

# 2015 Address Validation Test

Patrick J. Cantwell and Michael R. Ratcliffe

Presented at

Census Scientific Advisory Committee  
Fall Meeting

September 18, 2014

# Outline

1. A Plan for Address Canvassing
2. Information on part 1 of the Address Validation Test (AVT): the MAF Model Validation Test
3. Geographic activities
  - Imagery-based analysis
  - Partial Block Canvassing (PBC)
4. Information on part 2 of the AVT: PBC

# A Few Terms Used in This Presentation

- Master Address File (MAF)
- Delivery Sequence File (DSF)
- Address canvassing

# Six Steps of Address Canvassing

## In-Office Work

1. For specific special areas, apply alternative procedures
2. Automated: Use geographic and statistical approaches separately to pare list of blocks
3. Clerical: Integrate two approaches; which blocks need in-field work?
4. Consider geographic proximity; update list
5. Partial block canvassing, imagery: Determine, for each block on list, whether to canvass all or part of block

## In-Field work (6.)

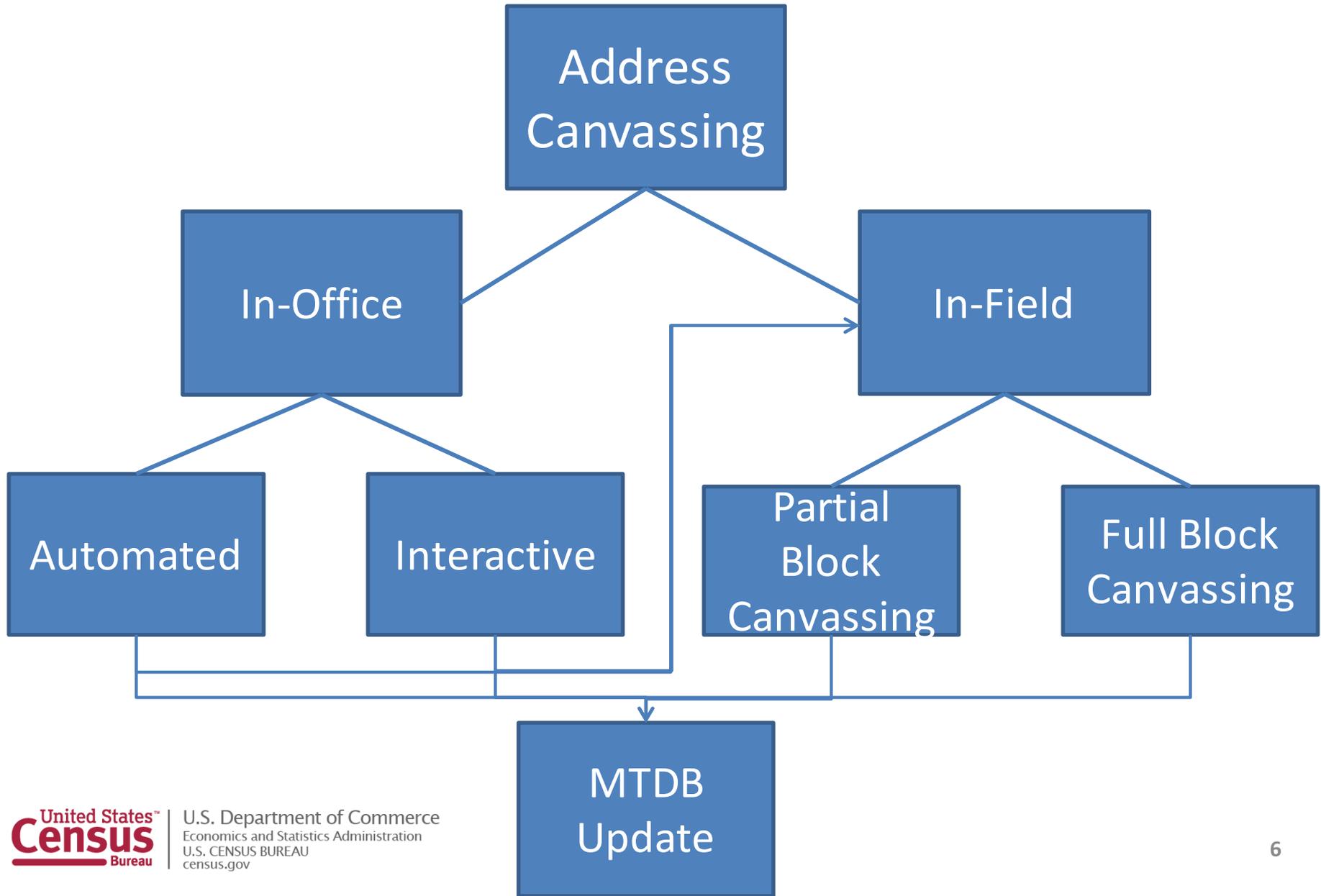
# During the Decade, Various Options

Between now and 2019, the six steps could be similar, but Step 3 can be broadened into several options:

Assign each block (or group) to one option:

- Do nothing now.
- Possible change? Check after the next DSF update.  
Continue to monitor; action later?
- Change likely? Pursue through in-office activity.  
Pursue through the partnership program.
- Change very likely? Send to field; in-field activity.

# A proposed view of address canvassing in years leading up to the 2020 Census



# Address Validation Test (AVT): Purpose

## Two parts to AVT

Part 1. Full-Block Canvassing. Assess our ability to use statistical modeling . . .

- to measure error in the MAF
