Recent Advancements in Statistical Modeling to Identify Address Updating Areas for the 2020 Census

Christine Gibson Tomaszewski, John L. Boies[1]
US Census Bureau, 4600 Silver Hill Road, Washington, DC 20323

Abstract
The 2010 Address Canvassing (AC) Operation was the second most expensive field operation in the 2010 Census at more than 450 million dollars in direct costs. Since the last Census research effort concluded in 2012, substantial gains have been made in both the predictive accuracy and resulting potential cost reduction of the statistical procedures being researched. A range of new independent variables from diverse sources including property tax records, the United States Post Office, and Federal agencies have resulted in substantial improvements in model fit. Better cost estimates of the 2010 AC operations have greatly improved the accuracy and utility of the cost/benefit analyses stemming from the statistical modeling.

Key words: address listing, modeling, statistical prediction, logistic regression, cost-benefit, cost reduction, decennial Census, 2010 Census, 2020 Census

1. Introduction

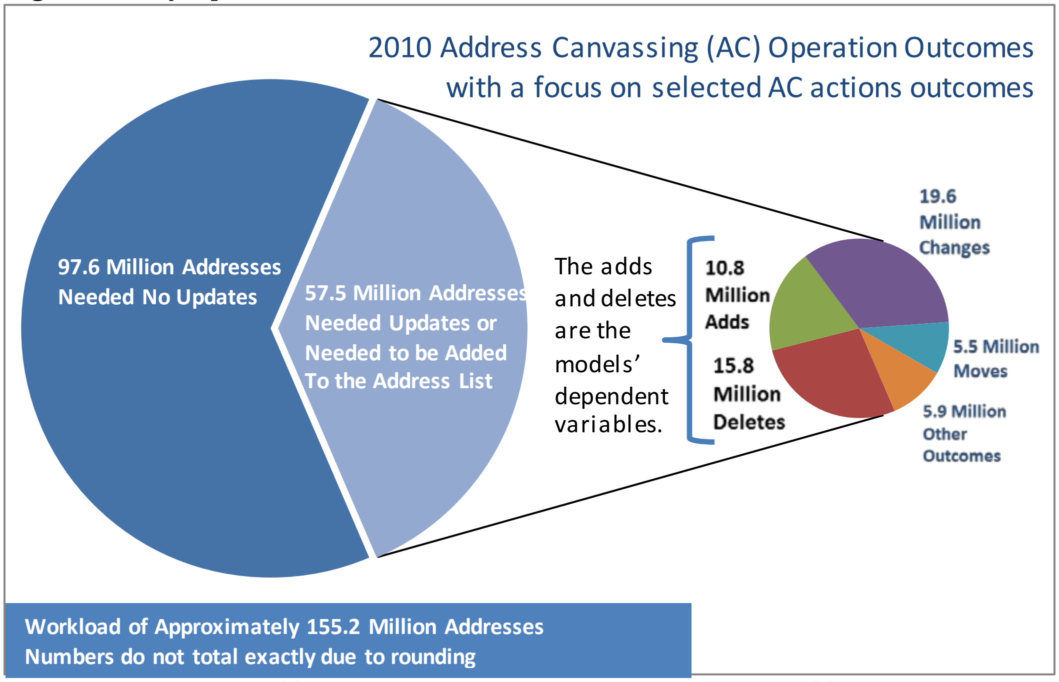
The 2010 Address Canvassing (AC) operation updated the U.S. Census Bureau's address list nationwide in preparation for the 2010 Census. It was the second most expensive single operation in the 2010 Census at more than 450 million dollars in direct costs and about 845 million in total cost (Holland, 2012). Between April and July of 2009, about 110 thousand field representatives worked over 10 million hours in the AC operation prior to the 2010 Census, while driving over 68 million miles. The 2010 AC operation required an extensive field infrastructure, which is unlikely to be funded during the next Census. Prior research (Boies, Shaw, and Holland, 2012), undertaken as part of the 2010 Census Program for Evaluation and Experiments (CPEX), indicated that Targeted Address Canvassing (TAC) was a feasible alternative. The work presented here is a direct extension of the CPEX TAC research. This paper expands on the microsimulation scenario guiding their research: "What if we used a statistical model developed from 2008 or earlier data to select areas for canvassing in 2009?" These predictions are compared to the actual address list updates by the 2010 Census AC operation.

