UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001



July 31, 2012

**2010 CENSUS PLANNING MEMORANDA SERIES**

**No. 216 (Reissue)**

MEMORANDUM FOR The Distribution List

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

Subject: 2010 Census Address Canvassing Targeting and Cost Reduction Evaluation Report

Attached is the revised 2010 Census Address Canvassing Targeting and Cost Reduction Evaluation Report. This report is being reissued to improve the readability of the charts. The previous version was issued in black and white making it difficult to interpret the charts.

If you have any questions about this document, please contact Kevin Shaw at (301) 763-1851.

Attachment

# 2010 Census Address Canvassing Targeting and Cost Reduction Evaluation Report

Final

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

John L. Boies, Kevin M. Shaw, Jonathan P. Holland
Decennial Statistical Studies Division

This page intentionally left blank

Table of Contents

List of Tables

List of Figures and Exhibits

This page intentionally left blank

**Executive Summary**

The purpose of this national level study is to examine the cost reduction and coverage impact that would result from conducting a targeted Address Canvassing operation. This study was conducted with 2009 vintage data, to simulate a 2010 Targeted Address Canvassing operation. As examined here, TAC is defined and studied as a way to identify geographic areas (in this analysis, Census 2000 current blocks) using statistical models to identify the most cost beneficial updates to the Master Address File. Statistical modeling is one approach to targeting blocks for Address Canvassing that may result in cost savings. The research presented here indicates that there is substantial potential for cost reduction using a model-based TAC approach, and make six recommendations based on our experiences conducting this research.

A nationwide Address Canvassing operation was used in 2009 to update the Master Address File in preparation for the 2010 Census. Other updating procedures involved using Administrative Record data from the United States Postal Service in the form of the Delivery Sequence File and address files from the Local Update of Census Addresses program. The workload for the 2010 Address Canvassing operation totaled over 150 million address records (mostly Living Quarters). While this was a valuable endeavor, it was also very expensive; with the Census Bureau incurring approximately 459 million dollars in execution costs (lister training, lister salary and benefits, lister mileage) and about an additional 400 million dollars in other costs (materials/equipment, infrastructure and contract costs, etc.).

Motivations for this study include:

1) Research on prior address/block canvassing operations (e.g., Dixon et al 2008) indicated that for many parts of the country the Master Address File did not need to be updated by Address Canvassing.

2) The 2010 AC operation at 459 million dollars was the second most expensive single item expenditure in the 2010 Census behind the Nonresponse Followup operation – this fact makes Address Canvassing a good candidate for cost reduction efforts. However, any cost reduction effort, will likely result in some coverage degradation.

3) TAC research is a central component of the Geographic Support System Initiative and the 2020 Decennial Census planning process.

In this exploration of the feasibility of using TAC for the 2020 Census two operational study questions were examined:

1) Is it possible to model the outcomes of the 2010 Address Canvassing operation based on a priori data?

2) Once some basic models predicting Address Canvassing outcomes are developed, can these statistical models be turned into useful tools to allocate Address Canvassing resources?

These questions were addressed using a scenario based micro-simulation using the 2010 Address Canvassing operation. The scenario examined was: "What if the Census Bureau had used statistical models to select blocks for canvassing in 2009?" Binomial logistic regression in SAS™ was the primary analysis tool used here.

The data used in our micro-simulation came from two sources. Address Canvassing outcome data are from the 2010 Census Address Frame Combination file. Demographic data and some address characteristic data are from the 2000 to 2008 Statistical Administrative Records System files. Two data sources were aggregated and merged by Census 2000 current block identifiers.

The study universe included only records that were in the Group Quarters Validation file with a valid Address Canvassing action code. This file contains the action codes resulting from the 2010 Census Address Canvassing operation after Geography Division processing to ready the Master Address File for 2010 Census operations. The study universe contains 5,809,915 Census 2000 current blocks (encompassing data from 155,167,767 address records; the vast majority of which are housing units).

Two types of independent variables were used in the statistical modeling – Block physical structure measures such as number of housing units, proportion of housing units in multi-unit structures, and residential complexity; and block social structure measures such as the variables that defined whether blocks had Black residents, Hispanic residents or children present.

The dependent variables in the analysis are based on the action codes from the 2010 Census Address Canvassing operation. The actions analyzed here are:

1) Adds– New addresses added to the Master Address File
2) Deletes – Does not exist in the block or not valid for future operations
3) Moves – Adds found to match a deleted Master Address File address from another block
4) Changes – Other types of updates to a Master Address File address
5) Any Action of Any Kind– Any of the above actions

Several findings regarding these action code distributions at the block level are of interest:

1) The most common number of actions per block is 0.
2) The second most common number of actions is 1.
3) More than 5,000,000 blocks have no moves (≅88 percent).
4) Over 4,000,000 blocks have no adds (≅75 percent).
5) There are no deletes or changes in over 3,000,000 blocks (≅75 percent).
6) Only about 1,700,000 blocks have no actions of any kind (≅30 percent).

These findings indicate that if adds were the only priority, a substantial proportion of blocks may not need to be canvassed (i.e., if the Census Bureau had perfect prediction capability, the Census Bureau might only have to canvass 25 percent of blocks to capture all the adds). However, any TAC procedure that excludes a substantial proportion of blocks will come at the cost of some type of coverage degradation.

To answer the two study questions, 11 logistic regression models were estimated. The modeled outcomes included blocks with more than one add action and more then one of any type of action (add, delete, move, or change). All models used the same two groups of independent variables. Block level predicted probabilities of the outcome variables were saved from each model.

There are four primary limitations to this study:

1) The purpose of the modeling process presented here is to demonstrate that useful models can be estimated, not to generate the "Best" or "Final" models.
2) This is not a theory testing exercise; the adjudication of competing theories of residential change is not part of this research.
3) This is a cross-sectional analysis, a snapshot of one time period. The value of the data for decision making declines over time.
4) The study is limited to examining Address Canvassing action codes, the updating of geographic features is not addressed here.

The modeling procedure produced a number of interesting findings. Among these findings is that the more housing units in a block, once controls for other variables are introduced, the lower the odds of there being most types of Address Canvassing outcomes. Also of note is that blocks with higher proportions of housing units in multi-unit buildings have much higher odds of adds but lower odds of deletes.

Using the saved predicted probabilities from the statistical models, the authors developed an EXCEL™ spreadsheet tool that allows the comparison of the Address Canvassing Cost and Benefit outcomes of different statistical models and provides a decision tool to assess different Address Canvassing strategies. The Address Canvassing Cost/Benefit tool indicates that there is significant potential cost reduction from a model-based TAC operation. For example, the 1+ Any Action model (logistic regression model predicting whether a block has one or more of any Address Canvassing action) produces potential savings of nearly 249 million dollars, at the cost of a gross undercoverage rate of 0.47 percent. This estimate assumes an Address Canvassing cost per block of about 79 dollars (≈459 million dollars/≈5.8 million blocks). An alternative per-block cost that takes into account the residential complexity of each block suggested an estimated savings are about 117 million dollars.

The data indicate that the answer to the first operational study question – "Is it possible to model the outcomes of the 2010 Address Canvassing operation based on a priori data?" is yes. The answer to the second question – "Once some basic models predicting Address Canvassing outcomes are developed, can these statistical models be turned into useful tools to allocate Address Canvassing resources?" is also yes.

Based on the author's collective experiences and observations over the lifecycle of this research, six recommendations are proposed:

1) **Dedicated Team:** Identify a full-time team of Statisticians, Information Technology specialists, and Geographers from the decennial directorate and other parts of the Census Bureau to promptly begin a program of research and testing that is consistent with the approach and recommendations presented here.

2) **New Data Sources:** Acquire and analyze new data sources.

3) **Database Construction:** Collect, construct, and maintain a nationwide database integrating geographic, address and demographic data from multiple sources.

4) **Data Modeling, Development, and Verification:** Develop, refine, and field test TAC statistical models using existing benchmark and new data.

5) **Cost Modeling:** Acquire the necessary data to develop and test Address Canvassing cost models at multiple levels of analysis.

6) **New Ways of Clustering Census Data:** Examine ongoing efforts and explore alternative methods of grouping data into problem-specific clusters of data unrelated to extant census geocoding definitions, e.g., Address Canvassing actions cluster along streets that span multiple blocks.

## 1. Introduction

This report presents the results 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 targeting address canvassing efforts are examined by doing a micro-simulation using outcomes from the 2010 Census AC operation. The "What if" simulation question used is: "What outcomes would be different if the Census Bureau selected 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 primary statistical tool used is logistic regression. This problem-solving effort is an Address Information Micro-Simulation (AIMS).

Multiple data sources are used as inputs for this evaluation. Data from the 2010 Census AC operation aggregated from the 2010 Census Address Frame Combination file (AF COMBO) – a combination of eight decennial census databases – and Administrative Record (AR) data from the Statistical Administrative Records System (StARS) form the core of the analysis. Although most of the data analysis focused on the two previously mentioned data sources, some exploration of alternative data sources, including recent Maryland Property Tax data was done. In general, the data indicated that a modeling approach has the possibility of yielding significant cost reduction for an AC operation. This potential is sufficient to recommend that a more substantial effort be undertaken to develop improved models, and to develop methodology to test, verify, and update these models in the field. It will be critical that the use of additional data sources as inputs into the modeling procedures be explored.

