary splits with a maximum depth of four levels. Since Model #1 used only one variable, the overall match score, I set the depth to one and allowed five groupings to be formed. One of the limitations of this analysis is that different parameters can produce different trees.

Table 4: Parameters for Decision Tree Analysis

Parameter	Model 1	Models 2,3 and 4
Minimum Observations in Each Leaf	30	30
Observations Required for Split Search	60	60
Maximum Number of Branches from Each Node	5	2
Maximum Depth of Tree	1	4
Splitting Rules Saved in Each Node	1	1
Surrogate Rules Saved in Each Node	0	0

Enterprise Miner® allowed the analysis to be specified by icons and arrows. Figure 2 shows the specification of this analysis. I started with an icon for my input SAS® data set. I made a row of icons for each of the four models. In the Data Set Attributes icon, I specified which variables were being modeled. In the Data Partition icon for each row, I specified 50 percent of the data for training, 30 percent for validation and 20 percent for testing using simple random sampling. The decision trees will show results for the training and validation. The test data is held in reserve for any future comparisons. In the Tree icon, I specified the parameters for each analysis. Enterprise Miner® also has an Analysis icon that allows the results of the four models to be compared.

Figure 2: Enterprise Miner® Analysis



When trying to identify duplicates, I wanted to maximize our detection while keeping to a minimum the number of false matches. Figures 3-6 show the decision trees for the four models. Figures 3 and 5 show the trees using the overall score. Figure 3 shows the overall match score only. This tree forms 5 groupings of scores. We can see that the concentration of duplicates generally decreases as the score gets smaller.

Figure 5 shows using the overall score plus the number of links to the other housing unit. Figure 5 shows that the number of links is able to help discriminate the duplicates with lower match scores. For this tree, the first break is based on the overall match score. For the path with lower overall scores, cases with two or more links between the units are then separated from the cases where there is only one link between the units.

Figures 4 and 6 show the trees using the six matching variables. Figure 4 shows using only the six variables. In this tree, the first break is based on age. As would be expected, links that agree on age are more likely to be duplicates than links that disagree on age or can't be compared. For both paths, we see that the next break is based on the last name comparison. This part separates out links with disagreeing last names (very low scores) since they contain lower concentrations of duplicates. For the rest of the tree, we see that the right path uses the comparison of day of birth to further discriminate. As would be expected, links that agree on day of birth are more likely to be duplicates than links that disagree or can't be compared. The remaining variable used in the tree is the comparison of first names that gives a similar result as the last name comparison.

Figure 6 shows the six variables plus the number of links to the other housing unit. The first break is the same as the tree for Model 2 in Figure 4 based on the age comparison. The left side of the tree has similar breaks to the left side of the tree for Model 2 in Figure 4. On the right side where the age was different or couldn't be compared, we see that the number of links between units was used next. Cases with two or more links between the units are then separated from the cases where there is only one link between the units. Further down the right side, we see similar results based on comparisons of first name, last name and day of birth as was done for Model 2 in Figure 4.

Figure 7 shows the cumulative density using output from the Analysis node. This allowed me to compare the density of duplicates that each model was able to identify.

The results from Figure 7 are:

- The overall match score and the six matching variables produce similar results when the number of links is not used. For this analysis, using the six variables individually does not produce different results than the overall score.
- Using the number of links helped improve the cumulative density. After processing 70 percent of the links, models using the number of links were able to identify approximately three percent more duplicates than models not using the number (82 percent vs. 79 percent). This difference may be important if duplicates are being contacted and resolved during the enumeration.

My goal was to maximize the detection while keeping to a minimum the number of false matches. Table 5 uses results from the Analysis node using the validation data

for Model #3, the overall score plus the number of links, to show the dilemma in trying to achieve both objectives. The table shows what happens if I send more and more links to be resolved. It increases our detection rate but it is at the cost of a higher rate of false matches. Time and resources available can determine how you balance these two factors.

