Model-Based Targeted Address Canvassing: A Simulation Based on the 2009 Address Canvassing Program

John L. Boies, Kevin M. Shaw, and Jonathan P. Holland[1]

US Census Bureau
4600 Silver Hill Road, Washington, DC 20233

Abstract

We use data from the 2009 Census Addressing Canvassing (AC) operation to conduct a "What If" simulation of a model based "targeted" AC program where census blocks are selected for canvassing based on their predicted probabilities of deviating from the master address file data. Covariates measuring block characteristics of two kinds, physical and social structure, were used to predict 11 different canvassing outcomes. The results indicate that both physical and social structure are important predictors of whether blocks warrant a visit by field staff. The research indicates that models predicting which blocks should be targeted for address canvassing can be developed and that this approach could result in substantial savings of time and money in preparation for the 2020 Decennial Census with a minimal effect on Census quality.

Keywords: targeted address canvassing, logistic regression, cost/benefit analysis, Decennial Census, 2010 Census, 2020 Census, coverage, cost reduction, statistical prediction Adminstrative Records, StARS.

I. Background and Motivation

This research is from the 2010 Census Program for Evaluations and Experiments (CPEX) study evaluating, at the national level, the utility of using a model-based methodology to target specific areas for Address Canvassing (AC) in preparation for the 2020 Census. The primary focus of this evaluation is the possible cost reductions that may result from concentrating AC efforts on areas of the nation that will yield the most cost effective updating of the Master Address File (MAF). The potential cost outcomes resulting from targeted AC efforts are examined using a micro-simulation based on the 2010 Census AC operation. The "What if" simulation question used is: "What outcomes would be different if the Census Bureau selected census blocks for AC based on a model using data on the blocks in the AC universe that were available at the time of the operation?" The statistical tool used is logistic regression. This problem-solving effort is an Address Information Micro-Simulation (AIMS).

The Census Bureau developed the MAF in the 1990s. The MAF is primarily maintained through a semiannual update provided by the United States Postal Service's (USPS)

[1]The authors 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.

Delivery Sequence File (DSF). Other sources of MAF maintenance include the Local Update of Census Addresses (LUCA) program, other canvassing/listing operations, and geographic partnership programs.

Before the 2000 and 2010 Decennial Censuses, major operations were fielded to update the MAF in preparation for enumeration. In 2000, Block Canvassing and Address Listing operations were carried out do a final pre-census update to the MAF. In 2010, the AC operation involved field staff visiting nearly every block in the 50 states, District of Columbia, and Puerto Rico to verify and update the MAF. This was the second most expensive single operation in the 2010 Census behind the non-response follow-up operation at about $845 million in direct and indirect costs (Holland 2012). The opportunities for reducing costs in such a large operation are manifest and thus are the primary target of the research reported here.

Two research questions guided this research: 1) *Is it possible to model the outcomes of the 2010 AC operation based on a priori data?* And 2) *Once some basic models predicting AC outcomes are developed, can these statistical models be turned into useful tools to allocate AC resources?*

II. Data

This project used data from two sources: the 2010 Address Frame Combination (AF COMBO) file and extracts from the 2000 to 2008 Statistical Administrative Records System (StARS). The 2010 Census AF COMBO file consists of eight groups of census files merged together at the address level using the Master Address File Identification (MAFID) number. The universe for all these files is the 50 U.S. states, the District of Columbia (DC), and Puerto Rico (PR). An extract of the 2010 Census AF COMBO file, was produced using only records with a non-missing action code on the AC action code variable with data from selected variables, then summarized as counts and averages at the Census 2000 current block level. The StARS database is composed of Administrative Record (AR) data collected from other federal agencies, including the Internal Revenue Service (IRS), Centers for Medicare and Medicaid Services, Department of Housing and Urban Development, Indian Health Service, and Selective Service System; as well as data from the Social Security Administration. These files provide demographic variables and a range of address quality measures. The universe for the StARS files is the 50 U.S. states and DC.

The two extract files, from the 2010 Census AF COMBO file and StARS, were merged together using Census 2000 current block identifiers into the Multi-Use Multi-Source (MUMS) file. After aggregating and merging these files into the MUMS file, there were 6,319,298 census block-level records. Removing duplicate blocks and blocks not in the study universe left 5,809,915 records available for this study.

As part of the quality control process, data on AC operation outcomes published in the 2010 Census AC Assessment Report (Table 11.7) were compared to summary data from the records used by this study. The comparisons shown in Table 1 indicate a nearly exact match between the distribution of action codes shown in Table 11.7 from the 2010 Census AC Assessment Report and the action codes used in this study.

[Table 1 about here]

The differences in the number of action codes germane to this research are small: five fewer Adds (action code "A"), one fewer Change ("C"), two fewer Deletes ("D"), and 30 fewer Verifications ("K"). Most of the differences between the 2010 Census AC Assessment totals and the totals in the study universe are likely the result of different Census geographies being combined to make the 2010 AF COMBO file.