## 2. Background

In the mid-1990s, the Census Bureau developed the MAF. Currently, address records on the MAF are linked to the Topologically Integrated Geographic Encoding and Referencing (TIGER) database, comprising the MAF TIGER Database (MTdb). The address data in the MTdb are primarily maintained through a semiannual update provided by the United States Postal Service's (USPS) Delivery Sequence File (DSF). Other sources of MAF maintenance include the Local Update of Census Addresses (LUCA) program, AC operations, and other canvassing/listing operations and geographic partnership programs. Additionally, the American Community Survey Time-of-Interview (ACS-TOI) and Demographic Area Address Listing (DAAL) provide updates to the MAF throughout the decade.

Census 2000 used three major approaches to improving the decennial MAF (DMAF) (Vitrano et al., 2004). Field operations, including Block Canvassing (BC), Address Listing (AL), and Update/Leave (U/L), updated the MAF for Census 2000. The operations were specific to the Type of Enumeration Area (TEA). Each TEA designates an area in the United States where particular enumeration methodologies will be used for census data collection. For example, in areas with successful mail delivery by the USPS and low rates of self-enumeration problems, Mailout/Mailback (MO/MB) is the most cost effective approach.

In Census 2000, field staff (listers and enumerators) conducted a 100- percent BC operation of predominantly city-style addressed areas later used for the MO/MB TEA. This operation had field staff take a list of addresses from the MAF to the field. The "dependent listing[1]" of addresses contained mostly city-style addresses with a house number, street name, and ZIP Code. Field staff could verify, add, delete, correct addresses, or make geographic code (Census Bureau state, county, block) corrections to this dependent listing. The field staff contacted every third housing unit to inquire about adjacent addresses and to identify "hidden" Housing Units (HUs). BC accounted for 51 percent of the total number of blocks and included 91,612,770 addresses. Of the 3,801,560 blocks canvassed, 31 percent did not receive any updates. This outcome indicated that many city-style address areas experience little change over time.

The Census 2000 AL operation created the initial address list for areas where the questionnaires would be delivered by hand during the U/L operation. These areas contained primarily non city-style addresses and were mutually exclusive from the BC universe (Vitrano et al, 2004). Field staff created an inventory of addresses by listing all residential addresses and simultaneously adding the physical location of addresses to decennial census maps with location designations known as map spots. Field staff canvassed door-to-door while identifying mailing addresses and map spots to create an address list for, largely, rural areas. This operation added 22 million HUs to the MAF (Ruhnke, 2002). Of the added HUs, about 40 percent matched addresses identified on the September 1998 DSF, the most recent DSF update available at the time. Over 73 percent of the added addresses were complete city-style addresses.

In the 2004 Census test, field staff updated the MAF by conducting a 100- percent AC operation in Queens, New York (Dixon et al, 2008). They canvassed each block and verified addresses by examining every structure and comparing addresses on the ground with those on their maps and address registers. Investigating, including interviewing a respondent where necessary, about every structure was a change from the Census 2000 BC. Field staff could add, change, or delete addresses, and designate an address as a HU or Other Living Quarters (OLQ).[2] These changes were documented and then used to update the MTdb.

The 2010 Census Research and Development Planning Group on Coverage Improvement recommended testing a targeted AC approach in the 2004 Census Test (Vitrano, 2003). The group agreed that a successful implementation of targeting would enable the Census Bureau to capture growth and correct duplication while minimizing the use of costly field resources. However, due to budget constraints the proposal was not executed. Without early decade testing, the effect of this significant change on AC could not be adequately studied in time for implementation in the 2010 Census.

In the 2006 Census Test, field staff used Hand-Held Computers (HHCs) to verify, update, add, and delete addresses in Travis County, Texas and the Cheyenne River Indian Reservation in South Dakota. They also collected or updated map spots and captured latitude and longitude coordinates data whenever a Global Positioning System (GPS) signal was available. The AC operation required field staff to canvass assigned census blocks looking for every place where people lived (or could live) or stayed and to make contact with every structure. Field staff

[1] Field staff use a preexisting address list to canvas an area in preparation for a census or survey.
[2] For this report the referent Housing Unit (HU) is used to refer to a living quarter with an address, i.e., a record in the data bases used here with a MAFID.

compared the living quarters found in the field to data from the MAF on the HHC (Dixon et al, 2008; Schneider et al, 2006).

For Travis County, nearly 80 percent of the 208,678 addresses required only verification. However, Field Representatives made corrections for 50 percent of the 3,015 addresses in the Cheyenne River Indian Reservation. This test indicated that targeting AC activities could be extremely effective. Under a targeting model, the Travis County site would require minimal updating, while the Cheyenne River Indian Reservation would benefit from a 100- percent canvassing approach.

In the 2008 Census Dress Rehearsal (DR), the field staff conducted AC in San Joaquin County, California, and a nine-county area surrounding Fayetteville, North Carolina (Dixon 2008). The AC operation required field staff to canvass assigned census blocks, contact every structure, look for all potential living quarters, and compare these living quarters on the ground to information on the HHC. Of the 679,886 addresses canvassed, field staff verified 64 percent of the addresses. Twenty-one percent required a change in one or more parts of the address. For both sites, an address's presence on the DSF as a residential unit was a good indicator of whether or not it was verified in the AC operation. For MO/MB addresses, 57 percent were verified. In Update/Leave areas, 28 percent of the addresses were verified while 24 percent received negative actions, indicating that the address list exhibited reduced coverage for more rural areas.

The AC operation for the 2010 Census covered all of the U.S. and P.R., except select areas of Alaska and Maine. Field staff added, deleted, and made corrections using HHCs or laptops. They also updated geographic information and collected GPS map spots. The total execution cost of the 2010 Census AC field operations, including quality control, was approximately 459 million dollars[3], excluding contract-related costs (Holland, 2012).

## 3. Methodology

### 3.1 Questions to be Answered

The primary research objectives for this study are to explore the cost reduction and coverage impact of a Targeted Address Canvassing (TAC), resulting from allocating AC resources based on modeled predictions of the most cost beneficial part of the nation to canvass. Thus, this study is guided by these two questions:

1) What targeting criteria will identify dynamic and stable census blocks, and provide for an alternative approach to full-scale AC?

2) What is the impact of our targeting approach on undercoverage, overcoverage, and operational costs?

[3] This figure was used for all cost reduction simulations in this report, and includes both the approximately 443.6 million dollar AC and 10.3 million dollar Large Block AC costs detailed in the 2010 AC Assessment (2012), as well as the 5.1 million dollar "Provide OCS/HHC Technical Support" cost in Holland (2012). The sum of 459 million dollars represents the total of the 2010 Census AC "Execution" cost category in Table 11 in Holland (2012).

At this stage of the research, there is no intention to develop models that best predict AC outcomes nor to develop the procedures for the allocation of these resources. The focus is to provide a general assessment of the potential of using model-based procedures to yield cost efficient methods of managing scarce census resources. Take note that it is not the objective of this study to test competing theories of residential change, nor to build the best possible models predicting residential change. The results presented here demonstrate feasibility and suggest great value for pursuing a model-based TAC program for the 2020 Census.

As is standard practice in scientific research,[4] two, more specific, operational research questions are derived from the two general research questions initially proposed in the study plan in order to better define the research presented here:

1) Is it possible to model the outcomes of the 2010 Address Canvassing operation based on a priori data?

2) Once some basic models predicting Address Canvassing outcomes are developed, can these statistical models be turned into useful tools to allocate Address Canvassing resources?

There are three general assumptions made in the research reported here:

1) The outcome of the 2010 Census AC operation is the "ground truth" regarding census relevant housing unit and address information;
2) That the relationships here will hold relatively stable over time; and
3) The AC outcome database as delivered to the Decennial Statistical Studies Division (DSSD) for this analysis is accurate.

## 3.2. Methods

### 3.2.1. Data

This project used information from a file combining data from two sources: the 2010 Address Frame Combination (AF COMBO) file and extracts from the 2000 to 2008 Statistical StARS. The 2010 Census AF COMBO file is a database constructed by the DSSD for use by statisticians at the Census Bureau to assist in assessing the 2010 Census (see Ward, 2011, Documentation for the 2010 Census Address Frame Combination File, version #1 for detailed information describing this file). It consists of eight groups of census files merged together at the address level using the Master Address File Identification (MAFID) number or an equivalent address record identifier (e.g., customer number). These groups of decennial files are: Pre-AC (PREAC); Census Evaluations and Experiments (CEE); Reject data (REJECT); Large Block (LB); Group Quarters Validation (GQV); 2000 COMBO (OC); Enumeration Universe (ENUM); and the January 2009 American Community Survey (ACS) MAF extract (MAFX) files. The universe for all these files is the 50 U.S. states, the District of Columbia (DC), and PR. The

[4] The first operational research question extends the first study plan question by taking the criteria identified in the research and using them to model the 2010 Census AC outcomes. The independent variables used in the modeling presented also predict dynamic and stable blocks and therefore identify alternative approaches to a full-scale AC. The second operational question addresses the second study plan question directly by providing a tool for illustrating the potential outcomes of the alternative targeting approaches. The tool developed as a part of this research allows a user to determine the cost and coverage effects for a full range of targeting methods.

GQV file[5] records contained in the 2010 Census AF COMBO file are the source of the outcome variables for our research. An extract of the 2010 Census AF COMBO file, including the GQV variables, was produced using only records with a non-missing action code on the GQV action code variable with data from selected variables, summarized as counts and averages at the Census 2000 current block level.

The StARS extract files were created by the Data Integration Division (DID) per DSSD's programming specifications. The StARS database is composed of ARs 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 variables on the number of persons in Census 2000 current blocks for categories of age, ethnicity, sex, and race. Additionally, totals for a range of address quality measures (e.g., StARS addresses not matched to the MAF), number of live persons, all persons (both deceased and alive), and other counts by Census 2000 current block for 2000 through 2008 are also in the files. The nine annual files (one for each calendar year) were merged by Census 2000 current block identifiers (Federal Information Processing Standard (FIPS) state code, FIPS county, tract, and block). The universe for these data files is also the 50 U.S. states, DC, and PR.

