Fielding a Targeted Address Canvassing Operation: Alternative Approaches to Moving from Predictive Statistical Modeling to a Cost Effective Address Canvassing Field Operation for the 2020 Census

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

Abstract

To date substantial effort researching how to use statistical modeling to reduce the costs associated with updating the Master Address File for the 2020 Census has yielded results suggesting that applying statistical modeling approaches to targeting geographic areas for selective address canvassing before the 2020 Census could generate significant cost reductions. However, the problem remains on how to move from good quality statistical models to an actual field operation. This paper details a range of alternative methods to operationalize statistical predictions into a cost effective field operation. The development of a cost/benefit continuum to array blocks into categories for different kinds of address updating, applying different geographic units to the problem, and linking workload determination to modeling outcomes will be explored.

1. Introduction and Background

The purpose of this paper is to lay out some of the ways the statistical modeling research that has been done to date by members of the US Census Bureau's Decennial Statistical Studies Division (DSSD) and Research and Methodology Directorate (ADRM) may be used to operationalize a Targeted Address Canvassing (TAC) operation for the 2020 Decennial Census. The 2010 Census Address Canvassing (AC) Operation cost about 459 million dollars in direct costs and approaching 850 million dollars in total costs (Holland, 2012), the second most expensive field operation in the 2010 Census. This has made this aspect of fielding the 2020 Census AC a prime target for cost reduction research and planning.

The central purpose of AC is to use data collected from the field to update and correct the Master Address File (MAF)—the Census Bureau's master list of living quarter addresses, as well as to update other geographic information, prior to the decennial census. An accurate MAF minimizes the possibility of Housing Units (HUs) being missed by the census (undercoverage) or of census forms being mailed to non-existent or duplicated addresses for living quarters (overcoverage). The MAF is a recent development for the decennial census, developed only in the late 1990s. The MAF is primarily maintained through a semiannual update provided by the United States Postal Service's (USPS)

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

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

Before the 2000 and 2010 Decennial Censuses, major operations were fielded to update the MAF in preparation for enumeration. In 2000, Block Canvassing and Address Listing operations were carried out to do a final pre-census update to the MAF. In 2010, the AC operation involved field staff visiting nearly every block in the 50 states, the District of Columbia, and Puerto Rico to verify and update the MAF. Not only were these operations expensive, but they also indicated that for much of the nation the MAF was accurate but for others it was substantially less accurate (Boies, Shaw, and Holland, 2012, 2013). These two facts led to an evaluation of the feasibility of using statistical modeling to predict where AC would be most useful to update the MAF. This initial evaluation (Boies, et al., 2012) concluded that it was feasible to apply statistical modeling techniques (in this case logistic regression) to the problem of where to target AC resources for the 2020 Census. The modeling outcomes described here and their application to 2020 AC operations is the direct outgrowth of this initial research on selective AC.

The initial AC research, as well as the current AC research, uses data collected from the 2010 AC Operation to model at the block level (2000 Census Tabulation Blocks in the initial evaluation, 2010 Census Tabulation Blocks for the current research) for a "What If" simulation. We examined the scenario "What if the Census Bureau had used statistical models using 2009 and prior data to select blocks for canvassing in the 2010 Census AC Operation?" The research carried out for the first evaluation and the most recent evaluation is reported in detail elsewhere (see Boies, et al., 2013; Tomaszewski and Boies, 2014).

In brief, the most recent research has made significant progress in developing statistical models that efficiently select 2010 Tabulation Blocks for AC. Indeed, two different Census teams have worked on this endeavor, with different goals and statistical tools and have produced substantially similar TAC solutions. Before moving into the operationalization of these modeling outcomes, we briefly describe the results of this research.

2. Extant Targeted Address Canvassing Modeling Outcomes

The most recent AC modeling efforts use data at the block level from the 2010 AC Operation (the source of dependent measures) along with data from other sources (for independent measures) that existed prior to the operation in keeping with the intent of the simulation approach to produce predicted lists of blocks most likely to need field work. The data for the independent variables are from various databases including the 2000 Census, MAF extracts, several census operations (e.g., the Group Quarters validation operation, pre-address canvassing MAF), the USPS DSF, and the Statistical Administrative Records System. The AC outcome data are directly from the AC operation. The outcomes primarily used here are Adds ("True Adds" which are new addresses identified in the field and "Reinstated Adds" which are addresses that were on the MAF but were not sent out for AC[2]) and deletes ("Double Deletes" which are addresses found to be non-existent). Independent measures can be classified into two

[2]Possibly coded as not housing units in 2009 Master Address File.

groups: 1) physical characteristics of the addresses or blocks, e.g., block size, address quality; and 2) social characteristics of the blocks, e.g., demographic make up of the block.