Table 5: Comparing False Match and Overall Detection for Model #3 (70 Percent and Higher)

Cumulative Percent of Links	False Matches	Overall Detection
70	18%	70%
80	24%	74%
90	29%	79%
100	32%	83%

Characteristics of Duplicates Outside the Enterprise Miner® Analysis

Using a cutoff of 1.45 detected 83 percent of the duplicates. This leaves 17 percent (1,083 of the 6,234 links) which I was unable to put together based on our computer matching. This section examines the characteristics of the missed links from the computer matching analysis. There are three factors that can hinder the use of computer matching to detect duplication. The first is data collection errors. These can be from reporting errors by the respondent or the wrong value being entered by scanning and/or the enumerator. The second is missing data. If a respondent chooses not to report information like month and day of birth, I have less information to determine if they were duplicated. A third is unreliable criteria used by different clerks in determining duplication. Bean (2001) evaluated the duplicate search by the clerks and determined that three percent were incorrectly linked (false matches) and five percent were missed so this is not as large a concern for this matching.

Table 6 shows the exact comparison of the first and last name fields for the missed links. I did this exact comparison twice. I first compared the first name fields and the last name fields. The second inverted the comparison to account for persons entering their first and last name in the wrong boxes on the census form. It shows for very few of these cases that both agree. Of the 534 cases where one agreed and one disagreed, 327 of them disagreed on the first name. A hand review of these cases showed that approximately 50 of them could benefit from nickname standardization.

Table 6: Exact Comparisons of First and Last Name Fields For Missed Duplicates

First Name & Last Name	Regular	Inverted
Both Disagree	434	972
One Agrees, One Disagrees	534	49
Both Agree	115	62

For the cases where both comparisons agreed from Table 6, I compared the month and day of birth. Table 7 shows that both the month and day of birth disagreed for most of the links in the regular comparison. This makes it very hard to determine that a case is a duplicate when the information is reported differently. For the inverted comparison, I did see more agreement for month and day of birth and the benefit of using an inverted comparison to detect these duplicates.

Table 7: Comparisons of Month and Day of Birth For Links when the Both Name Comparisons Agree

Month of Birth & Day of Birth	Regular	Inverted
Both Disagree	114	3
Both Agree	0	48
Other Combinations of Agree, Disagree and Not Reported	1	11
Total	115	62

Next I decided to focus on the comparison of month and day of birth for all of the missed links. Table 8 shows the impact of the data collection errors and missing data. For 428 of the links, the data were not reported on one or both records. For 317 of the links, the reported values for both month and day of birth disagree. For only 188 cases did the two comparisons agreed.

For the 188 cases in Table 8 where both agreed, I examined the first and last name comparisons. Table 9 shows that 140 of the 188 cases do not match exactly on first and last name in the regular comparison. Only a handful of the cases that disagreed on first name would have benefited from nickname standardization. For the inverted comparison, I did see more agreement of the name fields and the benefit of using an inverted comparison to detect these duplicates.

Table 8: Month and Day of Birth Comparison of the Missed Duplicates

		Day of Birth		
		Agree	Not Possible (One or Both Blank)	Disagree
Month of Birth	Agree	188	4	93
	Not Possible (One or Both Blank)	5	428	1
	Disagree	31	16	317

Table 9: Exact Comparisons of First and Last Name Fields For Links when Both Month and Day of Birth Agree (188 Cases)

First Name & Last Name	Regular	Inverted
Both Disagree	140	111
One Agrees, One Disagrees	48	29
Both Agree	0	48

Conclusions

This analysis showed almost no difference between using the overall match score and the six individual results in the Decision Tree modeling.

As in previous research, I saw that using the number of links between the housing units was able to help discriminate and identify higher concentrations of duplicates.

I also examined the characteristics of the duplicates missed by the computer matching. The inverted comparison of the name fields and nickname standardization showed some benefit. These duplicates had certain qualities that allowed them to be linked together by the clerks. Our challenge is to quantify these qualities in a sufficient enough density so that we can identify them by computer processing.

References

Bean, Susanne (2001), "ESCAP II: Accuracy and Coverage Evaluation Matching Error," Executive Steering Committee for A.C.E. Policy II (ESCAP II) Report 7, October 12, 2001.