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 (Maryland Property Tax data were also merged and aggregated to these files). The Census 2000 geography was used for this national-level research because this geography was common to all the source files. Only those records that were in the GQV files and had non-missing values on the GQV action code variable are in the MUMS file (the study universe[6]). Records with non-missing values for this variable included those sent out for the AC operation (from the CEE file with a "Y" code denoting Yes/Eligible on the variable defining the AC universe) plus those records "added" during the AC operation that were not duplicates of existing MAF addresses (New Adds).

There are 202,166,334 records (addresses with MAFIDs) in the 2010 Census AF COMBO file. The StARS files for all years (2000 through 2008) contained 5,784,862 unique records (only Census 2000 current blocks populated by at least one person according to the StARS database). After aggregating and merging these files into the MUMS file, there were 6,319,298 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) was compared to summary data from the records used by this study.

[5] The GQV file as used in the 2010 Census AF COMBO file is also known as the Initial Universe Control and Management (UC&M)/Group Quarters Validation (GQV) Universe file.
[6] Duplicate and uninhabitable addresses are included in this file.

Exhibit 1. Bitmap of Table 11.7 from the 2010 AC Assessment Report

**Table 11.7**
**The 2010 Census Address Canvassing Operation:**
**Results compared to the Census 2000 Block Canvassing operation**

| Final Address Actions | 2010 Census Address Canvassing | | Census 2000 Block Canvassing | |
|---|---|---|---|---|
| | Count* | Percent of total+ | Count* | Percent of total+ |
| Total | 156,703,156 | 100.00 | 97,894,639 | 100.00 |
| Add | 10,776,894 | 6.88 | 6,389,271 | 6.53 |
| New | 6,624,155 | 4.23 | 4,536,234 | 4.63 |
| Matches to Existing Record | 4,152,739 | 2.65 | 1,853,037 | 1.89 |
| Change | 19,608,785 | 12.51 | 2,295,168 | 2.34 |
| Move | 5,450,563 | 3.48 | 2,948,414 | 3.01 |
| Verify | 97,635,517 | 62.31 | 81,115,466 | 82.86 |
| Negative Actions | 21,143,737 | 13.49 | 4,972,041 | 5.08 |
| Does Not Exist (Double Delete only) | 15,819,921 | 10.10 | 4,452,888 | 4.55 |
| Duplicate | 4,085,556 | 2.61 | 154,869 | 0.16 |
| Nonresidential | 1,238,260 | 0.79 | 364,284 | 0.37 |
| Uninhabitable | 551,566 | 0.35 | 174,279 | 0.18 |
| Rejected Records | 1,536,094 | 0.98 | | |

*Counts and percentages are unweighted.
+Percentages may not sum to 100 due to rounding.
The Census 2000 Address Listing operation, an independent listing not depicted above, added 23,271,819 new Stateside and Puerto Rico records to the MTdb. Adds from Address Listing combined with Block Canvassing represent 25 percent of the total actions to update records on the MTdb.
Verify in this table means that the address was found in AC and there were no changes to the address component of the record.
Negative Actions and Uninhabitable in this table is the same as "Delete" category in Burcham, 2002.
Sources: GQV Extract Files, as defined by the matched MAFSRC and ACTION operation variables, GEO AC Listed Records Tally File, Ruhnke, 2002, and Burcham, 2002.

The comparisons shown in Table 1 indicate a nearly exact match between the distribution of action codes shown above in Table 11.7 from the 2010 Census AC Assessment Report and the action codes used in this study.

**Table 1. 2010 CPEX AC Targeting: Study Universe[7]**

| 2010 Census Post-AC Action Code for Selected Actions | Addresses from TAC Study Universe with Action Codes Used in this Study* | 2010 AC Assessment (Table 11.7)* |
|---|---|---|
| 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 |

*Counts of addresses with action codes are unweighted.
**Verified Addresses are included for reference.
Source: TAC MUMS File.

[7]In this report the delete action code was analyzed independent of other negative actions. Combining all the negative actions (i.e., deletes, duplicates, nonresidential, uninhabitable) together for analysis is also common. When conceptualizing the work for this evaluation, the focus was on the most prevalent unique AC actions (i.e., adds, deletes, changes, and moves). Later iterations of this research will include a broader range of AC outcomes, including negative actions not used here. However, including these actions in the analyses for this report will have only a marginal effect on the predicted coverage rates presented here.

The differences in the number of action codes germane to this report 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 total and the totals in the study universe are likely the result of using Census 2000 current blocks rather than the 2010 Census collection blocks used by the AC operation, and records that are in the GQV and CEE files but were not assigned an action code.

### 3.2.2. The Independent Variables

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.[8] Primarily the age, sex, race, and ethnic composition of neighborhoods are suggested as all contributing to residential changes over time. The extant literature, however, does not detail the expected outcomes of any specific demographic makeup on residential change. This necessitates that the researcher artfully tests demographic variables that cover a wide range of different demographic compositions in a community. Given that address-level changes within blocks are being modeled, it's expected that measures of residential structure of blocks may affect the outcome of any 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, an inventory of variables from our two primary data sources was used to test in the modeling process. Substantial testing and evaluation resulted in 11 variables being viable for use in the models presented here.[9] The variables selected to include here are not meant to produce the "best" model of residential change nor do they represent any attempt at testing competing theories of residential change. The variables used here generate models that are more than sufficient to meet the goals of this study.

The StARS file provided these social structural measures:

- Blocks with a population that is more than two percent[10] Black were scored 1, 0 otherwise—Blacks present blocks;
- Blocks with more than two percent of the people under the age of 19 were scored 1, 0 otherwise—Children present blocks;
- Blocks with more than two percent Hispanic origin population were scored 1, 0 otherwise—Hispanics present blocks;

---

[8] Because the timing of this research did not allow adequate time for Internal Revenue Service (IRS) approval for use of IRS-reported income data was not included in this research.

[9] Since the testing of competing theories and the development of a "best" model of AC outcomes are not primary objectives of this study, few resources were expended on these goals. Our criteria for which variables to include were primarily based on contribution to overall model fit. SAS's Max-rescaled adjusted $R^2$ was used to measure model fit. Model diagnostics were also used in this assessment, e.g., multi-collinearity.

[10] The 2 percent cutoffs for the Black, Hispanic, and Child block variables were determined through an iterative procedure with an upper bound of 40 percent and a lower bound of zero percent.

- Blocks were scored a 1 if there was less than two percent change in the mean population from 2007 through 2008 compared to the mean for 2004 to 2008, 0 otherwise—No Block Population Change; and
- The standard deviation of the yearly population of each block for 2005-2008 was also used—STD of Population.

The physical structure of the block was measured with six variables:

- the number of HUs (addresses) in a block;
- the proportion of HUs in a block that are in multi-unit structures of any size;
- the ratio of the number of addresses in a block that are in the StARS database that do not match to the MAF to the number that do have MAF matches;
- whether a block from the 2000, 2001, or 2002 StARS data files that matches to the COMBO file also matches to a Block from the 2008 StARS files (1=yes, 0 otherwise);
- if a block has more than 600 HUs, it was considered a "Large Block" for the purposes of this report and given a code of 1, otherwise it was coded a 0; and[11]
- a measure of residential complexity. This is an index based on the length in kilometers of the distance between unique addresses in a block and the number of HUs in a block. Larger values indicate greater complexity[12].

Five interaction terms using the large block variable are also used here. 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. Some other interactions were tested but did not contribute substantially to the modeling outcomes. If additional research is pursued, additional interaction terms will be examined.

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. While this was not unexpected given the nature of the data, because logistic regression is a very robust procedure, there is no expectation that this issue will significantly effect the interpretation of the data (Hosmer and Lemeshow, 2000).

[11] In other census documents and operations "Large Blocks" are defined differently, e.g., in the 2010 Large Block Address Canvassing operation large blocks were defined as having more than 1,000 HUs or 2,000 HUs (Chaar and Marquette, 2012, p16).

[12] Residential complexity measure equation:

$$\tau = \omega + 4 \times \gamma$$

$$\gamma = 3 + \sum_{i=1}^{\alpha} \sum_{j=1}^{\beta_i} \delta_{ij}$$

t = residential complexity measure
w = pre-AC Housing Unit Count.
$d_{ij}$ = distance (km) between map spots on the same street with the same parity.
a = number of street parity combinations
$b_i$ = one less than the number of housing units with map spots on street segment parity combination i.

