A Comparison of Methodologies for Classification of Administrative Records Quality for Census Enumeration

Darcy Steeg Morris*

Abstract

The use of administrative records - data collected by governmental or non-governmental agencies in the course of administering a program or service - for household enumeration may be one way to significantly reduce Census costs, particularly in nonresponse follow-up (NRFU). Administrative records suffer the complications of big data in that they are collected for purposes not related to Census enumeration; yet they contain a wealth of information relevant to Census enumeration. This work investigates different classification techniques for determining which administrative records are sufficiently reliable to use to achieve a Census enumeration that maintains data quality but reduces costs. In addition to the cost/quality tradeoff associated with using administrative records, we seek a methodology for using administrative records that strikes a balance between predictive power and model complexity. In this paper, we compare the use of logistic regression and machine learning techniques for extracting and synthesizing the most important enumeration information from a set of governmental and non-governmental data sources.

Key Words: Administrative Records, Census Enumeration, Classification, Supervised Learning

1. Introduction

A primary cost driver of the Decennial Census is the collection of data from households for which a self-response is not obtained. In the 2010 Census, the nonresponse follow-up (NRFU) operation included about fifty million addresses requiring up to six personal visits each, totaling over $2 billion. For purposes of planning the 2020 Decennial Census, the Census Bureau is researching ways to reduce costs of NRFU operations while maintaining data quality. One solution may be to use administrative records (AR) in lieu of personal visits. Government and commercial administrative records include sources from agencies and companies such as the Internal Revenue Service (IRS), Center for Medicare and Medicaid Services (CMS), U.S. Postal Service (USPS), and Targus. The classification problem presented in this paper aims to determine which housing units to enumerate using such data sources (and thus remove from NRFU fieldwork) and which to continue the usual NRFU operations. While administrative records can also be used to identify addresses that can be removed from the NRFU workload via a housing unit status determination (i.e. vacant or delete), this research focuses on using administrative records to determine the person count in occupied housing units.

By some definitions, the use of administrative records for Census enumeration can be thought of as a big data problem. Capps and Wright (2013) present two informal characteristics of big data: that they "come as byproducts of other primary activities without asking explicitly" and "come with unknowns (e.g. uses are less clear, data are less understood, data are of unknown quality, and representativeness is largely unknown)." Whether administrative records are collected as part of administering a government program (e.g.

*U.S. Census Bureau, Washington, DC 20233. This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

Medicaid) or part of a marketing strategy (e.g. Targus), the files used in this research have at least one thing in common: each have a data point associating a person with an address at some point in time. We want to pool this person-address information across sources and use statistical techniques to extract the best and most relevant person-address pairs for Census enumeration. However, the data suffers from potential bias: address information from different administrative sources need not coincide with the address desired to enumerate Census day population. Each source may have different definitions of address (or residence rules) and different collection dates. Nonetheless, there is valuable information to be attained from this wealth of data, and this paper investigates strategies to best extract the good address information with respect to Census enumeration.

Census Bureau researchers are studying a variety of methods for determining high-quality administrative records to use to curtail NRFU operations. Brown (2013) proposed creating and analyzing a composite dataset - a compilation of all person-place combinations found in any of the administrative record sources. This yields a universe of persons and addresses eligible for administrative record enumeration. In addition to the collection of person-place combinations, the data contains information as to the characteristics of those combinations, for example, which sources had those combinations and how many sources had those combinations. Under this setup, we are interested in using supervised learning classification techniques to make predictions regarding which person-place combinations to use for administrative records enumeration and which to ignore. Brown (2013) developed a logistic model that uses administrative records, housing unit, and geographic explanatory variables to estimate the probability that the composite data matched the person to their Census day address. Such an approach investigates the quality of the person-address combinations in the administrative records via a retrospective study of the 2010 Census data, where the 2010 Census outcome is treated as the "truth."

This paper extends the work of Brown by investigating alternative machine learning techniques for classification (classification trees and random forests) and also determining optimal binary predictions. This work investigates different classification techniques for determining which administrative records are sufficiently reliable to use to achieve a Census enumeration that maintains data quality but reduces costs. In addition to the cost/quality tradeoff associated with using administrative records, we seek a methodology for using administrative records that strikes a balance between predictive power and model complexity. In this paper, we compare the use of logistic regression and machine learning techniques for extracting and synthesizing the most important enumeration information from a set of administrative records.

2. Methodologies

2.1 Person-Place Models

The compilation of person-address pairs in 19 administrative record files from federal and commercial sources are matched to 2010 Census person-address pairs to define the dependent variable of interest:

$$y_{ih} = \begin{cases} 1 \text{ if person } i \text{ found in AR and 2010 Census at the same address } h \\ 0 \text{ otherwise} \end{cases}$$

We are interested in estimating the probability, $p_{ih} = P(y_{ih} = 1)$, that the 2010 Census and the administrative records composite data places the person at the same address. Broadly, the universe for the composite data is all unique combinations of persons and addresses found in the administrative record composite data for people who reside at a NRFU address;

however there are many details associated with building this dataset. For example, specific vintages of files are used (e.g. all filings for tax year 2009 are included), only records that can be person-validated to create a unique person identifier (PIK) are included, persons in the composite data are de-duplicated, and persons with data collected via a non-NRFU operation are removed from the composite data. Please see Brown (2013) for complete details. Note that a caveat of this approach is that data is only defined for those people and addresses present in the administrative record sources; thus any modeling is conditional on existence in the administrative records files.

In addition to housing unit and geography level predictors, the models include information about the source of the person-address combination. The primary administrative record explanatory variables are indicators for presence/absence of each particular source. For example, for each person-place pair, the variable "IRS 1040 Here" is equal to 1 if the IRS 1040 places that particular person at that particular address and 0 otherwise; while the "IRS 1040 Elsewhere" variable is equal to 1 if an IRS 1040 record places that particular person at a different address. These types of variables are created for all 19 sources. Please see the Appendix for a full list of the explanatory variables used in the logistic model, classification tree and random forest.