Recent work includes two logistic regression models predicting blocks with two or more adds per block and blocks with two or more adds or deletes per block. A benchmark model (called the "Perfect Model") was created that allows the comparison of the predictive performance of the predictive models to be compared to a model on perfect knowledge of which blocks contain 2010 AC outcomes (Tomaszewski and Shaw, 2013). The metrics for the "Perfect Model" represent the theoretical maximum (TMAX) efficiency that any modeling endeavor could achieve for predicting which 2010 blocks might require AC. The modeling was done on the approximately 6.6 million 2010 Census Tabulation Blocks containing addresses at the start of the 2010 AC Operation (out of the 11.2 million total number of 2010 Tabulation Blocks).

The modeling performed to date has produced a range of difference outcomes. While the most recent efforts have produced models with more than adequate fit statistics (e.g., logistic regression models predicting blocks with two or more adds or deletes per block have an Area Under the Curve of 0.84), there is still room for improvement in their ability to select blocks for canvassing when compared to the TMAX from the Perfect Models. Additionally, all the modeling endeavors to date using 2010 Census Tabulation Blocks are inefficient predictors of where deletes occur (Tomaszewski and Boies, 2014).

However, while there is strong potential for substantial cost reduction resulting from applying a statistical model based solution to the TAC problem, there are several limitations to the current research that must be solved before a statistical model based TAC program could be implemented.

All the model outcomes result in substantial overcoverage due to deleted addresses in 2010 that were identified in blocks that were not flagged for targeting by the model. These excess addresses could result in a substantial increase (as much as 20%) in the workload for the census Non-Response Follow Up (NRFU) operation. Each HU in the NRFU operation cost about as much to process as did each block in the AC operation (Walker, Winder, Jackson, and Heimel, 2012).

Another limitation on the current research is that, as of the writing of this paper, the data used are already five years old. The next AC operation will not occur until, at the earliest, 2019. While the models may be adequately predictive of what happened in 2009, we have no data to indicate if the MAF content in 2019 will have the same relationship to the field that it had in 2009. The analysis to date also does not shed significant light on the question as to how dispersed, or conversely, how concentrated the targeted blocks should be to ensure an efficient AC operation. In the 2010 AC Operation, mileage costs made up about 17% of the $459 million direct costs (Holland, 2012). If the targeted blocks are very close to each other then field staff can efficiently work them with minimal travel costs, but widely dispersed blocks could be considerably more costly. Until sufficient knowledge is garnered regarding the clustering of targeted blocks, the most efficient implementation of a 2020 TAC solution remains a question.

The last limitation to be discussed here is the issue of blocks empty of addresses eligible for 2010 AC prior to the operation but had addresses added during the operation. There are a total of about 11.16 million 2010 Census Tabulation Blocks. Table 1 shows that the

distribution of AC outcomes of primary interest here (91.0% of adds and 100.0% of deletes) are entirely contained in Partition A, blocks containing at least one pre-AC HU. Partition C, all water blocks, contain no pre-AC addresses and no AC outcomes. Partition B however, contains about 1.3 million True Adds and Reinstated Adds (about 650,000 of each). These blocks are not part of the modeling universe because there is no address level data for these blocks prior to the 2010 AC Operation. Therefore current modeling results do not include these blocks and their associated AC outcomes.

Table 1. Distribution of Address Canvassing Outcomes in the Universe of 2010 Census Tabulation Blocks

		Attribute (in millions)[1]					
Partition[2]		Blocks	Housing Units	Add Actions[3]		Negative Actions[4]	
				True	All	Double Deletes	All
A	1+ Housing Units Pre-AC	6.6 (59.0)	144.8 (100.0)	6.1 (91.0)	9.5 (87.9)	15.8 (100.0)	21.7 (100.0)
B	ZERO Housing Units Pre-AC: Land	4.0 (36.1)	0.0 (0.0)	0.6 (9.0)	1.3 (12.0)	0.0 (0.0)	0.0 (0.0)
C	ZERO Housing Units Pre-AC: Water Only	0.5 (4.9)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
TOTAL		11.2 (100.0)	144.8 (100.0)	6.7 (100.0)	10.8 (100.0)	15.8 (100.0)	21.7 (100.0)

[1]All values are expressed in millions. For each cell, the column percentage is provided in parentheses.
[2]U.S. and PR are divided into three mutually exclusive partitions/subsets for the statistical modeling research. These partitions and the first four attributes were defined/tabulated using only data available prior to the 2010 Census AC operation.
[3]The 2010 Census AC Add Actions are divided into two overlapping categories: "true" adds (addresses added in the field, that were not previously on the MAF) and "all" adds (the true add universe combined with the universe of add actions that were matched to addresses previously on the MAF but not identified as part of the dependent listing universe; largely due to missing geocodes).
[4]The 2010 Census AC Negative Actions are divided into two overlapping categories: "double deletes" (addresses that received two separate delete actions in the field) and "all" negative actions (the double delete universe combined with the universes of addresses identified as single deletes, duplicate, nonresidential, or uninhabitable).
Source: Internal DSSD Block-Level Dataset

Fortunately, there are no deletes or other actions in the approximately 210,000 blocks in Partition B. The undercoverage resulting from these blocks not being part of the modeling universe indicates that addressing the "Empty Block Issue" is essential to a successful reduced AC program for 2020.