It is reasonable to ask why we need to update the Census Bureau's address list. Figure 1 shows that of the roughly 155 million addresses worked during the 2010 AC operation, over 57 million needed updating or needed to be added to the Census Bureau's address list. In particular, as will be discussed further in the methodology section, we needed to

[1]Christine Gibson Tomaszewski and John L. Boies are Mathematical Statisticians in the Decennial Statistical Studies Division of the U.S. Census Bureau. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical and methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.

add nearly 11 million addresses to our address list and remove nearly 16 million addresses. AC adds were important to avoid missing people in the Census who lived in addresses we did not know about, while deletes were important to avoid incorrect mail-outs and increased expenses in follow-up operations.

Figure 1. Why Update the Census Bureau's Address List?



Source: 2010 Census Address Canvassing Operational Assessment (Address List Operations Implementation Team, 2012).

The research reported here is different from this prior research in several substantive ways. While the earlier research (Boies, et al., 2012) used 2000 census tabulation blocks as its primary unit of analysis, this research is based on 2010 tabulation blocks. This not only allows the use of additional independent variables in the analysis that are only available for 2010 tabulation geography, it also makes the modeling outcomes more meaningful for researching the implementation of a TAC solution for the 2020 Census. Nearly all 2010 to 2020 intra-census data and research will be carried out using 2010 tabulation blocks. Based on the results of the prior research, the data reported here are focused on a narrower range of outcome measures (two vs. eleven) and examined the explanatory power of a much wider range of predictor variables (> 2,000 versus < 40). We also developed new benchmarks to assist in evaluating the performance of the statistical models based on the outcomes of a “Perfect Model,” i.e., a set of coverage and workload predictions based on the assumption that we could perfectly predict where AC outcomes would have occurred in the 2010 AC operation.

2. Data

The primary source of the data used in this analysis is the 2010 Census Address Frame Combination File (the 2010 Combo file), which contains data for all fifty states, the District of Columbia, and Puerto Rico. The 2010 Combo file is comprised of eight extracts from the Master Address File (MAF)/Topologically Integrated Geographic

Encoding and Referencing (TIGER[2]) database (MTdb). (See Ward, 2011.) The 2010 Combo file offers snapshots of the MTdb before and after the 2010 AC operation. We considered two categories of data (physical characteristics and social characteristics) when developing the models presented here. Physical characteristics (derived from the Combo file) covers items such as address descriptions or housing unit (HU) counts aggregated to a block level. The social characteristics data (or social demographics, such as race, sex, age, and population) originated from the Statistical Administrative Records System (StARS). StARS is composed of Administrative Records data collected from other federal agencies, including the Internal Revenue Service, Centers for Medicare and Medicaid Services, Department of Housing and Urban Development, Indian Health Service, and Selective Service System. The models presented in this paper only use data on the block physical characteristics.

The universe of records examined for this research was tallied into 2010 tabulation blocks. While 2010 collection blocks were used to administer the AC operation, collection blocks are not used in other operations while 2010 tabulation blocks will be used throughout the decade; thus, 2010 tabulation blocks are appropriate for this research.

To address our scenario of "What if we used a statistical model developed from 2009 data to select areas for canvassing in 2010?" we chose to use binomial logistic regression to obtain a 'target'/'do not target' ('1'/'0') outcome probability for each census block in the country. We considered various types of AC action outcomes for dependent variables. AC outcomes later shown in Table 1 include adds (addresses not sent to AC), changes (address changes or updates), deletes (address deletion), duplicates (two or more addresses consolidated to one address), moves (an add outcome matched to a delete outcome resulting in the address being assigned to a different block), and verifies (no address change or update needed). The universe of the approximately 11.2 million 2010 census tabulation blocks in the United States and Puerto Rico contained about 6.8 million blocks with AC outcomes. The AC outcomes represent the differences between the contents of the MAF just prior to AC and the "ground truth" as determined by Census field staff during the AC operation. We chose AC adds and AC deletes as the AC outcomes to investigate as dependent variables for the logistic regression modeling applied to this scenario.

Just over 200,000 of the blocks with AC outcomes contain only AC outcomes that were unknown prior to the completion of the AC Operation - about 1.3 million AC adds. Hence, these blocks had no usable MTdb data prior to the 2010 Census AC Operation. Therefore, these blocks are considered 'empty blocks' for the purposes of the current research. Limited information is available for blocks with zero valid addresses before AC, so this paper excludes those empty blocks. (See Boies and Tomaszewski, 2014, for more information on empty blocks.) This leaves about 6.6 million 'non-empty' blocks containing data on just over 144 million addresses for use in this research. The study universe contains 9,490,804 AC adds in the 6,585,835 2010 census tabulation blocks.