While the extant literature describing the modeling of residential change for census address listing purposes is virtually non-existent, there is substantial research regarding the causes of residential change and development (Tauber, 2009; Schiwirian 1983). This literature indicates that neighborhood demographics are central to understanding the stability of local residential communities. Primarily the age, sex, race, and ethnic composition of neighborhoods are suggested as all contributing to residential changes over time. Given that address-level changes within blocks are being modeled, it is expected that measures of residential structure of blocks may affect the outcome of a listing operation. The size of blocks, types of housing units, and population density of blocks could all play a role in AC outcomes.

Based on these two general sources of residential dynamics, social structure and physical structure, we selected 11 variables as being viable for use in the models presented here. The variables selected were not meant to produce the "best" model of residential change nor do they represent any attempt at testing competing theories of residential change.

The StARS file provided block-level social structure measures: blocks with a population more than two percent Black were scored 1, 0 otherwise; blocks with more than two percent under the age of 19 were scored 1, 0 otherwise; blocks with more than two percent Hispanic origin were scored 1, 0 otherwise; blocks were scored 1. If there was less than two percent change in the mean population for 2007-2008 from the mean for 2000-2004, 0 otherwise; and the standard deviation of the population for 2005-2008.

The physical structure of the block was measured with six variables: the number of HUs (addresses); the proportion of HUs in multi-unit structures; the ratio of the number of addresses that are in the StARS database that do not match to the MAF to the number that do match; whether a block from the 2000, 2001, or 2002 StARS files that matches to the COMBO file matches to a block from 2008 StARS (1=yes, 0 otherwise); blocks with more than 600 HUs were coded 1, otherwise 0; and a measure of residential complexity.

Five interaction terms using the large block variable were also used. Large block interactions with the number of HUs, Hispanic, Black, and Children present blocks, and the standard deviation of the block's recent population are included in the models presented here.

The descriptive statistics for the independent variables used in the model procedures shown in Table 2 indicate that most of the independent variables are substantially skewed. This was not unexpected given the nature of the data, and because logistic regression is a very robust procedure, there is no expectation that this issue will significantly affect the interpretation of the data (Hosmer and Lemeshow, 2000).

[Table 2 about here]

We used simple dichotomous yes/no measures of AC outcomes in order to produce easy-to-interpret predictions of the likelihood of a block being included in an AC operation. The use of dichotomous dependent measures in conjunction with the appropriate statistical method, in this case logistic regression, will produce a predicted probability assignment for each block in the study universe (Hosmer and Lemeshow, 2000). Blocks can therefore be easily ranked by their predicted probability of containing outcomes of interest.

This research focuses on four AC action outcome codes: adds that did not duplicate existing MAF records, delete actions in the field that resulted in addresses being removed from the MAF (double deletes), changes, and moves. These outcomes are the final results of the entire 2010 Census AC operation. The action code summary statistics in Table 3 indicate that the most common action was "verified." Field staff verified over 97 million addresses in the 2010 Census AC operation.

[Table 3 about here]

The mean number of moves and new adds per block are both near 1 with similar standard deviations, about 9 and 8 respectively. The distributions for both of these variables are narrower than any of the other action codes. Changes are the most common individual action with an average of over 3 per block. This is not unexpected since the actions included in this category range from changing a unit designation from "A" to "1" to changing a street name and street directional. Deletes have a direct effect on gross overcoverage, and were therefore an area of interest.[2]

Figure 1 specifies that, for some action codes, the most common number per block is 0 with the next most common value being 1 per block. More than 5 million blocks (about 88 perent) have no moves and over 4 million blocks (about 75 percent) have no adds. Delete and changes are absent from over 3 million blocks (about 56 percent). It is very important to note, however, that most blocks have at least one of these four actions occurring in them. Only about 1.7 million blocks (about 30 percent) do not have any type of action, i.e., no adds, deletes, changes, nor moves. As the number of actions per block increases the decline in the number of blocks in each category, for all charted groups, is fairly shallow. Still, as Figure 1 indicates , the number of blocks with very large numbers of actions is small, e.g., for adds and moves, very few blocks have more than 100 actions, 4,771 and 6,571 respectively.

[Figure 1 about here]

These distributions have several implications for Targeted Address Canvassing (TAC) research. Most importantly, the fact that most blocks contain at least one action means that any allocation scheme that substantially reduces the number of blocks canvassed will also reduce the number of some type of actions being discovered, even if it retains all occurrences of a given action. Thus, some prioritization of action codes is required.

The models presented here use 11 dependent variables derived from the four actions previously described. Since this is an exploratory study, there are no strong reasons to select these coding schemes over any other except that they are a wide array of possible

[2] Undiscovered deletes may result in Census questionnaires being mailed to uninhabited or non-existent addresses and thus result in costly Nonresponse Followup (NRFU) visits.

measures of AC outcomes. Three measures of new adds are used: blocks with 1 or more, 5 or more, and 10 or more adds. Deletes were coded into 1 or more and 5 or more deletes per block. Changes were coded into 1 or more, 5 or more, and 10 or more changes. The last variable used encompassed all four acitons; 1 or more of any the four actions occurring in a block.

