



This document was prepared by and for Census Bureau staff to aid in future research and planning, but the Census Bureau is making the document publicly available in order to share the information with as wide an audience as possible. Questions about the document should be directed to Kevin Deardorff at (301) 763-6033 or [kevin.e.deardorff@census.gov](mailto:kevin.e.deardorff@census.gov)

June 29, 2012

## **2010 CENSUS PLANNING MEMORANDA SERIES**

**No. 207**

MEMORANDUM FOR      The Distribution List

From:                      Burton Reist *[signed]*  
                                 Acting Chief, Decennial Management Division

Subject:                    2010 Census Evaluation: Evaluation of Data-Based Extraction  
                                 Processes for the Address Frame

Attached is the 2010 Census Evaluation: Evaluation of Data-Based Extraction Processes for the Address Frame. The Quality Process for the 2010 Census Test Evaluations, Experiments, and Assessments was applied to the methodology development and review process. The report is sound and appropriate for completeness and accuracy.

If you have any questions about this document, please contact Justin Ward at (301) 763-8895.

Attachment

June 27, 2012

# 2010 Census Evaluation of Data-Based Extraction Processes for the Address Frame

U.S. Census Bureau standards and quality process procedures were applied throughout the creation of this report.

**FINAL**

Justin Ward

---

**Decennial Statistical Studies Division**



(This Page Intentionally Left Blank)

## Table of Contents

Executive Summary .....	vi
1. Introduction.....	1
2. Background.....	1
2.1 MAF Coverage Research .....	3
2.2 Prior Data Mining Research .....	3
2.3 Frame Assessment for Current Household Surveys (FACHS) Filter Rules Research .....	4
2.4 Related Evaluations and Assessments.....	5
3. Methodology.....	6
3.1 Question to be Answered.....	6
3.2 Decision Tree Modeling .....	6
3.3 Software.....	7
3.4 Data.....	9
3.4.1 2010 Census Address Frame COMBO File .....	9
3.4.2 2009 infoUSA File .....	9
3.5 Cost.....	11
4. Limitations .....	11
5. Results.....	12
5.1 Rules .....	12
5.2 Projections for 2010 and 2011.....	18
6. Conclusions and Recommendations .....	20
6.1 Conclusions.....	20
6.2 Recommendations.....	22
7. Acknowledgements.....	23
8. References.....	24
Appendix A: SAS Enterprise Miner 6.2 Report .....	27
Appendix B: WEKA 3.6.2 User Interface Screenshot.....	40
Appendix C: January 2009 ACS Universe Specifications.....	41
Appendix D: January 2010 ACS Universe Specifications .....	44
Appendix E: January 2011 ACS Universe Specifications.....	47

## List of Tables

Table 1. 2010 CPEX Data Mining: January 2009 ACS Master Address File Extract (MAFX) Distribution of Address Records by ACS Filter Rule .....	3
Table 2. 2010 CPEX Data Mining: WEKA Performance Summary.....	8
Table 3. 2010 CPEX Data Mining: January 2009 ACS Status by 2010 Census Post-Address Canvassing Status .....	11
Table 4. 2010 CPEX Data Mining: Rule 1 Validation .....	13
Table 5. 2010 CPEX Data Mining: Rule 1 Performance.....	13
Table 6. 2010 CPEX Data Mining: Rule 2 Validation .....	14
Table 7. 2010 CPEX Data Mining: Rule 2 Performance.....	14
Table 8. 2010 CPEX Data Mining: Rule 3 Validation .....	14
Table 9. 2010 CPEX Data Mining: Rule 3 Performance.....	15
Table 10. 2010 CPEX Data Mining: Rule 4 Validation .....	15
Table 11. 2010 CPEX Data Mining: Rule 4 Performance.....	16
Table 12. 2010 CPEX Data Mining: Rule 5 Validation .....	16
Table 13. 2010 CPEX Data Mining: Rule 5 Performance.....	17
Table 14. 2010 CPEX Data Mining: Rule 6 Validation .....	17
Table 15. 2010 CPEX Data Mining: Rule 6 Performance.....	18
Table 16. 2010 CPEX Data Mining: Projected Records Selected by Rule 2 and Rule 5 .....	19
Table 17. 2010 CPEX Data Mining: 2010 and 2011 Projection Performance of Rule 2 and Rule 5.....	19
Table 18. 2010 CPEX Data Mining: Performance for All Rules in 2009 .....	20

## List of Figures

Figure 1. 2010 CPEX Data Mining: National Housing Unit Counts by Year from 2009 National Estimate of Coverage.....	2
Figure 2. 2010 CPEX Data Mining: Decision Tree Modeling Result.....	7

## Executive Summary

U.S. Census Bureau decennial census, survey, and estimates programs work with subsets, known as extracts, of the Master Address File. These extracts are produced using a set of rules called filters. Filters attempt to maximize the number of valid Master Address File units, while minimizing the number of invalid units on the resulting extracts. These extracts provide the basis for the address frames used in census operations or the sample universes for current demographic household surveys. One such survey is the American Community Survey. The American Community Survey filter rules tend toward overcoverage (inclusion of invalid units) due to the higher difficulty of correcting undercoverage (exclusion of valid units) in field work. The 2010 Census Evaluation of Data-Based Extraction Processes for the Address Frame, also referred to as the Data Mining Evaluation, presents possible improvements to the American Community Survey filter rules following analysis using data mining techniques to answer the research question:

*How can the quality of the address frame be improved with a more scientific extract process?*

Data-based extraction processes, or more specifically data mining, provide a way to identify meaningful descriptive and predictive information from large datasets. Decision tree modeling, one tool of data mining, presented the best opportunity to generate improvements in the American Community Survey filter. To answer the research question, the 2010 Census Program for Evaluations and Experiments Data Mining Evaluation tested three software packages -- Waikato Environment for Knowledge Analysis 3.6.2, Salford Systems Classification and Regression Trees 6.0, and SAS Enterprise Miner 6.2 -- to create additional American Community Survey filter rules. These filter rules were measured by their ability to contribute to the coverage of the American Community Survey's Address Frame (the survey's sample universe), attempting to maximize inclusion of additional valid address records (reducing Type I Error – incorrectly excluding valid records) and minimizing the inclusion of invalid address records (Type II Error – erroneously including invalid records).

The analysis used the results of the 2010 Address Canvassing operation as ground truth to measure the validity of the American Community Survey filters on the January 2009 American Community Survey extract. Six new filter rules were produced, primarily from research conducted using two of the three software packages. Five of these rules (Rules 1 – 4, 6) only used the variables available from the Master Address File, while Rule 5 used the commercially available dataset, infoUSA, along with the same variables from the Master Address File.

All six new filter rules are designed to augment the current American Community Survey filters:

**Rule 1.** Rule 1 selected records where the housing unit was flagged as eligible for the Demographic Area Address Listing<sup>1</sup>, flagged as residential on the Spring 2008 Delivery Sequence File<sup>2</sup>, and the source of the oldest operation on record existing on the Master Address File Operations table was one of six sources -- 1990 Address Control File<sup>3</sup>, 2000 Address Listing operation<sup>4</sup>, sent to 2000 Local Update of Census Addresses<sup>5</sup>, 2000 Rural Update/Leave operation<sup>6</sup> or 2000 Block Canvassing<sup>7</sup>.

**Rule 2.** Rule 2 selected records where the housing unit was flagged as eligible for the Demographic Area Address Listing and flagged as residential on the Spring 2008 Delivery Sequence File.

**Rule 3.** Rule 3 selected records where the housing unit was flagged as eligible for the Demographic Area Address Listing, flagged as residential on the Spring 2008 Delivery Sequence File, the source of the oldest operation on the record was the 1990 Address Control File, and eligible for 2010 Local Update of Census Addresses.

**Rule 4.** Rule 4 selected records where the housing unit was flagged as eligible for the Demographic Area Address Listing, flagged as residential on the Spring 2008 Delivery Sequence File, and the source of the oldest operation on the record was the 1990 Address Control File.

**Rule 5.** Rule 5 selected records where the housing unit was flagged as eligible for Demographic Area Address Listing, flagged as residential on the Spring 2008 Delivery Sequence File, and present on the infoUSA file.

**Rule 6.** Rule 6 selected records where the housing unit was not flagged as eligible for the Demographic Area Address Listing and not flagged as residential on the Spring 2008 Delivery Sequence File.

---

<sup>1</sup> A post-Census 2000 program that coordinates various operations related to the review and automated update of the geographic content of the TIGER® database and the addresses in the Master Address File.

<sup>2</sup> A U.S. Postal Service (USPS) computer file containing all mailing addresses serviced by the USPS.

<sup>3</sup> The residential address list used by the Census Bureau to label questionnaires, control the mail response check-in operation, and determine the Nonresponse Followup workload for the 1990 census.

<sup>4</sup> A Census 2000 field operation to develop the address list in areas with predominantly noncity-style mailing addresses.

<sup>5</sup> A Census 2000 program, established in response to requirements of Public Law 103-430, that provided an opportunity for local and tribal governments to review and update individual address information or block-by-block address counts from the Master Address File and associated geographic information in the TIGER® database.

<sup>6</sup> A method of data collection in which enumerators canvassed assignment areas to deliver a census questionnaire to each housing unit. At the same time, enumerators updated the address listing pages and Census Bureau maps.

<sup>7</sup> A Census 2000 field operation to ensure the currency and completeness of the Master Address File within the mailout/mailback area.



Two benchmarks were used to measure the effectiveness of each rule:

- an improvement in the number of valid address records to the American Community Survey sampling frame (where address validity is determined by the 2010 Census Address Canvassing operation outcome) at a rate of 0.25 percent or greater, and
- a ratio of valid to invalid added address records, Type I/II Error ratio, of 4:1.

Rules 1, 2, and 5 showed the greatest potential for improvement in the American Community Survey filter rules. Rule 1 improved the number of valid American Community Survey address records by 0.56 percent while adding address records to the sampling frame extract with a Type I/II Error ratio of 3.20:1. Rule 2 improved the number of valid records by nearly double the amount of Rule 1 at a rate of 1.12 percent, but at a reduced Type I/II Error ratio of 2.57:1. Rule 5 included data from infoUSA and produced a 0.97 percent improvement in valid records, while adding records at a 3.13:1 Type I/II Error ratio.

Based on these findings, the Decennial Statistical Studies Division provides the following recommendations:

**1. Profile Rule 2 records, and if validated apply Rule 2 to the July 2012**

**American Community Survey Extract:** The units affected by Rule 2 should be further profiled through a joint Decennial Statistical Studies Division and American Community Survey Office partnership to garner any additional information about their characteristics. This task should identify geographic (clustering/dispersion across block, tract, city, etc.) and physical (single/multiple unit, group quarters/housing unit class, etc.) distributions, as well as final Census 2010 status (final validity, vacant/occupied, population count, etc.). If favorable results are obtained from profiling, the July 2012 American Community Survey filter rules should be updated to include Rule 2, given the potential to reduce gross undercoverage and total error shown by the analysis. Rule 2 only uses variables currently available on the Master Address File, which allows quick implementation.

- 2. Identify new Administrative Record data sources:** Rule 5 shows that the 2009 infoUSA file proved valuable in confirming the accuracy of valid addresses on the Master Address File (where address validity is determined by the 2010 Census Address Canvassing operation outcome). Confirmation of an address on many different, independent data sources, and the various permutations of these data sources, only increases the potential use of future data mining research efforts.

3. **Use SAS Enterprise Miner:** SAS Enterprise Miner should be given primary consideration in continuing data mining research. The software is included in the Census Bureau's site license, has the capacity to work with the large datasets necessary for census research, and has an easy-to-use user interface. The other packages tested here did not provide all of these benefits.
4. **Continue Data Mining Research:** Due to the success of the research here, address frame data mining research using the 2010 Census Address Canvassing data should continue. A data mining evaluation is currently part of the planned 2020 Census projects. With additional resources, Master Address File filter rules could be further improved.
5. **Conduct Verification:** In order to ensure accuracy, develop a system to test (field and office) additional Master Address File filter rules prior to, or in parallel with, the inclusion of the American Community Survey sample universe.

## 1. Introduction

The goal of the 2010 Census Program for Evaluations and Experiments (CPEX) Evaluation of Data-Based Extraction Processes for the Address Frame, also referred to as the Data Mining (DM) Evaluation, was to explore the use of data mining on the Master Address File (MAF) to refine the extraction process of the address frame for future censuses and current surveys. Data mining is a set of statistical tools including decision trees, regression models, clustering algorithms, and neural networks that provide predictive models. The Decennial Statistical Studies Division (DSSD) used data mining software to produce models that predict address validity, and then evaluated the predictions using the results from the 2010 Census Address Canvassing (AC) operation as indicators of whether or not an address was valid (ground truth).

## 2. Background

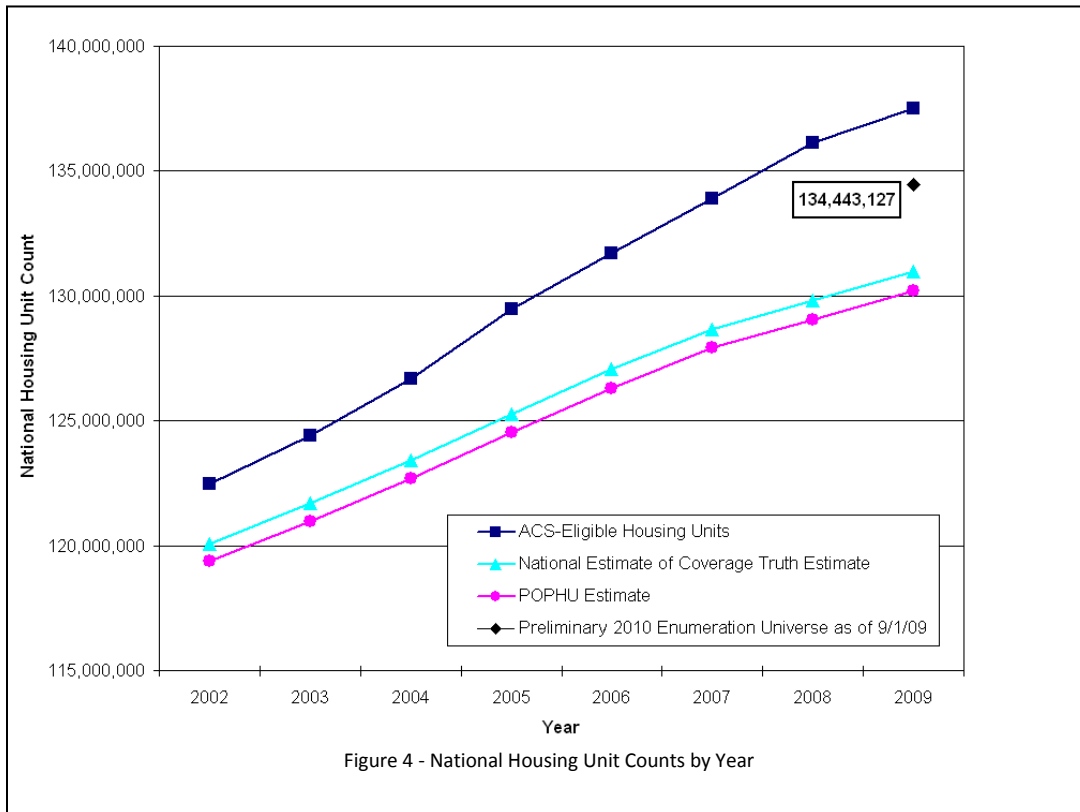
The MAF is a computer file of every address and physical/location description known to the U.S. Census Bureau, including geographic locations. As of early 2009, there were approximately 180 million units on the MAF. In practice, U.S. Census Bureau decennial census, survey, and estimates programs typically work with subsets of the MAF, known as extracts; produced with sets of rules known as filters. The goal of the filters is to maximize the number of valid addresses and minimize the number of invalid addresses on the resulting extracts, which become the address frames in census operations or the sample universe for current demographic household surveys. Filter rules rely on categorical variables such as when a unit was added to the MAF, its residential status, and outcomes from past field operations to determine whether or not an address is valid for an extract.