AC adds can be separated into two types: 'matched adds' and 'new adds.' Matched AC adds were addresses that were not on the dependent list sent out to the AC operation but were later matched to existing records at headquarters (HQ); in other words, we had

[2] TIGER is the mapping/spatial portion of the database, which contains street and other feature details, as well as mapspots, or geographic coordinates, corresponding to the location of MAF addresses.

received the addresses from some source, usually from the Delivery Sequence File (DSF) of the United States Postal Service (USPS). Most of these records were not sent out to the AC operation since they were ungeocoded (in other words, there was no corresponding block of any type in the MTdb). New AC adds were addresses that were not on the dependent AC list and for which we had no prior knowledge of their existence. While some data and software is available for assigning blocks to ungeocoded DSF addresses so that these matched adds can be located outside of a canvassing operation and are available for subsequent Census operations (see Tomaszewski, 2013), the new AC adds would not have been available for further Census operations without the 2010 AC operation. Therefore, new AC adds are of higher priority for listing purposes, as matched AC adds may be locatable through less expensive HQ operations. Both types of AC adds represent potential missing people in the Census if those addresses and the people living there are not counted in subsequent operations. Consequently our models predict the presence of both types of AC adds.

We also modeled AC deletes alongside AC adds. AC deletes are addresses (from the dependent canvassing list) removed in the field and verified as a delete by a second field representative (thus sometimes referred to as double deletes). Missed AC deletes from the non- selected areas represent potential increases in expenditures for later operations, such as the 2010 Non-Response Follow-Up (NRFU) operation conducted to enumerate census non-respondents. If AC deletes are not identified prior to NRFU, this could result in wasted resources in a more expensive operation than AC. Therefore, canvassing cost avoidance should not be achieved by delaying expenses to later operations.

Table 1 shows the distribution of selected AC outcomes. It indicates that, with the exception of AC verifies, AC outcomes are heavily skewed with actions being concentrated in relatively few blocks. Data presented in Boies, et al. (2012) showed that, for most AC actions, the substantial plurality of outcomes occurred in a small handful of 2000 Census tabulation blocks. For 2010 tabulation blocks, the concentration of AC actions of interest remains in relatively few blocks.

Table 1. Select Address Canvassing Outcome Statistics per 2010 Tabulation Block

Address Canvassing Outcome	Minimum	Mean	Maximum	Skewness	Standard Deviation	Sum	N
All Adds	0	1.44	3,675	46.48	9.44	9,493,487	6,586,835
New Adds	0	0.92	3,674	74.28	6.57	6,064,638	6,586,835
Matched Adds	0	0.52	1,578	56.75	6.04	3,428,849	6,586,835
Verifies	0	14.82	1,970	8.04	31.66	97,635,142	6,586,835
Changes	0	2.98	2,995	34.01	17.26	19,608,740	6,586,835
Moves	0	0.83	1,193	33.91	6.70	5,450,557	6,586,835
Duplicates	0	0.62	790	34.79	4.82	4,083,890	6,586,835
Deletes	0	2.40	4,215	45.14	13.40	15,812,653	6,586,835

Source: 2009 TAC research database.

The data from the Perfect Adds (all adds) Model shown in Table 2, which orders the 11.2 million 2010 tabulation blocks by all 10.8 million AC adds, shows that about four and a half million adds (41.8 percent of all adds) occurred in 0.7 percent of the blocks (75,879) containing about five percent of the addresses. The methodology of the Perfect Adds Model is discussed in the next section. About three-fifths of all AC adds (and 10 percent of addresses) were in 211,194 blocks (1.9 percent). Just over four-fifths

(8,653,706) of all AC adds (and 20 percent of addresses) were in only 590,349 blocks (5.3 percent).