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

| Independent Variable | Minimum | Maximum | Mean | Standard Deviation | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| Number of HUs/Block[13] | 1 | 13,138 | 26.707 | 67.425 | 20.149 | 1,367.920 |
| Proportion of Multi-Units/Block | 0 | 1 | 0.071 | 0.2 | 3.189 | 9.598 |
| Hispanic Population >2 Percent of Block Count | 0 | 1 | 0.402 | 0.49 | 0.399 | -1.841 |
| Black Population>2 Percent of Block Count | 0 | 1 | 0.392 | 0.488 | 0.488 | -1.804 |
| Child Population>2 Percent of Block Count | 0 | 1 | 0.719 | 0.45 | -0.974 | -1.051 |
| Population Change <2 Percent from 2004-2008 | 0 | 1 | 0.239 | 0.426 | 1.225 | -0.500 |
| Ratio of HUs in Block with No MAF Match | 0 | 628.5 | 0.081 | 1.143 | 199.270 | 67,502.520 |
| Blocks with more than 600 Units | 0 | 1 | 0.002 | 0.045 | 22.039 | 483.723 |
| Standard Deviation of Population 2005-2008 | 0 | 9,137.94 | 4.413 | 23.059 | 88.837 | 17,936.640 |
| StARS Block Mismatch | 0 | 1 | 0.028 | 0.165 | 5.724 | 30.764 |
| Residential Complexity | 13 | 64,478.62 | 49.881 | 101.389 | 83.183 | 38,137.074 |
| Large Block Interaction Terms | | | | | | |
| Number of HUs | 0 | 13,138 | 2.003 | 50.827 | 44.593 | 4,299.510 |
| Hispanic Present Blocks | 0 | 1 | 0.002 | 0.044 | 22.550 | 506.484 |
| Black Present Blocks | 0 | 1 | 0.002 | 0.044 | 22.882 | 521.597 |
| Children Present Blocks | 0 | 1 | 0.002 | 0.044 | 22.391 | 499.335 |
| STD of Block Population | 0 | 9,137.94 | 0.391 | 19.751 | 135.518 | 33,063.780 |

Source: SAS Modeling Output.

### 3.2.3. The Dependent Variables

The primary purpose of this research is to produce models that will predict which decennial census blocks (for this study Census 2000 current blocks) should be included in an AC operation. Simple dichotomous yes/no measures of AC outcomes rather than continuous measures are used in order to produce easy-to-interpret predictions of the likelihood of a block being included in an AC operation. Continuous measures of AC outcomes, e.g., number of adds in a block, would likely be better for theory-building purposes. However, the resulting analysis would be less easily translatable into measures useful for policy making than simple dichotomous outcome measures. 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.

[13] There are 8,262,363 Census 2000 tabulation blocks (including water blocks) in the 50 states, DC, and PR. Only those with at least one 2010 AC action code and thus, an address record, are included in the study. Blocks from StARS data contain at least one AR-based person for any given year for 2000 through 2008.

The research presented here focuses on four AC action outcome codes: adds that did not duplicate existing MAF records, double deletes, changes, and moves; all from the GQV files (the GQV files contain the post-AC outcomes, applied to the MAF by the Geography Division (GEO). These outcomes are the final results of the entire 2010 Census AC operation. Nearly all addresses in the GQV files received an action code from Address Canvassing; the ones of interest here are "A" for new adds, "D" for double deletes, "C" for address changes, "M" for moves, and "K" for verified addresses. These codes are used in this analysis because they have the most significant potential impact on census coverage outcomes. 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. 2010 CPEX AC Targeting:**
**Summary Statistics of GQV Action Codes at the Block Level (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 |

*Not used in any models.
Source: TAC MUMS File.

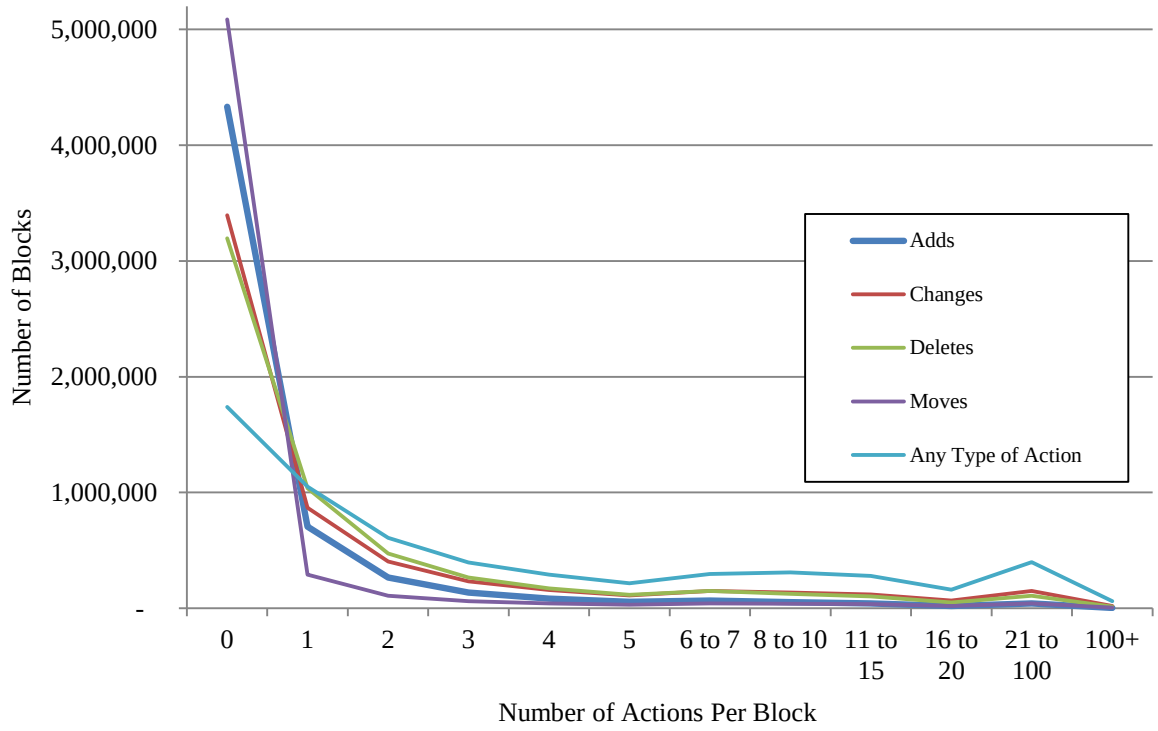
The mean number of moves and new adds per block are both near one with similar standard deviations, about 9 and 8 respectively. The distributions for both of these variables are narrower than any of the other action codes. Changes are the most common individual action with an average of over 3 per block. This is not unexpected since the actions included in the "C" category range from changing a unit designation from "A" to "1" to changing a street name and street directional. Similarly, deletes have a direct effect on gross overcoverage, and were therefore also an area of interest.[14]

Some questions may exist regarding the distribution of actions across blocks, e.g., do all blocks contain actions other than verifications ("K"s), or are most actions concentrated in just a small number of blocks? This distribution has significant implications for the potential of using modeling procedures to allocate AC resources. For example, if actions are uniformly distributed across blocks then all blocks should be given equal priority for AC resources (the standard deviations in Table 3 indicate that is not the case). On the other hand, if all or most actions are concentrated in just a few blocks—often called dynamic blocks—then targeting which blocks to include in an AC operation would easily offer substantial savings (the data in Table 3 also show that this is not the case).

[14] Undiscovered deletes may result in census questionnaires being mailed to uninhabited addresses and subsequent costly Nonresponse Followup (NRFU) visits.

Figure 1 indicates that, for some action codes, the most common number per block is zero with the next most common value being one per block. More than 5 million blocks have no moves (about 88 percent) and over 4 million blocks have no adds (about 75 percent). Deletes or changes are absent from over 3 million blocks (about 56 percent). It is very important to note 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 Table 4 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 at, 4,771 and 6,571 respectively.

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




Source: TAC MUMS File.

**Table 4. 2010 CPEX AC Targeting: Action Codes by Block**

| Actions Per Block* | Adds | Changes | Deletes | Moves | Any Type |
|---|---|---|---|---|---|
| 0 | 4,332,894 (74.6 %) | 3,395,056 (58.4 %) | 3,196,563 (55.0 %) | 5,087,206 (87.6 %) | 1,738,479 (29.9 %) |
| 1 | 705,581 (12.1 %) | 867,264 (14.9 %) | 1,037,931 (17.9 %) | 290,540 (5.0 %) | 1,050,932 (18.1 %) |
| 2 | 266,860 (4.6 %) | 404,333 (7.0 %) | 472,545 (8.1 %) | 109,697 (1.9) % | 609,001 (10.5 %) |
| 3 | 136,910 (2.4 %) | 232,015 (4.0 %) | 264,880 (4.6 %) | 61,828 (1.1 %) | 396,026 (6.8 %) |
| 4 | 82,946 (1.4 %) | 157,426 (2.7 %) | 171,137 (2.9 %) | 42,862 (0.7 %) | 289,636 (5.0 %) |
| 5 | 54,650 (0.9 %) | 110,963 (1.9 %) | 116,500 (2.0 %) | 30,250 (0.5 %) | 216,821 (3.7 %) |
| 6 to 7 | 67,310 (1.2 %) | 150,455 (2.6 %) | 150,261 (2.6 %) | 43,306 (0.7 %) | 297,802 (5.1 %) |
| 8 to 10 | 54,017 (0.9 %) | 135,045 (2.3 %) | 125,345 (2.2 %) | 40,274 (0.7 %) | 310,194 (5.3 %) |
| 11 to 15 | 42,036 (0.7 %) | 120,674 (2.1 %) | 101,752 (1.8 %) | 35,952 (0.6 %) | 280,533 (4.8 %) |
| 16 to 20 | 20,629 (0.4 %) | 66,067 (1.1 %) | 50,541 (0.9 %) | 19,374 (0.3 %) | 160,901 (2.8 %) |
| 21 to 100 | 41,311 (0.7 %) | 149,745 (2.6 %) | 107,290 (1.8 %) | 42,055 (0.7 %) | 398,741 (6.9 %) |
| 100+ | 4,771 (0.1 %) | 20,872 (0.4 %) | 15,170 (0.3 %) | 6,571 (0.1 %) | 60,849 (1.0 %) |
| Total | 5,809,915 (100.0 %) | 5,809,915 (100.0 %) | 5,809,915 (100.0 %) | 5,809,915 (100.0 %) | 5,809,915 (100.0 %) |

*Note this column represents any combination of actions, e.g., 2 adds plus 100 moves or 51 of each.
Source: TAC MUMS File.