III. Methodology

To answer the first research question, "*Is it possible to model the outcomes of the 2010 AC operation based on a priori data?*" we focused on building a number of basic models adequate to demonstrate the feasibility of a modeling approach. The micro-simulation concept was implemented by using data that existed prior to the 2010 Census AC operation to model the AC outcomes. This part of the study relied on SAS™'s PROC LOGISTIC to estimate the models. Although a range of models were tested with different versions of the dependent variables as well as a range of independent measures, only 11 models were selected for presentation here. They are not meant to be the definitive models of AC outcomes, but rather, representative of the range of reasonable approaches to answer the first study question.

The primary criterion for selection of the independent variables was the Max-rescaled R^2 goodness of fit measure calculated by the SAS procedure. Those variables that did not produce an observable (approximately greater than 0.001) improvement in that measure were not pursued further in the modeling process. However, since this exercise is a proof of concept rather than a model testing exercise, no rigid application of a systematic variable selection process was attempted (e.g., stepwise procedures). The ones chosen for presentation here simply represent examples of the range of reasonable measures tested. Predicted probabilities from each of the 11 models were saved.

The Max-rescaled R^2 for each model was used to assess our models' goodness of fit. This is the ratio of likelihood (L) of the intercept-only model to the likelihood of the estimated model rescaled, so it has the same range as the ordinary least square (OLS) R^2 estimator, 0 to 1. However, it only approximates the meaning of the OLS goodness of fit measure.[3]

The second study question, "*Once some basic models predicting AC outcomes are developed, is it possible to turn these statistical models into tools useful for allocating AC resources?* is addressed by using the saved predicted probabilities from the logistic models to assess the potential savings and costs of selecting some census blocks for canvassing over others. Based on these predicted values, an AC Cost/Benefit (AC/CB) spreadsheet tool was developed to assist in the evaluation of the cost and coverage tradeoffs. The creation of this tool involved three basic steps: 1) The individual block predicted probabilities are rounded to the nearest centile; 2) Evaluation variables including number of HUs, action outcomes, and AC cost were then summed by the centiles; and 3) Cumulative sums of these variables were then calculated

This AC/CB tool can be used to estimate the approximate cost reduction a specific selection of census blocks may have produced in the 2010 Census AC operation, and detail the coverage trade-off resulting from the lost adds, deletes, changes, or moves. The

[3] For a more explanation and sources see: http://support.sas.com/onlinedoc/913/docMainpage.jsp

simulated cost reduction and concurrent coverage degradation is a function of which model's predicted probabilities are chosen for selecting blocks.

IV. Results

The results shown in Table 4 for the 11 models estimated for this study have Max-rescaled R^2 values ranging from a low of 0.079 for the model predicting census blocks with 1+ Moves to a high of 0.243 for the 5+ Deletes model, with an average of about 0.168. Given that logistic regression does not maximize the fit of the estimated parameters and the skewed distributions of most of the independent variables in the model, the fit of models presented here is acceptable for the primary objectives of this study. The odd-ratios shown in Table 4 are calculated at the mean of the continuous variables, e.g., number of HUs per block, or at the value of 1 for the 0/1 measures, e.g., blocks with children present (Child-Present Blocks).

[Table 4 about here]

Because some variables are part of interaction terms with the variable for large blocks, i.e., the number of HUs/block, Children-, Blacks-, and Hispanics-present blocks, and standard deviation of the last 5 years of block population, two odds ratios are reported for these measures. One ratio is for blocks with more than 600 HUs, "large blocks," and the other ratio is for those with 600 or fewer HUs.

The reported odd-ratios for the block physical structure variables indicate some counterintuitive results. Logic would predict that the more HUs there are in a block, the greater the odds of a block having some type of AC operation outcome. However, at the mean number of HUs/block (26.707) the odds ratios for blocks that are not large are above 1.0 only for models predicting blocks with 5+ Deletes and the two "Moves" models. In all other cases the odd-ratios are below 1.0. In the case of large blocks, the odds ratios are above 1.0 only for the 5+ Moves model. The reduction in the odds ratios for most models for the non-large block estimate is substantial, as much as about 63 percent for the 1+ Any Action model at the mean number of HUs/block compared to blocks, at a theoretical minimum of 0.[4] For large blocks, the reduction in the odds ratio is even greater, as much as 75.8 percent in the case of the 1+ Any Action model. A possible reason for the counterintuitive result for the number of HUs is inclusion of the multi-unit, residential complexity, and large block variables in the analysis. At the mean value of the multi-unit measure, the odds ratios are positive for all of the Adds models. For the two "Deletes" and the two "Moves" models, the odds-ratios are of similar magnitude.