- to define the address canvassing workloads

Part 2. Partial-Block Canvassing (PBC) Approach

- Investigate how the review of aerial imagery and other sources can work with statistical modeling
- More coming, Mike's slides

# AVT, Part 1: Field Implementation

- In the field Sept. 2 – Dec. 18, 2014
- National sample of 10,100 blocks,  $\approx$  1.04 million housing units; stratified by size of block, with oversampling of larger blocks
- Verify, update, add, or delete addresses on the dependent list
- Compare predictions from statistical models to results in the field

# Geographic Activities

- Imagery-based methods
- Partial block canvassing

# Partial Block Canvassing and Imagery-Based Analysis

- The Partial Block Canvassing component of the Address Validation Test relies upon a predecessor operation using imagery to detect change, compare numbers of housing units visible in imagery to numbers of addresses in the Master Address File (MAF).
- In this portion of the presentation, we will discuss the in-office, imagery-based review, and then Partial Block Canvassing.

# Targeted Address Canvassing Research, Model, and Area Classification (TRMAC) Goals

- Identify geographic areas that do not need to be canvassed in the field (i.e., in-office improvement and validation) and areas for inclusion in the in-field canvassing universe.
- Focus of the in-office effort is on decreasing the in-field canvassing by identifying areas of stability and as well as areas in which the Master Address File (MAF) can be updated and maintained using the USPS Delivery Sequence File (DSF), local government address sources, and other sources.
- Identify, obtain, and manage data needed to support this activity and related review and decision-making efforts through the decade.