These distributions have several implications for 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 logistic regression 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 measures of AC outcomes. Three measures of new adds are used: blocks with one or more adds, blocks with 5 or more adds, and blocks with 10 or more adds. Deletes were coded into two variables, one or more deletes per block and five or more deletes per block. Changes were coded into one or more, five or more, and ten or more. One or more adds, deletes, moves, or changes in a block measures the presence of any action. The data in Table 5 indicate no obvious problems with the distributions of any of the variables that might affect our choice of logistic regression for our analysis.

**Table 5. 2010 CPEX AC Targeting: Summary Statistics for the 11 Dependent Variables (N=5,809,915)**

| Dependent Variable | Sum of Blocks | Mean | Standard Deviation |
|---|---|---|---|
| 1+ Adds | 1,477,021 | 0.254 | 0.435 |
| 5+ Adds | 284,724 | 0.049 | 0.216 |
| 10+ Adds | 122,698 | 0.021 | 0.144 |
| 1+ Deletes | 2,613,352 | 0.449 | 0.498 |
| 5+ Deletes | 666,859 | 0.115 | 0.319 |
| 1+ Changes | 2,414,859 | 0.416 | 0.493 |
| 5+ Changes | 753,821 | 0.129 | 0.336 |
| 10+ Changes | 394,565 | 0.068 | 0.252 |
| 1+ Any Action | 4,071,436 | 0.701 | 0.458 |
| 1+ Moves | 722,709 | 0.124 | 0.330 |
| 5+ Moves | 217,782 | 0.038 | 0.189 |

Source: SAS Modeling Output.

### 3.2.4. Statistical Methodology

To answer the first question — "Is it possible to model the outcomes of the 2010 Address Canvassing operation based on a priori data?" — the research focused on building a number of basic models that were 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, 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 the coefficient 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). No exhaustive or systematic search for the best dependent measures was attempted. The ones chosen for presentation here simply represent examples of the range of reasonable measures tested. For each of the 11 models predicted probabilities were saved in a SAS system file.

For most research using statistical modeling, the data are from a sample, representative or convenience. However, the data used here are from a census and thus the inferential statistics-based evaluation measures like Chi Square tests, Wald statistics, and similar tools are not applicable. The estimates presented here are "true" population parameters rather than sample-based estimates. One standard output that SAS provides from PROC LOGISTIC is the odds ratio for each independent variable. These ratios are the antilog of the estimated logit coefficients (Hosmer and Lemeshow, 2000).

The odds ratio assesses the risk of an event conditional on the presence of a factor (an exposure to a treatment or a characteristic like being male). This ratio is a relative measure of risk, telling us how much more likely it is that someone who is exposed to the factor under study will develop the outcome as compared to someone who is not exposed. Odds are a way of presenting these probabilities. Specifically, the odds of an event happening is the probability that the event will happen divided by the probability that the event will not happen (www.blackwellpublishing.com/specialarticles/jcn_10_268.pdf).

Odds ratios offer the convenient property of being above 1 if the effect of an independent variable is positive and below 1 if the effect is negative. The reported odds ratio estimate is the change in the odds of an outcome for each change in one level of the respective independent variable. The amount the odds ratio is above or below 1 is the proportionate change in the odds of an event resulting from a one unit increment in the independent variable, e.g., an odds ratio of 0.75 represents a 0.25 decrease in the odds of an event for each one unit increase in an independent variable. An odds ratio of 21 indicates that a one unit increase in an independent variable increases the odds of the occurrence 21 times.[15]

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.[16]

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 regression models to describe the potential savings and costs of selecting some blocks for canvassing over others. PROC LOGISTIC produces a predicted probability of each block being a "1" or positive outcome based on the specific values of the independent variables for that block. These predicted probabilities can then be used to allocate AC resources to blocks with the higher probabilities of producing useful AC outcomes. This would be the ideal method to prioritize work given a pre-existing budget for an AC operation. However, as indicated by Figure 1 and Table 4, the distribution of useful AC outcomes is such that dropping any substantial number of blocks from the AC operation will have some negative affect on MAF coverage.

[15] Here is an example (derived from a Wikipedia entry) of odds ratios:
A study was done of 2,000 people, 1,000 each males and females, comparing their favorite food. 900 males preferred fried chicken to pizza, 100 males did not. Only 200 females preferred fried chicken.
-The odds of males preferring fried chicken to pizza is 900:100 or 9:1.
-The odds of females preferring fried chicken to pizza is 200:800 or 1:4.
The odds ratio of men preferring chicken over pizza compared to women is 9:1/1:4 or 9/(1/4). A ratio of male odds to female odds of 36.

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

An AC Cost/Benefit (AC/CB) spreadsheet tool was developed to assist in the evaluation of the cost reduction compared with coverage tradeoffs using the modeling results to allocate AC resources. Here is a step-by-step description of the tool:

1) The individual block predicted probabilities are rounded to the nearest percentile.
2) Evaluation variables including number of HUs, action outcomes, per-block cost estimate, and several block characteristics were then summed by these predicted probability categories.
3) Cumulative sums of the evaluation variables were then calculated (See Appendix A for sample screenshot of AC/CB spreadsheet tool).

This AC/CB tool can then be used to estimate the approximate cost reduction a specific selection of blocks may have made in the 2010 Census AC operation, and detail the coverage trade-off resulting from the lost adds, moves, or deletes. The simulated cost reduction and concurrent coverage degradation is a function of which model's predicted probabilities is chosen for selecting blocks. The results section explains how this operates in practice.

### 3.3. Costs

The cost for this study was entirely incurred by staff at Census Bureau Headquarters (HQ). This study spanned a period of approximately two years, with an estimated cost of about 700,000 dollars. This amount accounts for six employees, including overheads, who worked on the study in some capacity over the project lifecycle. For two employees, at different points in time, this project was their only assignment; while other employees had competing priorities while working on the project. With the amount of data necessary to conduct these types of data modeling and micro-simulation analyses, a small amount of money was estimated to cover the additional hard disk storage capacity necessary.

## 4. Limitations

There are four primary limitations to this study:

1) The purpose of the modeling process presented here is to demonstrate that useful models can be estimated, not to generate the "Best" or "Final" models.
2) This is not a theory testing exercise. The report does not adjudicate competing theories predicting residential change.
3) This research is based on cross-sectional data, i.e., the 2010 Census AC Operation. This means that there is no certainty that conclusions reached based on these data will remain valid in the future. While the results will be of great assistance to research and planning purposes, the value and validity of the data will decline as time passes.
4) The study is limited to examining AC action codes, the updating of geographic features is not addressed here.

## 5. Results

### 5.1 Addressing Operational Question 1: Data Modeling

The results shown in Table 6 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 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. Additional research should improve the fit of the models, and therefore the quality of the model predictions.

The odd-ratios shown in Table 6 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 6. 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 |

[*]Odds ratios are calculated at the mean for continuous variables, at the value of 1 for 0/1 variables.
[**]Not Available
Source: SAS Modeling Ouptut.

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," the other ratio is for those with 600 or fewer HUs.

The reported odd-ratios for the block physical structure variables indicate some counter intuitive 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 one. In the case of large blocks the odds ratios are above one only for the five or more 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.[17] 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 counter intuitive 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 (0.071) the odds ratios are positive for all of the Adds models, 1.067 , 1.049 , and 1.030 for the 10+ Adds, 5+ Adds, and 1+ Adds models respectively, as well as for the Changes and 1+ Any Action models (1.092, 1.085, and 1.073 for 10+ Changes, 5+ Changes, and 1+ Changes models respectively, and 1.034 for the 1+ Any Action model). For the two Deletes and the two Moves models the odd-ratios are less than one and 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, 1.064 for the 1+ Moves and 0.986 for the 5+ Moves models. In order to test the robustness of the effects of this residential complexity measure, the measure was decomposed into its unweighted components, the number of pre-AC addresses and the distance between structures, in a block and substituted these variables into the models. The predicted outcomes did not change measurably, however, some of the odds-ratios for the demographic variables weakened. Additionally, several of the models could not be estimated using the SAS logistic regression algorithm without substantial changes to the modeling parameters. Hence, the residential complexity measure was sufficient for the purposes of this report.

The effect for the large block measure (large blocks that are less than 2 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 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

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

complexity measure probably account for the counter intuitive effect of the number of HUs. This observed change in a relationship between two variables resulting from adding additional variables to the analysis is a good example of a statistical phenomenon known as suppression (Pearl, 2000).

The ratio of MAF non-matched addresses to MAF matched addresses in a block also has mostly negative effects (although relatively weak) 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, and, in the case of the 1+ Deletes model the value is calculated as 1.000. The ratios are near 1 for the 1+ Deletes and 1+ Any Action models. The StARS block mismatch measure has mostly weak and negative effects except for the 5+ Deletes, 1+ Deletes, and 1+ Any Action models with odds ratios of 1.400, 1.801, and 1.335 respectively.

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 substantially 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 persons of Hispanic origin present, the odds ratios are lower than for smaller blocks for the Adds models: 0.563 compared with 0.999 for the 10+ Adds model, 0.493 compared with 0.946 for the 5+ Adds model, and 0.437 compared with 0.866 for the 1+ Adds model. The pattern is similar for the two percent Black blocks. In the case of the large blocks with Children the effects are weaker than in smaller blocks for two of the Adds models, 0.725 compared with 0.696 and 0.788 compared with 0.767 for the 10+ Adds and 1+ Adds models. However, the effect is somewhat stronger for 5+ Adds model, 0.628 compared with 0.736.