The residential complexity measure has a strong positive effect on the likelihood of actions for all models except the two "Moves" models. The odd-ratios at the mean complexity value (49.881) range from a substantial 15.310 to a low of 1.731 for the 1+ Any Action and 10+ Changes, models respectively. The effect for the two "Moves" models is considerably weaker. The effect for the large block measure (large blocks that are less than two percent Hispanic, Black, or children and have a population standard deviation of 0) is also quite impressive. For these blocks, the odds ratios range from 26.977 for the 10+ Adds model to a low of 1.797 for the 1+ Deletes model. For the "Changes" models, the effect is negative, ranging from 0.001 for the 10+ Changes model

[4]There are no census blocks with zero HUs in the study universe.

to 0.927 for the 5+ Changes model. Note that very few blocks have a population standard deviation of zero. However, it is still valid to interpret these coefficients as indicating that large blocks with these demographic indicators have a higher relative likelihood of experiencing nearly all types of actions than do other blocks. The large block, multi-unit, and residential complexity measure probably account for the counterintuitive effect of the number of HUs.

The ratio of MAF non-matched addresses to MAF matched addresses in a block also has mostly negative effects on the action outcome likelihoods with odds ratios as low as 0.744 for the 10+ Adds models. For the rest of the models, the values differ little from 1.

For the block social structure variables, the outcomes are less consistent than the physical structure measures. The presence of greater than two percent Blacks, Hispanics, and children in a block variables for blocks with less than 600 units, all have a consistently negative effect in the "Adds" models, with Hispanics present blocks having the weakest effects. Blocks with Children-present has the strongest effect on these models, reducing the odds by as much as 30 percent. All three measures increase the odds of blocks having moves. In the case of the Child-present measure, this increase is as much as 63 percent (odds ratio of 1.628) in the 1+ Moves model. The presence of Blacks in a block increases the odds of there being five or more deletes in a block (53 percent).

For large blocks, the effects of these three variables sometimes deviate considerably from their effects in smaller blocks. For large blocks with more than two percent Hispanic, the odds ratios are lower than for smaller blocks for the "Adds" models. The pattern is similar for the two percent Black blocks. In the case of large blocks with Children, the effects are weaker than in smaller blocks for two of the "Adds" models. However, the effect is somewhat stronger for the 5+ Adds model. The odds ratios for the large blocks with more than two percent Hispanic population are greater than 1.0 for the 1+ Deletes models. In the "Changes" models the effects are positive for the 5+ Changes model (the same direction as for smaller blocks) but negative for the 1+ Changes model (for large blocks, 1.562 compared with 0.933 for smaller blocks). The direction is the same but the effect is weaker for the 1+ Any Action model. The direction is also the same for the 5+ Moves models, but stronger in smaller blocks. The odds ratios for large blocks with more than two percent children are in the same direction as in smaller blocks in all three of the "Adds models", the 5+ Deletes, 5+ Changes, and the 1+ Moves models. In these latter models, the magnitude of the effects is noticeably larger for large blocks compared with smaller blocks. The effects are the opposite (positive) for large blocks with children than for smaller blocks in the case of the 1+ Deletes, 1+ Changes, and 1+ Any Action models. The largest difference was for the 1+ Any Action model with the large block odds ratio being 2.132 compared to the smaller block ratio of 0.692. The odds ratio for the 10+ Changes was not calculable because of the paucity of large blocks with more than two percent children and 10+ changes in the study universe.

For the population variables, no population change and standard deviation of the population from 2005 through 2008, the effects are varied. With the exceptions of the 1+ Adds and 1+ Any Action models, blocks with no population change have a negative effect in all the models with the strongest effects for the 5+ Deletes and 1+ Deletes models. The odds ratios for the standard deviation of the population, for both large and smaller blocks, are near 1.0 for most of the models. The biggest effect is for the 5+ Adds model with a reduction in the odds of a block having 5+ adds of only four percent.

There are five important takeaways from the logistical regression results. More HUs in a block, once adjusted, reduces the odds of AC outcomes. Blocks with higher proportions of HUs in multi-unit buildings have higher odds of adds but lower odds of deletes. More complex blocks have much higher odds of having AC outcomes. Blocks with more than two percent Hispanics, children, or Blacks present reduces the odds of adds. Large blocks with less than two percent children, Hispanics, or Blacks present have substantially larger odds of most types of outcomes.

The predicted probabilities produced by the models just described are used to create a prototype management tool. Table 5 summarizes the outcomes from one scenario using the predicted probabilities from each of the 11 models presented in this study. For this scenario, the primary criterion used was a gross undercoverage of about 0.5 percent resulting from excluding some census blocks from the operation. All the blocks at or below the predicted probability cut-off point corresponding to a gross undercoverage rate of about 0.5 percent were used to calculate the estimates in Table 5.[5] The coverage estimates are based on the number of adds or other actions lost by not canvassing the blocks at or below the probability cut off (in this case 0.5 percent). Table 5 provides the estimated gross undercoverage, gross overcoverage, net coverage, total error, number of blocks excluded, number of HUs excluded, percent of blocks excluded, percent of adds, deletes, changes, and move actions lost, and the expected cost reduction resulting from not canvassing the selected group of blocks.[6]

[Table 5 about here]

The data shown in Table 5 indicate that there is significant potential cost reduction to be had in an AC operation if blocks are targeted for canvassing based on model-informed allocation procedures. For example, the 1+ Any Action model produces potential savings of nearly $249.8 million at the penalty of introducing a gross 0.47 percent gross undercoverage. The $250 million estimate assumes that all census blocks cost the same to canvass, in this case about $79 per block.

