An Overview of Census Bureau Efforts to Assess Address List Coverage and Quality

Robin A. Pennington, Kevin M. Shaw, Michael Ratcliffe
U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233

Abstract:
Address Canvassing was an operation in advance of the 2010 Census that created the address list for the census in most parts of the United States. Analysis of Address Canvassing results showed that much of the work from the operation resulted in no changes to the address list. The Census Bureau hopes to identify in advance of Address Canvassing for 2020 which areas require address list updates. This is a complex modeling and operationalization problem. This paper will outline some of our plans for this effort.

Keywords: 2020 Census, Statistical Modeling, Reduced Workload

1. Introduction

From a certain perspective, a project to determine where housing is changing and where it is not seems an easy one, with obvious answers. Everybody knows where changes are happening in their cities or towns. Surely this information could be obtained somehow to determine where the address list needs to be updated in advance of the 2020 census. It's the details that make this problem much more complex than it initially appears.

The overall goals of our efforts are both to determine the appropriate methodology to identify geographic areas with poor address coverage and to test and operationalize that methodology to support a decennial census. The ideal outcome is a reengineered canvassing operation for the 2020 Census that realizes a sizable decrease in both workload and cost over the prior census, while maintaining an acceptable level of accuracy.

2. Background

We will start with a little census history that is relevant for this research topic. It has only been since 1970 (Census History) that an address list was created for mailing questionnaires in the decennial census, rather than having everyone enumerated in person at their doorstep, so to speak. The first time that the address list created for the previous census was used as the starting point for the subsequent census was when the 1990 Address Control File became the basis of the Master Address File (MAF), which was used as the building block for the Census 2000 address list. The MAF has been maintained since that time by the Census Bureau's Geography Division. For Census 2000, areas designated for the Mailout/Mailback method of questionnaire delivery and response were defined using geographic characteristics. In general, urban areas were areas where questionnaires were mailed according to the address list developed for that census. The areas outside those areas were included in other methods of address list creation or update and delivery of questionnaires or enumeration.

Shortly after Census 2000, the Census Bureau's Geography Division merged the MAF and the mapping database, known as the Topologically Integrated Geographic Encoding and Referencing System (TIGER), into one connected database. Based on some lessons

The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. Note: Substantial portions of this work first appeared as contributions by the authors to internal Census Bureau documents. The text may appear in future publicly available Census Bureau documents.

learned from Census 2000, a monumental effort to update this MAF/TIGER database (MTdb) was undertaken in advance of the 2010 Census. This work primarily updated and corrected the maps, which in turn should have made the address list more accurate, if not more complete. The updated MTdb was the starting point for the 2010 Census Address Canvassing operation, which created the enumeration universe for the 2010 Census by having field staff canvass most of the ground in the United States. The Address Canvassing listers updated the address list and collected map spot coordinates for structures containing residential units. The coordinates were collected by Global Positioning System (GPS) when possible.

This decade the Geography Division has undertaken an initiative to work more closely with governmental partners to obtain updates between 2010 and 2020. This project is called the Geographic Support System Initiative (GSS-I). At the same time, various evaluations emanating from the 2010 Census showed that in many areas of the country, Address Canvassing listers made no updates in the field other than collecting the map spot coordinates. In particular, almost 30% of blocks had no updates from this operation other than the collection of GPS coordinates (Boies, Shaw, Holland). If only adds to the address list are considered, the percentage of blocks with no actions from Address Canvassing goes as high as 74.6% (Boies, Shaw, Holland). Future changes in the map spots are not expected for units that were mapped well in 2010 Census operations, particularly those that were given GPS coordinates. Therefore, it is believed that much of the country does not require the same level of updating that was applied in the 2010 Census. Preliminary research suggests that it will be possible to pre-identify accurately those areas that require updating.

Companion papers to this overview describe the work on this research objective that is being done under the auspices of the 2020 Census Research and Testing program and the GSS-I (Pritts, Johnson; Tomaszewski, Boies; Boies, Tomaszewski; Young, Johnson). These efforts are coordinated but with some separate goals. Our testing efforts will determine how aligned the efforts will become in the future.

The objective of the 2020 Research and Testing project is the creation of a model that predicts errors in the MAF, known as the MAF Error Model (Young, Johnson). The general idea is to use other files and variables – whatever may be available, at any level of geography – to model and predict where the MAF will have coverage errors. This effort is complicated to a large extent by the fact that coverage error on the MAF changes over time and also changes with inputs to the MAF. For example, it is not enough just to predict housing unit growth, as the growth may be completely captured by the United States Postal Service and conveyed on their Delivery Sequence File (DSF) of addresses. Similarly, growth or decline may be captured by local governments that provide their files in a partnership effort with the Census Bureau through the GSS-I. Any determination of where canvassing should occur to rectify coverage error must go hand in hand with information about what sources may be available for capturing the change.