Fay, Robert (2002), "Probabilistic Models for Detecting Census Person Duplication," American Statistical Association, Proceedings of the Survey Research Section.

Kostanich, Donna (2002), "A.C.E. Revision II: Design and Methodology," DSSD A.C.E. Revision II Memorandum Series # PP-51, December 31, 2002.

Mule, Thomas (2001), "ESCAP II: Person Duplication in Census 2000," Executive Steering Committee for A.C.E. Policy II (ESCAP II) Report 20, October 11, 2001.

_____ (2002), "A.C.E. Revision II Results: Further Study of Person Duplication" DSSD A.C.E. Revision II Memorandum Series # PP-51, December 31, 2002.

Yancey, William (2002), " BigMatch: A Program for Extracting Probable Matches from a Large File for Record Linkage" U.S. Census Bureau report, March 6, 2002.

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Thomas Mule
U.S. Census Bureau
Building 2, Room 2505
Suitland, MD
301 763 8322
Vincent.t.mule.jr@census.gov

This report has undergone a more limited review than official Census Bureau publications. The results and conclusions expressed are those of the author and do not necessarily indicate concurrence by the Census Bureau. This report is released to inform interested parties of research and to encourage discussion.

Figure 3: Decision Tree for Model #1, Overall Match Score

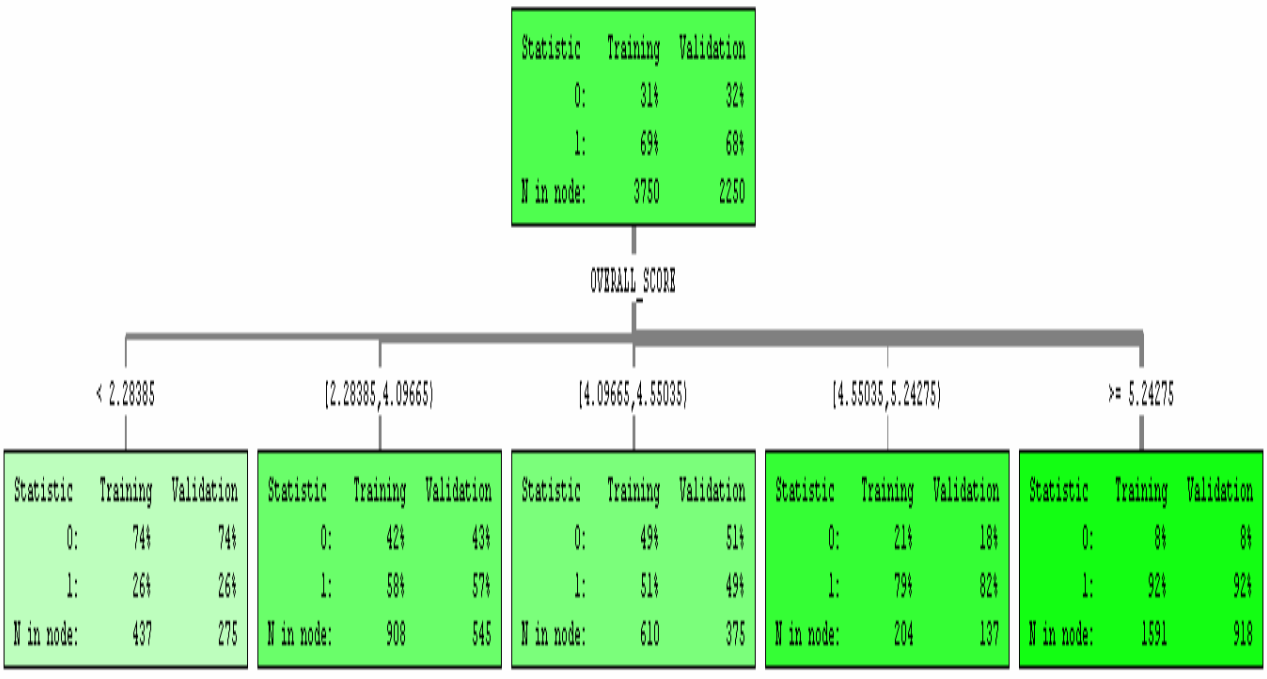


Figure 4: Decision Tree for Model #2, Six Matching Variables

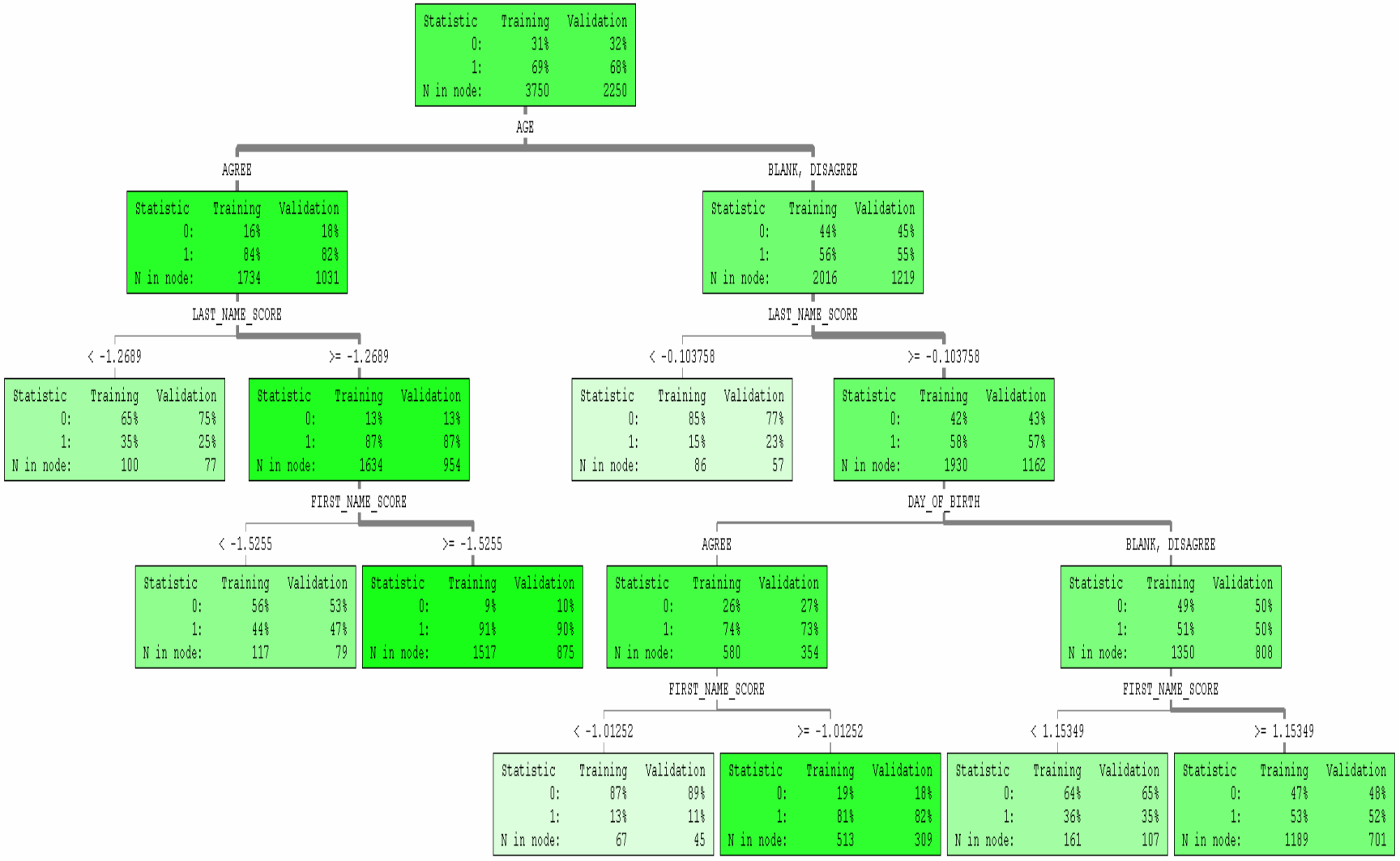


Figure 5: Decision Tree for Model #3, Overall Match Score Plus the Number of Links Between Units