However, it is very unlikely that all blocks require the same amount of resources, e.g., staff time and mileage. Some blocks take more time for field staff to list while other require more travel time. At the time of the writing of this paper, there were no readily available block-level cost estimates. Consequently, the residential complexity measure was used to create a more specific block-level cost estimate. This measure assumes that the relationship between block cost and complexity was linear, i.e., blocks with complexity measures two times larger than other blocks cost twice as much to list. Working with this assumption, the $459 million cost estimate for AC (Holland, 2012) was allocated to each block according to its complexity score as a proportion of the sum of all the blocks' complexity scores. This produced the estimated cost per block (est. dollars/block) savings shown in Table 5. For comparison, a saving estimate using a simple random sample of blocks for TAC is also presented. For the 1+ Any Action model, the savings range from a low of $45.30 million for a random selection of blocks,

[5] Because the summaries were done by percentile the gross undercoverage values are above or below the stated 0.5 percent value.

[6] For this study, gross undercoverage is the "Lost Adds"/All Positive AC actions (133,471,779 HUs (Chaar and Marquette, 2012)). Gross overcoverage is the "Lost Deletes"/All Positive AC actions. Net coverage is gross overcoverage - gross undercoverage and total error is the absolute value of the sum of gross undercoverage and the absolute value of gross overcoverage.

$117 million for the estimated per block cost estimate, to about $249.8 million for the average block cost estimate; all with a gross HU undercoverage rate of 0.47 percent. This table also shows the percent of adds, deletes, changes, and move actions that would have been lost had blocks been targeted for canvassing based on this model.

Table 5 makes it clear that the possible cost reduction and coverage tradeoffs are very dependent on which model is chosen to drive the block targeting plan. The 1+ Moves model yields an expected savings of only $115.7 million, based on the average cost per block cost estimate, only about twice the savings from a random approach. Using the estimated per block cost, the savings is expected to be lower than the random method, nearly $10 million less, $46.1 million compared with $55.6 million. The fact the expected outcomes of the block selection is so strongly affected by the model selected indicates the importance of the model building and selection process to this endeavor.

Data derived from the modeling process lend themselves well to graphical representations. Figure 2 plots the cost/benefit curves for the two cost estimates as well as the random selection curve (a straight line in this case) derived from the 1+ Any Action model. This representation allows the user to easily determine the expected

[Figure 2 about here]

cost reduction and associated coverage degradation to be estimated for a range of cost/benefit requirements. For example, a gross undercoverage cutoff of 0.3 percent yields an expected savings of just under $200 million for the average cost per block estimate and about $100 million for the estimated per block calculation.

The curves in Figure 3, also derived from the 1+ Any Action model, show the percent loss of adds, deletes, changes, and moves by cost reduction (cost per block measure). This curve allows the user to estimate the loss of specific action outcomes corresponding to specific levels of cost reduction. For example, at a cost reduction requirement of $300 million, the expected loss is about 14 percent of adds, 18 percent of changes, 19 percent of deletes, and 24 percent of moves.

[Figure 3 about here]

For some this may be too great a coverage degradation for the expected cost reduction. Instead, at a cost reduction point of $150 million the predicted loss of adds would be only 4 percent, with change and delete losses of about 8 percent, and lost moves at about 12 percent. This level of savings is still substantial, but with significantly fewer lost actions.

V. Conclusions

The distributions of AC actions presented here suggest two important conclusions. First, substantial rewards can be garnered if census blocks with no address canvassing outcomes can be identified prior to an AC operation. If only add actions were considered, then as much as 89 percent of total AC costs could be saved if those areas could be perfectly modeled identified (Tomaszewski and Shaw, 2013). Second, because about 70 percent of all blocks have some type of AC action, it will be difficult to garner substantial savings by excluding blocks from the AC operation without sacrificing some AC outcomes. While procedures or processes could be developed to mitigate some of this

degradation, e.g., continual updating of the MAF prior to the operation might reduce the number of lost moves or deletes, it is likely that no modeling effort will be able to perfectly predict where all types of outcomes might be found.

It is possible to model the outcomes of the 2010 AC operation based on a priori data. Albeit, the models presented here only scratch the surface of the available data that could be used, and only a narrow range of models and modeling methods were tested. Despite these limits, they do provide useful predictive power and some interesting information regarding block-level residential changes. The modeling outcomes presented in this report yielded some interesting results about residential dynamics, but can they be used for managing future census operations? Our answer to the second research question is that once some basic models predicting AC outcomes are developed, these statistical models can be turned into useful tools to allocate AC resources. The cost/benefit techniques developed here show good potential for allocating Census Bureau resources.