3. Using Statistical Models for Targeted Address Canvassing

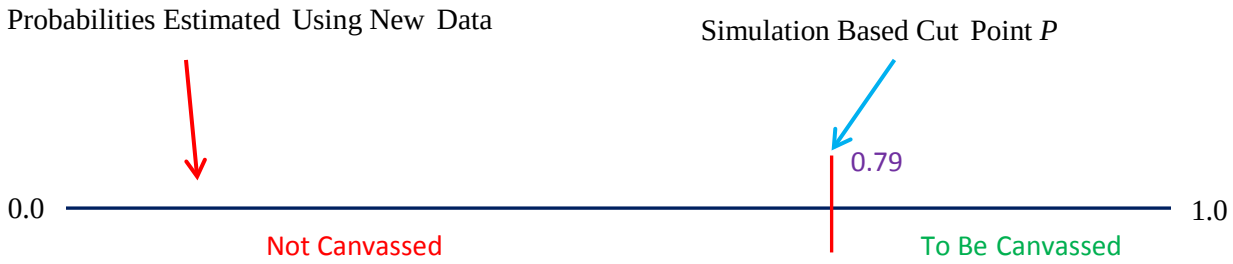
The statistical modeling has produced a number of models that predict which 2010 Census Tabulation Blocks contained Adds for the 2010 AC Operation. While the relative strengths and weaknesses of the coefficients used in producing the models are interesting in their own right, the estimated coefficients are the primary tool in determining the areas to be listed in a future listing operations. A linear programming algorithm can combine these model coefficients from the with data current to the listing operation to calculate predicted probabilities of a block containing the specific modeled AC outcomes. For example, if a list of 2010 Census Tabulation Blocks based on a model predicting one or more adds was needed to conduct a field AC operation in December 2014, we would need a dataset containing the variables used in the final model for each Tabulation Block and as current as possible to the proposed field AC operation. Using those variables in the following equation we could then calculate *P*—the Predicted Probability of a Block Containing one or more adds, as follows:

$$P = \frac{e^{intercept + b_1 * v1 + b_2 * v2 + b_3 * v3 \ldots.}}{1 - (e^{intercept + b_1 * v1 + b_2 * v2 + b_3 * v3 \ldots.})}$$

Where P	= Predicted Probability of a Block Containing one or more adds
b_1, b_2, b_3	= Logistic Regression Coefficients from Model
v1, v2, v3	= Current values for variables in 2010 AC Outcome model

Figure 1 provides a graphical representation of this procedure. In this example the simulation based *P* is 0.79 (a value that might correspond to canvassing only 20% of HUs being canvassed). Once new data are applied to the simulation based model parameter and the blocks are arrayed by the new predicted probabilities, blocks would be selected for canvassing if their predicted probability was equal to or above 0.79. Blocks with values less than this would not be canvassed.

Figure 1. Arraying the Blocks by Calculated Predicted Probabilities



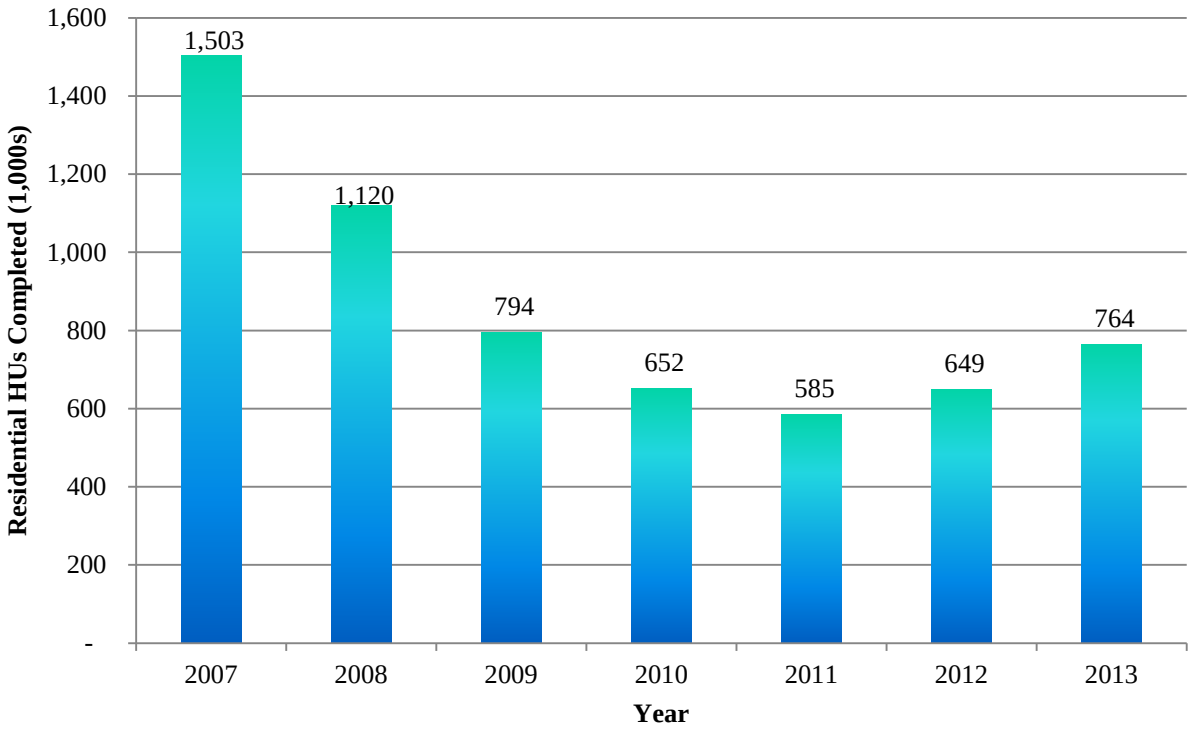
The longer the period between when the data used to generate the simulation based model *P*s is collected and when the field operation is carried out and the greater the time between when the operation *P* are calculated and when the field operation occurs, the lower the likely efficiency of the selected block list. There is, however, considerably more to be done before we can count on successfully performing this task for any real field implementation of a statistical model based reduced AC solution.