The odds ratios for the large blocks with more than two percent Hispanic population are greater than 1 for the 2+ Deletes models, 1.068 and 1.179 for the 5+ Deletes and 1+ Deletes models respectively. 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 2+ 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 2+ Moves models. In these latter models the magnitude of the effects is noticeably larger for large blocks compared with smaller blocks (2.015 compared with 1.559 and 2.138 compared with 1.628 in the 5+ Moves and 1+ Moves models respectively). 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 change 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 universe.

For the population variables, no population change and standard deviation of the populations 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, odds ratios of 0.665 and 0.738 respectively. The odds ratios for the standard deviation of the population, for both large and smaller blocks, are near 1.000 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 (odds ratio of 0.960).

There are five prominent outcomes from this initial attempt at modeling AC outcomes:

1) More HUs in a block, once other variables are controlled, reduces the odds of AC outcomes.
2) Blocks with higher proportions of HUs in multi-unit buildings have higher odds of adds but lower odds of deletes.
3) More complex blocks have much higher odds of having AC outcomes.
4) Blocks with more than 2 percent Hispanics, children, or Blacks present reduces the odds of adds.
5) Large blocks with few children, Hispanics, or Blacks present have substantially larger odds of most types of outcomes.

Using additional data sources and further refining of the statistical modeling, will likely result in at least some of these conclusions changing, but the research presented here does provide a foundation for continued research.

## 5.2 Addressing Question 2: A Useful Management Tool

The predicted probabilities produced by the models just described to estimate changes in the outcomes of the 2010 Census AC Operation had these models been used to select blocks in 2009 are used to create an prototype management tool for TAC management. The SAS logistic regression procedure uses the following equation to calculate the predicted probabilities used here (See Appendix B for the estimated coefficients used in the calculation for the 1+ Any Action Model):

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

Where:

$P$ = predicted probability

$a$ = estimated intercept

$b$ = estimated coefficient

$X$ = value of variable x

Table 7 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 in this sample scenario was a gross undercoverage of no more than about 0.5 percent resulting from excluding some blocks from the operation. All the blocks at or below this predicted probability cut-off point corresponding to a gross undercoverage rate of about 0.5 percent were used to calculate the estimates in Table 7.[18] 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 7 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 a selected group of blocks.[19]

[18] Because the aggregations were done by percentile the actual gross undercoverage cutoff values are somewhat above or below the stated 0.5 percent value.

[19] For this study, gross undercoverage is the "Lost Adds"/All Positive AC actions (including verifications, 133,471,779 HUs at the time of this writing(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. Only "Lost Deletes" are included in the gross overcoverage calculations, since the duplicate, nonresidential and uninhabitable AC actions (in total about 5.7 million HUs) were not the focus of this research.

**Table 7. 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 |

Source: TAC MUMS File.

The data shown in Table 7 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 dollars at the price of a gross undercoverage rate of 0.47 percent and a loss of about 31 million HUs. The 250 million dollar estimate assumes that all blocks cost the same to canvass, in this case about 79 dollars per block. However, it is very unlikely that all blocks require the same amount of resources, e.g., staff time, mileage expenses.

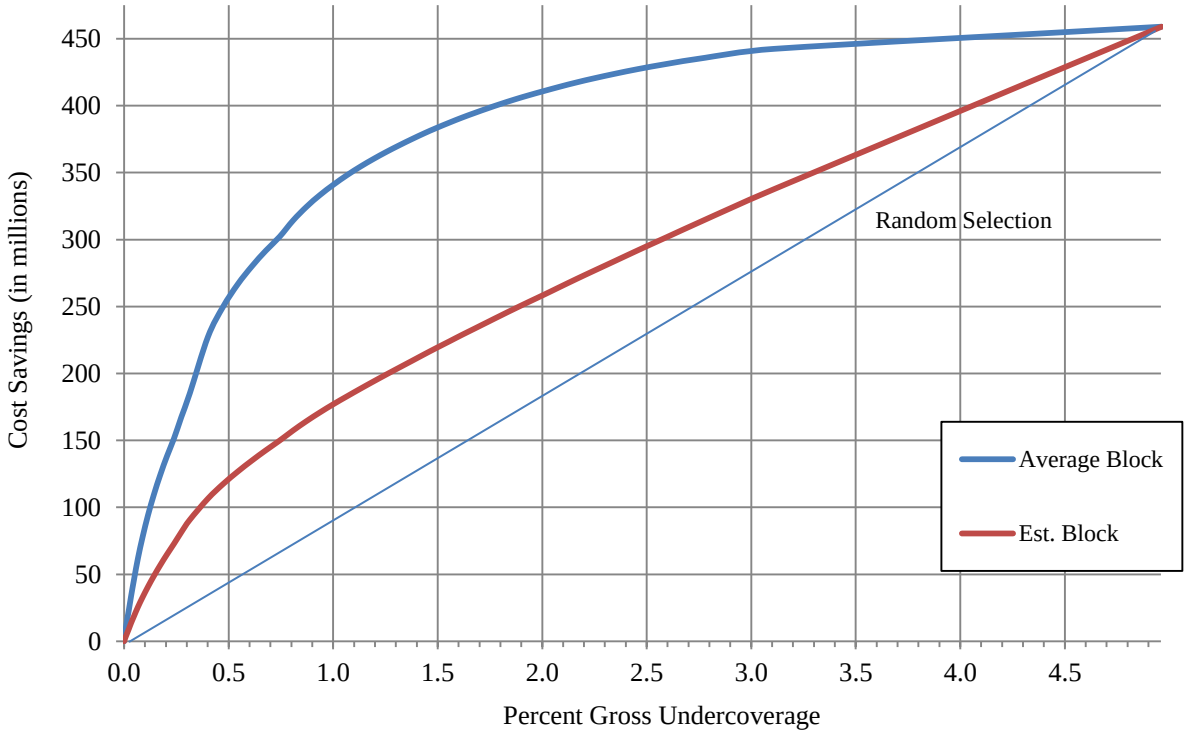
Some blocks will likely take more time for field staff to list while others may require more travel time. At the time of the writing of this report, there were no readily available block-level cost estimates. Consequently, the residential complexity measure was used to create a more specific block-level 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 dollar 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 block's complexity scores. This produced the estimated cost per block (est. dollars/block) savings shown in Table 7. 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 dollars for a random selection of blocks, 117 million dollars for the estimated per block cost estimate, to about 249.8 million dollars 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 assigned based on this model (the random selection procedure will yield a different block selection, so these values will not be the same for that procedure).

Table 7 makes it clear that the possible cost reduction and coverage tradeoffs, however, 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 dollars based on the average 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 dollars less, 46.1 million compared with 55.6 million. The fact the expected outcomes of the block selection is so strongly effected 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/coverage 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

**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)**

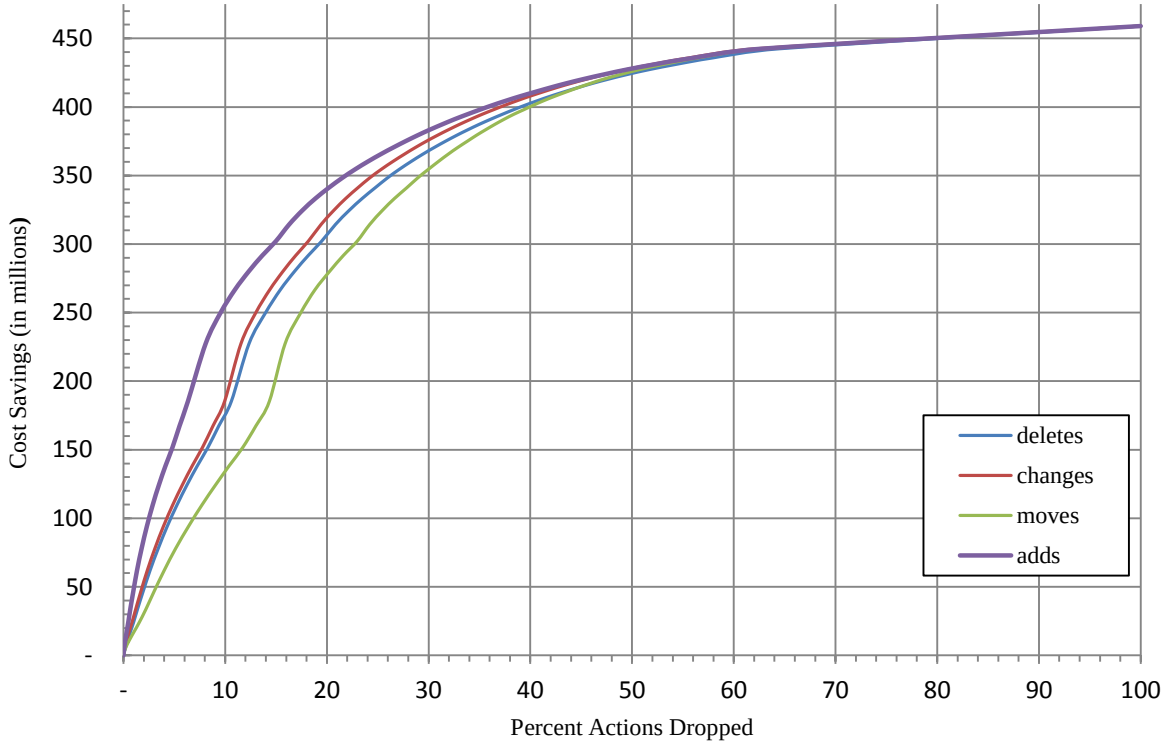



Source: TAC MUMS File.

allows the user to easily determine the expected cost reduction and associated coverage degradation to be estimated for a range of cost/coverage requirements. For example, a gross undercoverage cutoff of 0.3 percent yields an expected savings of just over 200 million dollars for the average cost/block estimate and about 100 million dollars 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 (based on the average cost/block measure). With this curve, the user can estimate the resulting reduction in specific action outcomes corresponding to specific levels of cost reduction. For example, if it is determined that a cost reduction requirement of 300 million dollars is most appropriate, the expected loss is about 14 percent of adds, 18 percent of changes, 19 percent of deletes, and nearly 24 percent of moves.