2.1.1 Logistic Regression

Logistic regression analysis of the composite dataset with the binary outcome and predictors described in the previous section has been researched and documented in Brown (2013). Standard logistic regression predictions $\hat{p}_{ih}$ are determined at the person-level, aggregated to the housing unit level, and used to classify the administrative records quality of a given housing unit. Brown (2013) incorporates two stages of modeling to obtain the person-level predictions. First, administrative record source-level models are fit to determine the predicted probability of an administrative records/Census address match for a given individual and a given source. Second, these predicted probabilities are used as independent variables in the person-place model. One goal of this research is to balance model complexity and predictive power, thus the added value of the first-stage predicted probabilities as independent variables in the second stage logistic regression was investigated and determined to not have significant impact. Results presented in this paper use a simpler version of the method used in Brown (2013); specifically, only the second stage logistic regression without controls for the predicted probabilities from source-level regressions.

2.1.2 Classification Trees and Random Forests

Logistic regression is just one of many methods for predicting binary outcomes in the context of a classification problem. In order to assess the impact of model assumptions associated with logistic regression on our predictions, the main results from the logistic regression are compared with those obtained from a classification tree (Breiman et. al., 1984) and a random forest (Breiman, 2001). The classification tree analysis is carried out using the *rpart* package in R and the random forest analysis is carried out using the *randomForest* package in R. The same explanatory variables are used in the classification tree and random forest as were used in the logistic regression. The classification tree offers a much simpler and intuitive summary of prediction that may provide additional value for Census production purposes and can be compared to the arguably more complex logistic regression and random forests in an accuracy/complexity tradeoff.

We are interested in the class probabilities obtained from the classification tree and random forest and, in particular, (1) how they compare to the predicted probabilities estimated from the logistic regression, (2) how they translate into an optimal cutpoint to determine

binary predictions, and (3) how the classification of housing units based on the optimal cutpoint differs from that derived from the logistic regression. Answers to these questions provide insight as to how sensitive the classification is to modeling assumptions. All methods are trained on a 1% sample of the 2010 Census NRFU housing unit universe and tested on a different 1% sample.

2.2 Housing Unit Level Predictions

The goal of this work is to determine the best classification of a housing unit's administrative records quality for enumeration based on individual level predicted probabilities of having an administrative record/Census address match. Because the classification techniques are carried out at the person-level, the person-level predicted probabilities $\hat{p}_{ih}$ need to be aggregated within a housing unit in order to make a decision about the quality of the administrative records for a given housing unit. Various aggregation metrics have been investigated, including:

1. the *minimum* of individual predicted probabilities, $f_1(\hat{p}_{ih}) = \min(\hat{p}_{1h}, \ldots, \hat{p}_{n_h h})$
2. the *mean* of individual predicted probabilities, $f_2(\hat{p}_{ih}) = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{p}_{ih}$

where n_h is the number of administrative records individuals in housing unit h. The minimum corresponds to the $\hat{p}_{ih}$ for the person in the housing unit for which we have the lowest confidence, while the mean corresponds to our average confidence in the set of individuals in the housing unit; thus using the minimum as a housing-unit aggregation metric corresponds to a more conservative approach[1]. The administrative records housing unit roster is defined as the aggregate of all individuals associated with a given administrative records address, and each address has an associated predicted probability of having an administrative records/Census address match.

2.3 Housing Unit Level Classification

In order to determine which housing units to enumerate using administrative records, we want to classify each housing unit as predicted to have good or bad quality administrative records. We can define a binary prediction based on some cutoff c, so that:

$$\hat{y}_h^c = \begin{cases} 1 \text{ if } f(\hat{p}_{ih}) \geq c \\ 0 \text{ otherwise} \end{cases}$$

where we use administrative record enumeration if $\hat{y}_h^c = 1$. That is, we define the binary prediction $\hat{y}_h^c$ based on the value of a function of $\hat{p}_{ih}$ (e.g. the minimum) and use administrative record enumeration if $f(\hat{p}_{ih})$ exceeds some cutoff, c. In this paper we define the predicted probability for a given housing unit as $\hat{p}_h = f_1(\hat{p}_{ih})$, so that the housing unit level predicted probability is the minimum person-level predicted probability for a given housing unit.

We have a choice about how to define whether we have observed good quality administrative records for a housing unit. This measure of observed success compared to predicted success will serve as the basis for classification. Some possibilities for classifying the housing unit as having good quality administrative records include: (1) the administrative record

[1] See Brown (2013) for results from the logistic regression analysis using both the minimum and the mean person-level predicted probabilities

population count and the 2010 Census count match, (2) all individuals in the housing unit have good quality observed administrative records, and (3) both (1) and (2) are true. We define all of the above only for the subset of addresses that were non-discrepant (e.g. not imputed in 2010, not a proxy observation in 2010, etc.). In this research we will define observed success as a population count match:

$$y_h = \begin{cases} 1 \text{ if } (\text{AR count for housing unit } h) = (\text{2010 Census count for housing unit } h) \\ 0 \text{ otherwise} \end{cases}$$

Since the classification depends crucially on the cutoff c, an objective function to optimize needs to be chosen to determine the optimal cutoff. Some examples of optimality criteria include:

1. *minimize* the *Euclidean distance* from $(0, 1)$ to ROC Curve (Metz, 1978):

$$\arg\min_c \sqrt{\left(\underbrace{1 - \frac{\sum_{h=1}^{n} I(y_h = 0 \;\&\; \hat{y}_h^c = 0)}{\sum_{h=1}^{n} I(y_h = 0)}}_{\text{Specificity}}\right)^2 + \left(\underbrace{1 - \frac{\sum_{h=1}^{n} I(y_h = 1 \;\&\; \hat{y}_h^c = 1)}{\sum_{h=1}^{n} I(y_h = 1)}}_{\text{Sensitivity}}\right)^2}$$