Shown in Figure 1 from the 2009 National Estimate of Coverage (NEC) report (Kephart, 2010), the American Community Survey (ACS) sampling frame in 2009 resulted in about 137 million valid housing units. In general, the ACS filter rules produce an extract that tends toward overcoverage because overcoverage (the erroneous inclusion of invalid records) can be accounted for during field work for the survey and also in the controlled raking<sup>8</sup> procedures, while undercoverage (exclusion of valid records) is less likely to get corrected during the field work or other procedures. Figure 1 shows that the number of ACS-Eligible Housing Units (HUs) is greater than the Preliminary 2010 Enumeration Universe, and even greater than the Population Division Housing Unit (POPHU) estimate.

---

<sup>8</sup> Procedure used to improve the relation between the sample and the population

**Figure 1. 2010 CPEX Data Mining: National Housing Unit Counts by Year from 2009 National Estimate of Coverage**



The ACS filter is updated annually based on new information (research findings, etc.). For 2009 (Bates, 2009), the basic theory behind the filter was to include addresses that can be categorized into one or more of the following six classes:

- Valid Census 2000 addresses
- Count Question Resolution (CQR) adds and reinstatements.
- Postcensus Delivery Sequence File (DSF) adds from the United States Postal Service (USPS) regardless of geocoding status, but restricted by block-level Address Characteristic Type (ACT) and Census 2000 Type of Enumeration Area (TEA) Codes.
- Census deletes that persist on the DSF.
- Demographic Area Address Listing (DAAL) adds.
- Special Census and Census Test adds.

**Table 1. 2010 CPEX Data Mining: January 2009 ACS Master Address File Extract (MAFX) Distribution of Address Records by ACS Filter Rule**

Existing ACS Filter Rule	Addresses*	Percent**
Valid Census 2000 address .....	115,728,143	64.22
CQR adds and reinstatements .....	4,724	0.00
Post-Census DSF adds .....	19,265,867	10.69
Persistent Census deletes .....	1,535,508	0.85
DAAL adds .....	338,677	0.19
Special Census and Census Test adds .....	325,315	0.18
Invalid for ACS .....	42,996,615	23.86
Total .....	180,194,849	100.00

\*Counts and percentages are unweighted.  
\*\*Percentages may not sum to 100 due to rounding.  
Source: January 2009 ACS MAFX.

Both the DSSD and the Demographic Statistical Methods Division (DSMD) have evaluated how well the ACS filter produces an extract that represents the ground truth for survey and decennial census operations. Some of the previous research findings are summarized below.

## 2.1 MAF Coverage Research

Since 2002, the DSSD has produced a NEC report annually as part of the Address Coverage Improvement and Evaluation Program (ACIEP). The coverage estimates in these reports are produced from comparisons to the Population Division's (POP) annual HU estimates adjusted using the Dual-System Estimate (DSE) from the Census 2000 Housing Unit Coverage Survey (HUCS). The most recent report was released November 30, 2010, containing net coverage estimates for 2009. For the ACS frame, there was 5.19 percent net overcoverage in 2009 (Kephart, 2010). For the 2010 Census AC projected frame (the ACS frame without ungeocoded records, i.e., those lacking a block designation), the report estimated 1.10 percent net overcoverage. These net coverage estimates indicate that each frame might benefit from a different data mining model, to minimize coverage error.

## 2.2 Prior Data Mining Research

Data mining, also known as knowledge discovery, is a way to identify meaningful descriptive and predictive information from large data sets. Researchers specify a training data set, which is a representative sample of the larger data set used to build the models. The training data set contains the outcome variable, so models can be evaluated based on predicted outcome compared with the actual outcome. When analysis on the training data set generates models, these models are validated on new larger data sets. For data mining research on the MAF, the outcome variable identified for the research here was whether or not the address was valid on the ground (determined a valid Census address at the time of the Enumeration extract).

The DSSD conducted data mining analyses on the MAF during 2006 and 2007 using the Salford Systems CART v5.0 data mining software. The goal of the work was to predict the validity

status of MAF records based on a selection of predictor variables from the MAF. Those analyses used DAAL field work from September 2005 through February 2006 as the indicator of ground truth since it was the most recent field work available. This sample was not necessarily representative of the entire MAF. The same addresses were pulled from the July 2005 Master Address File extract (MAFX) to form the training data set used to build the predictive models. These models were evaluated against the July 2006 MAFX based on predicted validity rates and cost ratios of undercoverage and overcoverage. The final models were chosen because they minimized undercoverage, since undercoverage (not including records that should be in the frame) was considered more problematic than overcoverage (including records that should not be in the frame). However, overcoverage is still a concern for decennial census operations, with potential implications for quality and cost.

### **2.3 Frame Assessment for Current Household Surveys (FACHS) Filter Rules Research**

During 2008, DSMD produced filter rule research findings as part of the FACHS program (Martin and Loudermilk, 2008). Their report uses data from the National Evaluation Sample (NES), a nationally representative sample of 5,722 tabulation blocks. Field representatives canvassed each block and then classified each address on the MAFX as valid (existing, habitable) or invalid (nonexistent, uninhabitable, nonresidential, or duplicate). These classifications were considered “ground truth” for their analysis of the January 2007 ACS MAFX.

One issue of interest to DSMD for their current surveys frame was the delivery status of the record on the DSF from the United States Postal Service (USPS). There are two categories: Included in Delivery Statistics (IDS) and Excluded in Delivery Statistics (EDS). The IDS records are addresses to which the USPS delivers mail while EDS records are addresses that do not receive mail delivery (some may be newly constructed housing units). DSMD analyzed IDS/EDS status when the record first appeared on the DSF as compared to the most recent DSF and then compared that to the “ground truth” validity. Of those initially and most recently IDS, 19.3 percent were invalid on the ground. Of those changing from EDS to IDS, 8.9 percent were invalid. For records remaining EDS, 49.4 percent were invalid. In their report, the authors recommended excluding DSF records that remain EDS for an extended period of time from the current surveys frame.

DSMD also explored the approximately 1.5 million records on the MAF that were classified as residential on the latest DSF, but were identified as deletes (invalid HUs) during Census 2000 operations. Of these census deletes, approximately 35 percent were invalid on the ground. DSMD recommended further research into these records to better identify valid HUs.

The reports on filter rules research produced several major categories of addresses that should be considered for inclusion in a sample frame:

1. **DSF Adds.** When new records appear on the DSF they often have not yet been built or occupied. This category contributes heavily to the overcoverage for ACS. There is a flag on the DSF to indicate that mail is being delivered to the address, but limiting the extract to those units that are receiving mail has been unacceptable to ACS implementers.

Updates to the MAF from the DSF occur every 6 months; thus a unit that should be included in the sample for a given year may be excluded because the DSF has it flagged as not yet receiving mail. The latest Frame Assessment for Current Household Surveys Filter Rules Research (FACHS-FRR) results indicate that addresses flagged as not receiving mail for the previous 6 DSF cycles (3 years) are invalid at a rate of 64.1 percent.

2. **Census deletes that persist on the DSF:** In past censuses there have been substantial numbers of addresses in the DSF that were marked as invalid by decennial census operations. These addresses are estimated by the FACHS-FRR report to be invalid at a rate of about 35 percent and thus contribute considerably to overcoverage as the decade progresses if they are included in the extracts. The report suggests that these units are probably new construction that was not occupied until after census enumeration. This study also attempted to categorize the deletes in a variety of ways such as urban/rural status, structure size, delivery point type, etc., but found no clear predominant category.
3. **Duplication Zones:** The ACS designates certain areas as “duplication zones” where there is a high probability that recent adds to the DSF may duplicate existing addresses. These areas are primarily rural mail delivery areas that have undergone recent Emergency 911 (E911) conversion. Thus, new city-style addresses in the DSF may duplicate existing rural-style addresses in the MAF. DSMD plans to continue refining the definition of the duplication zone along its current lines. The Census Bureau may want to consider exploring rates of invalid addresses regardless of the cause at the county level to attempt to establish reliable localized criteria for improving coverage.
4. **Erroneous Filter Exclusions.** Addresses that were valid according to the NES but rejected by the filtering rules constitute 1.7 percent of all valid HUs. While the FACHS-FRR could not find any patterns to the categories of these addresses they did recommend further research after the 2010 universe has been finalized.

## 2.4 Related Evaluations and Assessments

There are many 2010 evaluations and assessments that are related to the work performed here:

- Evaluation of Address Frame Accuracy and Quality
- Study of Address Canvassing Targeting and Cost Reduction
- Evaluation of Small Multi-Unit Structures
- Evaluation of Address List Maintenance Using Supplemental Data Sources

### 3. Methodology

#### 3.1 Question to be Answered

How can the quality of the address frame be improved with a more scientific extract process?

To answer this question, benchmarks to measure improvement were set. As no previous benchmarks were available to compare the results of new rules, benchmarks were determined based on reasonable assumptions of acceptable performance. Any new filter rule was measured by its ability to contribute to the coverage of the ACS Address Frame (the survey's sample universe), attempting to maximize inclusion of additional valid records (reducing Type I Error – incorrectly excluding valid records) and minimizing the inclusion of invalid records (Type II Error – erroneously including invalid records). The benchmark for any new filter rule needed to increase the number of valid addresses (reduce gross undercoverage) by 0.25 percent or greater, while adding these records in a ratio of valid addresses to invalid addresses (Type I/II Error Ratio) of 4:1 or better. In addition, gross overcoverage and total error were included in the composition of the final report as complementary performance measures.

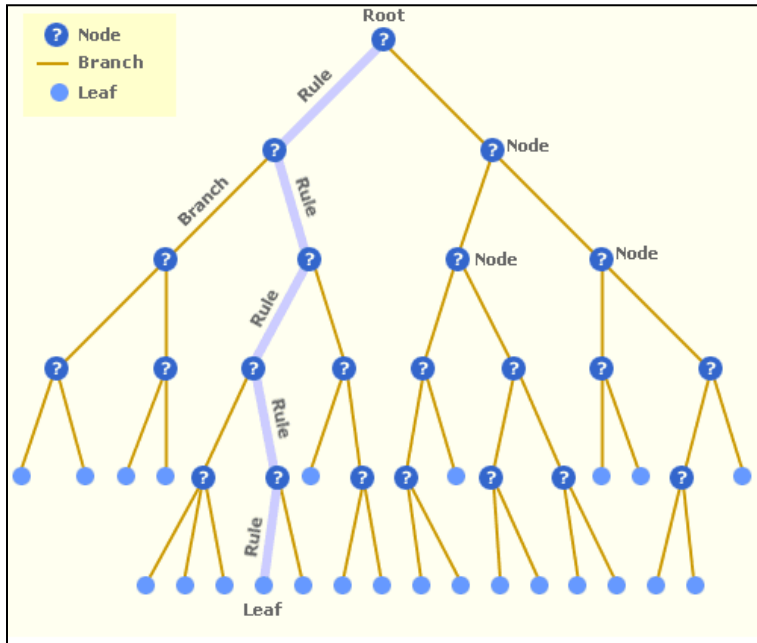
#### 3.2 Decision Tree Modeling

Data mining includes a set of statistical tools to create predictive models such as neural networks, decision trees, clustering algorithms, and regression modeling. We found decision tree modeling to be the most effective tool to analyze and improve the filter rules. Decision trees partition large amounts of data into smaller segments by applying a series of rules which split the data into pieces until no more splits can occur. The purpose of partitioning the data by these rules is to create isolated subsets in which the designated target variable has a lower diversity of values than the overall sample population. For instance, if the data have a target variable that has values of yes or no with an overall distribution of values of 60 percent yes and 40 percent no, then decision tree modeling may be able to generate a model that creates a subset of the sample population that has a distribution of the target variable with 90 percent yes and 10 percent no.

Figure 2 gives an example of the results from using decision tree modeling. The root represents the beginning of the tree where no subsets of the sample population have occurred. At each node, the data split into two or more subsets categorized by specifications given in the branch. Each leaf represents a subset of data that cannot be split anymore based on the criteria of the decision tree process. By looking at the distribution of the target variable amongst the leaves, the model can determine the best rule to predict the target variable.



**Figure 2. 2010 CPEX Data Mining: Decision Tree Modeling Result**



© 2008 SAS Institute, Inc.

### 3.3 Software

Three software packages were used for this evaluation: Salford Systems Classification and Regression Trees (CART) 6.0, Waikato Environment for Knowledge Analysis (WEKA) 3.6.2, and SAS Enterprise Miner (EM) 6.2. Along with decision tree modeling, all of these software packages can produce predictive regression models, neural network diagrams, and other multivariate analytical outputs.

WEKA 3.6.2 is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. WEKA is free software available under the GNU General Public License. While WEKA can perform decision tree modeling, the software imposes several restrictions on the format and size of the data. First, WEKA can only handle two types of files: comma separated values (csv) and attributes-relations file format (arff). Given that most datasets used for this evaluation were created in SAS, WEKA required additional time to modify datasets into a functional format. Second, based on each personal computer (pc) and/or server installation, WEKA imposes a strict limitation on the size of the dataset that can be used as an input into the program. After stress testing the software, the WEKA software reached a performance threshold using a dataset of 150,000 records with 167 variables, at a file size of 56,042 kilobytes (KBs). These tests were performed on a pc with an Intel Core 2 Duo Central Processing Unit (CPU) at 2.33 Gigahertz (GHz) and 3.25 gigabytes (GBs) of Random Access Memory (RAM). Table 2 provides a comparison of model building and run times for different file sizes. WEKA is also available for Linux and server installations, which would likely yield improved performance measures.

Records	Variables	File Size (KB)	Model Building Run Time (sec)	Total time (sec)	Algorithm	Status
104,456	167	32,112	39.44	413	CV <sup>*</sup> =10	Completed
146,684	167	56,042	64.38	718	CV <sup>*</sup> =10	Failed
146,684	167	56,042	47.28	518	Default	Completed
207,139	167	74,671	N/A <sup>+</sup>	N/A <sup>+</sup>	Default	Failed

<sup>\*</sup>Cross validation algorithm, ten fold.  
<sup>+</sup>Software abended.  
Source: DSSD 2010 Decennial Census Memorandum #O-A-35.

Salford Systems CART 6.0 automatically sifts large, complex databases, searching for and isolating significant patterns and relationships. This discovered knowledge is then used to generate predictive models for applications. CART suffered from the similar limitations as WEKA, in that the software was limited to an installation on a personal computer with the same specifications as above. Again, Linux and/or server installation could overcome these limitations, but for this report only the pc version was feasible. In light of both WEKA and CART suffering from similar technical limitations, there was only limited pursuit of CART for the research here.

SAS EM 6.2 is a SAS module included in the SAS license held by the U.S. Census Bureau. Enterprise Miner is a powerful data mining tool used for pattern discovery and predictive modeling. For this evaluation, the work focused on using SAS EM's predictive modeling capabilities, specifically decision tree modeling. When compared to WEKA, SAS EM offers several distinct advantages.

SAS EM does not suffer from the same restrictions that were found when using the WEKA and CART software packages. By installing SAS EM on an Egenera Blade server with four 6-core CPUs at 2.4 GHz and 192 GB of RAM, it did not suffer from any input file size restrictions. SAS EM successfully performed using a national SAS dataset of approximately 190 million records and 280 variables, at 71 GB. The model building portion of the analysis took 37 hours, 24 minutes, and 28 seconds (134,668 seconds). Adding the time for data loading and partitioning, the total run time in SAS EM was 39 hours, 57 minutes, and 17 seconds (143,838 seconds). By having a national dataset as the input file, SAS EM can also prepare the dataset for use in the decision tree modeling tool where it can perform its own cleaning, sampling, partitioning, and analysis, while WEKA required the dataset to be cleaned and the sample chosen before running the software. By running the program with JAVA through a web browser, SAS EM also gives the user a high quality interface to edit and run analyses.

### 3.4 Data

This project used data from five sources: the 2010 Census Address Frame Combination (2010 Census AF COMBO) file (described below), the 2009 infoUSA file, the January 2009 ACS MAFX, the January 2010 ACS MAFX, and the January 2011 ACS MAFX.