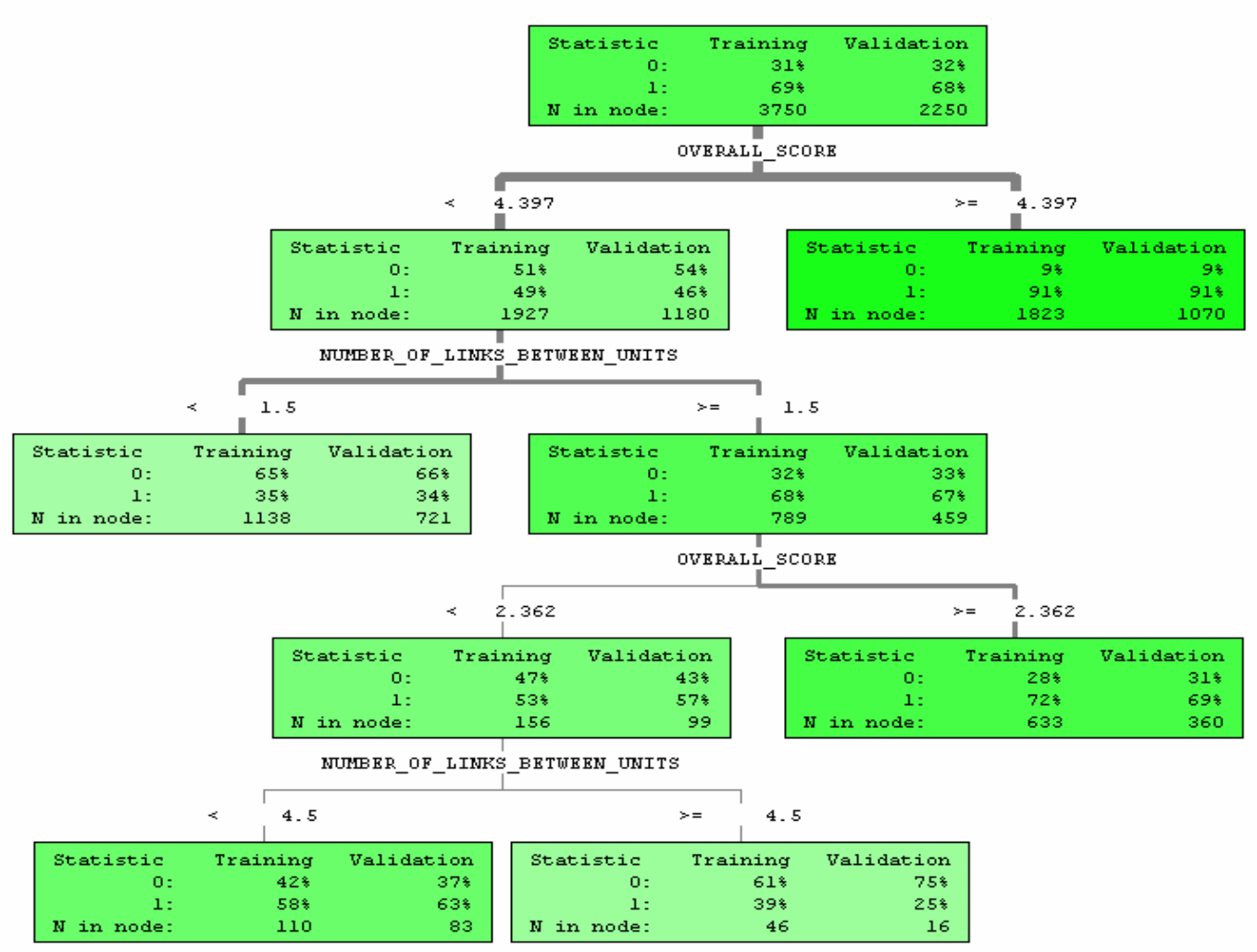


Figure 6: Decision Tree for Model #4, Six Matching Variables Plus the Number of Links Between Housing Units