While the Census Bureau as an agency will determine the criteria and definition for poor coverage quality, we also have to be cognizant of how to translate these decisions into efficient and cost-effective field work that does not sacrifice quality. Therefore, the second major prong of this effort is determining how best to handle the areas where high coverage error is known or suspected. For areas that are very remote or isolated, the best means of accomplishing this task may be to leave the list updating until the time of the census itself and to combine the task with questionnaire delivery or with enumeration.

For other areas where errors on the address list could significantly impact the more costly follow-up field procedures, especially where adaptive design or other strategies are being used, it might be best to canvass to receive updates in advance of the mail-out of census materials. In still other areas, work with a trusted governmental partner may be sufficient for creating an updated address list in advance of a mail-out. The costs and benefits of these various strategies in relation to the coverage error model results need to be assessed.

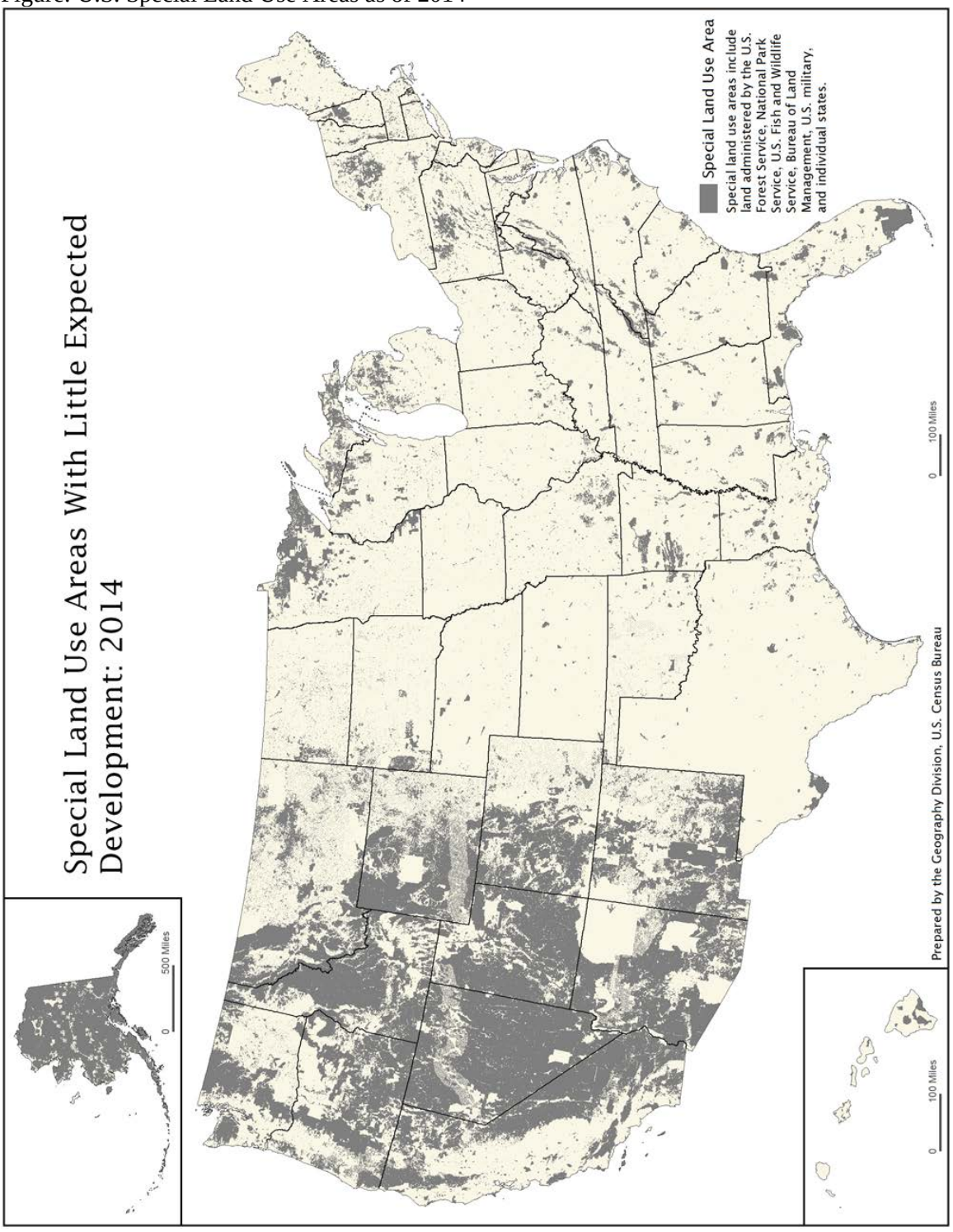
3. **Geographic Indicators**

Within the GSS-I, we have developed a continuum of how we expect to make decisions about which areas require canvassing. We use as major categories of predictors: Current State, Change Detection, Predictive Change, and 2010 Census Baseline. These are related to Quality Indicators (QIs) that the Geography Division is creating to assess quality of information in the MTdb. These defined QIs measure the quality of MTdb address and road data. They also measure its completeness by census tract and contribute to an overall assessment of MTdb data quality through the comparison of census tracts (and other geographic areas, as needed) based on their quality evaluation. The evaluation is used to identify geographic areas in which to focus partnership and update activities and contribute to a process for determining where to conduct a reengineered address canvassing.