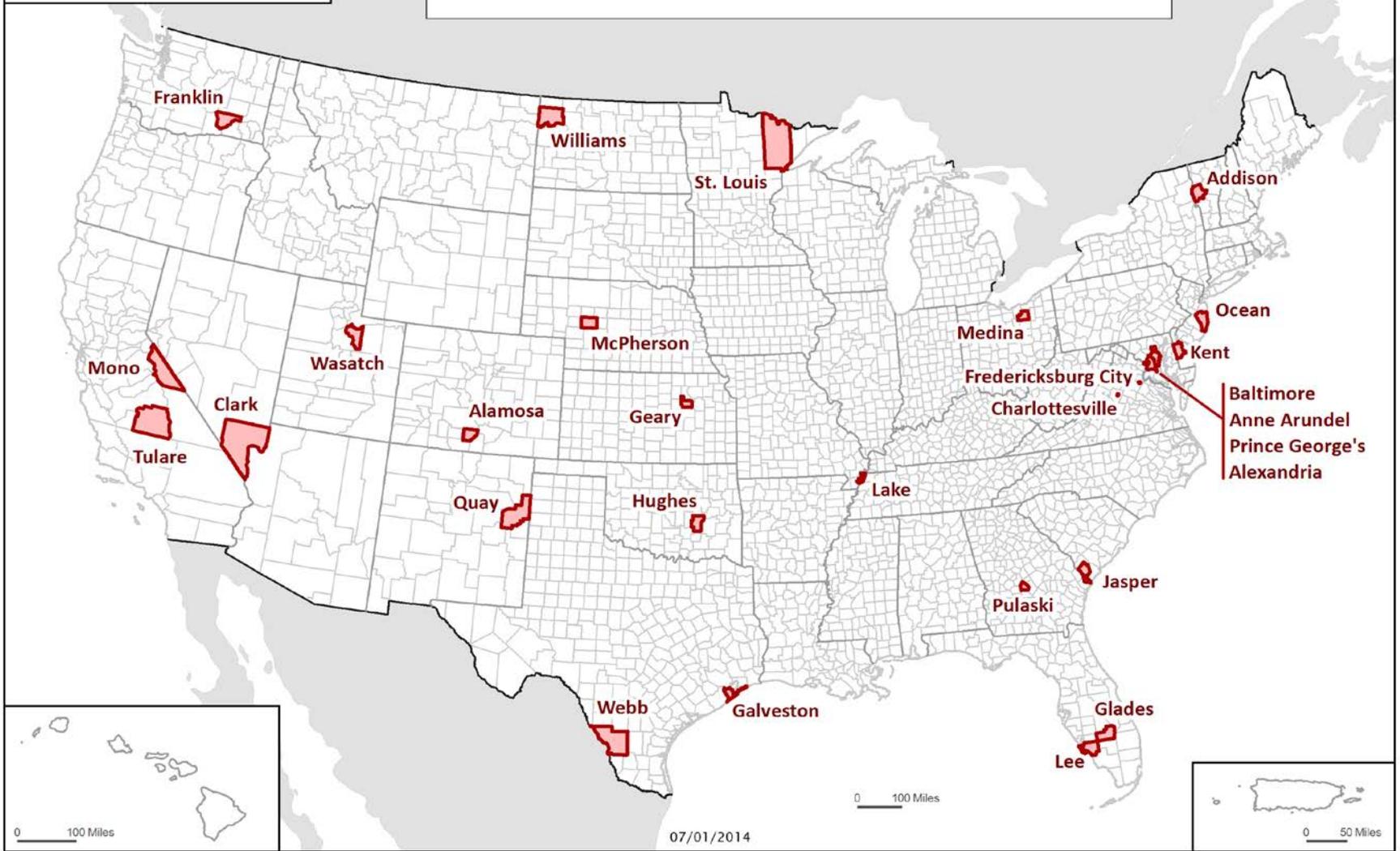
# TRMAC Interactive Review Processes

- Assess current imagery to 2010 vintage of imagery to identify change
- Assess current imagery in comparison to current housing unit information to identify coverage & geocoding issues
- Assess imagery (and parcels, if available) for likelihood of stability & future change
- Identify obvious errors in our data

# MAF-to-Imagery Review Pilot Project, Spring 2014

- 21,924 blocks (0.2% nationally) were reviewed
  - 11,286 reviewed twice to compare results between individual reviewers
- Review occurred in 29 counties selected for several characteristics, including:
  - Geographic Support System-Initiative (GSS-I) participation
  - MAF housing unit change
  - Population Estimates Program housing unit change
  - MAF Model Validation Test (MMVT) blocks
  - special land uses
  - urban vs. rural