4. The Road to a Statistical Model Based Targeted Address Canvassing Solution

This section lays out a number of processes and possible methods to move from an initial set of statistical models predicting where AC Outcomes occurred in the 2010 AC Operation, to an efficient AC operation in preparation for the 2020 Census, which optimizes cost reduction and quality preservation. The steps and processes outlined here are not US Census Bureau policy, represent only conceptual components of ongoing research and evaluation programs, and are presented only for the purposes of discussion. This section is organized around some exemplary issues and problems and possible solutions.

As indicated above, already the data used to create the extant models is five years old and aging by the minute. Very importantly, there are reasons to believe that conditions on the ground are changing in ways that may invalidate the existing models. For example, the Geography Division's Geographic Systems Support Initiative (GSSI) is likely to have changed the quality of the MAF in ways that may affect the impact many of the MAF based independent variables will have on AC outcomes. Further, the rates of new residential construction are considerably different now than they were during last few years of last decade. The number of new HU completions, shown in Figure 2, for 2007 was nearly double that of 2013, the most recent year for which complete data are available.

Figure 2. Residential Housing Unit Completions in the United States from 2007 to 2013



Source: (US Census Bureau, 2014)

While the downward trend seems to be reversing over the last few years, there is no guarantee that reversal will continue. These kinds of changes in the MAF environment

indicate that the collection of additional field data are necessary for the successful implementation of a statistical based TAC solution. The aging data issue, as well as the "Empty Block Problem," and overcoverage limitations are primarily problems that can be addressed by data and research. The two central aspects of this research program are: 1) the continued collection of data, database construction and data integration, and statistical modeling of 2010 AC Outcomes and 2) the collection and analysis of "ground truth" similar to the data collected for the 2010 AC Operation from at least one and possibly more field tests. There is one planned field test, the Address Validation Test (AVT), scheduled for the Fall of 2014, that will canvass in a fashion similar to the 2010 AC Operation. This test encompasses 10,000 Census 2010 Tabulation Blocks selected using a sample designed deal with the highly skewed distributions of the AC outcomes (e.g., adds and deletes mostly occur in just a handful of blocks) and many of the independent variables. In addition to the sample blocks, a purposefully selected pilot test sample of 100 "Empty Blocks" will be canvassed to explore paths to solving the "Empty Block Problem" using field collected data. Census Bureau employees made a significant effort to ensure that this operation produces a data collection outcome similar to the 2010 AC Operation.

The continued data collection and modeling efforts detailed in Johnson and Pritts (2014) include the further acquisition and integration of Administrative Record (AR) data into existing databases as well as the integration of additional data from census sources. The approach used for this project is to create a "string of pearls database" (see Johnson and Pritts, 2014, for details of this database structure) that maximizes flexibility and minimizes resource consumption. The multiple databases that have been acquired for use in this project are individually prepared for use by extracting variables of interest, aggregating to the preferred unit of analysis (blocks or addresses), and subsetting the records to match the universe of interest (e.g., 50 states, the District of Columbia, and Puerto Rico). Then these string of databases, or "pearls," are added to the existing analysis database as needed or as they become available. In addition to supporting the further analysis of 2010 AC Operation data this aspect of the project will facilitate research using the data from the AVT and possible future field tests to validate and verify existing models as well as exploring new and better models.

To date the most accepted solution to the aging data problem is to acquire more recent "ground truth" from field tests, but for the other two limitations discussed above, empty blocks and potential overcoverage, the primary tool in the short run will be further analysis of 2010 AC outcome data. As already noted, all of the current modeling endeavors that are efficient at predicting adds result in not being able to identify an unacceptably large number of addresses that should be deleted but were not identified because they reside in the non-selected blocks. While we can expect the ongoing GSSI to result in fewer expected deletes for the 2020 AC Operation, there is no reason to believe it will be 100 percent effective.