#### 3.4.1 2010 Census AF COMBO File

The 2010 Census AF COMBO file is a database constructed by DSSD, for assessing the 2010 Census (Ward, 2011a). The file combines eight groups of census files merged at the address level based on corresponding address-level record identifiers. The eight input files are: the 2010 Pre-Address Canvassing (Pre-AC) MAFX<sup>9</sup>, Census Evaluations and Experiments (CEE) files<sup>10</sup>, the 2010 AC Reject files<sup>11</sup>, the Large Blocks file<sup>12</sup>, the 2010 Group Quarters Validation (GQV) files<sup>13</sup>, the 2000 Combo files, the 2010 Enumeration Universe (EU) files<sup>14</sup>, and the January 2009 ACS MAFX. Since the goal of the data mining process was to predict AC results, the dependent variables used from the COMBO file were restricted to the 2010 Pre-AC MAF variables. The data mining algorithm used the Delivery Specific Address Flag (DSAF) variable from the 2010 EU files as the target variable. This vintage of the DSAF variable provided the most recent result of 2010 AC validity status (at the time this evaluation was conducted). Only records located in the United States were used from these files in the analysis because the 2009 infoUSA file (described in the next section) did not include records in Puerto Rico.

#### 3.4.2 2009 infoUSA File

infoUSA is a commercial database comprised of household-level data used mainly for direct marketing purposes. The company that produces this database defines a household as every unique address / last name combination in the file. infoUSA collects data from phone directories and business sources and then validates the data with phone interviews. The infoUSA file used in this evaluation was slightly modified using the 2010 Pre-AC MAFX. The Data Integration Division (DID) matched the infoUSA file to the MAF, per the specifications DSSD provided (Clark, 2009). DID used a probabilistic matching process with blocking by 3-digit ZIP code. DID performed all necessary passes to match city-style, rural route, and Post Office (PO) Box addresses. They first attempted to match addresses at the unit level, and then attempted to match addresses at the Basic Street Address (BSA) level. For addresses associated with a multi-unit structure, the matching process included at least two passes. They first determined whether the

---

<sup>9</sup> Files containing units as they existed on the MAF prior to the 2010 Census AC operation.

<sup>10</sup> Files containing units as they were recorded in the field during the 2010 Census AC operation prior to processing by the Geography Division.

<sup>11</sup> Files containing units that were rejected by the Geography Division after processing.

<sup>12</sup> Files containing units as they were recorded in the field during the 2010 AC operation for only large blocks.

<sup>13</sup> Files containing units as they existed on the MAF after the 2010 Census AC operation.

<sup>14</sup> Files containing units as they existed on the MAF after the 2010 Census GQV operation.

address matches for the specific apartment unit. If the first pass failed, the second or subsequent pass determined if the address matched at the BSA level. The process created two new variables, MATCH\_PASS\_UNIT (indicates apartment unit match) and MATCH\_PASS\_BSA (indicates BSA level match). In cases where a record on the infoUSA file matched to multiple addresses on the MAF, DID gave higher precedence to cases where the ACS Delivery Flag (ACSAF) on the MAF was not equal to zero (unit was eligible for the ACS). After DSSD received the matched infoUSA file from DID, the file required substantial additional processing in order for the file to be usable in the data mining software.

The infoUSA dataset contained 298 variables with the possibility that a Master Address File Identification number (MAFID) had multiple entries. The matching process caused duplication of MAFIDs because the infoUSA database provides data on people that may have been found to no longer live at an address (historical residences for persons). The variable PRIMARY\_IUSA\_REC was used to denote the primary record for each unique infoUSA address at the unit level. In order for the data to be usable for data mining, the infoUSA data needed to be collapsed into an address-level file as opposed to its original household-level structure. By using the PRIMARY\_IUSA\_REC variable, a transformation of the data was:

- 1) If there was only one record with PRIMARY\_IUSA\_REC with a value of Y, then that was the only record saved for the corresponding MAFID.
- 2) If there was more than one record per MAFID with PRIMARY\_IUSA\_REC with a value of Y, then for each variable the data were consolidated into a new variable. Examples of transformations were minimum or maximum values, mean values, and various tabulations.
- 3) If there was no record with PRIMARY\_IUSA\_REC with a value of Y, then all records were used to consolidate the variables into a single record.

The process created an infoUSA file with one record per MAFID, and reduced the number of variables from 298 to 106 and records from about 361 million to approximately 123 million.

Table 3 provides the distribution of Census and ACS (January 2009) validity statuses for the approximately 180 million records on the 2010 Census EU files.

**Table 3. 2010 CPEX Data Mining: January 2009 ACS Status by 2010 Census Post-Address Canvassing Status**

	Invalid for ACS*	Valid for ACS*	Total*
Invalid for Census .....	37,735,008	16,406,589	54,141,597
Valid for Census <sup>+</sup> .....	5,410,540	120,642,711	126,053,251
Total .....	43,145,548	137,049,300	180,194,848

\*Counts are unweighted.

<sup>+</sup>Excludes 8,405,549 valid Census units that were not present on the January 2009 ACS MAFX.

Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.

By using the results from AC as ground truth, Table 3 shows that the January 2009 ACS sample universe contained 5,410,540 falsely invalid records (Type I Error) and 16,406,589 falsely valid records (Type II Error). Not accounted for in Table 3 are approximately 8.4 million addresses that were not present on the January 2009 ACS MAFX but were valid Census units on the 2010 Census EU files. These records are not included because this analysis focuses on the ability of the filter rules to correctly identify valid records currently on the MAF. Therefore, the net coverage results in this report are calculated using the total number of “Valid for Census” records. The goal of this study was to use data mining methods to find a rule or set of rules to offer some correction to these Type I and II Errors. Using the stated benchmark, a 0.25 percent or greater increase in valid records translates to an addition of approximately 340,000 or more valid records that were previously invalid for ACS.

### 3.5 Cost

This evaluation incurred costs from both U.S. Census Bureau Headquarters (HQ) staff and contractor staff. This evaluation spanned a period of approximately two years, with an estimated cost of about \$515,733. This amount accounts for three federal employees, including overheads, working on the evaluation in some capacity over the project lifecycle. This amount also includes contract costs totaling \$242,754. The contract work was performed by Sabre Systems, Inc. The final incurred contract costs were \$25,601 less than the contract award, or about 10 percent under budget. Lastly, a small amount of costs were incurred to license software that was not under an existing U.S. Census Bureau license. These licensing costs totaled about \$16,000.

## 4. Limitations

- Validating models with ground truth measures depends upon the quality and accuracy of defined ground truth. In this study, the data resulting from the 2010 AC operation were considered ground truth. Errors in field work and processing will affect the integrity of the data mining models and any extrapolation of the results.
- Any undercoverage measure calculated for the rules generated could not include units that are not on the MAF.

## 5. Results

### 5.1 Rules

The results of this evaluation generated six MAF filter rules, complementary to the existing MAF filter rules.

The first set of data mining models used a dataset created by merging 2010 Pre-AC variables and the target variable (DSAF) from the 2010 EU files, by MAFID. WEKA used a training dataset of approximately 180,000 randomly selected records to create the decision tree model. The model used the DAAL Address Flag (DAALAF) variable in all the rules. DAAL is a post-Census 2000 program that coordinates various operations related to the review and automated update of the geographic content of the TIGER<sup>®</sup> database and the addresses in the MAF. The definition of the DAALAF variable is:

Set DAALAF = 1 (YES) if:

1. Valid ACS (ACSAF != 0) or
2. Valid address (unitstat = 1) or
3. Nonexistent units (unitstat = 4) and source of DAAL or FACHS (MAFSRC in {061,062,065,066,012,063,064}) or
4. Demolished units (unitstat = 2) and MAFSRC from 3 above or
5. Provisional adds (unitstat=5) and an entry in the street name field.

The first four rules that merited further analysis and validation were:

**Rule 1.** *If DAALAF = 1 and DSFSR08 = 1 and FIRSTSRC in {02,09,13,19,26}, then DSAF = Y*

If the HU was flagged as eligible for the DAAL, flagged as residential on the Spring 2008 DSF, and the source of the oldest operation on record existing on the MAF Operations (MAFOP) table was one of six sources -- 1990 Address Control File (ACF), 2000 Address Listing (AL) operation, sent to 2000 Local Update of Census Addresses (LUCA), 2000 Rural Update/Leave (U/L) operation or 2000 Block Canvassing (BC) -- then it was a valid unit for Census operations.

**Rule 2.** *If DAALAF = 1 and DSFSR08 = 1, then DSAF = Y*

If the housing unit was flagged as eligible for the DAAL and flagged as residential on the Spring 2008 DSF, then it was a valid unit for Census operations.

**Rule 3.** *If DAALAF = 1 and DSFSPR08 = 1 and FIRSTSRC = 02 and LUCAAF = Y, then DSAF = Y*

If the housing unit was flagged as eligible for the DAAL, flagged as residential on the Spring 2008 DSF, the source of the oldest operation on the record was the 1990 ACF, and eligible for 2010 LUCA, then it was a valid unit for Census operations.

**Rule 4.** *If DAALAF = 1 and DSFSPR08 = 1 and FIRSTSRC = 02, then DSAF = Y*

If the housing unit was flagged as eligible for the DAAL, flagged as residential on the Spring 2008 DSF, and the source of the oldest operation on the record was the 1990 ACF then it was a valid unit for Census operations.

Rules 1, 3, and 4 are subsets of Rule 2, but were still validated to investigate if they would exhibit an improvement in the Type I/II Error Ratio. These rules were validated using the full January 2009 ACS MAFX consisting of approximately 180 million MAF records.

	Invalid for ACS*	Valid for ACS*	Total*
Invalid for Census .....	221,421	3,296,141	3,517,562
Valid for Census .....	708,940	95,979,001	96,687,941
Total.....	930,361	99,275,142	100,205,503

\*Counts are unweighted.  
Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.

	Type I/II Error Ratio *	3.20
Improvement/ Degradation <sup>1</sup>	Gross Undercoverage <sup>++</sup> .....	-0.56%
	Gross Overcoverage <sup>++</sup> .....	0.18%
	Total Error *	-0.39%

\*Counts and percentages are unweighted.  
<sup>+</sup>Percentages may not sum to Total Error due to rounding.  
<sup>1</sup>All valid EU addresses, less those not present on the January 2009 ACS MAFX, were used as the denominator (126,053,251).  
Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.

Table 4 shows that Rule 1 selected 100,205,503 records with 96.49 percent (96,687,941/100,205,503) accurately classified as valid for Census. Of the selected records, 708,940 records were valid for Census operations while being invalid for the ACS. From Table 5, if Rule 1 were implemented with the 2009 ACS filter, it would result in a 0.56 percent decrease in the gross undercoverage, while adding 221,421 records that were deemed invalid for Census operations. Rule 1 does not meet the benchmark ratio of 4:1, but exceeds the benchmark

for reducing undercoverage by 0.25 percent or greater. Table 5 shows that total error would decrease by 0.39 percent after factoring in the additional overcoverage of the new rule.

	Invalid for ACS*	Valid for ACS*	Total*
Invalid for Census .....	548,783	4,425,078	4,973,861
Valid for Census .....	1,410,548	106,140,432	107,550,980
Total.....	1,959,331	110,565,540	112,524,841

\*Counts are unweighted.  
Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.

	Type I/II Error Ratio * .....	2.57
Improvement/ Degradation <sup>1</sup>	Gross Undercoverage *+ .....	-1.12%
	Gross Overcoverage *+ .....	0.43%
	Total Error * .....	-0.68%

\*Counts and percentages are unweighted.  
<sup>1</sup>Percentages may not sum to Total Error due to rounding.  
<sup>1</sup>All valid EU addresses, less those not present on the January 2009 ACS MAFX, were used as the denominator (126,053,251).  
Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.

Table 6 shows that Rule 2 selected 112,524,841 records with 95.58 percent (107,550,980/112,524,841) accurately classified as valid for Census. By adding Rule 2 to the existing ACS filter, the ACS frame would see an increase of 1,410,548 valid records or 1.12 percent. With these records, 548,783 records that were invalid for both operations would be added giving the rule a Type I/II Error Ratio of 2.57:1. By relaxing Rule 1’s conditions on the variable FIRSTSRC, Rule 2 increases the improvement of valid records to 1.12 percent, nearly double the improvement of Rule 1 and more than quadruple the benchmark of 0.25 percent. The increase of this percentage comes at the cost of a reduced Type I/II Error Ratio from 3.20:1 to 2.57:1. However, the new rule also decreases total error by 0.68 percent.

	Invalid for ACS*	Valid for ACS*	Total*
Invalid for Census .....	18,975	1,824,076	1,843,051
Valid for Census .....	31,157	67,826,773	67,857,930
Total.....	50,132	69,650,849	69,700,981

\*Counts are unweighted.  
Source: January 2009 ACS MAFX and 2010 Census AF COMBO file



	Type I/II Error Ratio * .....	1.64
Improvement/ Degradation <sup>1</sup>	Gross Undercoverage *+ .....	-0.02%
	Gross Overcoverage *+ .....	0.02%
	Total Error * .....	0.00%
<small>*Counts and percentages are unweighted.  <sup>+</sup>Percentages may not sum to Total Error due to rounding.  <sup>1</sup>All valid EU addresses, less those not present on the January 2009 ACS MAFX, were used as the denominator (126,053,251).  Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.</small>		

Table 8 shows that Rule 3 selected 69,700,981 records with 97.36 percent (67,857,930/69,700,981) accurately classified as valid for Census. With only 31,157 previously invalid records being added by Rule 3, the improvement to gross undercoverage of 0.02 percent is far below the benchmark of 0.25 percent. In addition to the low improvement percentage, the Type I/II Error Ratio of Rule 3 is also well below the benchmark ratio at 1.64:1.

	Invalid for ACS*	Valid for ACS*	Total*
Invalid for Census .....	82,445	1,828,859	1,911,304
Valid for Census .....	205,413	67,845,210	68,050,623
Total .....	287,858	69,674,069	69,961,927
<small>*Counts are unweighted.  Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.</small>			

Table 10 shows that Rule 4 selected 69,961,927 records with 97.27 percent (68,050,623/69,961,927) accurately classified as valid for Census. Rule 4 added 205,413 records to the 2009 ACS filter reducing gross undercoverage by 0.16 percent. With the 82,445 invalid records that are added by Rule 4, the Type I/II Error Ratio is 2.49:1. As we compare Rule 4 to Rule 3, the addition of LUCA status in Rule 3 greatly diminishes its ability to predict valid units. Rule 4 results in an improved Type I/II Error Ratio and reduction in gross undercoverage. The ratio of Rule 4 is similar to the 2.57:1 of Rule 2, but the 0.16 percent decrease in gross undercoverage of Rule 4 falls well below the 1.12 percent rate gross undercoverage improvement of Rule 2.

	Type I/II Error Ratio * .....	2.49
Improvement/ Degradation <sup>1</sup>	Gross Undercoverage <sup>*+</sup> .....	-0.16%
	Gross Overcoverage <sup>*+</sup> .....	0.07%
	Total Error * .....	-0.10%

\*Counts and percentages are unweighted.  
<sup>+</sup>Percentages may not sum to Total Error due to rounding.  
<sup>1</sup>All valid EU addresses, less those not present on the January 2009 ACS MAFX, were used as the denominator (126,053,251).  
Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.

After generating the first 4 rules, the infoUSA data was included in the dataset for analysis. The target variable remained the same, the DSAF on the EU file. As opposed to the previous analyses, these models used SAS EM to construct the filter rules. By using this software, a much larger training dataset could be used to build the decision tree model. The partitions of the data for the model were a random sample of 40 percent for training, 30 percent for validation, and 30 percent for testing. The model produced one rule for determining valid HUs, and a second for determining invalid HUs. By construction, this dataset has an inherent variable for any rule that will be found from the model. The main analytical file for these sets of rules only kept records that merged with the infoUSA file. It follows, any rule will be conditioned on each record's presence on the infoUSA file.

**Rule 5.** *If DAALAF = 1, DSFSR08 = 1, and IUSA\_FLAG = 1, then DSAF = Y*

If the HU was flagged as eligible for DAAL, flagged as residential on the Spring 2008 DSF, and present on the infoUSA file then it was a valid unit for Census operations.

	Invalid for ACS*	Valid for ACS*	Total*
Invalid for Census .....	392,022	3,687,204	4,079,226
Valid for Census .....	1,228,780	101,976,047	103,204,827
Total .....	1,620,802	105,663,251	107,284,053

\*Counts are unweighted.  
Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.