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




Source: TAC MUMS File.

For some this may be too great a coverage degradation for the expected cost reduction. Alternatively, at a cost reduction selection of 150 million dollars the predicted loss of adds would be only 4 percent, with changes and deletes being about 8 percent, and moves at about 12 percent. This level of savings is still substantial, but at significantly fewer lost actions. Curves similar to these for each of the 11 models were also derived. These figures, while similar in shape to the examples shown in Figures 2 and 3, vary substantially in the conclusions the user would draw from them (See Appendix C for examples from other models.), which is not unexpected considering the variety of results indicated by the 0.5 percent gross undercoverage rate cutoff examples shown for each model in Table 7.

## 6. Related Studies

Some of the 2010 Census Evaluations and Assessments related to the work done here include:

- Address Canvassing Assessment Report
- Address Canvassing Quality Profile
- Evaluation of Address Frame Accuracy and Quality
- Address List Maintenance Using Supplemental Data Sources Evaluation

## 7. Conclusions and Recommendations

### 7.1. Conclusions

The research discussed here indicates that few blocks have substantial quantities of AC actions. Only 4,771 blocks had more than 100 adds recorded (0.1 percent of all blocks) and about 109,000 blocks (1.9 percent) had more than 10 adds. Nearly 75 percent of blocks had no add actions at all. However, the research done for this report did not examine whether blocks with add actions (or any other type of action) might cluster together into nodes of residential dynamism. Especially noteworthy for this research, only 30 percent of blocks did not experience any of the types of AC actions examined here (adds, deletes, changes, or moves).

The distributions of AC actions presented here suggest two important conclusions. First, substantial rewards can be garnered if blocks with high likelihoods of having no AC actions can be identified prior to an AC operation. If the emphasis is on add actions, then as much as 75 percent total AC expenditures could be saved if those blocks or areas could be correctly identified. Second, because about 70 percent of all blocks have some sort of AC action occurring in them, it will be difficult to garner substantial savings by excluding blocks from the AC operation without sacrificing some AC action 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.

The data in this report provide clear answers to the two operational research questions posed earlier in this report. Regarding the first question: "Is it possible to model the outcomes of the 2010 Address Canvassing operation based on a priori data?" **The research presented here indicates that it is possible to adequately model the outcomes of the 2010 operation with pre-existing data.** Albeit, the models presented here are relatively crude in the sense that they 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 binomial logistic regression procedure used here is only one of a range of modeling approaches suitable for the available data and research questions. Some examples of other techniques that might be used include multinomial logistic regression, OLS regression predicting the proportion of HUs in a block with action outcomes, or causal modeling procedures that estimate more complex systems of variables.

Still, the models here provide interesting results. The relationship between the number of HUs in a block and the prevalence of AC actions is more complex than expected. The presence of multi-unit structures, complexity of the blocks, and large numbers of HUs in a block (more than 600 units) all play an important positive role in the probability of AC actions being present. When these variables are controlled, the effect of block size becomes negative. Moreover, the social structure of a block is important to predicting AC actions. The presence of minority populations, and children in a block seem to reduce the chances that a block has added units, but increase the chances of deleted units. A stable population, not surprisingly, decreases the chances of almost all types of actions. Despite this and other new information garnered from the research done here, much work remains. For example, there are numerous new sources of data that may indicate stability within a block that have yet to be explored, e.g., satellite imagery, traffic pattern changes, building permits, and local or regional economic 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: "Once some basic models predicting Address Canvassing outcomes are developed, can these statistical models be turned into useful tools to allocate Address Canvassing resources?" is yes. **The logistic regressions produced predicted probabilities that offer significant utility for allocating resources for future AC operations.** The cost/benefit techniques developed here for using these probabilities show good potential for allocating scarce Census Bureau resources.

The AC/CB tool used to make Table 7 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 pick a savings goal, coverage goal, or various other types of cutoff points and readily see expected outcomes. The results from the sample scenario presented in Table 7 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 perform better at preserving some other types of outcomes (e.g., deletes) than others.

In conclusion, additional research on model-based decision rules to assist in allocating scarce U.S. Census Bureau resources for AC operations should be pursued. The initial modeling results presented here indicate that the potential for cost reduction is great, upwards of 250 million dollars or more. In FY2011, the Census Bureau began to receive funding for the Geographic Support System Initiative (GSS-I), an integrated program of improved address coverage, continual spatial feature updates and enhanced quality measurement. The GSS-I supports using a TAC operation for the 2020 Census. The research presented here significantly furthers this core mission of that initiative. Further, programs like the ACS and other demographic surveys could also benefit from modeling block level changes in residential structure, providing additional justification for expanding the research here. Additionally, the data garnered from the modeling process itself regarding the causes of residential change could prove to be useful input for enhancing small area estimation and refining population estimates used as survey controls. Finally, this research is directly related to modeling MAF errors. The modeling results presented here address block characteristics that can inform the causes of the divergence of the MAF from the ground truth in the field.

## 7.2. Recommendations

The research presented here indicates that the potential rewards from pursuing a TAC modeling program are extensive. The results here are sufficient to justify that this research continue and be expanded. Based on the author's collective experiences over the lifecycle of this research there are several observations that heavily influence the recommendations made in this report: the 2010 Census AC operation data is already three years old as of the writing of this report; new data will have to be collected to keep any research up-to-date with changes in data availability and data collection technology; changes in the MAF and TIGER resulting from the GSS-I will have to be integrated into any new modeling research; and using dedicated expert personnel is needed to ensure that databases and analyses are completed in a timely fashion with appropriate quality controls. Regarding the final two observations in the previous statement, over 18 months in this study were absorbed acquiring access to the StARS data and building the reference databases. Based on these observations, and pursuant to the goals of the GSS-I and the 2020 Census, there are six recommendations:

1) **Create a Dedicated Team:** Identify a full-time team of Statisticians, IT Specialists and Geographers from the decennial directorate and other parts of the Census Bureau to continue the current AIMS-based TAC research efforts. A dedicated team is necessary to reduce the latencies experienced in this project – acquiring, building and integrating administrative record files and reference databases. Emphasis should be on fully exploring, in collaboration with the 2020 Census stakeholders, the current data for improved measures explaining residential change as well as the testing of additional modeling techniques including OLS regression, latent variables analysis, and other more advanced techniques. All research and data acquired and integrated by this team should be made as widely available to the rest of the Census Bureau as feasible. This team should also proceed to expand the project consistent with the subsequent recommendations.

2) **New Data Sources:** New sources of data, including satellite imagery (specifically vegetation indexes and similar measures), traffic flow patterns, AR data from local governments (e.g., property tax, utility construction, road construction) as well as from federal sources, should be acquired through appropriate procedures from originators in the Census Bureau (e.g., Research and Methodology (R&M) Directorate, other Decennial Directorate divisions) and from external sources (e.g., other Federal agencies, state or local governments and businesses). The dedicated team should evaluate these data sources for their usefulness to the TAC program and other 2020 Census research and planning projects.

3) **Database Construction:** A continuing program of data integration, database construction, and maintenance should begin. Central to this effort would be a government records-based dataset of demographic and address characteristics starting with data for 2009. These data should be enhanced with the inclusion of data from other sources including current surveys, commercial databases, local government data, and MAF data. Final 2010 Census HU validity status and population counts should be added to the evaluation criteria. Special care should be taken to minimize inappropriate duplication of databases and to ensure that all databases created by the team are available to the rest of the Census Bureau.

4) **Data Modeling, Development, and Verification:** A program of model development, model refining, and model verification using existing data sources, new data sources and field testing should begin. Predictions from the modeling endeavors should be office validated, tested using data available at the Census Bureau (existing survey data as well as simulated and estimated data), and field tested to maximize the utility of the models both for planning and for scientific purposes.

5) **Cost Modeling:** Acquire the necessary data to develop and test AC cost models at multiple levels of analysis. Without an accurate way to estimate the costs of AC, either at the block or address level, decisions regarding the allocation of AC resources, even high quality residential activity predictions, will not be able to adequately optimize the costs and benefits of a Targeted AC operation.

6) **New Ways of Clustering Census Data:** Examine ongoing efforts and explore alternative methods of grouping data into problem specific clusters of data unrelated to extant census geocoding, e.g., AC actions cluster along streets that span multiple blocks, tracts, or even counties.

## 8. Acknowledgements

This report is the product of the efforts of many colleagues. These people include Mayra Garcia, one of the first persons to work on this project, Justin Ward and Kevin M. Shaw, whose fine work on the 2010 Census AF COMBO file made the data analysis in the report possible, as well as the comments and support of the rest of the Census Evaluations Branch in DSSD: Matthew Virgile, Kathleen Kephart, Christine Tomaszewski, and Nancy Johnson. Title 26 IRS data were used for a large portion of the research in this report. Throughout the process to identify, build-out and secure a separate office to conduct this research using Title 26 data, we owe many thanks to Kevin M. Shaw, Jennifer Reichert, David Whitford, Dan Weinberg, Tom Mesenbourg, Susan Boyer, John Fisher, Bob Drew and others. Additionally, it was necessary to stand up a secure server to conduct this research. For this activity, we also owe many thanks to Claude Jackson, Kevin M. Shaw, Jennifer Reichert, David Whitford, James Trigg, Curtis Broadway, Pam Mosley, Arnold Jackson and numerous others in the Decennial System and Contracts Management Office and Information Technology and Security Office. Without the establishment of a secure office and server, this research would not have occurred. Additionally, we thank Nicholas Doner from GEO for his help in matching the Maryland Property Tax data to the MAF. We also appreciate the comments and support provided by DSSD's senior management, David Whitford and Jennifer Reichert. We thank Deborah Wagner and Chris Boniface and others in DID for their detailed work tabulating and delivering data from StARS. Lastly, we thank our two summer interns, Jay Spry and Andrea Gensler, who made several valuable contributions to the project.

