

# Algorithms for Including Administrative Data to Address NRFU Efforts

Presentation to the Census Scientific Advisory Committee

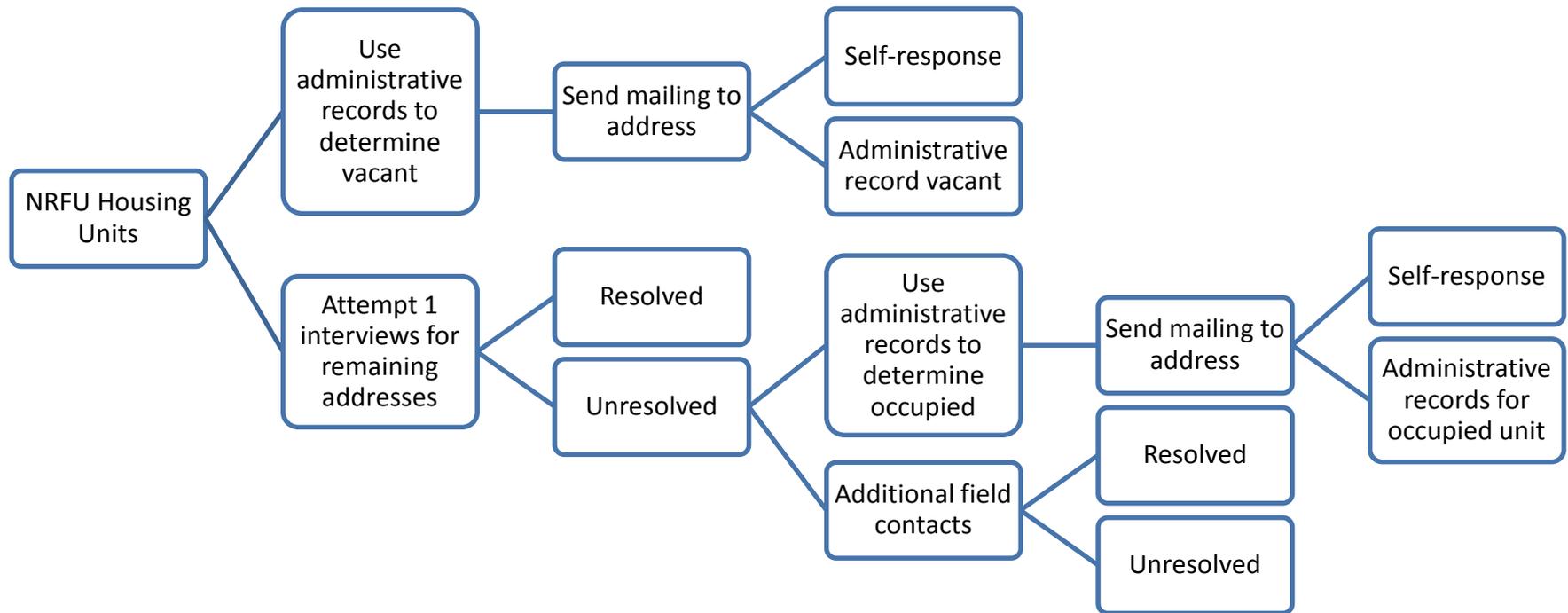
March 30, 2017



# Outline

- 2016 Nonresponse Followup (NRFU) contact strategy with administrative records (AR) determination
- Administrative Record Methodology for Occupied, Vacant and Non-Existent/Delete addresses
  - Highlight changes from November 2015 virtual meeting
- Characteristics for administrative record enumeration
- Research findings
- Proposed 2018 End-to-End Census Test NRFU Contact Strategy

# 2016 Contact Strategy



# Administrative Records Data Sources

## Vacancy Determination

- United States Postal Service Information
  - USPS Undeliverable-as-Addressed (UAA) reasons for census mailings made around April 1
    - Vacant
    - No Such Number, No Such Street
    - Others
  - Delivery Sequence File
- Other Sources
  - Internal Revenue Service (IRS) 1040 filings
  - IRS 1099 information returns
  - Medicare Enrollment Database
  - Indian Health Service Patient Database
  - Third-party Veterans Service Group of Illinois (VSGI) files