2. *minimize* the *Manhattan distance* from $(0, 1)$ to ROC Curve:

$$\arg\min_c \left[\left(\underbrace{1 - \frac{\sum_{h=1}^{n} I(y_h = 0 \;\&\; \hat{y}_h^c = 0)}{\sum_{h=1}^{n} I(y_h = 0)}}_{\text{Specificity}}\right) + \left(\underbrace{1 - \frac{\sum_{h=1}^{n} I(y_h = 1 \;\&\; \hat{y}_h^c = 1)}{\sum_{h=1}^{n} I(y_h = 1)}}_{\text{Sensitivity}}\right)\right]$$

where the ROC curve plots the sensitivity versus the false positive rate ($1-$specificity), n is the number of housing units and $I(.)$ is the indicator function. We focus attention on minimizing the Euclidean ROC distance since this corresponds to a joint decision to maximize the true positives and minimize the false positives.

3. Results

3.1 Housing Unit Level Classification

3.1.1 Distribution of Housing Unit-Level $\hat{p}_h$

Figure 1 presents the distributions of $\hat{p}_h$ separately for those housing units with a population count match ($y_h = 1$) and those without a population count match ($y_h = 0$). The densities are estimated via kernel density estimation for the logistic regression and random forest where $\hat{p}_h$ are continuous in nature, and via a histogram for the classification tree where the $\hat{p}_h$ are discrete. All three methods show a separation in the distribution of housing unit level predicted probabilities by housing unit population count match status. For example, in the random forest analysis, the distribution for those housing units which have an observed population count match between the administrative records and 2010 Census is skewed to the left (with the exception of some mass at very low $\hat{p}_h$); while the distribution for

those housing units which do not have an observed population count match between the administrative records and 2010 Census is skewed to the right. Given these distributions, a cutoff needs to be chosen to determine binary predictions; a point at which all housing units with $\hat{p}_h$ greater will be used for administrative records enumeration and all housing units with $\hat{p}_h$ less than will not. The primary interest of this work is to obtain good predictions of which housing units to enumerate using administrative records, thus the separation in the distributions at large values of $\hat{p}_h$ is particularly promising.

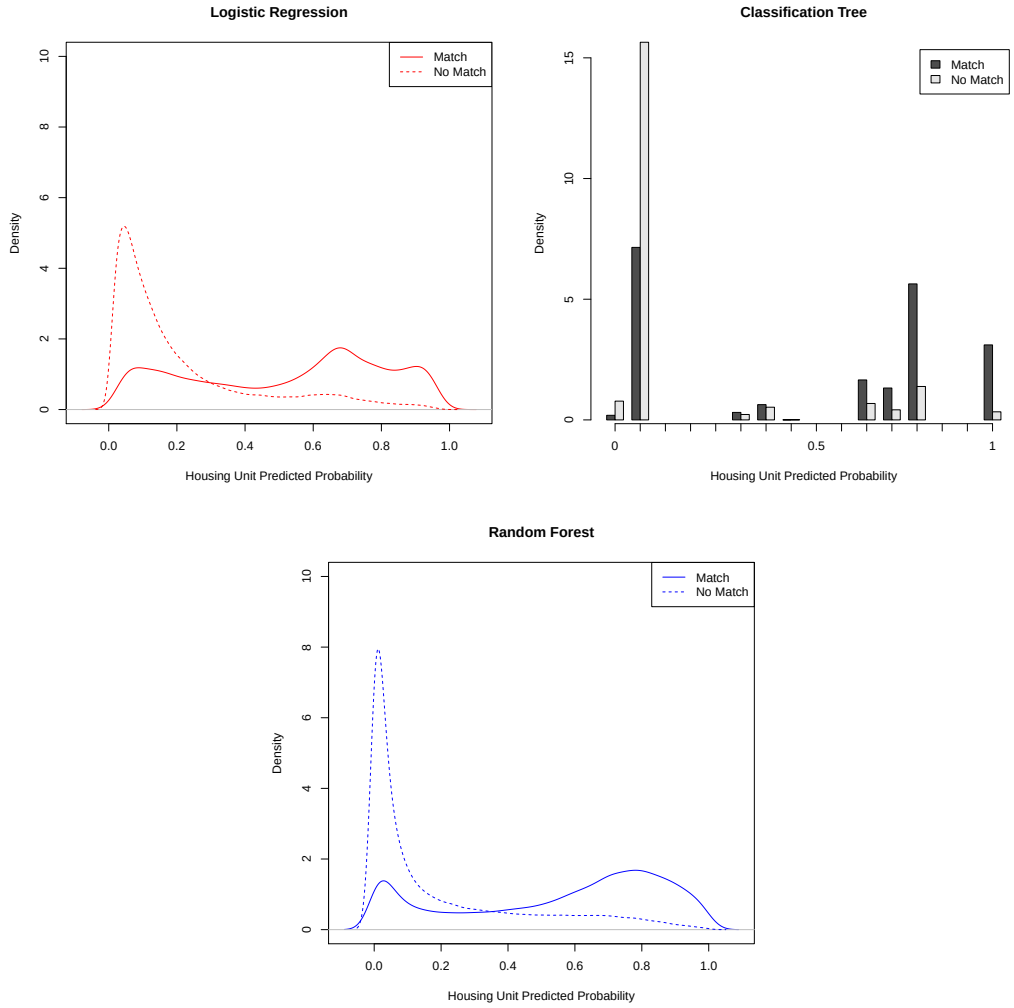

Figure 1: Distribution of Housing Unit Predicted Probabilities for Logit, Tree and Random Forest

3.1.2 Optimal Cutoff and ROC Curve

Figure 2 displays the ROC curve and the point corresponding to the optimal cutoff based on the minimizing ROC distance for each of the three methodologies (along with the 45^o line and the ideal ROC point of $(0, 1)$). The ROC curve for the logistic regression and random forest show very similar predictive power. While the discrete ROC function for the classification tree falls below both the random forest and the logistic regression ROC curve, at the optimal cutpoint the misclassification measures are somewhat comparable.