# TRMAC Interactive Review Test Counties



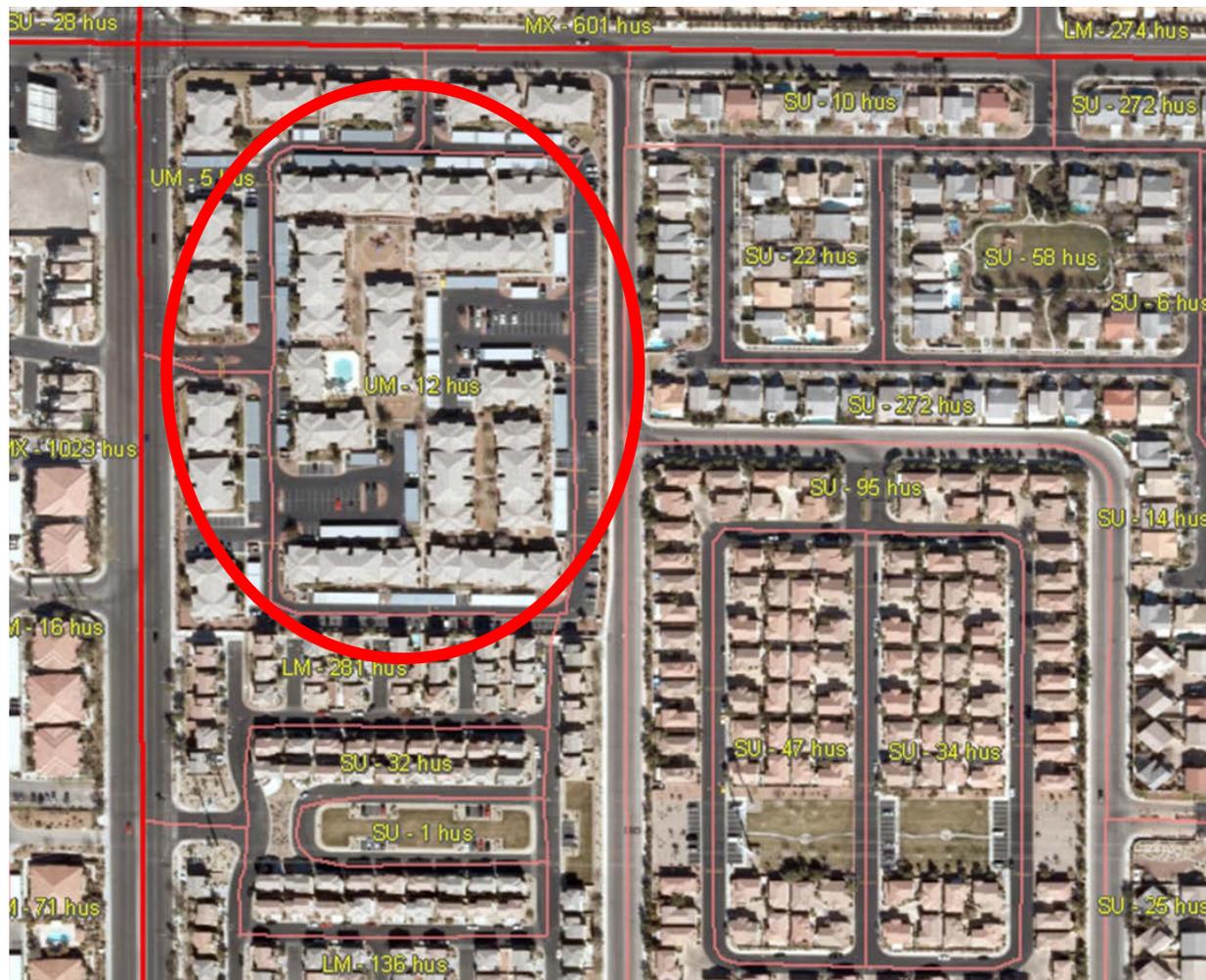
# Stability: Built Out Blocks

- 69% of reviewed blocks were classified by the reviewers as “built out”
- Assuming the same pattern applied to all blocks nationally this would be roughly 7,697,000 blocks



# Blocks with general data errors

- 8.2% of reviewed blocks had general data errors – where the MAF counts didn't match what was on the ground. In this example, the MAF reported 12 housing units for the highlighted block. The block contains 12 multi-unit buildings.
- Skewed toward easy to observe errors, and blocks with few housing units (often small blocks)



Imagery can be used to detect change, but does not necessarily indicate the full extent of change. We can see that what may be a multi-unit building has been built on the site of a former parking lot. We cannot know from imagery, however, the number of units in the building or whether it's residential, commercial, or both.