Address QI scores are established through analyzing selected characteristics (known as sub-QIs) such as: the existence of a house number, street name, and ZIP Code; whether the address is used by the USPS for mail delivery; and whether the address can be accurately coded to an existing block within the MTdb. Sub-QIs are applied to each address, and the scores are aggregated to create an overall census tract score. A score is recalculated as updates are made to the MTdb. Similarly, road QI scores are built from characteristics at the individual road segment level and then aggregated for a score of each census tract. Address QI scores measure consistency of data as the means to demonstrate confidence in MTdb address data. The scores are high or increase when addresses in the census tract indicate a combination of factors such as multiple years of consistent DSF history, existence of a housing unit for an address, or completeness of city-style addresses with associated block codes. Scores are low or decrease when there is an inconsistent or short history of DSF updates for addresses in the census tract or when there are missing, inconsistent, or multiple locations assigned to an address. We also examined 2010 Census address counts compared to current DSF counts at the block and tract levels to create additional measures of stability across time.

The Census Bureau has also created some maps and other indicators as initial inroads into working this multi-pronged problem. In the following figure, we show a map of Special Land Use Areas where address updates are not expected. The areas shown on this map are U.S. Forest Service, National Park Service, U.S. Fish and Wildlife Service, Bureau of Land Management, U.S. Military grounds, and some additional areas identified by individual states. It may very well be that housing units exist on these lands, such as housing for staff who manage or oversee these areas, but the responsible governing authority may be able to provide this information.

Figure: U.S. Special Land Use Areas as of 2014



4. Statistical Modeling of Coverage and Coverage Errors

One of the primary goals of the 2010 Census Address Canvassing (AC) operation was to identify additions and deletions to the census address frame prior to the questionnaire delivery phase of the census, as these frame errors can have a direct impact on census coverage. As part of the 2010 Census Evaluations, Census Bureau statisticians began research on how regression models could be used to redesign the decennial canvassing operation. Using only data available prior to the 2010 AC operation, each census block was assigned a probability value from several binary logistic regression models denoting the likelihood of each block containing an erroneously excluded or included housing unit (i.e., an add or delete address action code from canvassing) (Boies, Shaw, Holland, 2012). With those modeled probabilities and the 2010 AC results on add or delete actions, the researchers evaluated the efficacy of their statistical models through a microsimulation.

Since 2010, various Generalized Linear Models have been studied as part of the Agency's Targeted Address Canvassing (TAC) research: Logistic (logit), and Negative Binomial and Poisson regression models. Logistic regression was chosen to model binary response data, while negative binomial and Poisson were chosen to model count response data. Given the high frequency of census blocks with no add or delete actions in the 2010 AC operation – the impetus for this research - zero-inflated versions of the Negative Binomial and Poisson models were developed to achieve better model fits. In the table below we present results from two of the top-performing logistic regression models available as of July 2014. Please see Tomaszewksi and Boies, 2014 for discussion of the derivation of these results.

The first logit model presented in this table, an Adds-only model, predicts the presence of two or more Add actions from the 2010 AC operation. As expected, this model performs better than the second model, an Adds and Deletes model, for capturing Add actions. For example, in 2009, at the 20 percent HU canvassing level (approximately 29.0 million HUs), the Adds model captures about 30 percent of all Add actions and 34 percent of Delete actions, while the Adds and Deletes model captures 25 percent of all Add actions and 41 percent of Delete actions. As just quantified, the Adds and Deletes model captures approximately an additional seven percent of Delete actions. Both models identify a very small percentage of blocks for canvassing – less than four percent of all blocks for both models - at each of the 5%, 10%, and 20% HU canvassing levels. These and other models have been assessed at numerous canvassing and coverage degradation levels. The current models have used available data as independent variables (predictors). In particular, available data refers to files and data that existed prior to the 2010 Census Address Canvassing and that were thought to be potentially related to address coverage. Additional variables are now being processed that reflect geographic aspects of addresses or blocks, such as proximity to recent housing change, stability of the block over time, and indicators of quality (e.g., locatability, mailability), which are anticipated to be even stronger predictors of change.

