

## **Master Address File (MAF) Model Validation Test (MMVT) Component of the Address Validation Test**

The Address Validation Test has 2 components—the MMVT and the Partial Block Canvassing (PBC) component. The purpose of this test is to collect address data in a dependent address listing operation for a sample of areas. We will use these address data to help determine the 2020 Address Canvassing workloads needed for the operational design decision.

### **Background on the MMVT in the Address Validation Test**

The purpose of the MMVT in the Address Validation Test is to assess our ability to use statistical modeling (1) to measure error in the MAF during the decade, and (2) to identify areas for the 2020 Address Canvassing operation. In the test, we will locate, update, add, or delete addresses in a dependent address listing operation for a national sample of blocks, excluding Alaska, Hawaii and Puerto Rico.

The project associated with the MMVT in the Address Validation Test is Project 3.101, MAF Error Model and Targeted Address Canvassing. In addition to the MMVT component of this test, the Partial Block Canvassing component in the Address Validation Test uses geographic approaches for identifying areas experiencing change not reflected in the MAF, and to identify portions of blocks where change is likely, thus enabling canvassing of partial blocks. The purpose of both components is to develop a plan for Address Canvassing for the 2020 Census. Information about the PBC component is found in other documents.

The statistical models to be tested are of two types, as described below. Each type works with available auxiliary data (MAF or DSF data, administrative records, geographic imagery data, economic summary statistics, etc.) to try to predict a measure for each block in the country. The measures are different, as the two approaches have different purposes.

1. To measure error in the MAF, we use a distributional model that estimates the quality of the MAF at any point in time. It predicts the number of "adds" (or deletes, changes, moves, etc.; or some combination of these potential errors) in a block. Based on the auxiliary data, it fits a statistical distribution (such as a negative binomial distribution) to represent the distribution of these adds. From this distribution, we can predict the average (expected) number of adds, an expected range of adds, the probability that the block will have, say, at least 10 adds, etc.
2. To identify areas that will most need canvassing in the 2020 Census, we use a logistic regression model to suggest (1) what proportion of the blocks we should target (include in the canvas), and (2) which blocks they are. The model predicts the probability that a block will have, e.g., at least five adds. Based on the auxiliary data, it uses a logistic regression model to assign a "propensity" or predicted probability of having at least five adds. With a propensity assigned to each block, we can order the blocks from highest (most likely to have five adds; we want to canvass these blocks) to lowest (least likely; we do not want to canvass these blocks). Then we can decide on a threshold propensity (above: canvass; below: do not canvass) based on cost, the loss in quality from the blocks we do not canvass, or some other criteria.

Each approach comprises an entire set of models. For example, for the second set of models, one could develop a model that assigns probabilities that the block has at least five adds, or at least one add, or at least one delete. Further, different input variables yield different models. It should be noted that each type of model--distributional or logistic regression--could be used for each of the stated purposes of the test, that is, measuring the quality of the MAF, and identifying areas to canvass.

To assess the performance of the models, we will conduct appropriate statistical analyses. For the distributional models, among other analyses, we will compare the number of errors observed within each block in the field test to the expected number based on the model. For the logistic regression models, we will study the results of the predictions. For example, for a specified range of predicted error propensities, in what percent of blocks did the predicted error occur? For both types of models, we will assess their performance predicting blocks that would be selected for a targeted canvassing given a specified level of targeting. For the set of blocks targeted by a model, we will tabulate and compare the level of errors captured in those blocks, as well as the level of errors found in the complementary set of blocks.

During the test, we will collect operational data as well to help estimate the cost of conducting the address canvassing.

This component of the Address Validation Test will thus provide information to help address two of the basic high-level questions as we plan the address canvassing operation before the 2020 Census, specifically,

- What level and location of targeted address canvassing will we undertake?
- How can we best balance cost and quality associated with a targeted address canvassing?

## **Sample Design**

When planning the sample, we stratified all blocks according to size, that is, the number of valid addresses in the block based on the MAF. Research has demonstrated that in the 2009 Address Canvassing operation, block size was one of the better predictors of the number of adds and deletes found in the block. We then sampled from all the strata, except for the stratum containing blocks with no addresses. In an attempt to include more blocks with adds and deletes, we oversampled (relative to proportional sampling) within the several strata of larger blocks, and under sampled within the strata of smaller blocks. In all, we selected 10,000 blocks from these strata, which includes an estimated 1.04 million addresses.

To avoid expending large resources on blocks with no addresses, where we expect to find few adds, we added to the sample a small number (100) of blocks with no addresses that were adjacent or close to blocks already selected for sample. We are currently conducting additional research to provide information on the treatment of blocks with no addresses in an address canvassing operation.

## **Procedures for Data Collection**

Field work for the MMVT will start on September 2, 2014 and end on December 19, 2014.

We will use the current general procedures of the Demographic Area Address Listing (DAAL) program to conduct the field work in the MMVT. The DAAL program is a dependent address listing operation that verifies, updates, adds or deletes addresses in selected blocks in the MAF for the national, demographic surveys that the Census Bureau conducts.

We will maximize, to the extent possible, the use of experienced field representatives (FR) — those who have worked or are currently working in the DAAL program— to conduct the field work in the MMVT. However, we will hire temporary staff to work in areas for which a current FR is not assigned. The existing staff will use their current laptops loaded with the Address Listing and Mapping Instrument (ALMI) to conduct the work for the MMVT. We will provide the newly hired staff a refurbished laptop to list addresses in the field. The staffing plan calls for approximately 640 experienced DAAL FRs, 320 FRs that have demographic survey experience but are new to DAAL, and 210 new hires to the DAAL program.

FRs will follow the same general procedures currently used in the DAAL program, with slight modifications to address some of the limitations with using the ALMI for the purposes of this operation. The ALMI allows the user to take certain actions that are not applicable for this test. As an example, the ALMI launches the Group Quarters Automated Instrument for Listing (GAIL) module to conduct the enumeration of a newly discovered Group Quarters, which for this test is not necessary. The training covers these types of modifications to the procedures.

The quality check for the MMVT follows the same procedures as Listing Check in the DAAL program. The goal of Listing Check is to verify FRs completed blocks accurately. A separate FR staff from production will conduct Listing Check. Experienced DAAL FRs will conduct the majority of the Listing Check workload. However, where we do not have an experienced DAAL FR in an area, we will use separate new-to-DAAL experienced FRs or new hires to conduct the quality check. This only occurs in limited areas of the country.

The plan for Listing Check is to randomly select and check one completed block for each FR working in the operation. The expected workload will be approximately 1,210 blocks, which is the equivalent of the estimated number of FRs working in the operation. This number may increase or decrease depending on the number of FRs used in the operation. For blocks with fewer than 35 housing units, the plan is to check all units. For blocks with 35 or more housing units, to design workloads efficiently, the plan is to check 35 of them. If a block (1) had no addresses prior to the MMVT data collection and (2) has no addresses added by the FRs, then this block is not eligible for Listing Check.

## Project Schedule

Operation	Start Date	Finish Date
Production Listing	September 2, 2014	December 5, 2014
Listing Check	September 12, 2014	December 19, 2014