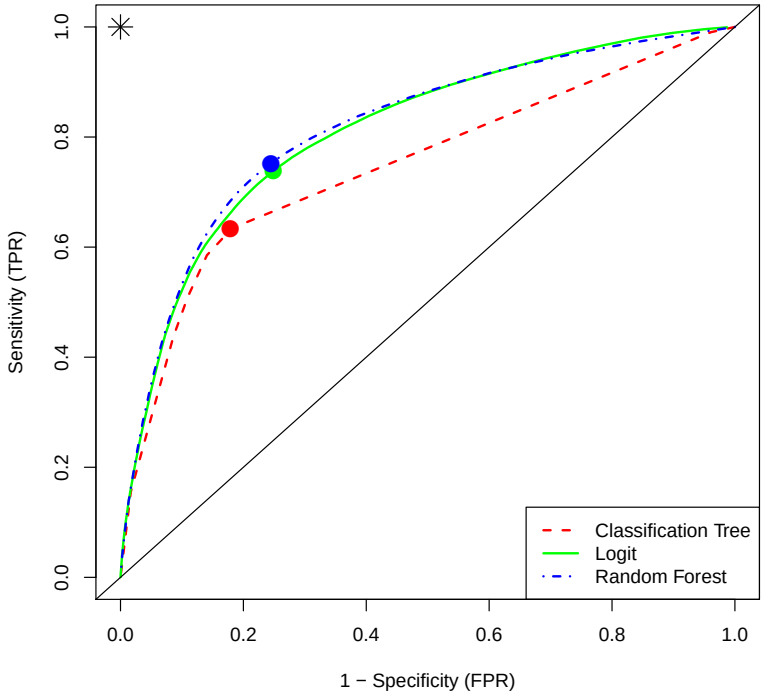

Figure 2: ROC Curve and Optimal Cutoff for Logit, Tree and Random Forest

Table 1 reports the corresponding false positive rate (FPR), false negative rate (FNR), optimal cutoff (c), proportion misclassifed, proportion of housing units in composite dataset selected for administrative record enumeration, and coverage ratio for housing units selected for administrative records enumeration (total population count from AR/total population count from 2010 Census) for the three methods. Figure 3 graphically presents the false positive and false negative rates through the distributions of $\hat{p}_h$.

Table 1: Results at Optimal Cutoff by Method

Method	FNR	FPR	c	Proportion Misclassified	Proportion in AR Enumeration*	Coverage Ratio**
Logistic Regression	.262	.249	.27	.25	.39	1.016
Classification Tree	.367	.179	.31	.23	.31	1.050
Random Forest	.258	.233	.32	.24	.38	1.000

* Of housing units with administrative records.

** For housing units selected for administrative records enumeration.

Overall, the random forest and logistic regression achieve similar false positive and false negative rates at the cutoff corresponding to minimizing the ROC distance. The classification tree performs better in terms of the false positive rate, but has a false negative rate larger than both the logistic regression and the random forest. The implication of this result depends on the penalty associated with false positives and false negatives, which will be discussed in Section 3.3. All three methods imply using administrative record enumeration