Table 2. Simulated 2010 Census Targeted Address Canvassing "Perfect Census Block Model" for Maximizing Add Capture Rate (All Adds)

		Perfect Adds Model					
HU count		Blocks		Add Capture Rate		Delete Capture Rate	
%	(millions)	Count	%	(millions)	%	(millions)	%
5	7.2	75,870	0.7	4.5	41.8	1.5	9.4
10	14.5	211,194	1.9	6.5	60.3	2.7	17.1
20	29.0	590,349	5.3	8.7	80.3	4.8	30.3
30	43.5	1,112,799	10.0	9.9	91.5	6.6	41.5
40	58.0	1,817,521	16.3	10.6	98.1	8.1	51.2
50	72.4	3,150,034	28.2	10.8	100.0	9.5	59.8
60	86.9	4,750,893	42.6	10.8	100.0	10.7	67.8
70	101.4	6,352,417	56.9	10.8	100.0	12.0	75.9
80	115.9	7,956,596	71.3	10.8	100.0	13.3	83.9
90	130.4	9,556,078	85.7	10.8	100.0	14.6	92.0
100	144.9	11,155,486	100.0	10.8	100.0	15.8	100.0

Source: 2010 Combo file.

From the TAC standpoint, this is good news, indicating that there is a substantial opportunity for reducing costs by only selecting blocks most likely to have AC outcomes of interest for field listing rather then canvassing all 11 million blocks. In the perfect world, field staff could have visited just over five percent of all blocks with addresses (20 percent of all addresses) and still recovered the vast majority of AC adds. This could have delivered a substantial cost reduction (close to a possible 95 percent reduction in block workload or 80 percent reduction in HU workload).

3. Methodology

The purpose of the TAC research is to develop an array of blocks (or other geographic designators) ordered from those that are the least likely to require fieldwork to those most likely to require fieldwork to update the MTdb prior to the 2020 Census. Because the goal is to predict a 'target'/'do not target' ('1'/'0') outcome for each block, we chose logistic regression for our modeling tool. This procedure is easy to use and very robust (Hosmer and Lemeshaw, 1989) and produces predicted probabilities admirably suited for this purpose.

The research we discuss here focused on two types of AC outcomes for dependent variables: AC adds and AC deletes. AC adds represent potential missing HUs and people in the Census if those addresses and the people living there are not counted in subsequent operations. Missed AC deletes from the non- selected areas represent potential increases in expenditures for later operations. We modeled the occurrence of two or more AC adds per block, which is referred to as the 'Adds Model' in this paper. The 'Adds and Deletes Model' is where we modeled the occurrence of at least two or more adds or deletes per block:

- two or more AC adds or
- two or more AC deletes or
- one or more AC add and one or more AC delete.

There were 1.0 million blocks containing at least two AC adds and 2.3 million blocks containing at least two AC adds or deletes

Given the priority of predicting AC adds, since many of them did not exist on the MTdb before the 2010 AC operation and therefore have no address level information, we chose block-level modeling using the address information from existing HUs in a block to predict whether adds occurred in the block.. We chose the block as our unit of analysis because preliminary research indicated census blocks would be more efficient for selective canvassing purposes than census tracts (Tomaszewski and Shaw, 2013). Address-level modeling is possible for AC deletes, but is not feasible for AC adds since the majority of them lacked pre-2009 address-level data.

We considered a wide range of independent variables for this analysis. More than 2,000 different measures were analyzed before we settled on the models presented here. Given the surfeit of data, a systematic approach to sifting through the available information was implemented.

The primary data reduction was done prior to the model building process. First, we examined the distributions of the available variables. Many of the variables were categorical variables at the address level that were aggregated to the block level. These categorical variables were coded as dummy variables. In many cases there were so few addresses in many of the categories (a few hundred or thousand out of more than 144 million addresses sent to the 2010 AC operation) that they would have little discernable impact on the dependent measures. These variables or categories did not make it any further in the data analysis. Some were rejected outright, while others were combined with related categories.

The next data reduction step was to remove variables from consideration because they were uncorrelated with any of the dependent measures. Variables with Pearson Product Moment Correlations of less than ±0.10 with the count or mean number of adds and deletes in a block were excluded from further investigation. This substantially reduced the number of variables under consideration for the model building process. We grouped the measures into theoretically meaningful blocks of variables (e.g., race and ethnicity measures, population counts) and used forward and backward selection procedures to create the most parsimonious models. Criteria for inclusion or exclusion were increment in max-rescaled R^2, change in census overcoverage, and reduction in workload compared to our primary benchmark, the “Perfect Model.” At the time of this writing, we have developed two adequate models, but have not yet created parsimonious “best” models.

To evaluate our models, we created several benchmark ‘Perfect Models’ (see Tomaszewski and Shaw, 2013). To construct these Perfect Models, we ordered all the 2010 tabulation blocks by counts of 2010 AC actions starting with the blocks with most AC outcomes through the blocks with the least.. The Perfect Adds Model maximizes the adds capture rate by ordering the 2010 tabulation blocks by all AC adds. It is the relevant benchmark for the Adds Model discussed in this paper.