Table 13. 2010 CPEX Data Mining: Rule 5 Performance		
	Type I/II Error Ratio * .....	3.13
Improvement/ Degradation <sup>1</sup>	Gross Undercoverage *+ .....	-0.97%
	Gross Overcoverage *+ .....	0.31%
	Total Error * .....	-0.66%
<small>*Counts and percentages are unweighted.            †Percentages may not sum to Total Error due to rounding.  <sup>1</sup>All valid EU addresses, less those not present on the January 2009 ACS MAFX, were used as the denominator (126,053,251).            Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.</small>		

Rule 5 selected 107,284,053 records with 96.20 percent (103,204,827/107,284,053) accurately classified as valid for Census. Of the selected records, about 1.2 million records would be correctly added to the 2009 ACS filter, for an improvement of 0.97 percent. The invalid records accounted for 392,022 of the total number of records giving Rule 5 a Type I/II Error Ratio of 3.13:1. Rule 5 differs from Rule 2 only in that Rule 5 includes a flag to indicate the presence of an address on the infoUSA file. This additional flag decreased the effectiveness by 0.15 percent, but increased the Type I/II Error Ratio to 3.13:1, which is a marginal decline compared to the 3.2:1 ratio of Rule 1. Table 13 shows that Rule 5 has a similar effect on total error as Rule 2, with a decrease of 0.66 percent.

Previous rules predicted “Valid for Census” records. Rule 6 used the records that do not qualify for Rule 2, and predicted that these records were invalid for Census.

**Rule 6.** *If DAALAF = 0 and DSFSR08 = 0, 2, or 3, then DSAF = N*

If the HU was not flagged as eligible for the DAAL and not flagged as residential on the Spring 2008 DSF, then it was not a valid unit for Census operations.

Table 14. 2010 CPEX Data Mining: Rule 6 Validation			
	Invalid for ACS*	Valid for ACS*	Total*
Invalid for Census .....	27,993,566	1,658,223	29,651,789
Valid for Census .....	494,197	419,193	913,390
Total .....	28,487,763	2,077,416	30,565,179
<small>*Counts are unweighted.            Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.</small>			

Rule 6 selected 30,565,179 records with 97.01 percent (28,487,763/30,565,179) accurately classified as invalid for Census. Of the total number of records, approximately 1.7 million previously valid ACS records were selected and correctly predicted to be invalid records. In contrast, the rule incorrectly selected 419,193 records that were valid for both Census and ACS. The Type I/II Error Ratio is 0.25 which means for about every 4 records Rule 6 correctly identifies as invalid, one falsely invalid record occurs. The changes in undercoverage and

overcoverage are different than previous rules because Rule 6 removes records currently valid for ACS instead of adding records to the ACS Address Frame. Table 14 shows that the removal of the records selected by Rule 6 would reduce gross overcoverage by 1.32 percent at the cost of increasing gross undercoverage by 0.33 percent. Since undercoverage is more costly to correct for in a decennial census or current survey (historically), the benchmarks considered for the results of the first five rules will not apply in the same manner to Rule 6.

<b>Table 15. 2010 CPEX Data Mining: Rule 6 Performance</b>		
	Type I/II Error Ratio * .....	0.25
Improvement/ Degradation <sup>1</sup>	Gross Undercoverage *+ .....	0.33%
	Gross Overcoverage *+ .....	-1.32%
	Total Error * .....	-0.98%
<small>*Counts and percentages are unweighted.            †Percentages may not sum to Total Error due to rounding.  <sup>1</sup>All valid EU addresses, less those not present on the January 2009 ACS MAFX, were used as the denominator (126,053,251).            Source: January 2009 ACS MAFX and 2010 Census AF COMBO file.</small>		

## 5.2 Projections for 2010 and 2011

Each of the previous rules was validated using the 2009 January ACS MAFX. In order to understand the impact of any new rule, we made projections of the rule results for 2010 and 2011. infoUSA files for the projected years were not readily available, so the infoUSA records used with the 2009 ACS MAFX were merged with the 2010 and 2011 ACS data to apply Rules 2 and 5. We believe the projections would be more accurate if the vintage of the infoUSA corresponded with the vintage of the ACS MAFXs used, but using the available data will provide a rough projection of the impact of Rule 5. The ACS extracts did not contain the DAALAF flag, so it was recreated from appropriate variables on the extracts.

By applying Rule 5, the data showed which records were predicted as valid. The ratio of Rule 5 validated against the 2010 AC operation (3.13:1) was then applied to predict how many valid and invalid records would be added if the rule was implemented in 2010 and 2011. For the projections, Rule 5 was modified to include each previous year's DSF Spring status, DSF Spring 2009 and DSF Spring 2010 for 2010 and 2011 respectively (instead of using the status of the Spring 2008 DSF). This allows the rule to filter new records added from the DSF for each year.

**Table 16. 2010 CPEX Data Mining: Projected Records Selected by Rule 2 and Rule 5**

		Rule 2		Rule 5	
		2010*	2011*	2010*	2011*
Currently Valid ACS Records <sup>1</sup>		136,234,120	136,594,285	136,234,120	136,594,285
Records Selected by Rule	Total	116,311,949	117,701,908	82,996,609	83,130,713
	ACS Valid	112,514,500	114,119,904	79,960,081	80,541,814
	ACS Invalid	3,797,449	3,582,004	3,036,528	2,588,899

\*Counts are unweighted.

<sup>1</sup>Valid ACS Records were determined using the January 2010 ACS MAFX specifications (Bates, 2010a) and the January 2011 ACS MAFX specifications (Bates, 2010b).

Source: January 2010 ACS MAFX, January 2011 ACS MAFX, and 2009 infoUSA file.

Table 16 shows that for the January 2010 ACS MAFX and the January 2011 ACS MAFX, Rule 5 selected approximately 83 million records of a possible 136 million records in both years. Rule 2 selected about 116 million records in 2010 and almost 118 million records in 2011. Records that were originally classified as invalid for ACS totaled 3,797,449 records for Rule 2 and 3,036,528 records for Rule 5 in 2010. In 2011, records that were classified as invalid for ACS totaled 3,582,004 records for Rule 2 and 2,588,899 records for Rule 5.

**Table 17. 2010 CPEX Data Mining: 2010 and 2011 Projection Performance of Rule 2 and Rule 5**

		Rule 2		Rule 5	
		2010*	2011*	2010*	2011*
Previously Invalid ACS Records		3,797,449	3,582,004	3,036,528	2,588,899
Projection	True ACS Valid	2,733,738	2,578,642	2,301,291	1,962,047
	False ACS Valid	1,063,711	1,003,362	735,237	626,852
Improvement/ Degradation <sup>1</sup>	Gross Undercoverage <sup>+</sup>	-2.01%	-1.89%	-1.69%	-1.44%
	Gross Overcoverage <sup>+</sup>	0.78%	0.73%	0.54%	0.46%
	Total Error	-1.23%	-1.15%	-1.15%	-0.98%

\*Counts and percentages are unweighted.

<sup>+</sup>Percentages may not sum to Total Error due to rounding.

<sup>1</sup>All valid ACS universe addresses were used as the denominator: 136,234,120 for 2010 and 136,594,285 for 2011.

Source: January 2010 ACS Extract, January 2011 ACS Extract, and 2009 infoUSA file.

Table 17 provides projections of valid and invalid records using the Type I/II Error Ratio of 2.57:1 for Rule 2 and 3.13:1 for Rule 5. For 2010, it is projected that for Rule 2 approximately 2.7 million addresses would be correctly added causing a 2.01 percent decrease in gross undercoverage. Similarly in 2011, it is projected that Rule 2 would identify about 2.6 million additional addresses, garnering a 1.89 percent decrease in gross undercoverage. Of the 3,036,528 previously invalid records in 2010, it is projected that Rule 5 would classify 2,301,291

additional valid addresses, thus reducing gross undercoverage by 1.69 percent. For 2011, we project that Rule 5 would correctly categorize 1,962,047 of 2,588,899 address records causing a reduction in gross undercoverage of 1.44 percent. By applying Rule 2 to the January 2010 and January 2011 ACS MAFX, Table 17 shows that it would reduce total error by 1.23 percent and 1.15 percent respectively. Table 17 also shows that Rule 5 reduces total error for 2010 and 2011 by 1.15 percent and 0.98 percent respectively.

## 6. Conclusions and Recommendations

The results shown in the previous section indicate that data mining techniques have the potential to enhance current ACS filter rules to generate a more complete ACS Address Frame for sampling.

### 6.1 Conclusions

The analyses here evaluated six new potential filter rules for ACS. Two independent software packages, WEKA and SAS EM, constructed similar decision tree models reinforcing the importance of the DAAL status variable and the Spring 2008 DSF residential status of records in predicting valid records on the frame. Rule 1 (WEKA), Rule 2 (WEKA), and Rule 5 (SAS EM) all used at least these two variables to predict the validity of a record.

	Type I/II Error Ratio	Improvement/Degradation <sup>1</sup>		
		Gross Undercoverage** (%)	Gross Overcoverage** (%)	Total Error* (%)
Rule 1 .....	3.20	-0.56	0.18	-0.39
Rule 2 .....	2.57	-1.12	0.43	-0.68
Rule 3 .....	1.64	-0.02	0.02	0.00
Rule 4 .....	2.49	-0.16	0.07	-0.10
Rule 5 .....	3.13	-0.97	0.31	-0.66
Rule 6 .....	0.25	0.33	-1.32	-0.98

\*Counts and percentages are unweighted.  
 \*\*Percentages may not sum to Total Error due to rounding.  
<sup>1</sup>All valid EU addresses, less all AC new adds, were used as the denominator (126,053,251).  
 Source: 2010 CPEX DM Tables 4, 6, 8, 10, 12, and 14.

From Table 18, Rule 1 shows a reduction of gross undercoverage in the 2009 ACS Address Frame of 0.56 percent while Rule 2 showed a reduction of 1.12 percent. However, the increased gross undercoverage reduction gained by Rule 2 was at the cost of a reduced Type I/II Error Ratio of 2.57, while Rule 1 maintained a ratio of 3.20:1. Both gross undercoverage reduction rates were above the benchmark of 0.25 percent, but both ratios fell short of the 4:1 benchmark

established at the beginning of the evaluation. Both ratios indicate an increase in gross overcoverage, which is more manageable within the ACS survey processes than an increase to gross undercoverage. The decrease in gross undercoverage by 0.56 percent for Rule 1, and even more so the 1.12 percent for Rule 2, may outweigh the cost of increased overcoverage. The decennial census and the ACS have operations and processes that can identify erroneous inclusions which are simpler to adjust for than undercoverage and its associated erroneous exclusions.

Rule 5 showed a reduction in gross undercoverage in the 2009 ACS Address Frame of 0.97 percent with a Type I/II Error Ratio of 3.13:1. The inclusion of infoUSA records into the data mining process refined Rule 2 to increase the accuracy of valid addresses while maintaining an improvement rate of nearly 1 percent (0.97). It is projected that Rule 2 and Rule 5 would cause a 0.68 percent and 0.66 percent, respectively, reduction in total error. Rule 6 showed the largest decrease in total error with a value of 0.98 percent. However, this change occurred with a decrease in overcoverage and an increase in undercoverage which is the opposite of the observed changes in Rules 1 through 5.

When projected on the ACS MAFX for 2010 and 2011, Rule 2 and Rule 5 showed great potential to contribute additional (not previously identified) valid addresses. Since each rule utilized the Spring 2008 DSF variable, we used the updated DSF variables (Spring 2009 and Spring 2010 respectively) when projecting results for 2010 and 2011. In 2010 and 2011, Rule 2 projected to reduce gross undercoverage by 2.01 percent and 1.89 percent, respectively. Rule 5 is projected to reduced gross undercoverage by 1.69 percent in 2010 and 1.44 percent in 2011. These rates could be improved by using synchronized infoUSA vintages along with the updated ACS extracts.

The data mining process revealed that the presence of a record on the infoUSA dataset increased the accuracy of the model's prediction of valid HUs. Other variables appeared in the decision tree, but did not have the required accuracy to be considered as potential rules. In its raw form, the infoUSA data contain information about individuals and their corresponding addresses. This structure allows for multiple entries for one address. To be used in conjunction with any MAFX, the dataset must first be collapsed into a dataset with unique entries for each address (MAFID). The process used in this evaluation could be improved upon, creating the chance for infoUSA variables to take on more prominent roles in predicting valid addresses. The infoUSA dataset also contains a large amount of missing values for certain variables. The data mining algorithm rejects variables for analysis when a large portion of the values are missing. The process to reduce the dataset may also be improved upon if imputation techniques were applied to simulate values for those that are missing. This may allow more variables from infoUSA to affect each rule's accuracy in predicting valid addresses.

## 6.2 Recommendations

Based on these findings, the DSSD provides the following recommendations:

1. **Profile Rule 2 records, and if validated apply Rule 2 to the July 2012**

**American Community Survey Extract:** The units affected by Rule 2 should be further profiled through a joint DSSD and ACSO partnership to garner any additional information about their characteristics. This task should identify geographic (clustering/dispersion across block, tract, city, etc.) and physical (single/multiple unit, group quarters/housing unit class, etc.) distributions, as well as final Census 2010 status (final validity, vacant/occupied, population count, etc.). If favorable results are obtained from profiling, the July 2012 ACS filter rules should be updated to include Rule 2, given the potential to reduce gross undercoverage and total error shown by the analysis. Rule 2 only uses variables currently available on the Master Address File, which allows quick implementation.

2. **Identify new Administrative Record data sources:** Rule 5 shows that the 2009 infoUSA file proved valuable in confirming the accuracy of valid addresses on the MAF (where address validity is determined by the 2010 Census AC operation outcome). Confirmation of an address on many different, independent data sources, and the various permutations of these data sources, only increases the potential results of future data mining research efforts.
3. **Use SAS Enterprise Miner:** SAS EM should be given primary consideration in continuing data mining research. The software is included in the Census Bureau's site license, has the capacity to work with the large datasets necessary for census research, and has an easy-to-use user interface. The other packages tested here did not provide all of these benefits.
4. **Continue Data Mining Research:** Due to the success of the research here, address frame data mining research using the 2010 AC data should continue. A data mining evaluation is currently part of the planned 2020 Census projects. With additional resources, MAF filter rules could be further improved.
5. **Conduct Verification:** In order to ensure accuracy, develop a system to test (field and office) additional MAF filter rules prior to, or in parallel with, the inclusion of the ACS sample universe.



## **7. Acknowledgements**

This report is the product of numerous persons. I wish to thank Sonja Clark for her work on the study plan and early analyses. Also, I owe many thanks to Kevin Shaw, Jennifer Reichert, and David Whitford for their guidance and comments throughout the evaluation; and to Kevin Shaw for his methodological direction throughout the project. Thanks to Arti Khatwani, of Sabre Systems, Inc., for her work with WEKA, the infoUSA file, and the numerous analyses she conducted for this evaluation. Thanks to Jack George for his assistance and help in validating and profiling the data mining results. Thanks to James Gibbs for his fine leadership of the Sabre Systems, Inc. contract. Thanks to Valeria Baker, Ann Dimler and Jennifer Reichert for their work as the Contracting Officer (CO), Contracting Officer's Representative (COR) and Task Manager (TM) respectively. Thanks to Larry Bates for his overall support of our efforts and providing the specifications for the ACS filters. Thanks to Deb Wagner, Aneesah Williams, Damon Smith, and Teresa Schellhamer in DID for their help in matching and delivering the infoUSA file. And lastly, many thanks to Claude Jackson for his endless and invaluable IT support.

## 8. References

Address List Operations Implementation Team (2011), “2010 Census Address Canvassing Operational Assessment,” 2010 Census Planning Memorandum Series No. 168, January 17, 2012.

Bates, Lawrence (2009), “Editing the MAF Extracts and Creating the Unit Frame Universe for the American Community Survey (2009 Main Phase),” DSSD American Community Survey Memorandum Series, January 5, 2009.

Bates, Lawrence (2010a), “Editing the MAF Extracts and Creating the Unit Frame Universe for the American Community Survey (2010 Supplemental Phase),” DSSD American Community Survey Memorandum Series, June 4, 2010.

Bates, Lawrence (2010b), “Editing the MAF Extracts and Creating the Unit Frame Universe for the American Community Survey (2011 Supplemental Phase),” DSSD American Community Survey Memorandum Series, December 29, 2010.

Bauder, Mark, D. H. Judson (2003) “Administrative Records Experiment in 2000 (AREX 2000) Household Level Analysis,” U.S. Census Bureau, April 17, 2003, page i.

Boies, John L., Kevin M. Shaw, Jonathan P Holland (2011), “DRAFT 2010 Census Program for Evaluations and Experiments (CPEX): Study of AC Targeting and Cost Reduction, Report,” 2010 Census Program for Evaluations and Experiments, November 29, 2011.