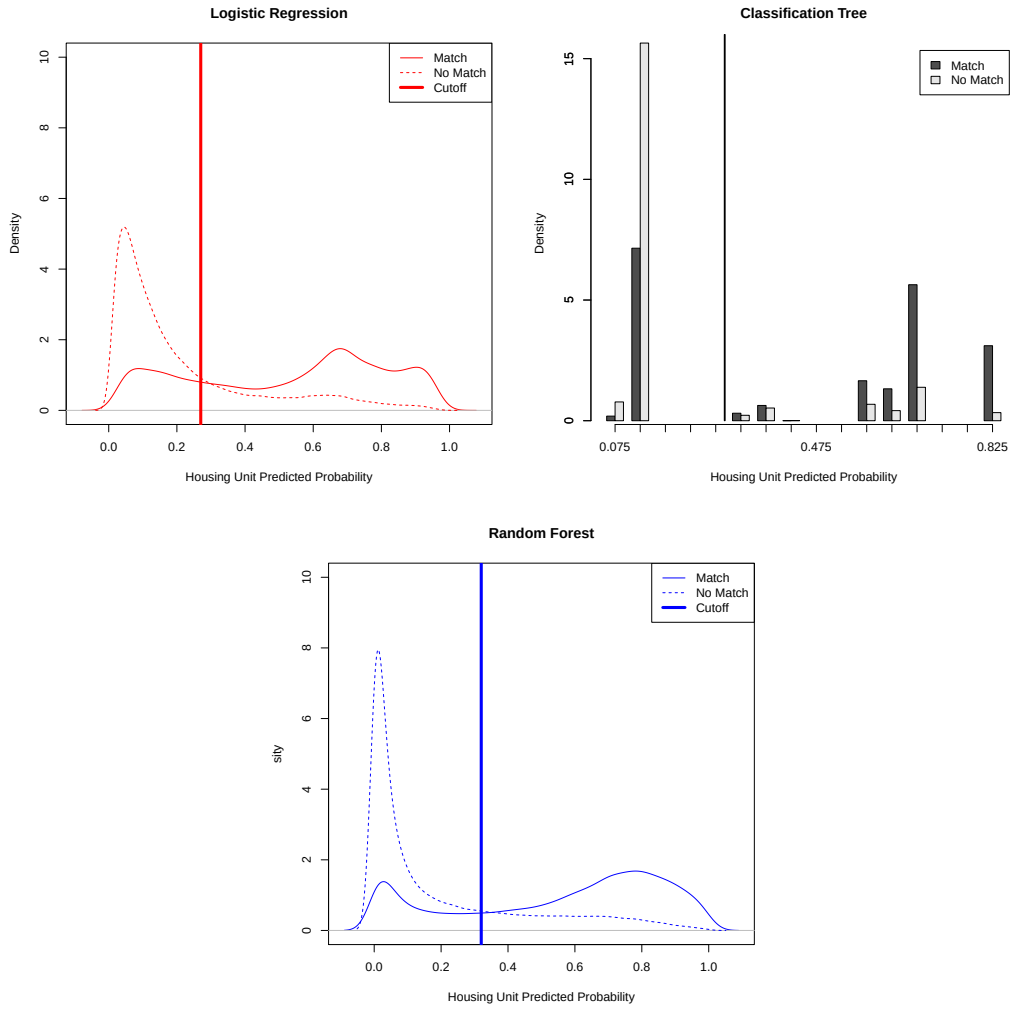

Figure 3: Distribution of Housing Unit Predicted Probabilities with Optimal Cutoff

for between 30 and 40% of the housing units for which we have administrative records.

Overall, the levels of agreement of housing unit level predictions of administrative records quality is high between all three methods. The highest level of agreement is between the random forest and logistic regression, where 92% of the predictions agree; while the classification tree agreement of predictions with both the random forest and the logistic regression is 88%.

Of the 12% of housing units predictions that differ between the logistic regression and the classification tree, about 83% of these housing units are deemed appropriate for administrative record enumeration by the logistic regression but not by the classification tree. This result is consistent with the high false negative associated with the classification tree approach, but only about 35% of these housing units are indeed a population count match. Similarly, of the 12% of housing units predictions that differ between the random forest and the classification tree, about 80% of these housing units are deemed appropriate for administrative record enumeration by the logistic regression but not by the classification tree, where about 39% of these housing units are a population count match.

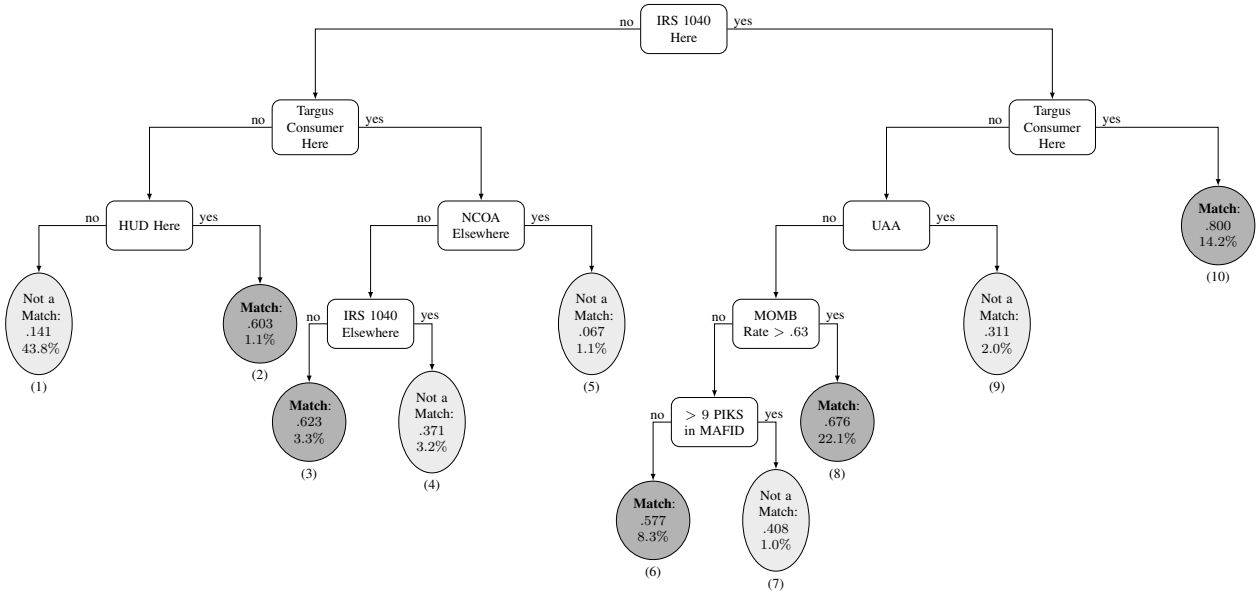

Figure 4: Classification Tree - Person Level, Majority Rules, Predicted Probability of an Address Match

3.2 Classification Tree: Tree Diagram and Comparison of Important Predictors

Figure 4 displays the classification tree grown on the 1% sample of NRFU housing units. Darker terminal nodes denote a predicted administrative record/Census address match and terminal node numbers are displayed below the nodes. Predicted address match probabilities are displayed below the binary prediction (Match/Not a Match) followed by the percentage of individuals from the 1% sample that follow that path in the classification tree. Note that the default class assignment for classification trees in the *rpart* package in R is based on majority rules, thus the tree in Figure 4 assigns final predicted outcomes based on the majority class at each terminal node. Cost-complexity pruning is used to prune the tree: the subtree at which the cross-validated error rates do not decrease by more than twice the standard error is selected. It is important to note that a more complex tree would yield greater prediction accuracy. However, we choose this tree pruning criteria since we are interested in a less complex tree for purposes of defining rules.

Trees naturally lend themselves to simple interpretation in the form of rules for classification. In this tree of 10 terminal nodes, we classify each person as a predicted administrative/Census address match or not based on the path they follow in the tree. For example, about 44% of individuals are predicted to not have an address match (Node 1: $\hat{p}_i = .141$) because their IRS 1040 return, Targus Consumer record, and Housing and Urban Development (HUD) record did not have that person at the given address; while about 22% of individuals are predicted to have an address match (Node 8: $\hat{p}_i = .676$) because their IRS 1040 return had that person at the given address, even though the Targus Consumer record did not, but the address was not "Undeliverable as Addressed" (UAA), and the mail-out-mail-back rate (MOMB Rate) of the tract was greater than 63%.