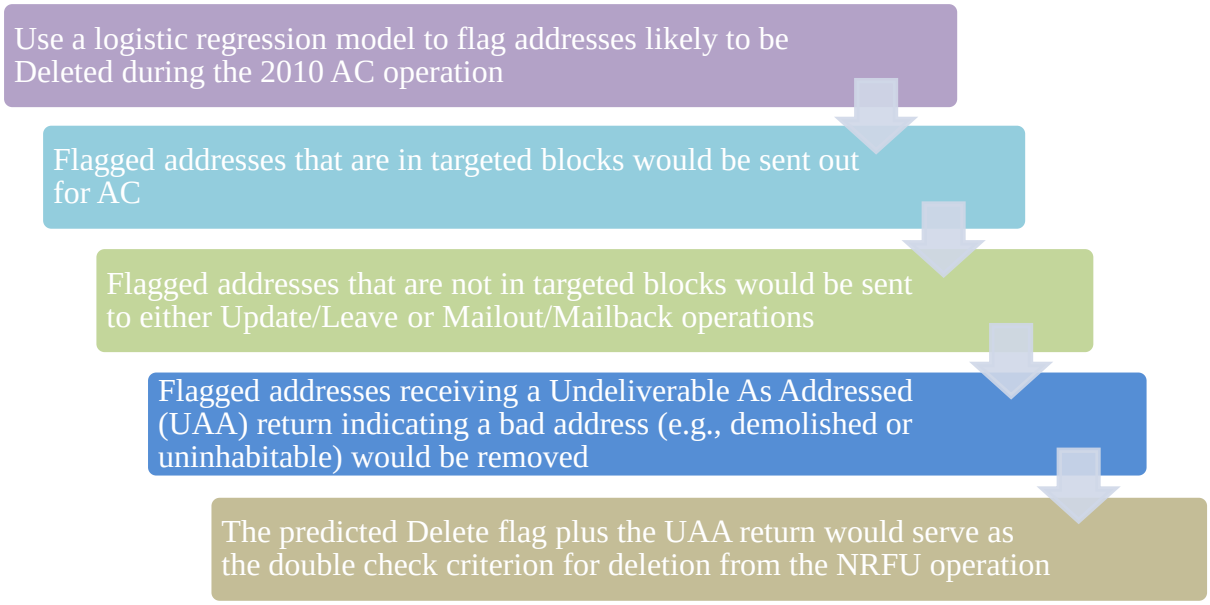
One way to handle the missed deletes is to focus on a reduced AC implementation that effectively harvests adds (focus on modeling areas containing adds) and hands the problem of undetected deletes to later census operations. The implications of doing this are on the research agenda for other 2020 Census research teams. Additionally, we are exploring some model-based possibilities to ameliorate the potential adverse outcomes of the undetected deletes and resulting overcoverage.

The extant research has focused on using generalized linear model (GLM) based (e.g., logistic regression) predictive models of block level AC outcomes. Since lost adds could result in a difficult census problem, undercoverage, they have been given the highest priority in this research to date. Very importantly, since the majority of adds are new addresses, no address level data exists for them in any census collected database, and while AR data has the potential to provide address level information for (new) added addresses, coverage for 2010 AC Operation adds is relatively low (Tomaszewski, 2013). This makes it necessary to use a level of analysis above the address (e.g., blocks or clusters of addresses) in order to efficiently predict where adds can be found. Deletes (as well as most other AC outcomes) have address level data in Census databases that could be used to predict which addresses are likely to be deleted during a field operation. We are now beginning to evaluate models predicting deleted addresses using address level data.

At the time of this writing, we have estimated a handful of simple logistic regression models using 2009 data to predict whether or not a pre-AC address was deleted as a result of the field operation. The study universe for this endeavor is the approximately 144 million addresses sent out for the dependent listing during the 2010 AC Operation. The current best model uses 34 variables from AR and census sources. The results are promising; the best model has a max-rescaled R^2 of 0.33 and an AUC of .686. On the other hand, at the predicted probability level of 0.5 the percent of false positives is high at 41.2%, or over 59 million addresses, at this probability level are falsely predicted to be deletes. The predictive efficiency of the model is, as yet, insufficient to yield useful results.

If the modeling endeavor is successful the prediction will have to be turned into a useful outcome. For the 2010 Decennial Census in order for an address to be deleted from the census address list it had to be flagged by at least two operations as an invalid address for census purposes (Walker, Winder, Jackson, and Heimel, 2012). Figure 3 outlines one approach to using a statistical prediction to remove excess records from the census address list. Central to this process is the use of the results of an initial mailing of either census forms or other postal communication as a precursor to a range of possible enumeration procedures.