## 9. References


Address List Operations Implementation Team (2012), "2010 Census Address Canvassing Operational Assessment," 2010 Census Program for Evaluations and Experiments Assessment, 2010 Census Planning Memoranda Series No. 168. January 17, 2012.

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

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

Bye, Barry V. (1997). Administrative Records Census for 2010 Design Proposal, Rockville, MD: Westat, Inc.

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.

Clark, Sonja, (2009), "2010 Census Program for Evaluations and Experiments Study Plan: Evaluation of Data-Based Extraction Processes for the Address Frame" 2010 Decennial Census Memorandum Series No. 6. August 13, 2009.

Czajka, John (1999). Can we count on administrative records in future U.S. Censuses? Presentation at the Bureau of the Census, December 15, 1999.

Dean, Jared and Alan Peterson (2005), "Updating the Master Address File: Analysis of Adding Addresses via the Community Address Updating System. "U.S. Census Bureau, 2005.

Devine, Jason and Kirsten K. West (2000), "Using County Level Housing Unit Estimates as a Benchmark for Comparison with the Decennial Master Address File. "U.S. Census Bureau, August 31, 2000.

Dixon, Kelly, Melissa Blevins, Robert Colosi, Amanda Hakanson, Nancy Johnson, Karen Owens, Matt Stevens, and Christine Gibson Tomaszewski (2008), "2008 Census Dress Rehearsal Address Canvassing Assessment Report. "2010 Census Program for Evaluations and Experiments, U.S. Census Bureau, April 15, 2008.

Garcia, Mayra, (2009a), Customer Requirements Document for StARS Housing Unit and Person Tallies for the Study of AC Targeting and Cost Reduction.

Garcia, Mayra, and Jonathan P. Holland (2009b), 2010 Census Program for Evaluations and Experiments Study: Study of AC Targeting and Cost Reduction, Study Plan. 2010 Decennial Census Memorandum Series No. 8.

Garcia, Mayra, and Jonathan P. Holland (2009c), Customer Requirements Document: Matching and Geocoding the Maryland Property Data.

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

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

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

Groves, Robert M. (2011), "Census: Learning Lessons from 2010, Planning for 2020," Prepared Statement, April 6, 2011.

Holland, Jonathan, P., Matthew Virgile, (2009), "2010 Census Program for Evaluations and Experiments Study Plan: Study of Automation in Field Data Collection for Address Canvassing," DSSD 2010 Decennial Census Memorandum Series #O-A-02, November 24, 2009, 2010 Census Planning Memorandum Series No. 65, August 12, 2010.

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.

Johnson, Nancy, (2011), "2010 Census Program for Evaluations and Experiments Study Plan: Evaluation of Address Frame Accuracy and Quality," DSSD 2010 Decennial Census Memorandum Series #O-A-3R, June 2, 2011, 2010 Census Planning Memorandum Series No. 146, June 14, 2011.

Judson, D.H., Popoff, Carole L., and Batutis, Michael (2001). An Evaluation of the Accuracy of U.S. Census Bureau County Population Estimation Methods. Statistics in Transition, 5:185-215.

Kennel, Timothy (2007), "Second Frame Assessment for Current Household Surveys: Comparing the Area Frame Listing to the Master Address File. "U.S. Census Bureau, October 2, 2007.

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

Owens, Karen (2011), "Project Charter: GSS Initiative Address Coverage Working Group," February 2, 2011.

Pearl, J. (2000) *Causality: Models, Reasoning and Inference,* Cambridge University Press. 2nd edition (2009).

Resnick, Dean (2010), "Current Records Linkage Research and Practice at the U.S. Census Bureau," 2010 Joint Statistical Meetings, August 1, 2010, Vancouver, B.C.

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

Ruhnke, Megan C. (2002a), "Census 2000 Evaluation F.2-The Address Listing Operation and Its Impact on the Master Address File. "U.S. Census Bureau, January 30, 2002.

Saintelien, H., S. Fifield, (2011), "Project Charter: GSS Initiative Quality, Assessments, & Evaluations Work Group," January 11, 2011.

Schneider, Glenn, Karen Owens, and Susan Perrone (2006), "2006 Census Test: AC Operational Assessment. "U.S. Census Bureau, 2006.

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

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

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

Tomaszewski, Christine, G. (2010), "2010 Census Evaluation Study Plan: Evaluation of Address List Maintenance Using Supplemental Data Sources," 2010 Decennial Census Memorandum Series No. 1, March 22, 2010.

Virgile, Matt (2012) "2010 Census Program for Evaluations and Experiments: Evaluation of Small Multi-Unit Structures Report," 2010 Census Program for Evaluations and Experiments, 2010 DSSD CPEX Memorandum Series #A-01, February 13, 2012, 2010 Census Planning Memorandum Series No. 175, February 24, 2012.

Vitrano, Frank (2003), "Recommendations for 2004 Census Test from the 2010 Research and Development Planning Group on Coverage Improvement. "2010 Census Planning Memoranda Series, U.S. Census Bureau, January 21, 2003.

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

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

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.

## Appendix

Appendix A: Screenshot of AC/CB Tool

| 1+ Any change model block "P" | # HUs | Adds by percentile | Cumulative sum of adds | Percent blocks dropped | Percent adds dropped | Percent Gross Under Coverage | Avg. Block Cost |
|---|---|---|---|---|---|---|---|
| 1.0 | 155,167,767 | 2,472,513 | 6,624,153 | 100.00 | 100.00 | 5.15 | 459.00 |
| .99 | 109,060,056 | 439,697 | 4,151,640 | 96.40 | 62.67 | 2.99 | 442.46 |
| .98 | 99,955,041 | 244,910 | 3,711,943 | 94.95 | 56.04 | 2.64 | 435.82 |
| .97 | 94,828,947 | 184,842 | 3,467,033 | 93.95 | 52.34 | 2.45 | 431.24 |
| .96 | 91,001,455 | 148,043 | 3,282,191 | 93.11 | 49.55 | 2.30 | 427.36 |
| .95 | 87,893,900 | 131,064 | 3,134,148 | 92.34 | 47.31 | 2.19 | 423.85 |
| .94 | 85,177,542 | 116,606 | 3,003,084 | 91.62 | 45.34 | 2.09 | 420.52 |
| .93 | 82,765,031 | 105,187 | 2,886,478 | 90.92 | 43.58 | 2.01 | 417.31 |
| .92 | 80,535,982 | 98,469 | 2,781,291 | 90.23 | 41.99 | 1.93 | 414.16 |
| .91 | 78,442,443 | 97,596 | 2,682,822 | 89.55 | 40.50 | 1.85 | 411.03 |
| .90 | 76,435,690 | 92,034 | 2,585,226 | 88.86 | 39.03 | 1.78 | 407.85 |
| .89 | 74,525,348 | 87,684 | 2,493,192 | 88.16 | 37.64 | 1.72 | 404.64 |
| .88 | 72,677,813 | 83,992 | 2,405,508 | 87.44 | 36.31 | 1.65 | 401.37 |
| .87 | 70,884,063 | 82,060 | 2,321,516 | 86.72 | 35.05 | 1.59 | 398.04 |
| .86 | 69,132,145 | 82,922 | 2,239,456 | 85.98 | 33.81 | 1.53 | 394.63 |

Appendix B: Example of SAS Output for Estimated Logistic Regression Coefficients for 1+ Any Action TAC Model

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -0.0919 | 0.00342 | 721.4749 | <.0001 |
| numhus_Sum | 1 | -0.0376 | 0.000215 | 30645.4405 | <.0001 |
| ghutypm_Mean | 1 | 0.4671 | 0.00627 | 5544.5756 | <.0001 |
| hisp | 1 | -0.2492 | 0.00233 | 11416.0113 | <.0001 |
| black | 1 | -0.0787 | 0.00232 | 1144.7373 | <.0001 |
| child | 1 | -0.3682 | 0.00279 | 17454.6579 | <.0001 |
| nopopchange | 1 | 0.0803 | 0.00270 | 881.2715 | <.0001 |
| nomafrate | 1 | -0.0369 | 0.00139 | 704.8045 | <.0001 |
| largeblocks | 1 | 2.0784 | 1.2290 | 2.8597 | 0.0908 |
| stdpop | 1 | 0.00399 | 0.000278 | 206.5857 | <.0001 |
| stars012 | 1 | 0.2892 | 0.00621 | 2168.3856 | <.0001 |
| effort | 1 | 0.0547 | 0.000173 | 100583.444 | <.0001 |
| lbbyhus | 1 | -0.0155 | 0.000828 | 350.9507 | <.0001 |
| lbhisp | 1 | 0.1805 | 0.8796 | 0.0421 | 0.8374 |
| lbblack | 1 | -2.2273 | 1.2225 | 3.3194 | 0.0685 |
| lbchild | 1 | 1.1252 | 0.7029 | 2.5624 | 0.1094 |
| lbstdpop | 1 | -0.00259 | 0.00120 | 4.6385 | 0.0313 |

Appendix C: Example AC/CB Figures (based on TAC MUMS File)