The current model evaluations examine HU cut points of 5, 10, 20, or 30 percent. At some point in the process, a cut point will be selected to create the areas selected for a 2020 field operation. The workload and coverage metrics for the Perfect Adds Model presented previously in Table 2 shows that, with perfect prediction, only 5 percent of the blocks in the model universe would be selected for AC at a HU cutoff of 20 percent (or

29 million HUs). This would result in an add capture rate of 80 percent and a delete capture rate of 30 percent.

The Perfect Adds and Deletes Model maximizes both the adds capture rate and the deletes capture rate by ordering the 2010 tabulation blocks by all AC adds and AC deletes. It is the relevant benchmark for the Adds and Deletes Model discussed in this paper. Table 3 shows that, with perfect prediction, only 2 percent of the blocks in the model universe would be selected for AC at a HU cutoff of 20 percent (or 29 million HUs). This would result in an add capture rate of 53 percent and a delete capture rate of 51 percent.

Table 3. Simulated 2010 Census Targeted Address Canvassing “Perfect Census Block Model” for Maximizing Add and Deletes Capture Rate (All Adds and Deletes)

		Perfect Adds and Deletes Model					
HU count		Blocks		Add Capture Rate		Delete Capture Rate	
%	(millions)	Count	%	(millions)	%	(millions)	%
5	7.2	23,376	0.2	1.9	17.9	3.0	18.7
10	14.5	74,503	0.7	3.4	31.2	5.0	31.7
20	29.0	260,253	2.3	5.7	52.6	8.1	51.1
30	43.5	564,964	5.1	7.4	68.6	10.5	66.1
40	58.0	1,000,654	9.0	8.6	80.1	12.3	78.0
50	72.4	1,594,525	14.3	9.5	88.6	13.8	87.3
60	86.9	2,395,477	21.5	10.2	94.6	14.9	94.1
70	101.4	3,454,774	31.0	10.6	98.6	15.6	98.5
80	115.9	5,587,473	50.1	10.8	100.0	15.8	100.0
90	130.4	8,374,961	75.1	10.8	100.0	15.8	100.0
100	144.9	11,155,486	100.0	10.8	100.0	15.8	100.0

Source: 2010 Combo file.

The cost-benefit analysis shown here examines trade-offs between potential costs (using workload estimates) as they relate to measures of quality (via address coverage estimates). Workload estimates were based on the number and percent of 2010 tabulation blocks and addresses included for a selected canvassing. As blocks vary in size and complexity, they will also vary in canvassing expense, thus the number of HUs canvassed will feed the potential costs. Coverage estimates in this report were based on the number and percent of adds and deletes found for a selected canvassing.

Evaluative measures of model fit discussed include Area Under the receiver operating characteristic Curve (AUC) and max-rescaled R^2. AUC (or Concordance from the SAS logistic output) varies from a low of 0.5 for a random selection curve to 1.0 for a perfectly fitting curve. The max-rescaled R^2 from the SAS logistic output is a measure of how well the model fits and how well the independent variables explain the dependent variable with a range from 0 (not at all) to 1 (perfect).

We also report on the overcoverage rates, as missed delete actions from the areas not selected represent potential increases in expenditures for later operations, such as the 2010 NRFU operation. Each HU in the NRFU operation costs about as much to process as did each block in the AC operation (see Walker, Winder, Jackson, and Heimel, 2012). Cost efficiencies for future canvassing operations is important, but those potential savings should not be achieved by transferring expenses to other operations.

4. Results

Table 4 comparing the Adds Model to Perfect Adds Model for four potential HU cut-off points indicates that the Adds Model still has room for improvement. Interestingly, the delete capture rate is higher than the adds capture rate in the Adds Model, indicating the variables that predict adds do a better job of predicting deletes.

Table 4. Comparison of the Adds Model to the Perfect Adds Model

		Adds Model			Perfect Adds Model		
% HUs	HU count (millions)	% Blocks	Add Capture Rate	Delete Capture Rate	% Blocks	Add Capture Rate	Delete Capture Rate
5	7.2	0.3	9.0	10.8	0.7	41.8	9.4
10	14.5	1.1	17.3	19.8	1.9	60.3	17.1
20	29.0	3.2	30.3	33.7	5.3	80.3	30.3
30	43.5	5.7	40.8	45.5	10.0	91.5	41.5

Source: 2009 TAC research database and 2010 Combo file.