Figure 1. Using Statistical Modeling Outcomes at the Address Level to Ameliorate Selective Address Canvassing Overcoverage



This proposal posits the use of a logistic regression model to flag addresses likely to be deleted during the 2010 AC operation and then count this model-based flag as the first flag in the double flag process to determine which addresses should remain valid in the census master address list. We will use a cost/benefit approach based on predicted probabilities similar to the one used for establishing the selected block list; a predicted probability cut point will be selected that results in an acceptable level of overcoverage without resulting in unacceptable undercoverage. In this process, all addresses flagged as potential deletes would be sent on to the regular census operations. If the address is in a selected AC block and is discovered to be a delete, the record will be deleted with no further follow-up. That is, this field outcome would be considered as the required second delete flag. Flagged addresses not in a TAC block would be sent to the normal operations, most likely some form of mail-out system, either of census forms or of post cards or something similar for the internet option. If the postal service provides an Undeliverable As Addressed (UAA) return indicating the address is likely to be a census acceptable delete, e.g., demolished or uninhabitable, this will constitute the second delete flag and the address will be deleted from the Census address list.

There are two primary benefits to this operation. First, it is likely to result in a substantial reduction in any overcoverage that might result from a reduced AC operation that uses 2010 Census Tabulation Blocks as its primary modeling unit. The second major benefit is that the modeling process may flag not just selective AC related overcoverage but also deletes that would be discovered in other operations. These deletes could then be deleted from census MAF after just one regular field or headquarters operation instead of the two previously required. While there is some cost reduction potential here, its amount is, as yet, unknown.

The 2010 AC Operation data will also be the primary source of data to examine the issue of the clustering of targeted blocks. As noted above, travel was a significant component of the direct cost of the 2010 AC Operation. Because far fewer blocks are likely to be

canvassed during a reduced AC operation the proportion of direct costs attributable to travel may be substantially higher. Consequently, the importance of developing an efficient methodology to allocate the TAC based workload is more important than in previous canvassing or listing operations. We have determined two significant issues that must be addressed. One is that of orphan blocks, selected blocks that are likely to have AC actions of interest but are located a substantial distance from other targeted blocks or from where the field staff live or work. Field staff cost on the average more than ten dollars/hour in 2009 and mileage costs were $0.55/mile. Therefore a block that is 60 miles from the nearest employee and other blocks could cost more than $110 dollars to list in travel and field time (120 miles times $0.55/miles and about five hours of travel and listing time; many blocks could be listed in less time), but only contain one added HU (Holland, 2012).

In a more urban setting, the selected blocks may be a few hundred yards apart or contiguous, resulting in substantially more efficient expenditures of travel dollars. Regarding orphan blocks, at least two questions must be answered: how many are there; and can we develop a way to substitute a less expensive block that is closer to the field staff's point of origin? If the number of orphaned blocks is trivial, then this problem becomes moot. If, on the other hand, the number of orphan blocks is substantial, then it would be desirable to be able to substitute blocks that are closer to listers that were likely have similar yields of adds. The use of a statistical procedure that produces block predictions of the number of AC actions in a block (e.g., Zero Inflated GLM procedures) could provide intelligence as to which non-orphan blocks might be a suitable alternative to the more expensive orphaned block. Indeed, using this information could also provide alternatives to blocks that were excessively expensive (compared to expected AC action returns) because of distance between addresses in the block, terrain, or road conditions.

The second issue is the efficient handling of targeted blocks that are neither orphaned nor essentially contiguous. In these cases, the inefficiency results from the distance between the blocks and the distance between the addresses within the block. Again, research on the degree of clustering of targeted blocks is essential in understanding how to efficiently assign workloads to these blocks. The Residential Complexity Index constructed for the Census Program for Evaluations and Experiments TAC report (Boies, et al., 2012) provides a good measure of the clustering of addresses within a block. In the short term, the biggest problem caused by these issues is for cost estimation. The travel time between blocks that are in the same area and within blocks causes substantial and yet unknown differences in the likely cost of any given TAC solution. Using this information to optimize the targeting lists could result in more efficient lists of blocks and further reduction in AC costs for 2020.

There are two main avenues to approach the Empty Block Problem. One is to eliminate, to the extent possible, blocks that cannot contain housing units. This rule based approach would be used to reduce the analysis universe for TAC modeling to only blocks that might have addresses. Obviously all water blocks are an especially easy target for this approach. Blocks that are 100% managed by the Bureau of Land Management cannot, in most cases, by law, have permanent residents and hence could be removed from the modeling universe. Overhead imagery might be used to scan large areas for signs of habitation (infrared signatures) and eliminate those uninhabited areas from consideration for modeling. An effective program to exclude uninhabitable (from a census standpoint) areas from consideration could improve the efficiency of the modeling endeavor