The AC/CB tool used to make Table 5 and Figures 2 and 3, provides an easy-to-use method for testing a range of scenarios regarding cost reduction and resulting tradeoffs that will likely occur in developing any prioritizing scheme for targeted AC. The "user" can select a savings goal, coverage goal, or various other types of cutoff points and readily see estimated outcomes. The results from the sample scenario presented in Table 5 also indicate that the AC/CB tool can be used to assess models and modeling techniques. Some models produce better cost reduction/coverage ratios than others. Other models might be better at preserving some types of outcomes (e.g., deletes) than others.

Based on the research presented here we have several recommendations for future action. First, there should be continued research on TAC and its implementation. This research should explore new sources of data, including traffic flow patterns, property tax, and other AR data to improve the modeling outcomes. More cost modeling research needs to be done in order to properly assess the research and development of TAC. And, since Census geography changes for each Decennial Census, new ways of clustering AC and other data must be developed in order to make this research applicable throughout the Decennial Census cycle. Note that, at the time of this writing, nearly all of these recommendations are being implemented to varying degrees at the Census Bureau.

Acknowledgements

This report is the product of the efforts of many colleagues including Jennifer Reicheart, Mayra Garcia, one of the first persons to work on this project, Justin Ward, Matthew Virgile, Kathleen Kephart, Christine Tomaszewski, and Nancy Johnson. I also thank the reviewers from the rest of the Census Bureau for their helpful comments.

References

Chaar, Ronia and RJ Marquette (2012). "2010 Census Program for Evaluations and Experiments: 2010 Census Address Canvassing Quality Profile" 2010 Census Planning Memoranda Series No. 184. April 4, 2012.

Holland, Jonathan (2012), "2010 Census Program for Evaluations and Experiments: "Study of Automation in Field Data Collection for Address Canvassing" Report, 2010 Census Evaluations and Experiments Memorandum Series #A-05, July 2012.

Hosmer, David, W. and Stanley Lemeshow (2000), *Applied Logistic Regression*. Wiley Series in Probability and Statistics.

Schwirian Kent P., (1983) "Models of Neighborhood Change", *Annual Review of Sociology*, 9:83-102.

Taueber, Karl E., (2009), *Residential Segregation & Neighborhood Change*, Aldine Transaction.

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 2020 Decennial Census R&T Memo Series #DCRT-R01 (May 21, 2013).

Ward, Justin (2012), "2010 Census Program for Evaluations and Experiments: "Evaluation of Data-Based Extraction Processes for the Address Frame" Report, 2010 DSSD CPEX Memorandum Series #A-04, June 27, 2012, 2010 Census Planning Memorandum Series No. 207, June 29, 2012.

Table 1. 2010 CPEX AC Targeting: Study Universe

2010 Census Post-AC Action Codes	Number of Addresses*	2010 AC Assessment*
Total	145,138,906	145,132,941
Adds	6,624,153	6,624,155
Changes	19,608,784	19,608,785
Double Deletes	15,819,919	15,813,921
Moves	5,450,563	5,450,563
Verified Addresses	97,635,487	97,635,517

Table 2. 2010 CPEX AC Targeting: Descriptive Statistics for Independent Variables (N=5,809,915)

Independent Variable	Minimum	Maximum	Mean	Std Dev.	Skewness
Number of HUs/Block	1	13,138	26.707	67.425	20.149
Proportion of Multi-Units/Block	0	1	0.071	0.2	3.189
Hispanic Population >2 Percent of Block Count	0	1	0.402	0.49	0.399
Black Population>2 Percent of Block Count	0	1	0.392	0.488	0.488
Child Population>2 Percent of Block Count	0	1	0.719	0.45	-0.974
Population Change <2 Percent from 2004-2008	0	1	0.239	0.426	1.225
Ratio of HUs without MAF Match	0	628.5	0.081	1.143	199.270
Blocks with 600+ Units	0	1	0.002	0.045	22.039
Std Dev. of Pop. 2005-8	0	9,137.94	4.413	23.059	88.837
StARS Block Mismatch	0	1	0.028	0.165	5.724
Residential Complexity	13	64,478.62	49.881	101.389	83.183
Large Block Interactions					
Number of HUs	0	13,138	2.003	50.827	44.593
Hispanic Present Blocks	0	1	0.002	0.044	22.550
Black Present Blocks	0	1	0.002	0.044	22.882
Children Present Blocks	0	1	0.002	0.044	22.391
STD of Block Population	0	9,137.94	0.391	19.751	135.518

Table 3. 2010 CPEX AC Targeting: Summary Statistics of Block Level AC Action Codes (N=5,809,915)

Number of Action Codes	Sum	Minimum	Maximum	Mean	Standard Deviation
New Adds	6,624,153	0	2,336	1.140	8.188
Changes	19,608,784	0	5,626	3.375	23.421
Double Deletes	15,819,919	0	7,459	2.723	16.199
Total Any Action	47,503,419	0	7,858	8.176	37.644
Moves	5,450,563	0	2,279	0.938	9.048
Verified Addresses*	97,635,487	0	5,464	16.805	39.313