The models presented here represent progress since the first TAC research was published. Table 5 compares the Adds Model Outcomes for the current Adds Model with the 2012 CPEX TAC Report Adds Model.[3] The modeling outcomes presented here offer improvements over the work presented in the CPEX TAC report (Boies, et al., 2012). The current Adds Model provides a better model fit, assessed via the max-rescaled R^2 and the AUC. The max-rescaled R^2 has doubled, while the AUC has increased slightly. Unfortunately, the number of variables included in the model has increased as well.

Table 5. Comparison of the Adds Model from the Census Program for Evaluations and Experiments (CPEX) TAC Report to the Current Research

	CPEX TAC Report Adds Model	Current Adds Model
Max-rescaled R^2	0.13	0.26
Area under the Curve	0.72	0.80
Independent Variables	16	52

Source: 2009 TAC research database and 2010 Combo file.

The Adds and Deletes Model presented in Table 6 fits better than the Adds Model. The max-rescaled R^2 of the Adds and Deletes Model is higher at 0.42, while its AUC is 0.84. There are also more variables (61) in the Adds and Deletes Model than in the Adds Model (52). There are more variables available for modeling that correlate well with AC deletes than with AC adds. The distribution of adds and deletes among the 2010 tabulation blocks does not always overlap, as adds and deletes often occur in different tabulation blocks. Of the 3.5 million blocks that contain AC adds or AC deletes, only 1.1 million (32 percent) contain both AC outcomes. Selecting for variables predicting adds will thus help explain the difference in model fit, and in the add capture rate and delete capture rate seen in Table 6. Unfortunately, despite the relatively good AUC performances of both Adds Model and the Adds and Deletes Model, the apparent add capture for both using the percent HU metric is still low, as there is a relatively constant linear association between the number of HUs and AC adds in a block. Cutoffs based on HUs reduce the apparent efficiency of the models.

[3] The TAC report did not present an Adds or Deletes model.

Table 6. Comparison of the Adds Model to the Adds and Deletes Model

		Adds Model			Adds and Delete Model		
% HUs	HU count (millions)	% Blocks	Add Capture Rate	Delete Capture Rate	% Blocks	Add Capture Rate	Delete Capture Rate
5	7.2	0.3	9.0	10.8	0.2	6.2	13.9
10	14.5	1.1	17.3	19.8	0.6	12.9	24.5
20	29.0	3.2	30.3	33.7	2.2	25.2	41.4
30	43.5	5.7	40.8	45.5	4.6	36.0	54.8

Source: 2009 TAC research database and 2010 Combo file.

The delete capture rate increases from 45.5 percent in the Adds Model to 54.8 percent in the Adds and Deletes Model. This is a definite improvement, but still not ideal, given the potential costs in a large number of lost deletes (7.1 to 8.6 million) continuing on to a future NRFU operation. However, Table 7 shows that the gap between the add and delete capture rates of the Adds and Deletes Model and its corresponding Perfect Model is not as vast as between the Adds Model and its Perfect Model.