It is important to note that only a subset of variables from the logistic regression (about 14%) are driving the classification tree results. The selected variables in the classification tree are consistent with the statistical and practical significance of the explanatory variables from the logistic model. Table 2 and Figure 5 shows the the odds ratios and statistical significance of the subset of predictors from the logistic regression that the classification

tree algorithm selects to use to form the tree.

Table 2: Logistic Regression Results for Key Variables in Classification Tree

Variable	Odds Ratio	z-score	p-value
IRS 1040 Here	4.53	176.53	.000
Targus Consumer Here	1.52	38.88	.000
Housing and Urban Development (HUD) Here	4.87	76.80	.000
National Change of Address (NCOA) Elsewhere	.360	-79.47	.000
IRS 1040 Elsewhere	.653	-42.78	.000
Undeliverable as Addressed (UAA)	.428	-71.83	.000
Mail-out-mail-back (MOMB) Rate	3.40	35.86	.000
$>$ 10 PIKS at address (MAFID)	.330	-59.92	.000

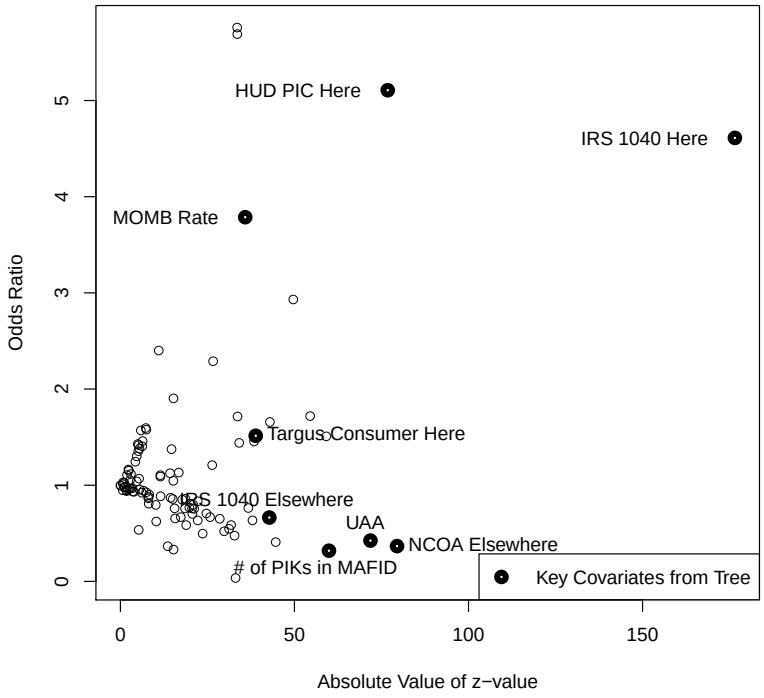

Figure 5: Logistic Regression Odds Ratios and z-values

Results from the random forest analysis also provide insight regarding the important variables for determining administrative record quality. The mean Gini impurity gain produced by each of the variables and mean decrease in classification accuracy after permuation of each of the variables used in the random forest are indicators of variable importance. Figure 6 shows the mean Gini impurity gain and mean decrease in classification accuracy for all variables. Not surprisingly, this set of variables with large mean Gini impurity gains and large mean decreases in classification accuracy are consistent with those selected from the classification tree and those with large practical and statistical significance from the logistic regression.

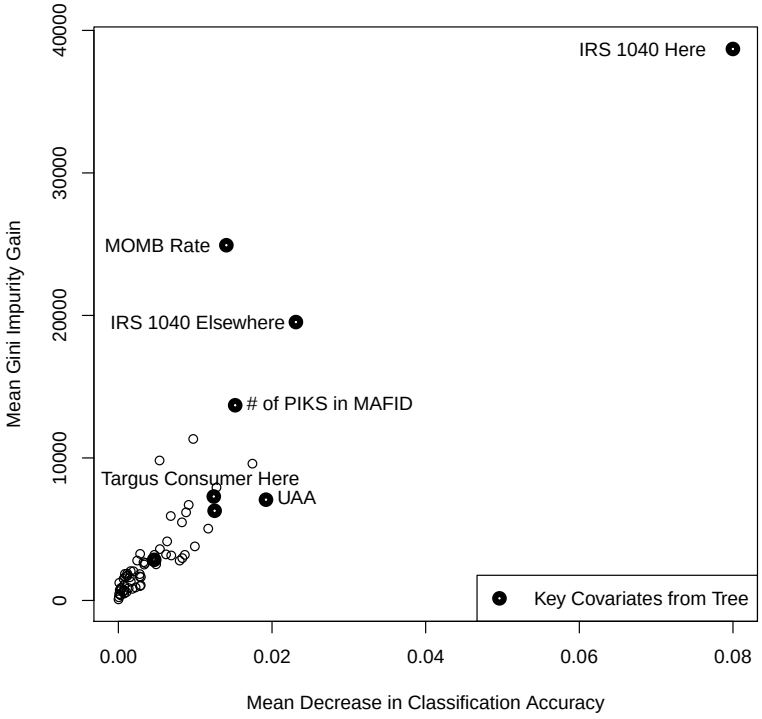

Figure 6: Random Forest Variable Importance

Rather than defining binary predictions via majority rules, we can classify each person as a predicted administrative/Census address match or not based on the path they follow in the tree *and* a chosen cutoff. For example, about 14% of individuals are classified into Node 1 with $\hat{p}_i = .80$ because their IRS 1040 return and Targus Consumer record placed that person at the given address. If $.80$ is larger than the chosen cutoff (e.g. based on a criteria of the ROC curve) then this person record would be deemed appropriate for enumeration purposes. Note however, the final decision is at the housing unit level, thus a housing unit roster based on the administrative records is deemed appropriate if all person record predicted probabilities exceed the chosen threshold.