Burcham, Joseph A. (2002), “Block Canvassing Operation,” U.S. Census Bureau, April 5, 2002, page i.

Clark, Sonja, (2009), “2010 Census Study Plan: Evaluation of Data-Based Extraction Processes for the Address Frame,” DSSD 2010 Decennial Census Memorandum Series #O-A-06 (August 2009), 2010 Census Planning Memorandum Series No. 64, December 4, 2009.

Clark, Sonja (2010), “2010 CPEX Evaluation of Data-Based Extraction Processes for the Address Frame – Sabre Systems Deliverable #1,” DSSD 2010 Decennial Census Memorandum Series, September 14, 2010.

Colosi, Robert and Amanda S. Hakanson (2007), “Using Data Mining to Define Valid Records on the MAF,” U.S. Census Bureau, May 7, 2007.

Dixon, Kelly, Melissa Blevins, Robert Colosi, Amanda Hakanson, Nancy Johnson, Karen Owens, Matt Stevens, and Christine G. Tomaszewski (2008), “2008 Dress Rehearsal Address Canvassing Assessment Report,” U.S. Census Bureau, April 15, 2008.

Garcia, Mayra (2009), “2010 Decennial Census Study Plan: Study of Address Canvassing Targeting and Cost Reduction,” DSSD 2010 Decennial Census Memorandum Series #O-A-08 (September 2009), 2010 Census Planning Memorandum Series No. 63, September 29, 2009.

Goldenkoff, Robert (2009), “2010 CENSUS: Efforts to Build an Accurate Address List Are Making Progress, but Face Software and Other Challenges,” Testimony Before the Subcommittee on Information Policy, Census, and National Archives, Committee on Oversight and Government Reform, House of Representatives, October 21, 2009.

Gordon, Judith J. (2009a), “Recommendations from 2010 Census: First Quarterly Report to Congress, August 2009 (OIG-19791-1),” August 14, 2009.

Gordon, Judith J. (2009b), “Reviews of 2010 Address Canvassing Operations. Including Activities Related to the American Recovery and Reinvestment Act,” March 6, 2009.

Holland, Jonathan, P., Matthew Virgile, (2010), “2010 Decennial Census Study Plan: Study of Automation in Field Data Collection for Address Canvassing,” DSSD 2010 Decennial Census Memorandum Series #O-A-02 (November 2010), 2010 Census Planning Memorandum Series No. 65, August 12, 2010.

Holland, Jonathan P (2011), “DRAFT 2010 Census Program for Evaluations and Experiments (CPEX): Evaluation of Automation in Field Data Collection in Address Canvassing Report,” 2010 Census Program for Evaluations and Experiments, June 1, 2012.

Johnson, Nancy, (2011), “2010 Decennial Census Study Plan: Evaluation of Address Frame Accuracy and Quality,” DSSD 2010 Decennial Census Memorandum Series #O-A-3R (June 2011), 2010 Census Planning Memorandum Series No. 146, June 14, 2011.

Kephart, Kathleen (2010), “National Estimate of Coverage of the Master Address File for 2009.” U.S. Census Bureau, November 30, 2010.

Mah, Ming-Yi and Dean Resnick (2007) “Preliminary Analysis of Medicaid Enrollment Status in the Current Population Survey,” Medicaid Undercount Project (SNACC), September 27, 2007.

Martin, Joel and Clifford Loudermilk (2008), "Frame Assessment for Current Household Surveys (FACHS) Filter Rules Research: 2008 Analytical Report," U.S. Census Bureau, October 9, 2008.

Ruhnke, Megan C (2002), "The Address Listing Operation and Its Impact on the Master Address File," U.S. Census Bureau, January 30, 2002, page i.

Russell, Chad E. (1992), "Results of the Precanvass Suppression Study," U.S. Census Bureau, April 13, 1992.

Stuart, Elizabeth, A., Judson, D.H. (2003) "An empirical evaluation of the use of administrative records to predict census day residency," 2003 Proceedings of the American Statistical Association, Section on Government Statistics, 2003.

Tomaszewski, Christine, G. (2010), "2010 Decennial Census Study Plan: Evaluation of Address List Maintenance Using Supplemental Data Sources," DSSD 2010 Decennial Census Memorandum Series #O-A-01 (March 2010), 2010 Census Planning Memorandum Series No. 68, April 5, 2010.

Vitrano, Frank A., Robin A. Pennington, and James B. Treat (2004), "Census 2000 Testing, Experimentation, and Evaluation Program Topic Report No. 8, TR-8, Address List Development in Census 2000," U.S. Census Bureau, March 2004, page ii.

Virgile, Matt, (2010), "2010 Census Program for Evaluations and Experiments Study Plan: Evaluation for Small Multi-Unit Structures," DSSD 2010 Decennial Census Memorandum Series #O-A-14 (January 2010), 2010 Census Planning Memorandum Series No. 66, March 15, 2010.

Virgile, Matt (2012), "2010 Census Program for Evaluations and Experiments: Evaluation of Small Multi-Unit Structures Report," 2010 Census Program for Evaluations and Experiments, February 24 2012.

Ward, Justin (2011a), "2010 CPEX Evaluation of Data-Based Extraction Processes for the Address Frame – Sabre Systems Deliverable #2," DSSD 2010 Decennial Census Memorandum Series, July 5, 2011.

Ward, Justin (2011b), "DRAFT 2010 CPEX Evaluation of Data-Based Extraction Processes for the Address Frame – Sabre Systems Deliverable #3," DSSD 2010 Decennial Census Memorandum Series, October 3, 2011.

## Appendix A: SAS Enterprise Miner 6.2 Report

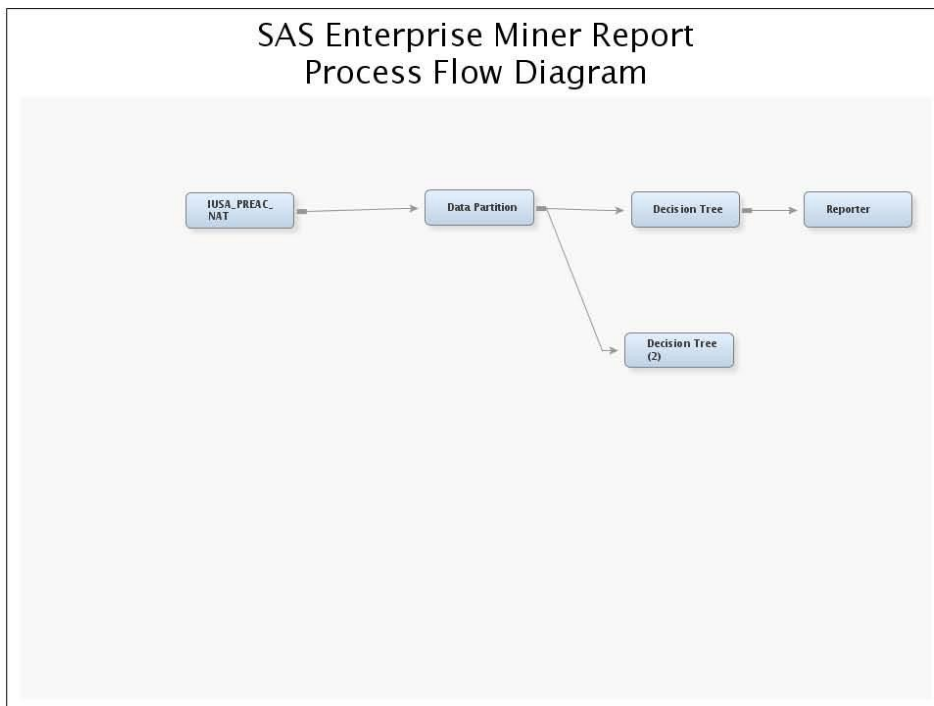
### SAS Enterprise Miner Report

1

User = ward0330  
Date = 10:36:22 November 01  
Project = DM 2  
Diagram = DM

Start Node = Report  
Node label = Reporter  
Nodes = PATH  
Showall = N

Format = PDF  
Graphics = GIF  
Style = LISTING



## SAS Enterprise Miner Report

### Node=IUSA\_PREAC\_NAT Summary

Node id = Ids  
 Node label = IUSA\_PREAC\_NAT  
 Meta path = Ids  
 Notes =

### Node=IUSA\_PREAC\_NAT Properties

Property	Value	Default	Property	Value	Default
Component	DataSource		IdentifyEmptyColumns	Y	
ApplyIntervalLevelLowerLimit	Y		IntervalLowerLimit	20	
ApplyMaxClassLevels	Y		Library	DM	
ApplyMaxPercentMissing	Y		MaxClassLevels	20	
CMeta	WORK.M1CEDQG9		MaxPercentMissing	50	
ComputeStatistics	N		MetaAdvisor	BASIC	
DBPassThrough	Y		NBytes	.	
Data	DM.IUSA_PREAC_NAT		NCols	289	.
DataSelection	DATASOURCE		NObs	-1	.
DataSource	IUSAPREACNAT		NewTable		
DataSourceRole	RAW		NewVariableRole	REJECT	
Description			OutputType	VIEW	
DropMapVariables	N	Y	Role	RAW	TRAIN
DsCreateDate	1632153095.2		Sample	D	
DsCreatedBy	ward0330		SampleSizeObs	10000	
DsId	IUSAPREACNAT		SampleSizePercent	20	
DsModifiedBy	ward0330		SampleSizeType	PERCENT	
DsModifyDate	1632153095.3		Scope	LOCAL	
DsSampleName			Segment		
DsSampleSize			Table	IUSA_PREAC_NAT	
DsSampleSizeType			VariableValidation	STRICT	
DsScope	LOCAL				

### Node=IUSA\_PREAC\_NAT Data Attributes

Attribute	Value	Attribute	Value
Data Label		Date Modified	15SEP11:15:58:37
Data Library	DM	Engine	SASDSV
Data Name	IUSA_PREAC_NAT	Number Columns	289
Data Size	.	Number Rows	-1

Attribute	Value	Attribute	Value
Data Type	VIEW	Role	RAW
Date Created	15SEP11:15:58:37	Segment	

## Node=IUSA\_PREAC\_NAT Variable Summary

Role	Level	Frequency Count	Name
TARGET	BINARY	1	enum_dsaf
REJECTED	BINARY	11	preac_bkcoordflg preac_colbikst preac_gqareacode preac_gqcontact preac_lozcpiclass preac_mailrdesc preac_mailstbidnd preac_maistnmsrc preac_mailzpclass preac_tabikst preac_tabstate
REJECTED	INTERVAL	8	family_id preac_gqid preac_mspid preac_oid preac_oidcb preac_oidtb preac_spid preac_survmfid
REJECTED	NOMINAL	76	EFFECT_PHONE_DT HH_ARR_DT HOME_SALE_DT PHONE_ACQUIS_DT PHONE_PUB_DT RESIDENCY_NEW RESIDENCY_OLD a_ref_dt max_find_income max_home_age max_residence_length min_find_income min_home_age ...
REJECTED	UNARY	30	A_Drop_Ind Addr_typ0 Addr_typ3 CSA_HOME_EQU_EST ROUTE_TYP_G avg_surname_cnt preac_acsaf preac_altcoordtyp preac_ctufvaf preac_ctufunrv preac_colbiksufx1 preac_colbiksufx2 preac_collms ...
INPUT	BINARY	68	A_Bus_File_Hit A_Deceased_Flg A_Fulfill_Flg A_Nocall_Flg A_Nomail_Flg A_Spouse_Ind A_Workathome_Flg Addr_typ2 Addr_typ4 Addr_typ5 CHILD_INDICATOR DEL_UNIT_SIZE DEMO_LV_IND DM_HIGH_TECH_IND ...
INPUT	INTERVAL	29	A_Married_CD_S A_age_max A_age_min A_sex_F A_sex_M Addr_typ1 CEN2000_GEO_MATCH_LV0 LAT_LONG_MTCH_LV0 MAFID USPS_DEL_SERV_C USPS_DEL_SERV_N avg_Occupancy_ct avg_adult_cnt avg_child_cnt ... count_N
INPUT	NOMINAL	66	A_Married_CD_M A_Married_CD_U CEN2000_GEO_MATCH_LV2 CEN2000_GEO_MATCH_LV4 CEN2000_GEO_MATCH_LVX HH_SIZE_1_2 LAT_LONG_MTCH_LV2 LAT_LONG_MTCH_LV4 LAT_LONG_MTCH_LVX RECENCY_DT RECENCY_NEW RECENCY_OLD ...

## SAS Enterprise Miner Report

### Node=Data Partition Summary

Node id = Part  
 Node label = Data Partition  
 Meta path = Ids => Part  
 Notes =

### Node=Data Partition Properties

Property	Value	Default	Property	Value	Default
Component	Partition		RandomSeed	12345	
ClassDistribution	Y		TestPct	30	
IntervalDistribution	Y		TrainPct	40	
Method	DEFAULT		ValidatePct	30	
OutputType	DATA				

### Node=Data Partition Variable Summary

Role	Level	Frequency Count	Name
TARGET	BINARY	1	enum_dsaf
REJECTED	BINARY	11	preac_blkcoordfig preac_colblkst preac_gqareacode preac_gqcontact preac_loczipclass preac_mailroesc preac_mailstblidnd preac_mailstnmsrc preac_mailzipclass preac_tabblkst preac_tabstate
REJECTED	INTERVAL	8	family_id preac_gqid preac_mspid preac_oid preac_oidcb preac_oidb preac_spid preac_survmfid
REJECTED	NOMINAL	76	EFFECT_PHONE_DT HH_ARR_DT HOME_SALE_DT PHONE_ACQUIS_DT PHONE_PUB_DT RESIDENCY_NEW RESIDENCY_OLD a_ref_dt max_find_income max_home_age max_residence_length min_find_income min_home_age ...
REJECTED	UNARY	30	A_Drop_Ind Addr_typ0 Addr_typ3 CSA_HOME_EQU_EST_ROUTE_TYP_G avg_surname_cnt preac_acsaf preac_altoordtyp preac_ctufvaf preac_ctufvuv preac_colbksufx1 preac_colbksufx2 preac_colims ...
INPUT	BINARY	68	A_Bus_File_Hit A_Deceased_Flg A_Fulfill_Flg A_Nocall_Flg A_Nomail_Flg A_Spouse_Ind A_Workathome_Flg Addr_typ2 Addr_typ4 Addr_typ5 CHILD_INDICATOR DEL_UNIT_SIZE DEMO_LV_IND DM_HIGH_TECH_IND ...
INPUT	INTERVAL	29	A_Married_CD_S A_age_max A_age_min A_sex_F A_sex_M Addr_typ1 CEN2000_GEO_MATCH_LV0 LAT_LONG_MTCN_LV0 MAFID USPS_DEL_SERV_C USPS_DEL_SERV_N avg_Occupancy_ct avg_adult_cnt avg_child_cnt ... count_N
INPUT	NOMINAL	66	A_Married_CD_M A_Married_CD_U CEN2000_GEO_MATCH_LV2 CEN2000_GEO_MATCH_LV4 CEN2000_GEO_MATCH_LVX HH_SIZE_1_2 LAT_LONG_MTCN_LV2 LAT_LONG_MTCN_LV4 LAT_LONG_MTCN_LVX RECENCY_DT RECENCY_NEW RECENCY_OLD ...