# Observed change: AdCan to Current Tracked vs Untracked in the MAF

	IR blocks	Percent of IR blocks	Assuming same pattern applied nationally...
Stable: Both IR observed and MAF contained no change	9,289	82%	9,148,000
Observed change in IR	664	6%	669,000
Tracked in the MAF	333	3%	335,000
Not tracked in the MAF	331	3%	335,000
MAF contained change and it was not observed in IR	1,133	12%	1,339,000

# Partial Block Canvassing

- Component of the Address Validation Test
- Fieldwork planned for December 2014
- 600-1,000 blocks
  - If possible, include blocks that also are in the MMVT 10,100 block sample
  - Blocks identified through MAF-to-Imagery Comparison process as well as other GEO review operations
- 25-30 professional staff (mostly geographers)
- Will use the Census Bureau's corporate Listing and Mapping Application (LiMA)

# Partial Block Canvassing Example: Imagery review identifies discrepancy between MAF and imagery; updates are clustered in a portion of the block



**Not all blocks are candidates for partial block canvassing. In this example, new units are scattered throughout the blocks.**



# How Blocks are Selected for PBC Test

- In-office review identifies inconsistencies between numbers of addresses in the MAF and housing units visible in imagery.
- Blocks in which updates are clustered in a relatively small area are candidates for PBC.
- Note that initial findings from the in-office review indicate that the vast majority of blocks (over 80%) are not exhibiting change, and are consistent when comparing the imagery against the MAF, and therefore would not need to be canvassed through either a full or partial block canvassing operation.

# Partial Block Canvassing Test Objectives

- Test ability to navigate to targeted area/coordinate using locational information produced by TRMAC in-office review staff.
- Collect specified information for use in comparison to information collected for the same block through full block canvassing.
- Collect metrics to measure efficiency, cost, etc.
- Identify potential issues affecting ability to conduct fieldwork and collect accurate information:
  - Is imagery required? What other tools/data are needed?
  - Should updates other than those specified be collected?
  - How do we limit the scope of work once in the field?

# Partial Block Canvassing Test Analysis

- Analyze results from fieldwork in comparison to expectations based on in-office review.
  - Did we collect the information we expected?
  - Did we find additional updates in the field?
- Compare results against GSS-I partner provided address updates (adds, changes, etc.) to confirm validity.
- For blocks in both the partial block and full block canvassing, compare results and assess reasons for differences, if any.
  - Did full block canvassers find additional updates, especially any that might not be detectable through in-office review? E.g., hidden units and E911 conversion changes.
- Prepare report, including recommendations relating to potential implementation.

# Questions for Committee

1. Through the decade, how can we properly assess new statistical models, when we won't have the "truth" until after the Address Canvassing operation in 2019?
2. How can we use economic data (e.g., changes in land values; data citing development) to predict change?
3. For partial block canvassing, should we consider blocks in which changes are clustered in several parts? Might we traverse the entire block, but canvass only part?

# AVT, Part 1: Field Implementation (cont.)

- Using current survey Field Representatives (FRs)
  - Experienced DAAL FRs, ≈ 640 FRs
  - Experienced FRs new to DAAL, ≈ 310 FRs
  - New hires, ≈ 240 FRs
- Listing Check – quality check of production work
  - Check one block for each FR
  - Check up to 35 addresses per block
- Demographic Area Address Listing (DAAL) procedures
  - Slight modifications to procedures

# AVT, Part 1: Analysis of Data

## 1. Selecting Blocks for Address Canvassing

- For the 10,000 blocks in sample
  - Order the blocks by prediction, e.g.,  $\Pr\{\geq 2 \text{ “adds”}\}$
  - Consider levels of canvassing: e.g., 10%, 20%, 30%, 40% of (weighted) housing units
  - Identify blocks selected (above the stated threshold)
- Compare outcomes from the field with these results

## 2. Measuring Error in the MAF

- Evaluate matrix of prediction x outcome

# Statistical Models: Logistic

- Logistic regression models
- $p = p(\mathbf{X}, \boldsymbol{\beta}) = \Pr\{\text{some occurrence of error}\}$   
e.g.,  $= \Pr\{5 \text{ or more "adds" in a block}\}$
- $\log\left(\frac{p}{1-p}\right) = f(\mathbf{X}, \boldsymbol{\beta})$
- Errors, e.g.,  $\geq k$  adds,  $\geq k$  deletes,  
 $\geq k$  adds or deletes