Table. Targeted Address Canvassing Statistical Modeling and Microsimulation Outcomes at selected Housing Unit Canvassing Levels using Census 2010 Tabulation Geography

Microsimulation Outcomes and Housing Unit (HU) Canvassing Level		Block-Level Statistical Models					
		Logistic Regression Models					
		Adds Model			Adds and Deletes Model		
In 2009, at a HU Canvassing Level		*Canvassing only the selected blocks yields …*					
Number of Housing Units[1] (millions)		Add Capture Rate[2]	Delete Capture Rate[3]	Percent Blocks Canvassed[4]	Add Capture Rate[2]	Delete Capture Rate[3]	Percent Blocks Canvassed[4]
5 Percent	7.2	9	11	0.3	6	14	0.2
10 Percent	14.5	17	20	1.1	13	25	0.6
20 Percent	29.0	30	34	3.2	25	41	2.2

[1]In total, there were approximately 145 million HUs eligible for canvassing in the 2010 Census Address Canvassing (AC) operation conducted in 2009. This is referred to as the dependent list count (U.S. and Puerto Rico).
[2]The Add Capture Rate refers to the percent of all Type A (true adds) and Type R (matched add) actions captured by canvassing *only* the selected blocks in 2009. The total number of Type A and R actions was about 10.8 million HUs.
[3]The Delete Capture Rate refers to the percent of all Type D (double deletes) actions captured by canvassing *only* the selected blocks in 2009. The total number of Type D actions was about 15.8 million HUs.
[4]The total number of 2010 Census tabulation blocks is about 11.2 million blocks (U.S. and PR).

Note: This table was originally presented in an internal Census Bureau document, but the results are derived from those presented in Tomaszewski and Boies, 2014.

5. Summary of Results

A key aspect of this research is measuring stability and consistency of housing unit and address counts as well as classifying census blocks based on land uses and types of housing units. In general, where stability and consistency can be identified between housing unit and address counts, there can be greater confidence in the MAF and in the ability to rely upon the DSF, local partner files, and other sources to update and maintain the MAF without having to canvass in the field. Identifying highly stable areas and areas with unique or non-residential land uses that can be maintained through in-office review and update helps narrow the types of geographic areas and number of housing units requiring canvassing in the field.

Address update processes have been implemented that are intended to minimize the areas where canvassing would be necessary by ensuring a complete, comprehensive MAF. These processes focus on partnering with local governments and researching other sources of addresses to update and maintain the MAF. Local government files also have

utility in maintaining and updating the TIGER road network and in coding new and existing MAF addresses to correct blocks.

Research on area classification demonstrates high levels of housing unit and address consistency throughout the United States. An automated comparison of the 2010 Census address list to the 2013 MAF demonstrated that housing unit counts in most census blocks in the nation have not changed between these two sources. The census blocks that have not changed account for a majority of the nation's housing units. Taken together with MAF updates using GSS-I Partnership Program files, results from a stability index developed for the DSF, and other activities, this suggests that many census blocks would not require canvassing in the field. Indications are that change detection through comparing the MAF to imagery and statistical modeling each can highlight areas to be canvassed or not canvassed. Initial results from the statistical modeling research demonstrate it presents a viable alternative to a full nationwide canvass. However, much modeling, integration and evaluation work remains. This methodology can only be further improved as a stand-alone methodology, and ideally made even more efficient, when integrated, tested and evaluated with other methodologies. We also note that this can contribute to significant cost savings. In summary, the identification of an optimal targeting solution for Address Canvassing is a prominent step towards a more efficient 2020 Census.

References

Census History at
http://www.census.gov/history/www/through_the_decades/overview/1970.html

Boies, John, Shaw, Kevin M., Holland, Jonathan: Address Canvassing Targeting and Cost Reduction Evaluation Report; www.census.gov/2010census/pdf/2010_Census_Address_Canvassing_Targeting_and_Cost_Reduction_Evaluation_Report.pdf 2010 Census Program for Evaluations and Experiments; Last accessed July 25, 2014

Boies, John, Tomaszewski, Christine; Fielding a Targeted Address Canvassing Operation: Alternative Approaches to Moving from Predictive Statistical Modeling to a Cost-Effective Address Canvassing Field Operation; JSM Proceedings 2014

Pritts, Mary, Johnson, Nancy; Designing an Adaptable Database for Model-Based Research; JSM Proceedings 2014

Tomaszewski, Christine, Boies, John; Recent Advancements in the Use of Statistical Modeling for Targeting Specific Geographic Areas for Address Canvassing in the 2020 Census; JSM Proceedings 2014

Young, Derek, Johnson, Nancy; Zero-Inflated Modeling for Characterizing Coverage Errors of Extracts from the U.S. Census Bureau's Master Address File; 2014 Submitted