## SAS Enterprise Miner Report

### Node=Decision Tree Summary

Node id = Tree  
 Node label = Decision Tree  
 Meta path = Ids => Part => Tree  
 Notes =

### Node=Decision Tree Properties

Property	Value	Default	Property	Value	Default
Component	DecisionTree		NSubtree	1	
AssessMeasure	PROFIT/LOSS		NodeRole	SEGMENT	
AssessPercentage	0.25		NodeSample	20000	
CV	N		NominalCriterion	PROBCHISQ	
CVNIter	10		Nrules	5	
CVRepeat	1		Nsurrs	0	
CVSeed	12345		NumInputs	1	
Criterion	DEFAULT		NumSingleImp	5	
Depth	Y		ObsImportance	N	
Dummy	N		OrdinalCriterion	ENTROPY	
Exhaustive	5000		Performance	DISK	
Freeze	N		Predict	Y	
Inputs	N		SigLevel	0.2	
IntervalCriterion	PROBF		Splitsize	.	
Kass	Y		Subtree	ASSESSMENT	
KassApply	BEFORE		TrainMode	BATCH	
LeafSize	5		UseDecision	N	
LeafId	Y		UseMultipleTarget	N	
Maxbranch	2		UsePriors	N	
Maxdepth	6		UseVarOnce	N	
MinCatSize	5		VarSelection	Y	
MissingValue	USEINSEARCH				

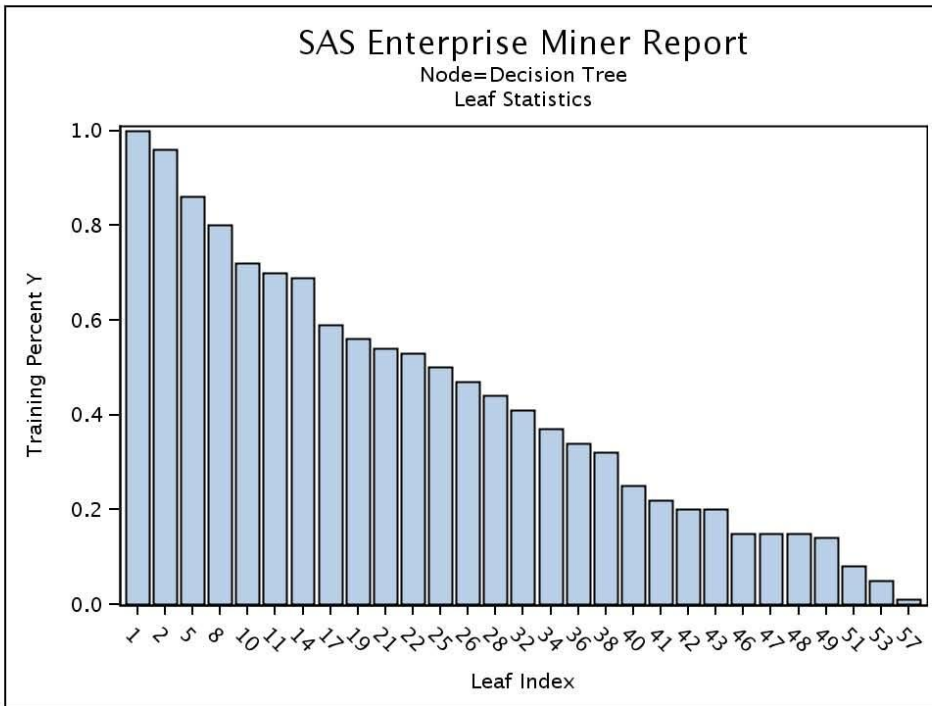
### Node=Decision Tree Variable Summary

Role	Level	Frequency Count	Name
TARGET	BINARY	1	enum_dsaf
INPUT	BINARY	68	A_Bus_File_Hit A_Deceased_Flg A_Fulfill_Flg A_Nocall_Flg A_Nomail_Flg A_Spouse_Ind A_Workathome_Flg Addr_typ2 Addr_typ4 Addr_typ5 CHILD_INDICATOR DEL_UNIT_SIZE DEMO_LV_IND DM_HIGH_TECH_IND ...
INPUT	INTERVAL	29	A_Married_CD_S A_age_max A_age_min A_sex_F A_sex_M Addr_typ1 CEN2000_GEO_MATCH_LV0 LAT_LONG_MTCH_LV0 MAFID USPS_DEL_SERV_C USPS_DEL_SERV_N avg_Occupancy_ct avg_adult_cnt avg_child_cnt ... count_N
INPUT	NOMINAL	66	A_Married_CD_M A_Married_CD_U CEN2000_GEO_MATCH_LV2 CEN2000_GEO_MATCH_LV4 CEN2000_GEO_MATCH_LVX HH_SIZE_1_2 LAT_LONG_MTCH_LV2 LAT_LONG_MTCH_LV4 LAT_LONG_MTCH_LVX RECENCY_DT RECENCY_NEW RECENCY_OLD ...

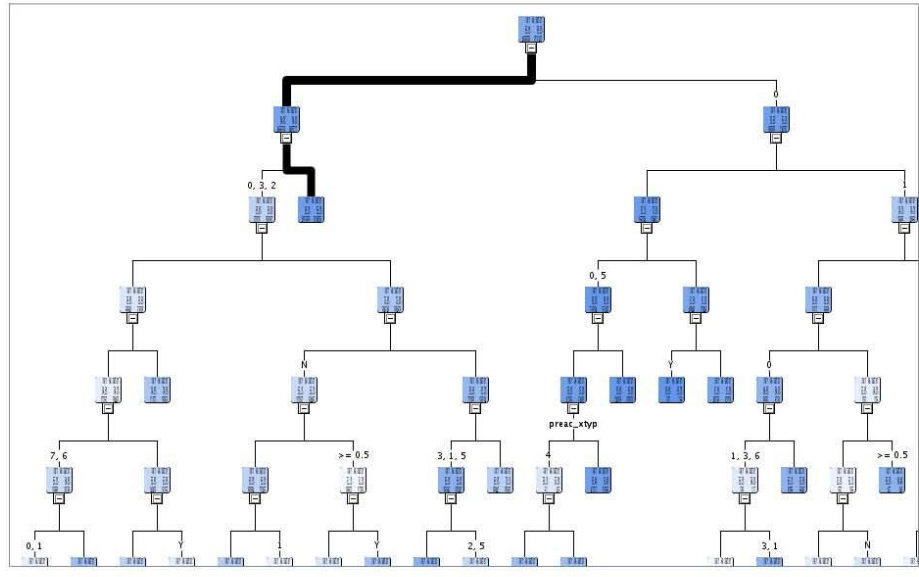
**Node=Decision Tree  
Model Fit Statistics**

Target=enum\_dsaf

Label of Statistic	Train	Validation	Test
Sum of Frequencies	49508583.00	37131437.00	37131439.00
Sum of Case Weights Times Freq	99017166.00	74262874.00	74262878.00
Misclassification Rate	0.05	0.05	0.05
Maximum Absolute Error	1.00	1.00	1.00
Sum of Squared Errors	4658828.11	3495854.85	3495213.99
Average Squared Error	0.05	0.05	0.05
Root Average Squared Error	0.22	0.22	0.22
Divisor for ASE	99017166.00	74262874.00	74262878.00
Total Degrees of Freedom	49508583.00	.	.

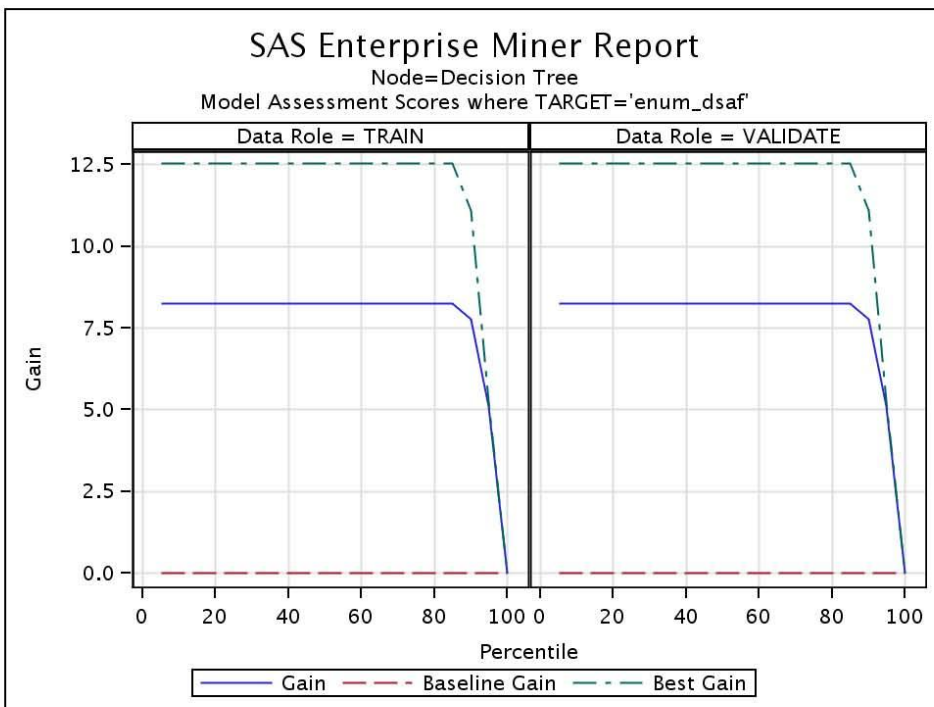
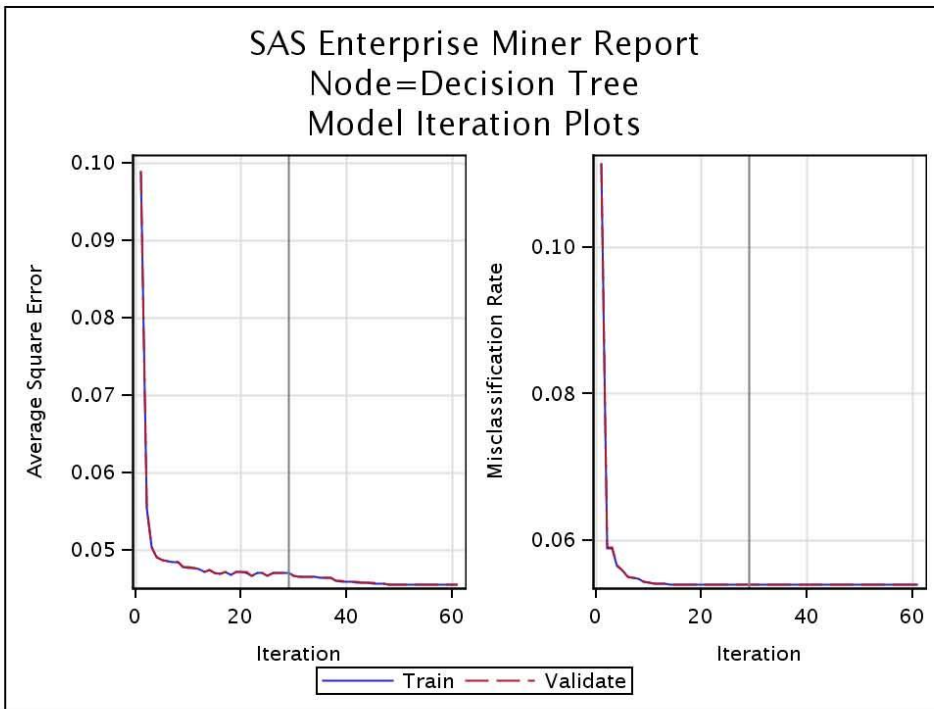


SAS Enterprise Miner Report  
Node=Decision Tree  
Tree Diagram



SAS Enterprise Miner Report  
Node=Decision Tree  
Icicle Plot

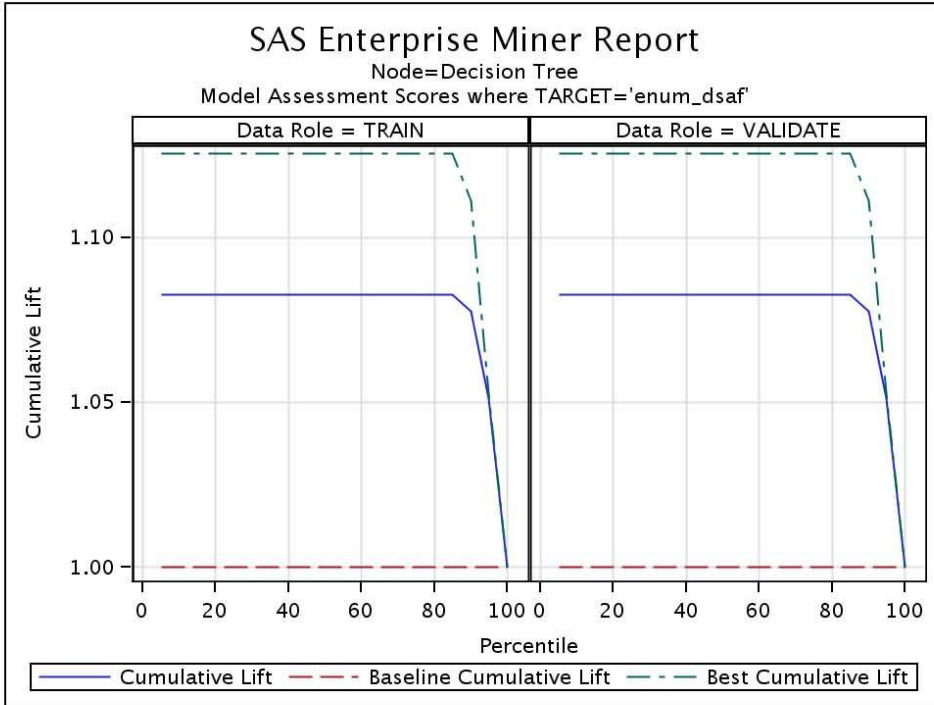




**SAS Enterprise Miner Report**

**Node=Decision Tree**

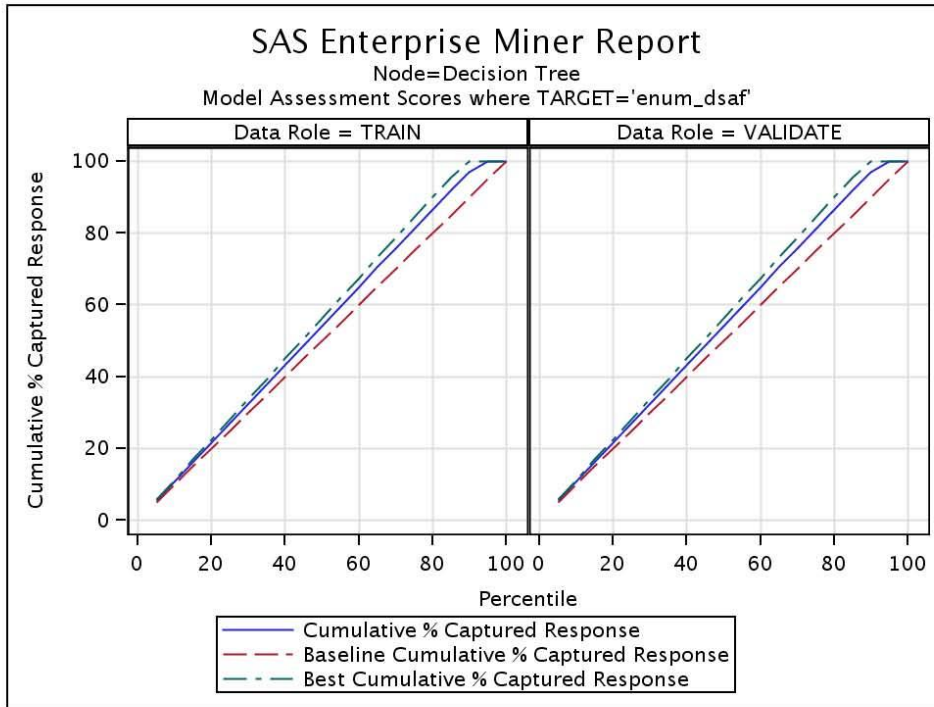
**Model Assessment Scores where TARGET='enum\_dsaf'**



**SAS Enterprise Miner Report**

**Node=Decision Tree**

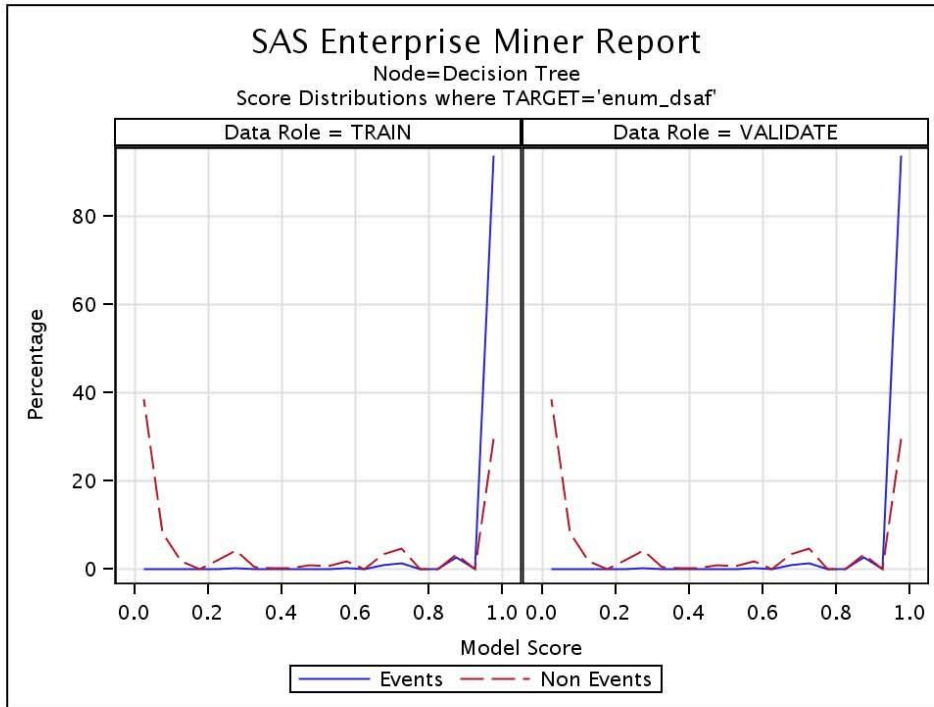
**Model Assessment Scores where TARGET='enum\_dsaf'**