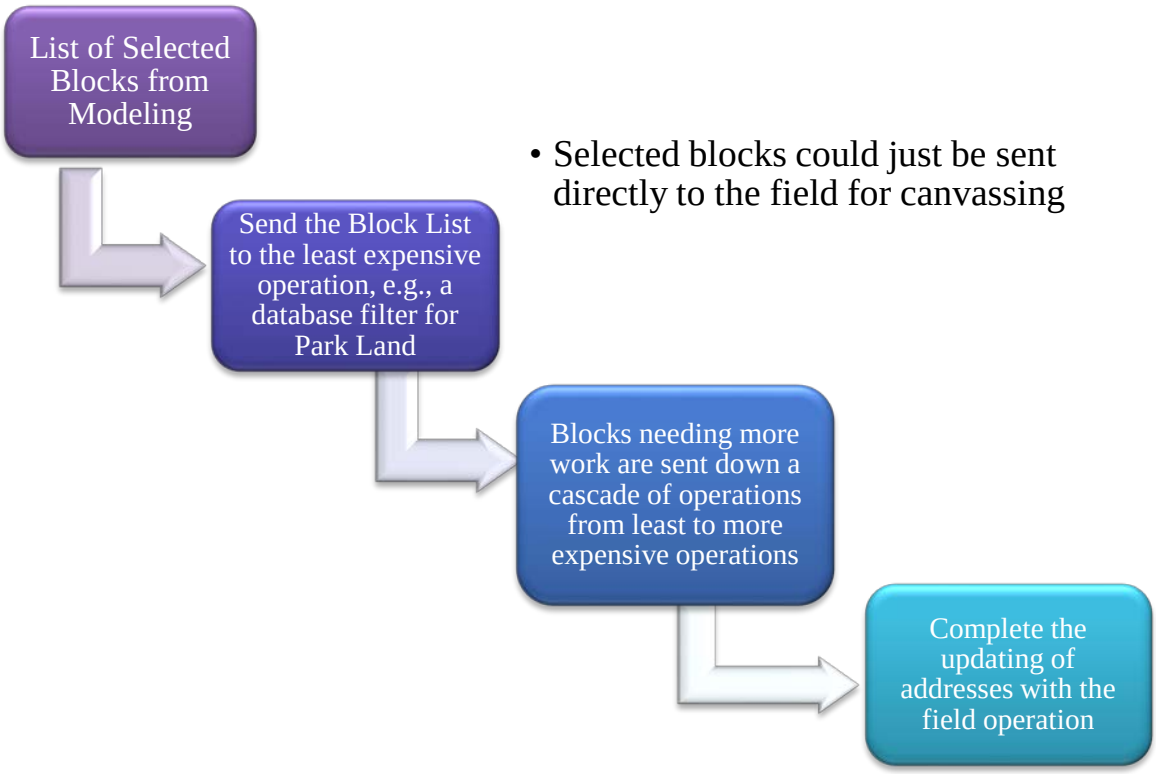
considerably by removing areas that otherwise might become candidates for false positive outcomes.

The second avenue to be followed is to add the possibly inhabitable empty blocks to the modeling universe along with block or other level indicators that might be correlated with the likelihood of there being adds in the block. Some indicators that are being collected and considered include whether the block was a Zero Population Block in the previous census, block size, and block land use. Zero population blocks from the prior census would be at the prior census block geography so there would be some inefficiencies in flagging the new census geographies. The size of the block and the block shape are two possible ways of indicating whether the block is large enough in area or large enough in dimension to contain at least one HU. Identifying blocks that are 100% commercial, agricultural, or state park land, may all be useful in predicting blocks without any adds. On the other hand, how close a block is to densely addressed blocks or proximity to a transportation hub may predict a greater likelihood of a block containing adds. Essentially the current modeling strategies would be applied to a new universe, containing both non-empty and empty blocks. Very importantly, if there are field tests after the AVT in 2014, empty blocks would have to be included in the sampled blocks.

The last aspect of moving from a statistical model providing predictions to an efficient 2020 TAC operation is integrating the statistical work into larger census operations. While there are many dimensions to this task, we will look at one of the core components: how the list of targeted blocks (or other possible geographies) is turned into the final workload input. In this case we examine the role the statistical model outcomes might play in the production of both a Census address file for pre-enumerations and the list of workload areas for an AC operation. Here, the integration of the selective AC modeling outcomes with the GSSI is highlighted.

The simplest approach to using the list of blocks in a limited AC operation is to create the list and provide it to the field staff to turn it into workload assignments. While this is simple and offers the possibility of substantial cost reduction compared to canvassing every area in the nation, various components of the extensive GSSI operation offer numerous opportunities for a more efficient operation, with possibly better quality outcomes. The simplified flowchart in Figure 4 indicates how an adaptive design could be used to pre-process the list of selected blocks to significantly reduce the ultimate number of blocks that need to be canvassed in the field. There is currently substantial indication that a modification in the existing Update/Leave (U/L) operation will be made to enhance its role in canvassing areas for errors in the address lists. This stems from the observation that it is less expensive to enter the field once than twice as was done in the 2010 census. Thus, the first step in this cascade of operations and filters is to remove blocks from the targeted list and send them directly to the U/L operation (or its 2020 successor operation). Next, blocks identified as being all or predominately National Parks, Military facilities, or other similar Federal uses could be either removed from further activity altogether or passed on to an operation equipped to contact Federal agencies to provide address update information. Blocks that were not targeted by the modeling endeavor, but were found in other aspects of the GSSI to be likely candidates for a field operation (e.g., indications from overhead imagery), could be added to the list of blocks to be canvassed.