# Statistical Models: Distributional

- Model based on a discrete distribution

$Y$  = no. of occurrences of some error

$p = \Pr\{Y = k; \mathbf{X}, \boldsymbol{\beta}\}$ , e.g., =  $\Pr\{Y = 5 \text{ “adds” in block}\}$

- E.g., Poisson, negative binomial; zero inflated:

$$Y' = \alpha 1_0 + (1-\alpha)Y, \quad 0 < \alpha < 1$$

- Errors, e.g.,  $k$  adds,  $k$  deletes,  $k$  adds or deletes

- From distribution, can summarize using

$E(Y; \mathbf{X}, \boldsymbol{\beta})$  or  $\Pr\{Y \geq k; \mathbf{X}, \boldsymbol{\beta}\}$ , etc.

# Selecting Blocks for Canvassing

1. Select a model
2. Determine a score (statistic, propensity) based on that model, as applied with  $X$ ,  $\beta$
3. Set a threshold, e.g., the number of addresses included in the canvas
4. Select (cumulate) blocks whose score is above the threshold

# Results of Statistical Modeling Applied to Data from 2009 Address Canvassing Operation

Source: Tomaszewski and Boies (JSM 2014)

Canvassing 20% of Housing Units	% of Blocks	Capture Rate of	
		Adds	Deletes
<p>“Perfect” Adds Model</p> <p><i>(Suppose we knew exactly where all the “adds” were <u>before</u> we canvass)</i></p>	5.3%	80.3%	30.3%

# Results of Statistical Modeling Applied to Data from 2009 Address Canvassing Operation

Source: Tomaszewski and Boies (JSM 2014)

Canvassing 20% of Housing Units	% of Blocks	Capture Rate of	
		Adds	Deletes
“Perfect” Adds Model	5.3%	80.3%	30.3%
Predictive Adds Model ( <i>logistic regr.</i> , $\Pr\{ \geq 2 \text{ “adds” } \}$ )	3.2%	30.3%	33.7%

# Results of Statistical Modeling Applied to Data from 2009 Address Canvassing Operation

Source: Tomaszewski and Boies (JSM 2014)

Canvassing 20% of Housing Units	% of Blocks	Capture Rate of	
		Adds	Deletes
“Perfect” Adds Model	5.3%	80.3%	30.3%
“Perfect” (Adds + Deletes) Model	2.3%	52.6%	51.1%

# Results of Statistical Modeling Applied to Data from 2009 Address Canvassing Operation

Source: Tomaszewski and Boies (JSM 2014)

Canvassing 20% of Housing Units	% of Blocks	Capture Rate of	
		Adds	Deletes
“Perfect” (Adds + Deletes) Model	2.3%	52.6%	51.1%
Predictive (Adds + Deletes) Model	2.2%	25.2%	41.4%

# Results of Statistical Modeling Applied to Data from 2009 Address Canvassing Operation

Source: Tomaszewski and Boies (JSM 2014)

Canvassing 20% of Housing Units	% of Blocks	Capture Rate of	
		Adds	Deletes
“Perfect” Adds Model	5.3%	80.3%	30.3%
Predictive Adds Model	3.2%	30.3%	33.7%
“Perfect” (Adds + Deletes) Model	2.3%	52.6%	51.1%
Predictive (Adds + Deletes) Model	2.2%	25.2%	41.4%

# Notes on Results of Statistical Modeling

1. Modeling with data from 2009 Address Canvassing operation; changes since then
2. Did not remove Federal lands, non-MO/MB
3. Did not have access to variables related to some geographic concepts, e.g., nearby growth, stability, address quality indicators
4. Have not yet determined “best” statistical models

# Potential Improvements to Models

1. Geographic variables: growth, stability, etc.
2. Blocks with no addresses
3. Other types of statistical models
4. Address-level models, especially for predicting “deletes”
5. Puerto Rico