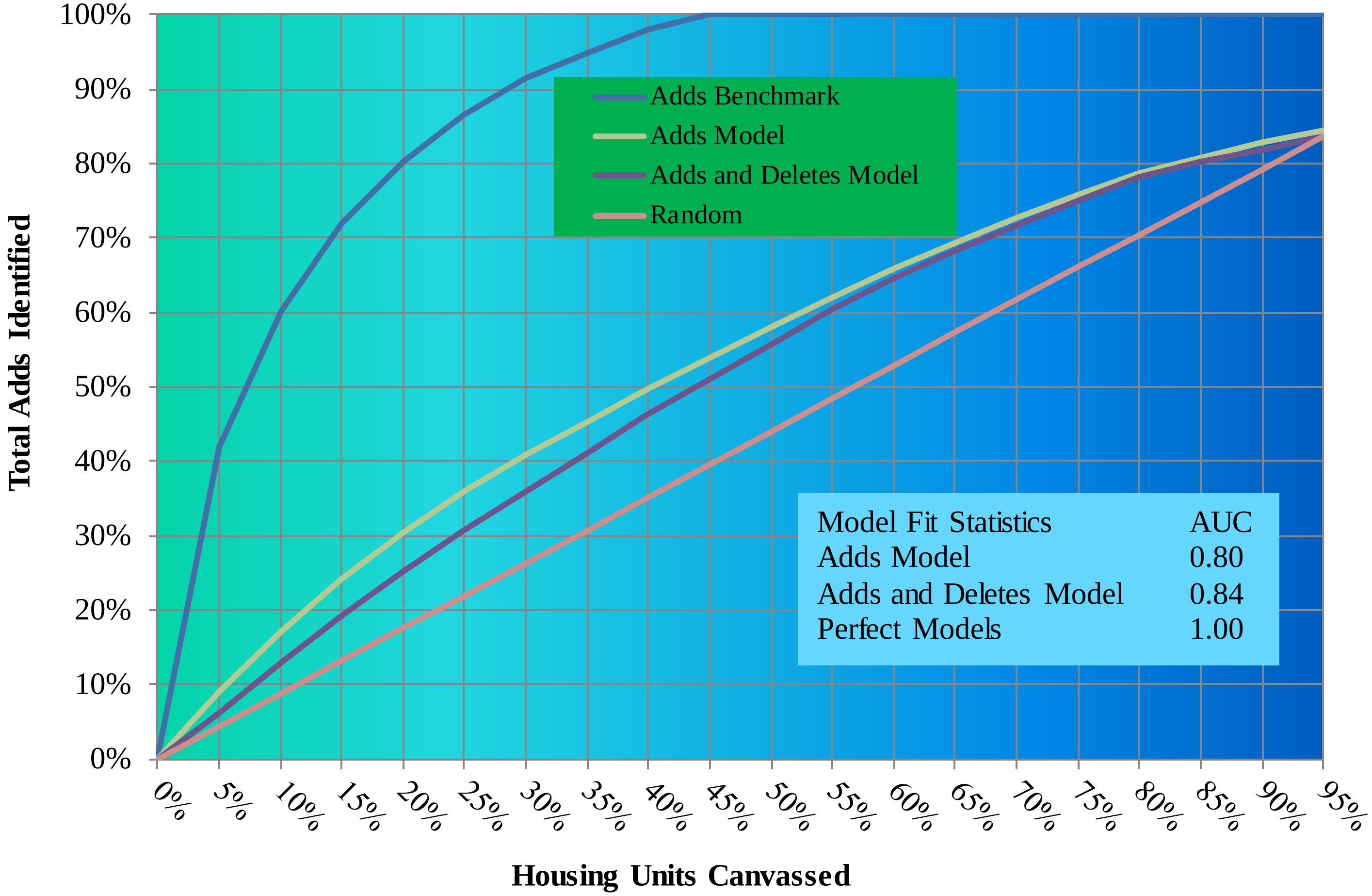
Table 7. Comparison of the Adds and Deletes Model to the Perfect Adds and Deletes Model

		Adds and Delete Model			Perfect Adds and Delete Model		
% HUs	HU count (millions)	% Blocks	Add Capture Rate	Delete Capture Rate	% Blocks	Add Capture Rate	Delete Capture Rate
5	7.2	0.2	6.2	13.9	0.2	17.9	18.7
10	14.5	0.6	12.9	24.5	0.7	31.2	31.7
20	29.0	2.2	25.2	41.4	2.3	52.6	51.1
30	43.5	4.6	36.0	54.8	5.1	68.6	66.1

Source: 2009 TAC research database and 2010 Combo file.

Figure 2 provides a cost/benefit comparison by plotting the percentage of HUs canvassed by percentage of AC adds found for each model. The fit of the two models can be seen visually in the middle two curves shown in Figure 2. While the Adds Model more closely fits the Perfect Adds Model better than the Adds and Deletes Model fits the Perfect Adds Model (the Adds and Deletes Model does a better job of fitting the Perfect Adds and Deletes Model; the Perfect Adds and Deletes Model is not shown in Figure 2), both models fit adequately as reflected by the AUC. Thus the blocks chosen in the Adds Model are in quite different order than the Adds and Deletes Model and more efficient at reducing the AC HU workload while optimizing for Adds.

Figure 2. Cost/Benefit Comparison for the Models – Coverage by Housing Units and Adds



Source: 2009 TAC research database.

This figure is useful not only for understanding the model fit, but also as a visual of the cost benefit trade-offs of selected canvassing. Depending on the acceptable levels of adds identified, a decision could be made to canvass larger or smaller percentages of HUs. Alternatively, a canvassing level could be chosen based on the available budget and the corresponding potential quality loss could be identified.