**SAS Enterprise Miner Report**

**Node=Decision Tree**

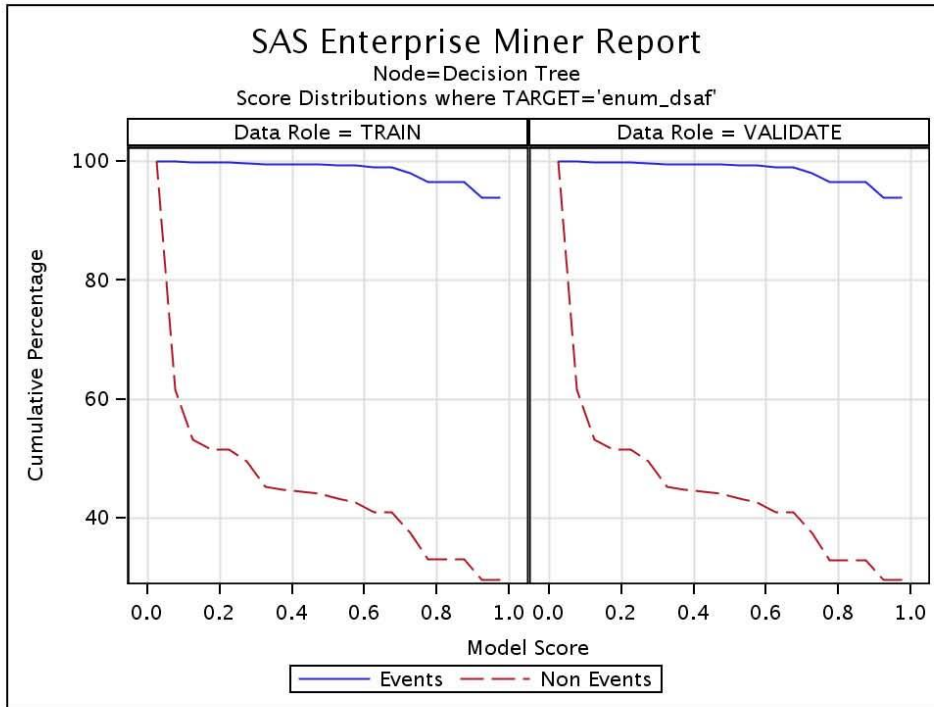
**Score Distributions where TARGET='enum\_dsaf'**



**SAS Enterprise Miner Report**

**Node=Decision Tree**

**Score Distributions where TARGET='enum\_dsaf'**



**Node=Decision Tree**  
**Score Distributions**

**Target Variable=enum\_dsaf Data Role=TRAIN**

Posterior Probability Range	Number of Events	Percentage of Events	Percentage of Nonevents	Cumulative Percentage of Events	Cumulative Percentage of Nonevents
0.95-1.00	41286142	93.8539	29.5718	93.854	29.572
0.90-0.95	0	0.0000	0.0000	93.854	29.572
0.85-0.90	1190041	2.7053	3.3877	96.559	32.959
0.80-0.85	0	0.0000	0.0000	96.559	32.959
0.75-0.80	311	0.0007	0.0014	96.560	32.961
0.70-0.75	637598	1.4494	4.6037	98.009	37.565
0.65-0.70	418298	0.9509	3.3211	98.960	40.886
0.60-0.65	0	0.0000	0.0000	98.960	40.886
0.55-0.60	134988	0.3069	1.7266	99.267	42.612
0.50-0.55	42190	0.0959	0.6537	99.363	43.266
0.45-0.50	43094	0.0980	0.8810	99.461	44.147
0.40-0.45	12601	0.0286	0.2926	99.490	44.440
0.35-0.40	10332	0.0235	0.3131	99.513	44.753
0.30-0.35	14356	0.0326	0.5521	99.546	45.305
0.25-0.30	78120	0.1776	4.2247	99.723	49.529
0.20-0.25	29337	0.0667	1.9819	99.790	51.511



**Target Variable=enum\_dsaf Data Role=TRAIN**

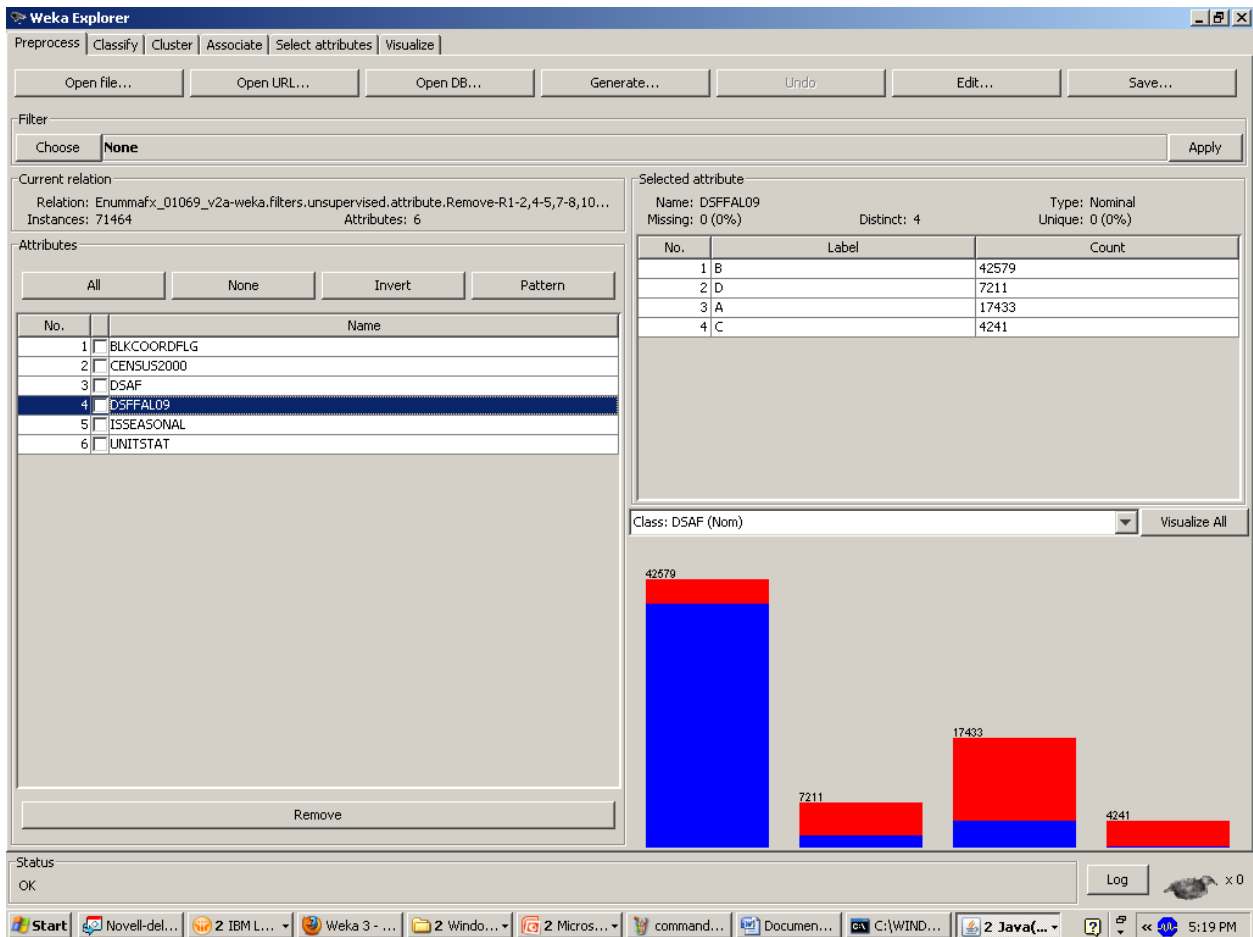
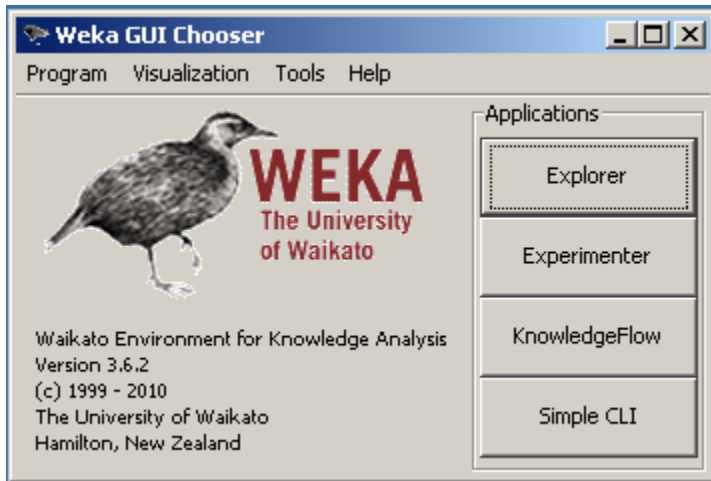
Posterior Probability Range	Number of Events	Percentage of Events	Percentage of Nonevents	Cumulative Percentage of Events	Cumulative Percentage of Nonevents
0.15-0.20	87	0.0002	0.0076	99.790	51.519
0.10-0.15	16169	0.0368	1.7092	99.827	53.228
0.05-0.10	40253	0.0915	8.2900	99.918	61.518
0.00-0.05	35888	0.0816	38.4818	100.000	100.000

**Target Variable=enum\_dsaf Data Role=VALIDATE**

Posterior Probability Range	Number of Events	Percentage of Events	Percentage of Nonevents	Cumulative Percentage of Events	Cumulative Percentage of Nonevents
0.95-1.00	30961878	93.8456	29.5735	93.846	29.574
0.90-0.95	0	0.0000	0.0000	93.846	29.574
0.85-0.90	892516	2.7052	3.3783	96.551	32.952
0.80-0.85	0	0.0000	0.0000	96.551	32.952
0.75-0.80	234	0.0007	0.0017	96.552	32.954
0.70-0.75	477581	1.4476	4.6125	97.999	37.566
0.65-0.70	315179	0.9553	3.3121	98.954	40.878
0.60-0.65	0	0.0000	0.0000	98.954	40.878
0.55-0.60	101630	0.3080	1.7280	99.262	42.606
0.50-0.55	31988	0.0970	0.6587	99.359	43.265
0.45-0.50	32476	0.0984	0.8813	99.458	44.146
0.40-0.45	9535	0.0289	0.3008	99.487	44.447
0.35-0.40	7713	0.0234	0.3071	99.510	44.754
0.30-0.35	10879	0.0330	0.5523	99.543	45.306
0.25-0.30	58733	0.1780	4.2309	99.721	49.537
0.20-0.25	22224	0.0674	1.9800	99.788	51.517
0.15-0.20	58	0.0002	0.0076	99.789	51.525
0.10-0.15	12213	0.0370	1.7100	99.826	53.235
0.05-0.10	30484	0.0924	8.2576	99.918	61.492
0.00-0.05	27033	0.0819	38.5077	100.000	100.000

End of Report

## Appendix B: WEKA 3.6.2 User Interface Screenshot



## Appendix C: January 2009 ACS Universe Specification

COLUMN DESCRIPTION	GEO NAME	ACS NAME
American Indian Tribal Subdivision (Census)	TRIBALSUBCE	AITSCCE
American Indian Tribal Subdivision (ANSI)	TRIBALSUBNS	AITSNS
Urban/Rural	UR	UR
Elementary School District	SDELMLEA	SDELEM
Secondary School District	SDSECLEA	SDSEC
Unified School District	SDUNILEA	SDUNI
Address Characteristic Type Code	ACT	ACT

**A-5. Determine which HUs are valid for ACS.** The Valid Unit Flag (VALDF) indicates whether or not a MAFID is eligible for interview by ACS. This section describes how to apply the “filter” to the **EDMAFX** in combination with other criteria to determine the final eligibility status of each MAFID for ACS.

1) **Set the ACS Universe Flag (ACSUNIV).**

For Puerto Rico counties (FIPST="72") set the value of ACSUNIV to the value stored in the ACSDEL variable.

For stateside counties (FIPST≠"72"), set ACSUNIV as follows (Note that the numbering scheme of the criteria (e.g. A-1, D-11, E-4, etc) refer to the ACS filter criteria in Attachment A.):

- a) Determine the blueline status for the county by looking up the fipst and fenty codes in the **BLUELINE** file and extracting the blstat variable. (Note that this only needs to be done once while processing a given county since all records in a file are in the same county.)
- b) Initialize ACSUNIV to "0".
- c) Determine if the record meets all of the general criteria for a good record. The general criteria are satisfied if all five of the following are true:
  - A-1 (UNTST is not "07" or "29") AND (SMAF is blank) AND (INVALIDSF is not "Y")
  - A-2 RESF is "1"
  - A-3 (ACACSTOI is not "D" OR UNTST is not "04") AND (AC04ACT is not "D") AND (AC06ACT is not "D") AND (GQV06ACT is not "D") AND (ACDRACT is not "D") AND (GQVDRACT is not "D")
  - A-4 At least one of the following three criteria are met:
    - A-4a. At least one of the following fields is not blank: HN1, MHN1, RRDES, RRID, BOXID, LOCDDES, GQNAME, MAPSP or
    - A-4b STRNM is neither blank nor "SAME" or
    - A-4c MSTRNM is neither blank nor "SAME"
  - A-5 OLQFLAG is not "3"

If a record does not meet A-1, A-2, A-3, A-4, and A-5 then set the ACSUNIV to "0" and go to step j. Otherwise continue with step d.

- d) Determine if the record is a Census 2000 record. A Census 2000 record meets either of the following two criteria:

B-1 CENSUS="Y" and CQRUSTAT="1"  
B-2 CENSURV="1"

If either B-1 or B-2 is true then set the ACSUNIV to "1" and go to step j. Otherwise, continue with step e.

- e) Determine if the record is an add or reinstatement from the 2000 Census Count Question Resolution (CQR) operation. A CQR add or reinstatement must meet at least one of the following two criteria:

C-1 ACCQR is "A" or "R"  
C-2 CQRUSTAT="1" and ACCQR is "G" or "H"

If either C-1 or C-2 is true then set the ACSUNIV to "2" and go to step j. Otherwise, continue with step f.

- f) Determine if the record is a post-Census add from the United States Postal Service (USPS) Delivery Sequence File (DSF). A post-Census DSF add meets ALL of the following conditions:

D-1 Neither DSF1 or DSF1A or DSF2 are ("1" or "3")  
D-2 PREVDEL="N"  
D-3 The record meets at least one of the following 5 criteria:  
D-3a BTEA is blank and ACT is blank and blueline status (from step a above) is not "OUT".  
D-3b ACT="MD" and blueline status (from step a above) is not "OUT".  
D-3c BTEA is "1", "6", "7", or "8"  
D-3d NEWLACS is "1", "2", or "3"  
D-3e ACT is "C1", "C2", "C3", "Z0", "B1", "B2", "B3", "MA", "MB", "MC", or "M3"

If one or more of the above 5 criteria are true then condition D-3 is satisfied

D-4 None of the following action codes are set to "A": ACDAAL, ACDAALGQ, ALMIAC, ACSC, AC04ACT, UL04ACT, AC06ACT, GQV06ACT, UE06ACT, GQE06ACT, NRFU06ACT, ACACSTOI, ACFACHS, ACDRACT, GQVDRACT  
D-5 DSF21 is either "1" or "3"  
D-6 DELPTTYPE is not "1", "2", "3", "4", "5", "6", "7", "8", "I", "J", "K", "L", "M", "N", "O", "P", or "X".