# Identifying AR Vacant Units

## Housing Unit Status Model

- Data: 2010 Census NRFU addresses
- 2010 Census status
  - 1: Occupied
  - 2: Vacant
  - 3: Nonexistent/Delete
- Explanatory variables:
  - UAA flag and reason (e.g., vacant, no such number) on first and second mailing
  - Consistency of UAA reasons by zip code
  - Indicators for presence of persons in AR sources at address
    - Indicators for presence of these persons at other addresses
  - ACS area-level estimates: % renters, % poverty, % black, etc.

# Identifying AR Vacant Units

## Distance Function

Which units to identify as vacant?

- Previously used optimization techniques to maximize vacant identification subject to constraints of misclassification rates
- Current approach uses Euclidean distance function to select cases with high vacant probability (near 1) and low occupied probability (near 0)

$$d_{vac} = \sqrt{(1 - \hat{p}_{vac})^2 + (0 - \hat{p}_{occ})^2}$$

- Given a specified threshold for the distance, all cases below that threshold are identified as AR vacant

# Identifying AR Vacant Units

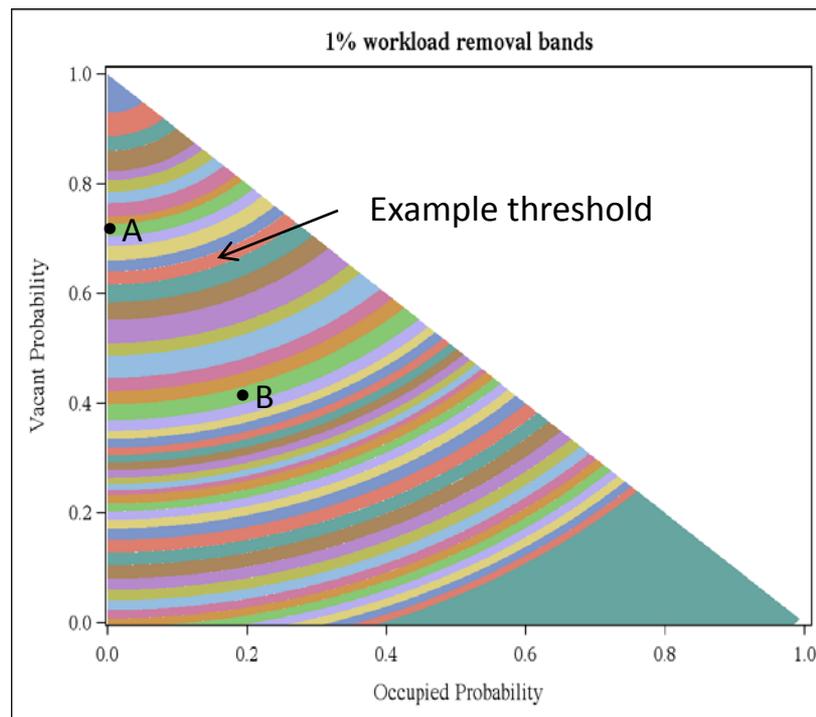
## Distance Function

The distance function can be visualized as successive bands of cases emanating from the point (0,1) in the top left corner

Each successive band represents an additional 1 percent of the NRFU workload

In this example, unit A is identified as AR vacant while unit B is not

Similar approach implemented for Non-Existent or addresses that need to be deleted



# Identifying AR Occupied Units

Can we reduce contacts for 101 Main Street?

1. Build a roster from administrative record sources
2. Check that multiple sources indicate the family lives at an address
3. Use statistical models to evaluate the roster
4. Decision for 101 Main Street
  - How likely is it that we are counting all of the people rostered from administrative records in the right place?
  - How likely is it that the household composition of the administrative record roster matches the Census?