Figure 1. 2010 CPEX AC Targeting: Number of Action Codes Per Block - Adds, Changes Deletes, Moves, and Any Type of Action

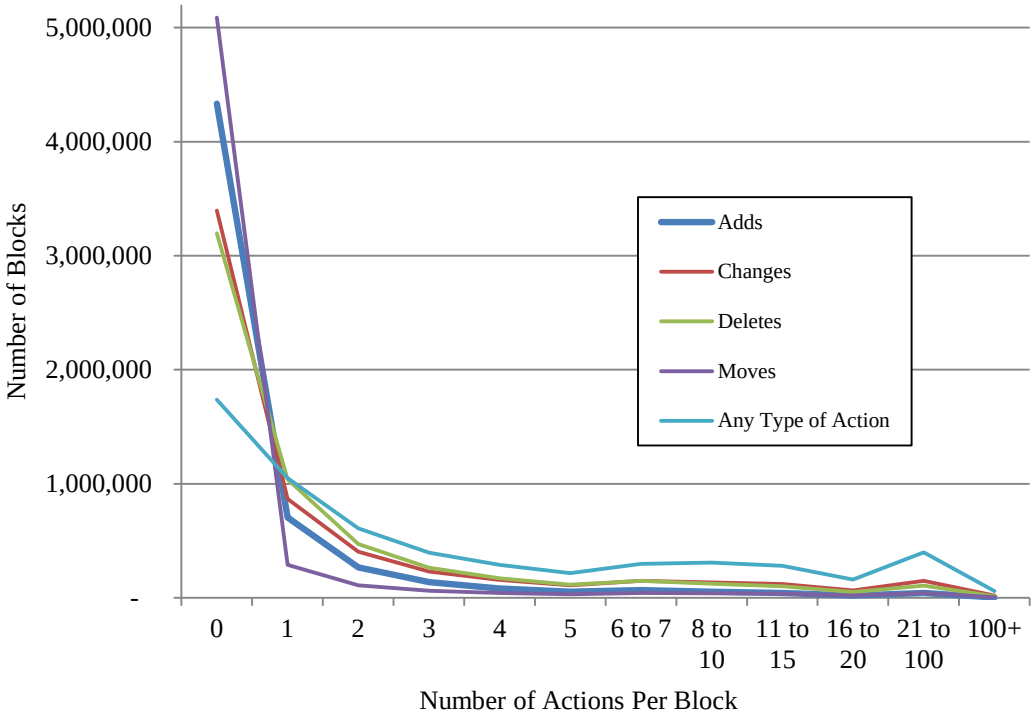

Table 4. 2010 CPEX AC Targeting: Odds Ratios from 11 TAC Models, Predicting Different Amounts and Types of Actions at the Block Level*

Dependent Variable	Model										
	10+ Adds	5+ Adds	1+ Adds	5+ Deletes	1+ Deletes	10+ Changes	5+ Changes	1+ Changes	1+ Any Action	5+ Moves	1+ Moves
Number of HUs	0.957	0.782	0.563	1.05	0.665	0.966	0.830	0.597	0.366	1.288	1.215
Number of HUs in Large Blocks (LBs)	0.723	0.607	0.464	0.688	0.488	0.779	0.670	0.497	0.242	1.011	0.958
Proportion Multi-Unit Structures	1.067	1.049	1.030	0.932	0.967	1.092	1.085	1.073	1.034	0.975	0.941
Complexity	1.766	2.394	3.845	2.061	3.962	1.731	2.255	3.923	15.31	0.986	1.064
Hispanics Present	0.999	0.946	0.866	0.931	0.801	1.448	1.201	0.933	0.779	1.493	1.081
Hispanics Present in LBs	0.562	0.493	0.437	1.068	1.179	0.006	1.909	1.562	0.934	1.640	1.337
Blacks Present	0.768	0.837	0.836	1.525	1.157	1.132	1.012	0.878	0.924	1.425	1.098
Black Present in LBs	0.447	0.435	0.588	0.924	0.911	0.000	0.932	1.191	0.100	0.857	0.867
Child Present	0.696	0.736	0.767	1.145	0.995	1.293	1.295	0.925	0.692	1.559	1.628
Child Present in LBs	0.725	0.628	0.788	1.104	1.540	NA**	1.244	1.359	2.132	2.015	2.138
No Population Change	0.881	0.929	1.109	0.665	0.738	0.896	0.795	0.906	1.084	1.063	0.897
Ratio of MAF Non-matches to Matches	0.744	0.834	0.941	0.986	1.000	0.969	0.966	0.977	0.997	0.900	0.958
LBs not Hispanic, Black, or Child Present	26.980	17.480	5.751	19.450	1.797	0.001	0.927	0.225	7.992	25.99	10.178
STD of Block Population	0.955	0.960	0.977	0.993	1.019	0.984	0.980	0.981	1.018	1.006	1.001
STD of Population in LBs	1.000	1.002	1.003	1.001	1.003	1.001	1.001	1.001	1.006	1.003	1.003
StARS Block Mismatch	0.916	0.963	0.884	1.400	1.801	0.908	1.097	1.066	1.335	0.692	0.819
Max-rescaled R Square	0.219	0.193	0.127	0.243	0.139	0.214	0.198	0.127	0.131	0.129	0.079