- D-7 ALTBASE is not "A"
- D-8 XTYPE is "0" or "9"
- D-9 DSFRT is not "R"
- D-10 NEWLACS is not "4" or "5"
- D-11 MZIPCLASS is not "U" and LZIPCLASS is not "U"

If all 11 conditions are satisfied then set the ACSUNIV to "3" and go to step j. Otherwise, continue with step g.

- g) Determine if the record is a Census delete that persists on the DSF. These records must meet ALL of the following conditions:

- E-1 PREVDEL="Y"
- E-2 CQRUSTAT="0"
- E-3 DELPTTYPE is either "A", "B", "C", "D", "E", "F", "G", or "H"
- E-4 DSF21 ="1"
- E-5 ALTBASE is not "A"
- E-6 DSFRT is not "R"
- E-7 NEWLACS is not "4" or "5"
- E-8 MZIPCLASS is not "U" and LZIPCLASS is not "U"

If all 8 conditions are true then set the ACSUNIV to "4" and go to step j. Otherwise, continue with step h.

- h) Determine if the record is an add from the Demographic Area Address Listing (DAAL) operation. These records must meet BOTH of the following criteria:

- F-1 ACLASTDAAL must be "A", "C", "M", "V", "G", or "H"
- F-2 NEWLACS is not "4" or "5"

If a record meets both F-1 and F-2 then set the ACSUNIV to "5" and go to step j. Otherwise continue with step i.

- i) Determine if the record is a Special Census or Census Test add. These records must meet all three of the following criteria:
  - G-1 One of the action codes (ACSC, AC04ACT, UL04ACT, AC06ACT, GQV06ACT, UE06ACT, GQE06ACT, NRFU06ACT, ACDRACT, GQVDRACT) must be "A", "C", "M", "V", "G", "H", or "K"
  - G-2 NEWLACS is not "4" or "5"

If a record meets G-1 and G-2 then set the ACSUNIV to "6".

- j) At this point the ACSUNIV flag is set to a value between "0" and "6". It should be "0" for records that either failed the general criteria (step c) or did not meet any of the criteria for inclusion in the ACS. It should be a value between "1" and "6" for records that passed step c and also passed either step d, e, f, g, h, or i.

## Appendix D: January 2010 ACS Universe Specification

12

- 7) **Puerto Rico only.** Reformat the KMHM field provided by GEO. This field should be stored in the files provided by GEO in x.y format. However, there are some cases where the data contains extra data, probably because of how it was collected in the field. We need to reformat these to look like "KM x HM y" in the edited MAFs.
- a) Remove all instances of 'K', 'M', and/or blanks from the KMHM field provided by GEO. Example, a KMHM value of "K28.0" should look like "28.0" and a KMHM value of "KM 1.8" should look like "1.8" after step 7a..

Then do either b, c, or d below based on the format of the remaining string.

- b) If the remaining string from step 7a is in "x.y" format then format KMHM to "KM x HM y". Example: A KMHM of "123.7" from GEO should look like "KM 123 HM 7" in the edited MAF.
- c) If the remaining string from 7a is formatted as "x" without the ".y" part then format KMHM as "KM x". Example: KMHM of "4" from GEO should look like "KM 4" in the edited MAF.
- d) If the first character of the remaining string from 7a is a decimal point (i.e, the string is formatted as ".y" without the x part) then format KMHM as "KM 0 HM y". Example: KMHM of ".6" from GEO should look like "KM 0 HM 6" in the edited MAF.

**A-9. Determine which HUs are valid for ACS.** The Valid Unit Flag (VALDF) indicates whether or not a MAFID is eligible for interview by ACS. This section describes how to apply the "filter" to the **EDMAFX** in combination with other criteria to determine the final eligibility status of each MAFID for ACS.

1) **Set the ACS Universe Flag (ACSUNIV).**

For Puerto Rico counties (FIPST="72") set the value of ACSUNIV to "0" if ENUMAF is "0". Otherwise set ACSUNIV to "1".

For stateside counties (FIPST≠"72"), set ACSUNIV as follows (Note that the numbering scheme of the criteria (e.g. A-2, B-1, C-9, etc) refer to the ACS filter criteria in Attachment A.):

- a) Determine the blueline status for the county by looking up the FIPST and FCNTY codes in the **BLUELINE** file and extracting the BLSTAT variable. (Note that this only needs to be done once while processing a given county since all records in a file are in the same county.)
- b) Initialize ACSUNIV to "0".

- c) Determine if the record meets all of the general criteria for a good record. The general criteria are satisfied if all five of the following are true:

- A-1 (UNTST is not "07" or "29") AND (SMAF is blank)
- A-2 RESF is "1"
- A-3 At least one of the following fields is not blank: STRNM, MSTRNM, MAPSP, or LOCDES
- A-4 At least one of the following conditions is true:
  - A-4a ACLUCA10 is not 'A'
  - A-4b UNTST is not "05"
  - A-4c Either ACADCAN10, ACLGBLK10, or ACGQV10 is not blank

If a record does not meet A-1, A-2, A-3, and A-4 then set the ACSUNIV to "0" and go to step f. Otherwise continue with step d.

- d) Determine if the record is in the 2010 Census enumeration universe.

- B-1 ENUMAF is not "0"

If B-1 is true then set the ACSUNIV to "1" and go to step f. Otherwise, continue with step e.

- e) Determine if the record is a post-Census DSF add. A post-Census DSF add meets ALL of the following conditions:

- C-1 Neither DSF1 or DSF1A or DSF2 are ("1" or "3")
- C-2 ADCANAF="0"
- C-3 The record meets at least one of the following 5 criteria:
  - C-3a ACT is blank and blueline status (from step a above) is not "OUT".
  - C-3b1 ACT is "C1", "C2", "C3", "Z0", "B1", "B2", "B3", "MA", "MB", "MC", or "M3"
  - C-3b2 ACT="MD" and blueline status (from step a above) is not "OUT".
  - C-3d NEWLACS is "1", "2", or "3"

If one or more of the above 4 criteria are true then condition C-3 is satisfied

- C-4 DSF24 is either "1" or "3"
- C-5 DELPTTYPE is not "1", "2", "3", "4", "5", "6", "7", "8", "I", "J", "K", "L", "M", "N", "O", "P", or "X".
- C-6 ALTBASE is not "A"

C-7 The record meets either of the following two criteria:

C-7a XTYPE is "0" or "9"

C-7b XTYPE="3" and NEWLACS is "1","2", or "3"

If either C-7a or C-7b are true then condition C-7 is satisfied.

C-8 DSFRT is not "R"

C-9 NEWLACS is not "4" or "5"

C-10 MZIPCLASS is not "U" and LZIPCLASS is not "U"

If all 11 conditions are satisfied then set the ACSUNIV to "3".

f) At this point the ACSUNIV flag is set to "0", "1", or "3". It should be "0" for records that either failed the general criteria (step c) or did not meet any of the criteria for inclusion in the ACS. It should be a value of "1" and "3" for records that passed step c and also passed either step D. or E.

2) **Set the initial value of the Valid Unit Flag (VALDF).** Set VALDF to "1" (valid) if the ACS Universe Flag (ACSUNIV) is not equal to "0" (zero). Otherwise, set VALDF to "0" (invalid).

3) **Invalidate special place (SP) records, group quarter (GQ) records, embedded housing unit records, and other "bad" addresses and update counts that will be kept in the Quality Assurance Statistics and References (QUASAR) dataset.** For each record on the MAF extract, determine if it falls into one or more of these categories:

TABLE OF MAF HU INVALIDATION TYPES	
<b>(1) Group Quarters (GQs), Special Places (SPs), and Transitory Locations</b>	
<b>Description:</b>	SPs, GQs, and transitory locations must be invalidated on the edited MAFs before the unit frame universe is created from them.
<b>How to Identify:</b>	ACSUNIV ≠ "0" <u>and</u> Group Quarters/HU Flag (GQHUF) ≠ "0"



## Appendix E: January 2011 ACS Universe Specification

look like "KM 0 HM 6" in the edited MAF.

**A-7. Determine which HUs are valid for ACS.** The Valid Unit Flag (VALDF) indicates whether or not a MAFID is eligible for interview by ACS. This section describes how to apply the "filter" to the **EDMAFX** in combination with other criteria to determine the final eligibility status of each MAFID for ACS.

- 1) **Set the ACS Universe Flag (ACSUNIV).** This flag is set differently for Puerto Rico records than it is for stateside records. Steps a through d should be used to set ACSUNIV for Records in Puerto while steps e through j should be used for stateside records.

For Puerto Rico counties (FIPST="72"), set ACSUNIV as follows (Note that the numbering scheme of the criteria (e.g. A-2, B-1, C-9, etc) refer to the Puerto Rico section of the ACS filter criteria in Attachment A.):

- a) Initialize ACSUNIV to "0".
- b) Determine if the record meets all of the general criteria for a good record. The general criteria are satisfied if all five of the following are true:

- A-1 (UNTST is not "07" or "29") AND (SMAF is blank)
- A-2 RESF is "1"
- A-3 At least one of the following conditions is true:

- A-3a ACLUCA10 is not 'A'
- A-3b UNTST is not "05"
- A-3c Either ACADCAN10, ACLGBLK10, ACGQV10, ACULUUE10, ACGQE10, ACETL10, ACINFOCOMM, ACNRFU10, ACFV10, or ACVDC10 is not blank

If a record does not meet A-1, A-2, and A-3, then set the ACSUNIV to "0" and go to step d. Otherwise continue with step c.

- c) Determine if the record is in the 2010 Census enumeration universe. A record is in the 2010 Census enumeration if it meets either of the following two criteria:

- B-1 ENUMAF is not "0" and none of the following action code variables are 'D': ACULUUE10, ACGQE10, ACETL10, ACNRFU10, ACFV10, and ACVDC10.
- B-2 Any of the following action code variables are set to 'A', 'C', 'M', 'V', or 'K': ACULUUE10, ACGQE10, ACETL10, ACINFOCOMM, ACNRFU10, ACFV10, or ACVDC10.

If either B-1 or B-2 is true then set the ACSUNIV to "1".

- d) At this point the ACSUNIV flag for the current Puerto Rico record should

be set to "0" or "1". You are done applying the filter to this Puerto Rico record so skip to step 2 (setting VALDF) below.

For stateside counties (FIPST ≠ "72"), set ACSUNIV as follows (Note that the numbering scheme of the criteria (e.g. A-2, B-1, C-9, etc) refer to the ACS filter criteria in Attachment A.):

- e) Determine the blueline status for the county by looking up the FIPST and FCNTY codes in the **BLUELINE** file and extracting the BLSTAT variable. (Note that this only needs to be done once while processing a given county since all records in a file are in the same county.)
- f) Initialize ACSUNIV to "0".
- g) Determine if the record meets all of the general criteria for a good record. The general criteria are satisfied if all five of the following are true:
  - A-1 (UNTST is not "07" or "29") AND (SMAF is blank)
  - A-2 RESF is "1"
  - A-3 At least one of the following fields is not blank: STRNM, MSTRNM, MAPSP, or LOCDES
  - A-4 At least one of the following conditions is true:
    - A-4a ACLUCA10 is not 'A'
    - A-4b UNTST is not "05"
    - A-4c Either ACADCAN10, ACLGBLK10, ACGQV10, ACULUUE10, ACRAUERUE10, ACGQE10, ACETL10, ACINFOCOMM, ACNRFU10, ACFV10, or ACVDC10 is not blank
  - A-5 Either ACRAUERUE10 is not "A" or DTRAUERUE10 is not 01AUG2010:00:00:00

If a record does not meet A-1, A-2, A-3, A-4, and A-5 then set the ACSUNIV to "0" and go to step k. Otherwise continue with step h.

- h) Determine if the record is in the 2010 Census enumeration universe. A record is in the 2010 Census enumeration if it meets one or both of the following two criteria:
  - B-1 ENUMAF is not "0" and none of the following action code variables are 'D': ACULUUE10, ACRAUERUE10, ACGQE10, ACETL10, ACNRFU10, ACFV10, and ACVDC10.
  - B-2 Any of the following action code variables are set to 'A', 'C', 'M', 'V', or 'K': ACULUUE10, ACRAUERUE10, ACGQE10, ACETL10, ACINFOCOMM, ACNRFU10, ACFV10, or ACVDC10.

If either B-1 or B-2 is true then set the ACSUNIV to "1" and go to step k.

Otherwise, continue with step i.

- i) Determine if the record is a post-Census DSF add. A post-Census DSF add meets ALL of the following conditions:

C-1 Neither DSF1 or DSF1A or DSF2 are ("1" or "3")

C-2 ADCANAF="0"

C-3 The record meets at least one of the following 5 criteria:

C-3a ACT is blank and blue line status (from step a above) is not "OUT".

C-3b1 ACT is "C1", "C2", "C3", "Z0", "B1", "B2", "B3", "MA", "MB", "MC", or "M3"

C-3b2 ACT="MD" and blue line status (from step a above) is not "OUT".

C-3d NEWLACS is "1", "2", or "3"

If one or more of the above 4 criteria are true then condition C-3 is satisfied

C-4 DSF26 is either "1" or "3"

C-5 DELPTYPE is not "1", "2", "3", "4", "5", "6", "7", "8", "I", "J", "K", "L", "M", "N", "O", "P", or "X".

C-6 ALTBASE is not "A"

C-7 The record meets either of the following two criteria:

C-7a XTYPE is "0" or "9"

C-7b XTYPE="3" and NEWLACS is "1", "2", or "3"

If either C-7a or C-7b are true then condition C-7 is satisfied.

C-8 DSFRT is not "R"

C-9 NEWLACS is not "4" or "5"

C-10 MZIPCLASS is not "U" and LZIPCLASS is not "U"

If all 10 conditions are satisfied then set the ACSUNIV to "3" and go to step k. Otherwise, continue with step j.

- j) Determine if the record is a post-2010 Census DAAL add. A record is a post-Census DAAL add if D-1 below is true.

D-1 The record meets at least one of the following 5 criteria:

D-1a ACDAAL is "A", "C", "G", "H", "M", "V" and DTDAAL is > 01AUG2010:00:00:00

D-1b ACDAALGQ is "A", "C", "G", "H", "M", "V" and DTDAALGQ is > 01AUG2010:00:00:00

- D-1c ALMIAC is "A", "C", "G", "H", "M", "V" and DTALMI is > 01AUG2010:00:00:00
- D-1d ACACSTOI is "A", "C", "G", "H", "M", "V" and DTACSTOI is > 01AUG2010:00:00:00
- D-1e ACFACHS is "A", "C", "G", "H", "M", "V" and DTFACHS is > 01AUG2010:00:00:00

If any one of D-1a through D-1e are satisfied then set the ACSUNIV to "5".

- k) At this point the ACSUNIV flag for the current stateside record should be set to "0", "1", "3", or "5". It should be "0" for records that either failed the general criteria (step g) or did not meet any of the criteria for inclusion in the ACS. It should be a value of "1", "3", or "5" for records that passed step g and also passed either step h, i, or j.
- 2) **Set the initial value of the Valid Unit Flag (VALDF).** Set VALDF to "1" (valid) if the ACS Universe Flag (ACSUNIV) is not equal to "0" (zero). Otherwise, set VALDF to "0" (invalid).
  - 3) **Invalidate special place (SP) records, group quarter (GQ) records, embedded housing unit records, and other "bad" addresses and update counts that will be kept in the Quality Assurance Statistics and References (QUASAR) dataset.** For each record on the MAF extract, determine if it falls into one or more of these categories:

TABLE OF MAF HU INVALIDATION TYPES	
<b>(1) Group Quarters (GQs), Special Places (SPs), and Transitory Locations</b>	
<b>Description:</b>	SPs, GQs, and transitory locations must be invalidated on the edited MAFs before the unit frame universe is created from them.
<b>How to Identify:</b>	ACSUNIV ≠ "0" and Group Quarters/HU Flag (GQHUF) ≠ "0"
<b>(2) MAF Addresses Known to Be "Bad"</b>	
<b>Description:</b>	Based on past processing and research, we are aware of specific MAFIDs that should not be considered valid HUs, though they appear as such on the MAF extracts. These are most often records flagged as HUs that are truly GQs.