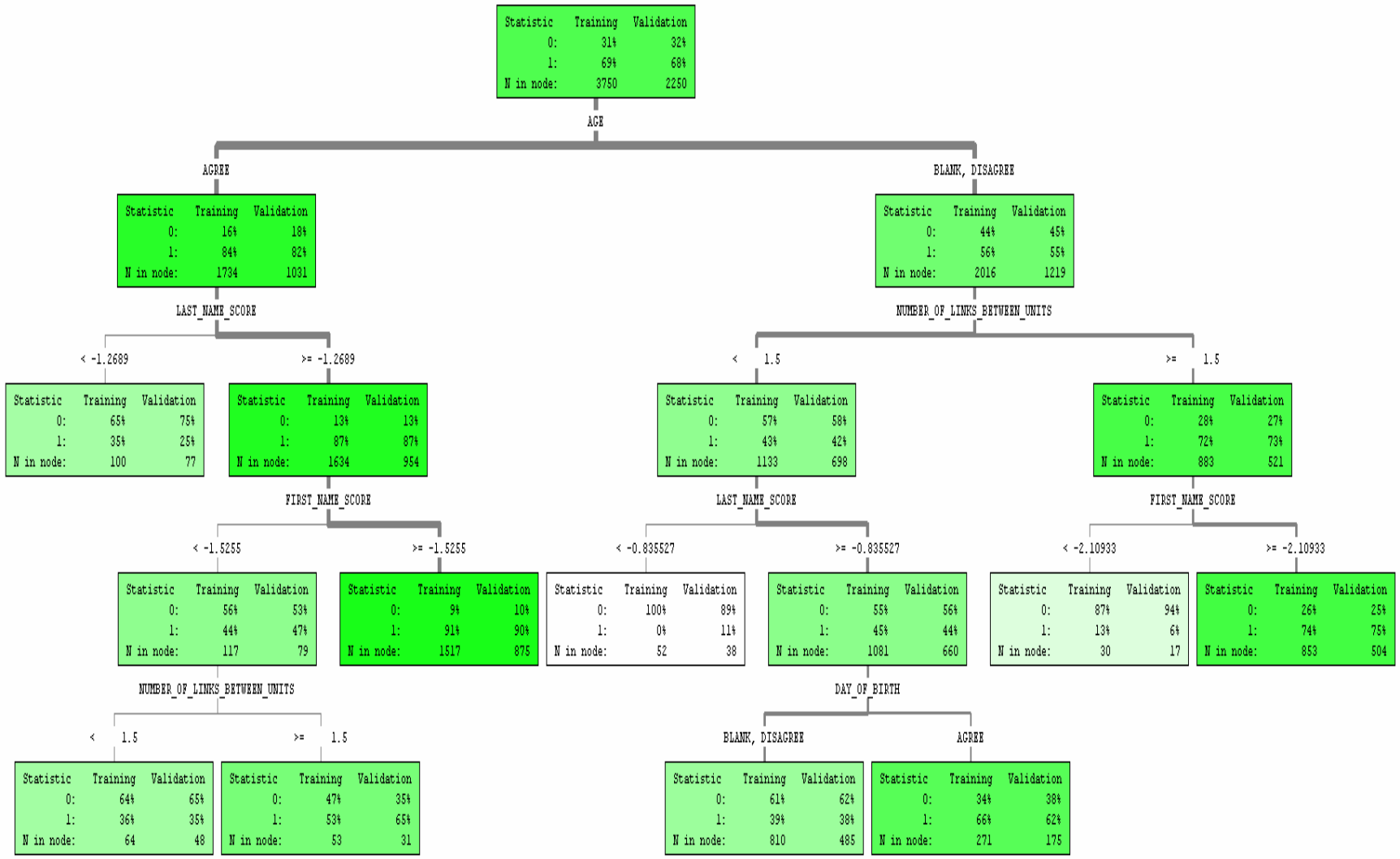


Figure 7: Cumulative Density of Decision Tree Models