# Identifying AR Occupied Units

## Data Sources

Core sources for occupied rosters:

- Internal Revenue Service
  - 1040 Individual Tax Returns
  - 1099 Information Returns
- Centers for Medicare and Medicaid Services Medicare Enrollment database
- Indian Health Service Patient database

Additional sources:

- Social Security Administration Number Identification (Numident) File
- Best Race and Hispanic Origin file from Center for Administrative Records Research and Applications (CARRA)
- United States Postal Service (USPS)
  - USPS Undeliverable-as-Addressed (UAA) reasons for census mailings made around April 1
  - Delivery Sequence File (DSF)
- Veterans Service Group of Illinois (VSGI) Name and Address Resource and TrackerPlus files

# Identifying AR Occupied Units

## Data Sources Continued

Additional sources being researched:

- Department of Housing and Urban Development
  - Computerized Home Underwriting Management System (CHUMS)
  - Public and Indian Housing Information Center (PIC)
  - Tenant Rental Assistance Certification System (TRACS)
- Selective Service System registration
- USPS National Change of Address (NCOA)
- Supplemental Nutrition Assistance Program (SNAP) and other state program participation data
- CARRA's Kidlink file
- CoreLogic Tax and Deed information

Will continue to research sources, and will finalize in September 2018

# Identifying AR Occupied Units

## Person-Place Model

- Binary logistic regression model
- Data: Person-place combinations across AR sources for 2010 Census NRFU addresses
- Dependent variable: 2010 Census status
  - 1: AR person-place pair is observed in 2010 Census
  - 0: AR person-place pair is not observed in 2010 Census
- Independent variables:
  - Properties of the address: AR count and composition, UAA, etc.
  - Indicators for presence of AR sources at address
  - Indicators for presence of AR sources at other addresses
  - ACS area-level estimates: % renters, % poverty, % black, etc.
- Model parameters applied to current vintage of AR data
- Minimum of person probabilities assigned to the housing unit

# Identifying AR Occupied Units

## Household Composition Model

- Multinomial logistic regression model
- Data: 2010 Census NRFU addresses
- Dependent variable: 2010 Census household composition
  - 0: Vacant/Delete (i.e., not occupied)
  - 1: 1 adult, 0 children
  - 2: 1 adult, 1+ children
  - 3: 2 adults, 0 children
  - 4: 2 adults, 1+ children
  - 5: 3 adults, 0 children
  - 6: 3 adults, 1+ children
  - 7: Other



# Identifying AR Occupied Units

## Household Composition Model

- Independent variables
  - AR household composition
  - Properties of the address: AR count, UAA, etc.
  - Indicators for presence of AR sources at address
  - Indicators for presence of AR sources at other addresses
  - ACS area-level estimates: % renters, % poverty, % black, etc.
- Model parameters applied to current vintage of AR data
- Predicted probability corresponding to the observed AR household composition assigned to the housing unit.

# Identifying AR Occupied Units

## Distance Function

- Which units to identify as occupied?
  - Previously used optimization techniques to maximize occupied identification subject to constraints of AR data quality
  - Current approach uses Euclidean distance function to select cases with high person-place probability (near 1) and high household composition probability (near 1)

$$d_{occ} = \sqrt{(1 - \hat{p}_{pp})^2 + (1 - \hat{p}_{hhc})^2}$$

- Given a specified threshold for the distance, all cases below that threshold are identified as AR occupied