5. Conclusions and Discussion

The Adds Model fits have improved since Boies, et al. (2012) released the CPEX TAC research. The Adds Model and the Adds and Deletes Model both fit well according to the AUC. These indicators bode well for 2020 cost efficiencies, as they indicate potential workload reduction that may allow reduction of infrastructure/overhead costs – not just direct costs. Unfortunately, the add and delete capture rates for both models is low, so further research is needed to improve the models. The low delete capture rate for both models may have implications for later operations, particularly given the high count of AC deletes.

In conjunction with exploring additional predictors in block level models including data on block economic characteristics using property tax and IRS data as well as geographic information, we plan to work on address level delete modeling to handle the issue of potential lost delete actions in non-selected blocks. Delete models could be used alongside Undeliverable as Addressed codes from the USPS to remove addresses from follow-up workloads. In Boies and Tomaszewski (2014), we explore some of the ways to ameliorate the deleterious effects of these lost deletes as well as examine the empty block problem touched on earlier here.

Acknowledgements

Many people contributed to the successful progression of this research. Our Census Evaluations Branch (Jonathan Holland, Nancy Johnson, Kevin Shaw, Matthew Virgile, and Justin Ward) offered feedback on the methodology, coding, statistical analysis, and supporting prose. Thanks to Magdalena Ramos for her facilitation and feedback. Thanks to Inez Chen and Jennifer Tancreto for their feedback. Many, many thanks to Claude Jackson for his endless and invaluable IT support.

References

Boies, John L., and Tomaszewski, Christine Gibson (2014), "Fielding a Targeted Address Canvassing Operation: Alternative Approaches to Moving from Predictive Statistical Modeling to a Cost Effective Address Canvassing Field Operation for the 2020 Census," presented at 2014 JSM.

Boies, John L., Kevin M. Shaw, and Jonathan P. Holland (2012), "2010 Census Program for Evaluations and Experiments (CPEX): Address Canvassing Targeting and Cost Reduction Report," DSSD 2010 Census Program for Evaluations and Experiments Memorandum Series A-09, July 12, 2012. http://www.census.gov/2010census/pdf/2010_Census_Address_Canvassing_Targeting_and_Cost_Reduction_Evaluation_Report.pdf

Address List Operations Implementation Team (2012), "2010 Census Program for Evaluations and Experiments (CPEX): 2010 Census Address Canvassing Operational Assessment," Washington, D.C., January 10, 2012.

http://www.census.gov/2010census/pdf/2010_Census_AC_Operational_Assessment.pdf

Holland, Jonathan P. (2012), "2010 Census Program for Evaluations and Experiments (CPEX): Evaluation of Automation in Field Data Collection in Address Canvassing Report," DSSD 2010 Census Program for Evaluations and Experiments Memorandum Series A-05, July 24, 2012. http://www.census.gov/2010census/pdf/2010_Census_Evaluation_of_Automation_in_Field_Data_Collection_in_Address_Canvassing_Report.pdf

Hosmer, Jr. David W. and Stanley Lemeshow (1989), Applied Logistic Regression. John Wiley & Sons, 1989.

Tomaszewski, Christine Gibson (2013), "2010 Census Program for Evaluations and Experiments: Evaluation of Address List Maintenance Using Supplemental Data Sources Report," DSSD 2010 Census Program for Evaluations and Experiments Memorandum Series A-06, February 20, 2013. http://www.census.gov/2010census/pdf/2010_Census_Evaluation_of_Address_Listing_Maintenance_Using_Supplemental_Data_Sources.pdf

Tomaszewski, Christine Gibson, and Kevin M. Shaw (2013), "Examining Census 2000 Tabulation Block and Tract Homogeneity of Census 2010 Address Canvassing Action Codes, for 2020 Census Targeted Address Canvassing Modeling," DSSD 2010 Decennial Census R&T Memorandum Series DCRT-R01, May 21, 2013.

Walker, Shelley, Susanna Winder, Geoff Jackson, and Sarah Heimel (2012), "2010 Census Program for Evaluations and Experiments (CPEX): 2010 Census Nonresponse Followup Operations Assessment," April 23, 2012. https://www.census.gov/2010census/pdf/2010_Census_NRFU_Operations_Assessment.pdf

Ward, Justin (2011), "Documentation for the 2010 Census Address Frame Combination File, Version #1," DSSD 2010 Decennial Census Memorandum Series D-19, August 24, 2011.

Reports in the 2010 Census Program for Evaluations and Experiments (CPEX) are located at http://www.census.gov/2010census/about/cpex.php.