Based on the ROC optimality criterion, the classification tree in Figure 7 represents the final individual-level decisions where the predicted classes at each terminal node are determined by their relationship with the optimal cutoff. In this scenario, nodes 4, 7, and 9 are now classified as an address match, which implies that all individuals who have an IRS 1040 at the given address (the entire right side of the classification tree) are classified as a predicted address match.

The discrete nature of the ROC curve obtained for the classification tree makes for easy interpretation of the tradeoff between misclassification error types. In fact, each level of predicted probabilities correspond to particular sets of paths down the tree. This information can be used to determine which paths to use as rules for administrative record enumeration. Table 3 presents a hierarchy of the decision rules based on predicted probability of an address match.

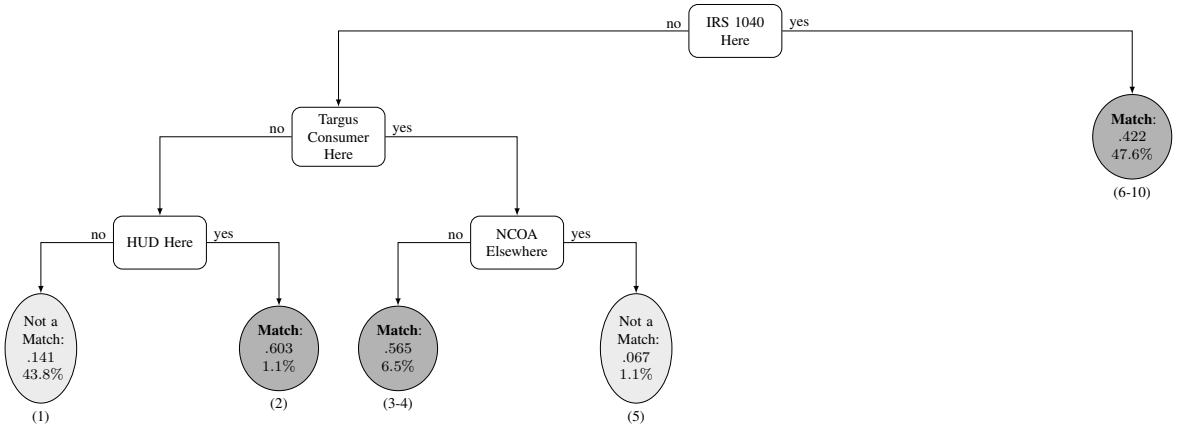

Figure 7: Classification Tree - Person Level, Optimal Cutoff, Predicted Probability of an Address Match

Table 3: ROC Curve Coordinates - Classification Tree

Nodes	Cutoff	FNR	FPR	Proportion in AR Enumeration
All	.067	.000	1.00	1.00
1,2,3,4,6,7,8,9,10	.141	.001	.961	.970
2,3,4,6,7,8,9,10	**.311**	**.367**	**.178**	**.313**
2,3,4,6,7,8,10	.371	.382	.168	.300
2,3,6,7,8,10	.408	.414	.141	.272
2,3,6,8,10	.577	.414	.141	.272
2,3,8,10	.602	.497	.107	.224
3,8,10	.623	.506	.104	.219
8,10	.676	.563	.086	.189
10	.800	.845	.017	.058
None	1.00	1.00	.000	.000

3.3 Weighting Misclassification Error Types

The objective functions for determining the cutoff described in Section 2.3 treat both types of misclassification error types equally. This assumption is likely not a a good one in this application. A false positive means that administrative record enumeration is used ($\hat{y}_h^c = 1$) for the housing unit when administrative record enumeration did not agree with Census enumeration ($y_h = 0$); while a false negative means that administrative record enumeration is not used ($\hat{y}_h^c = 0$) for the housing unit when administrative record enumeration did agree with Census enumeration ($y_h = 1$). In determining which administrative records are sufficiently reliable to use for Census enumeration, a false positive implies a loss of accuracy while a false negative implies a loss of cost savings. The consequences of these two types of error determine the cost/quality tradeoff.

Loss matrices, e.g. weighting of misclassification error types, can be incorporated in classification trees and random forests in two ways: (1) in determining the splits at each node and (2) in determining predicted outcomes from the terminal nodes. To compare to the logistic regression analysis in the weighted cases, we will assume varying degrees of importance of false negatives and false positives in the classification tree and random forest

analyses. We will focus specifically on using a loss matrix in determining the predicted outcome from the terminal node as this serves as the analog to the logistic analysis with weighting of misclassification error types.

Assuming varying degrees of importance of false negatives and false positives allows the decision-maker to directly influence the statistical methodology with their desired balance of cost and quality. A weight, w, can be introduced into the objection function to represent the decision-maker's opinion. For example, for the Manhattan distance objective function:

$$\arg\min_{c} \Big(w(\text{False Positive Rate}(c)) + \text{False Negative Rate}(c) \Big)$$

The weight, w, dictates how much more costly a false positive is compared to a false negative. In other words, how much do we value accuracy over cost savings? If we value accuracy more than cost savings, then $w > 1$. If we value cost savings more than accuracy, then $w < 1$. Table 4 shows optimal cutoffs and the corresponding false negative rate, false positive rate, proportion misclassified, proportion of housing units in composite data selected for administrative record enumeration, and coverage ratio for housing units selected for administratice record enumeration (total population count from AR/total population count from 2010 Census) for varying levels of relative importance.