Table 5. 2010 CPEX AC Targeting: Summary of Outcomes from Excluding Blocks at a Gross Undercoverage Rate of About 0.5 Percent

Selection Outcomes	Model										
	10+ Adds	5+ Adds	1+ Adds	5+ Deletes	1+ Deletes	10+ Changes	5+ Changes	1+ Changes	1+ Any Action	5+ Moves	1+ Moves
Percent Gross Undercoverage	0.97	0.66	0.54	0.52	0.58	0.75	0.52	0.54	0.47	0.42	0.60
Percent Gross Overcoverage	3.44	2.71	2.42	0.65	1.00	1.40	0.88	1.74	1.65	0.64	0.73
Percent Net Coverage Error	2.48	2.05	1.87	0.12	0.43	0.65	0.36	1.20	1.18	0.22	0.12
Percent Total Error	4,41	3.38	2.96	1.17	1.58	2.14	1.39	2.28	2.13	1.06	1.33
# of HUs Excluded (1,000s)	57,539	47,652	43,056	7,023	15,136	12,789	6,794	28,476	30,839	5,292	6,991
# of Blocks Excluded (1,000s)	3,899	3,126	2,769	1,482	2,284	2,365	1,610	3,238	3,162	1,399	1,464
Percent of Blocks Excluded	67.11	53.8	47.67	25.51	39.32	40.71	27.7	55.73	54.42	24.08	25.20
Percent Adds Lost	19.46	13.35	10.98	10.51	11.62	15.08	10.41	10.93	9.54	8.54	12.11
Percent Deletes Lost	29.04	22.9	20.39	5.45	8.46	11.77	7.39	14.67	13.96	5.40	6.13
Percent Change Lost	27.47	22.21	20.08	6.48	9.60	10.01	6.08	12.55	13.00	5.19	7.20
Percent Moves Lost	29.24	25.15	24.44	5.75	10.39	9.04	5.69	17.84	17.42	3.16	5.60
Cost Reduction (in millions):											
Avg. Cost/Block	308.1	246.96	218.79	117.11	180.46	186.84	127.16	255.81	249.81	110.51	115.66
Est. Cost/Block	192.1	153.47	133.18	43.44	72.30	74.17	46.11	114.79	117.00	43.08	46.10
Random Selection	89.33	61.26	50.40	48.23	53.35	69.21	47.80	50.19	45.30	39.20	55.61

Figure 2. 2010 CPEX AC Targeting: 1+ Any Action TAC Model Cost Benefit Curves, Gross Undercoverage Rate Compared with Two Cost Measures (Average Cost/Block, Est. Cost/Block)

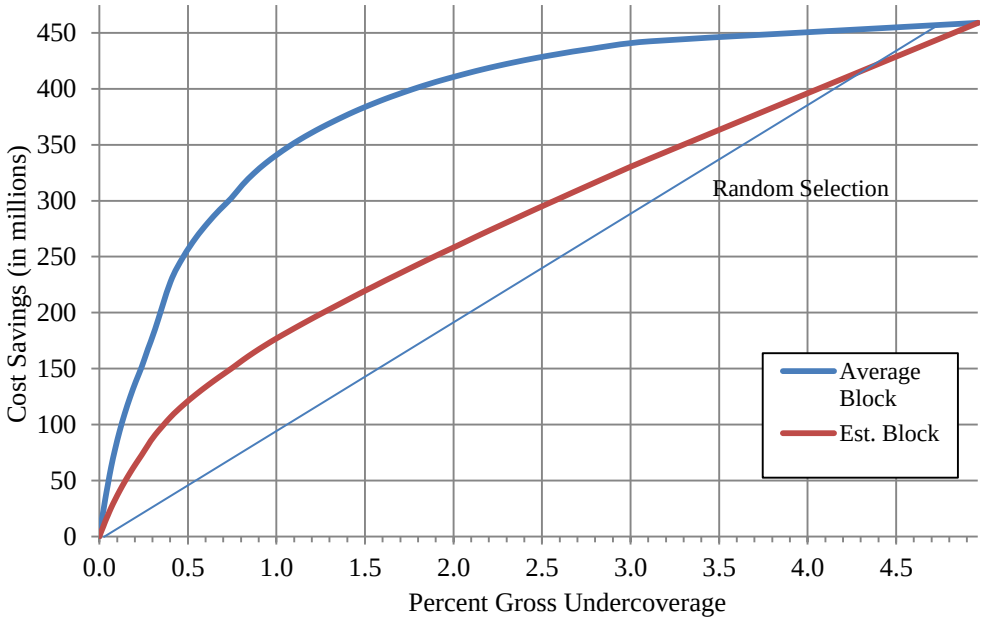

Figure 3. 2010 CPEX AC Targeting: 1+Any Action TAC Model, Cost Savings Compared with Lost Actions (Average Cost/Block)