Figure 2. A Simplified Path to Send Statistically Selected Blocks from Model Predictions to Field Listing



After that, blocks may be passed on to an operation that could use satellite imagery to detect changes in the block. If no changes were detected in the block then the block may be removed from further consideration. Addresses from the trusted Partnership Program could be compared to those in targeted blocks and if updates could be accomplished this way, the block would be removed from targeting (or if only some addresses were verified, the block could move to the next operation for completion). The ultimate outcome of this cascaded process would be the utilization of the data and operations resulting from the GSSI to further enhance the efficiency of the limited AC Operation. Only those blocks that could be more cost effectively updated in the field would be sent to the field while addresses from blocks verified or updated at headquarters could then be sent directly to the mail-out operations. This adaptive design process may substantially reduce the number of blocks requiring field operations, possibly enough to eliminate generous portions of the indirect costs of the operations (e.g., field offices, training programs, contracting). Very importantly, to implement such a suggestion would require significant research in the relative costs of the various candidate headquarters operations and their operational quality to ensure that the most cost-effective ordering of operations would be implemented.

5. Conclusion

To date the modeling of 2010 AC outcomes has had considerable success in developing models that efficiently predict where AC add outcomes might occur. This modeling endeavor has been less successful at dealing with identifying where HU address deletes might occur. This raises the issue of what to do about possible overcoverage of HUs

resulting from this inefficiency[3]. The extant modeling endeavors also do not include empty blocks in their study universe leaving the "Empty Block Problem" for future research. Other issues that are necessary to deal with, before a cost effective TAC operation can be fielded in 2020, include the level of clustering of targeted blocks, integrating the selective AC results with other census operations, and validating and verifying the reduced AC research results with more recent data. In this paper, we laid out and discussed some possible solutions to these issues. Further, the discussion here suggests some directions for continuing research to ensure an efficient operationalization of a statistical modeling based TAC solution for the 2020 Census.

Some of the areas suggested by this discussion for further research include:

Clustering of targeted blocks
Heterogeneity of addresses within blocks
Indicators of potential growth into previously unpopulated areas
Better predictors of address AC outcomes and MAF errors, e.g., address level indicators of socioeconomic status
GSSI operations suitable for replacing field work for targeted blocks
Between block cost heterogeneity

Not discussed here are a host of other issues that need to be addressed in addition to the ones examined here. The specifics of redesigned field operations to do a reduced AC workload efficiently as well as developing a cost estimation model that accurately represent these redesigned operations are two possibly important areas not discussed in this paper.

Acknowledgements

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

[3] 2000 Census Tabulation Blocks routinely contained both adds and deletes. 2010 Census Tabulation Blocks, however, do as often contain both adds and deletes in the same block. This means that any model that reduces both undercoverage and overcoverage to acceptable levels will likely increase the number of targeted blocks sufficient to raise doubts about the cost reduction returns of a TAC operation.

6. References

Boies, John L. and Christine Gibson Tomaszewski. 2014 (Forthcoming). Fielding a Targeted Address Canvassing Operation: Alternative Approaches to Moving from Predictive Statistical Modeling to a Cost Effective Address Canvassing Field Operation for the 2020 Census. In JSM Proceedings, Statistical Computing Section, Alexandria, VA: American Statistical Association.

Boies, John L., Kevin M. Shaw, and Jonathan Holland. 2013. Model-Based Targeted Address Canvassing: A Simulation Based on the 2009 Address Canvassing Program. In JSM Proceedings, Statistical Computing Section, Alexandria, VA: American Statistical Association.

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

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

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

Mazur, Christopher and Ellen Wilson. 2011. "Housing Characteristics: 2010," 2010 Census Briefs 7, October 2011. http://www.census.gov/prod/cen2010/briefs/c2010br-07.pdf

Prevost, Ron and Charlene Leggieri. .1999. "Expansion of Administrative Records Uses at the Census Bureau: A Long-Range Research Plan." Presented at the November 1999 Meeting of the Federal Committee on Statistical Methodology, November 1999.

Pritts, Mary and Nancy Johnson. 2014. Designing an Adaptable Database for Model-Based Research. . In JSM Proceedings, Statistical Computing Section, Alexandria, VA: American Statistical Association.

Tomaszewski, Christine Gibson, and John L. Boies. 2014 (Forthcoming). Recent Advancements in Statistical Modeling to Identify Address Updating Areas for the 2020 Census. . In JSM Proceedings, Statistical Computing Section, Alexandria, VA: American Statistical Association.

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

United States Bureau of the Census. 2009. "Introduction to the 2020 Census," Washington, D.C., June 18, 2009.

United States Bureau of the Census. 2010. “Strategic Plan for the 2020 Census,” Washington, D.C., Version 2.0, September 30, 2010.

United States Bureau of the Census. 2014. Building Permits Survey.

Walker, Shelley, Susanna Winder, Geoff Jackson, and Sarah Heimel. 2012. “2010 Census Program for Evaluations and Experiments: 2010 Census Nonresponse Followup Operations Assessment.” DSSD 2010 Census Planning Memoranda Series No. 190. https:// www.census.gov/2010census/pdf/2010_Census_NRFU_Operations_Assessment.pdf