Table 4: Results at Optimal Cutoff by Method with Weighting

w	c	FNR	FPR	Proportion in AR Enumeration*	Coverage Ratio**
Logit					
5	.74	.727	.036	.106	.974
2	.50	.416	.127	.262	.993
1	.30	.285	.227	.368	1.009
1/2	.16	.161	.405	.532	1.078
1/5	.04	.020	.846	.886	1.307
Classification Tree					
5	.80	.845	.017	.056	.953
2	.58	.414	.141	.272	1.032
1	.31	.367	.179	.313	1.050
1/2	.31	.367	.179	.313	1.050
1/5	.07	.000	1.00	1.00	1.462
Random Forest					
5	.76	.667	.046	.131	.968
2	.54	.390	.133	.273	.976
1	.36	.278	.212	.362	.993
1/2	.14	.169	.372	.507	1.059
1/5	.01	.026	.849	.886	1.337

* Of housing units with administrative records.

** For housing units selected for administrative records enumeration.

4. Summary of Comparison of Methodologies

Cost/Quality The overall goal of this research in the context of 2020 Census planning is to reduce costs of the NRFU operations while maintaining quality. The method of weighting of misclassification types for purposes of defining binary predictions allows the decision maker to have influence over this tradeoff between cost and quality. The weighting can result in very different conclusions, but provide a statistical grounding for defining and deciding on the level of cost savings and enumeration quality.

Complexity/Accuracy In this application, interpretability is important for purposes of implementing decisions for Census 2020 production. We have shown that classification trees achieve ease of interpretability without much of a loss in predictive power (particularly at low levels of the false positive rate), depending on the decision makers ideal balance of misclassification error types. While random forests and logistic regression may achieve better prediction accuracy than a classification tree, they lack the defined rule structure that a single classification tree exhibits. The classification tree analysis selects only a subset of the many possible predictor variables and optimally determines easily understood rules for determining the use of administrative records. While all three methods take all independent variables as inputs, the classification tree is the only of the three methods that does not subsequently rely on all independent variables for out-of-sample prediction. The decision maker may conclude that the sacrifice of predictive accuracy for a less complex methodology is warranted.

Strong Predictors All three methods uncover dominant explanatory variables for predicting administrative record/Census address matches in the person-place model. In fact, in one scenario, the classification tree for predicting administrative record/Census address matches in the person-place model relies on just 4 explanatory variables. Most notably, the address associated with the IRS 1040 return plays a prominent role in prediction. In fact, for one of the largest nodes in the classification tree based on the optimal cutoff, IRS 1040 address is the sole predictor. The level of the importance of the IRS 1040 address is also reflected in variable significance measures from the logistic regression and random forest.

Alternative Strategies A person-place model of the administrative record composite data is just one of many possible modeling techniques for determining an administrative record housing unit roster. Researchers are also investigating the use of housing unit-level models for determining occupancy status and housing unit counts. While the strong relationships in the person-place models suggest similar results may hold in housing unit-level models, a thorough analysis is left for future research.

This research builds an administrative records household roster by taking the union of all persons associated with a particular housing unit in any of the administrative sources, assuming an all-or-nothing approach (i.e. either all persons associated with an address are used to enumerate the housing unit or the address remains part of the NRFU operation). Researchers are investigating a variety of alternative ways to build a housing unit from administrative sources.

5. Appendix

5.1 Explanatory Variables in Person-Place Models

Geography and Housing Unit Level Explanatory Variables: mail-out-mail-back return rate, type of enumeration area (update/leave; military; urban update/leave indicators in logit), address characteristic type (indicators for all levels in logit), mobile or other housing structure, number of units in housing structure (2-4, 5-9, 10-19, 20-49, 50+ indicators in logit), Spring 2010 DSF deliverable flag, Spring 2010 DSF X flag, 6-Month periods since last DSF deliverable flag, never had DSF deliverable flags, had DSF deliverable flag every time since Fall 2008, MAF source (2000 LUCA address; post-2000 LUCA address; 2010 address canvassing address; 2010 decennial added address indicators in logit), replacement mailing type (block blanketed with second forms; targeted block, additional form sent; targeted block, additional form not sent indicators in logit), bilingual form, address type (business address; residential, excluded from delivery statistics indicators in logit), built after 2000, has location description in MAF, missing DSF route and MAF valid unit status.

Administrative Record Explanatory Variables: presence of UAA, same race/hispanic origin for all persons in housing unit, number of administrative record PIKs in the MAFID (indicators in logit), number of administrative record PIKs in MAFID that responded in non-NRFU operations (indicators in logit), indicators for presence of source at the given MAFID (sources: IRS 1040, IRS 1099, HUD CHUMS, HUD PIC, HUD TRACS, SSS, Medicare, IHS, NCOA, NY Snap, SSR, Experian-EDR, Experian-Insource, InfoUSA, Melissa, Targus-Consumer, Targus-Wireless, VASGI-NAR, VSGI-TRK, Texas SNAP, Targus-NAF, Corelogic, 2000 Census), and indicators for presence of source at a different MAFID (sources: IRS 1040, IRS 1099, HUD CHUMS, HUD PIC, HUD TRACS, SSS, Medicare, IHS, NCOA, NY Snap, SSR, Experian-EDR, Experian-Insource, InfoUSA, Melissa, Targus-Consumer, Targus-Wireless, VASGI-NAR, VSGI-TRK, 2000 Census).

REFERENCES

Breiman, L., (2001), "Random Forests," *Machine Learning*, 45 (1): 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., (1984), *Classification and Regression Trees*, Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Brown, J. D. (2013), "Synthesizing Numerous Pre-Existing Sources to Accurately Enumerate Nonresponding Decennial Housing Units," Census Report.

Capps, C., and Wright, T., (2013), "Toward a Vision: Official Statistics and Big Data," Amstat News.

Liaw, A. and Wiener, M., (2002), Classification and Regression by randomForest. R News 2(3), 18–22.

Metz, C.E., (1978), "Basic Principles of ROC Analysis," *Seminars in Nuclear Medicine*, 8(4): 283–298.

Therneau, T., Atkinson, B. and Ripley, B., (2013), rpart: Recursive Partitioning. R package version 4.1-3. http://CRAN.R-project.org/package=rpart