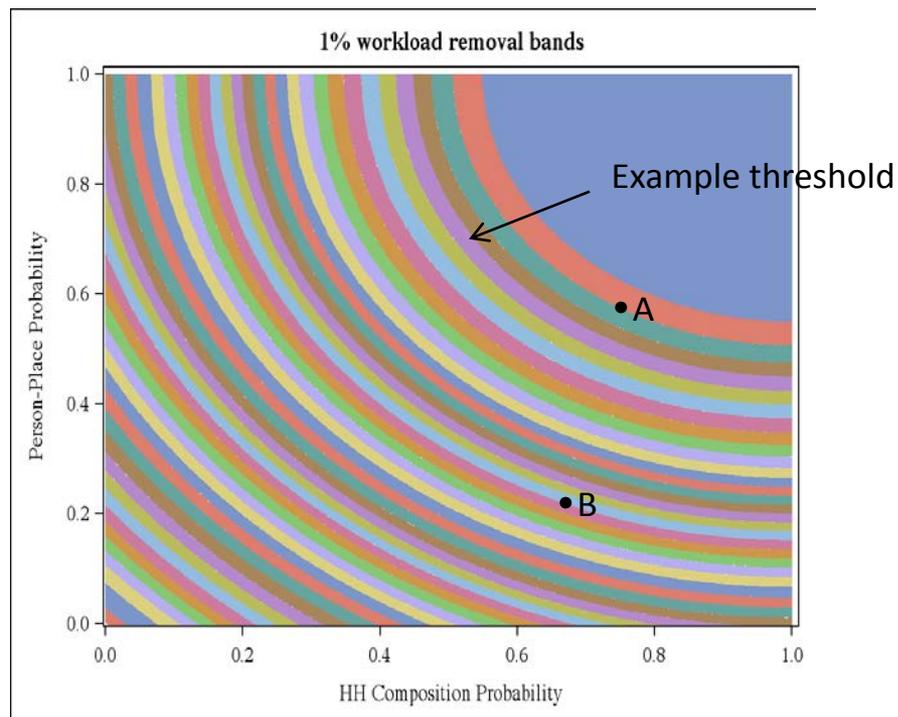
# Identifying AR Occupied Units

## Distance Function

The distance function can be visualized as successive bands of cases emanating from the point (1,1) in the top right corner

Each successive band represents an additional 1 percent of the NRFU workload

In this example, unit A is identified as AR occupied while unit B is not



# Characteristics

## AR Source Possibilities

### Age and Sex

- Census Numident

### Race and Hispanic Origin

- Prior Census, American Community Survey, or other Census Bureau responses
- Country of Origin (Census Numident)
- Census Bureau Best Race and Hispanic Origin file

### Relationship to Householder

- Census Bureau KIDLINK file

### Tenure

- Housing and Urban Development
- Tax and Deed Information

### Detailed Vacancy

- Housing and Urban Development

# Results

## 2016 Census Test Analysis

# 2016 Census Test

## Administrative record determination

- Occupied
- Vacant
- Delete

## 1-in-5 sample of administrative record cases received full fieldwork

- Compare our determinations to fieldwork results
- No address canvassing done before this test

# 2016 Census Test Results

## 1-in-5 Sample Evaluation Analysis

	Total	Occupied		Vacant		Delete		Unresolved	
	N	%	SE	%	SE	%	SE	%	SE
<b>AR Vacant</b>									
<b>Total</b>	715	21.1	1.3	42.8	1.6	20.7	1.2	15.4	1.2
<b>LA County</b>	36	19.9	2.3	43.2	2.9	22.5	2.3	14.4	2.1
<b>Harris County</b>	479	21.7	1.6	42.6	2.0	19.8	1.5	15.9	1.4
<b>AR Delete</b>									
<b>Total</b>	313	29.1	2.1	10.9	1.4	48.6	2.2	11.5	1.7
<b>LA County</b>	172	24.4	2.9	7.6	1.9	57.0	3.3	11.0	2.2
<b>Harris County</b>	141	34.8	3.2	14.9	2.1	38.3	2.7	12.1	2.5

# 2016 Census Test

## USPS UAA Reasons for AR Vacant and AR Delete Units Determined Fieldwork Occupied

### UAA Reasons for AR Vacant/Fieldwork Occupied Units

Reasons	Number	Percent
Vacant in both	86	57.0%
Vacant in one	29	19.2%
Any other reason twice	25	16.6%
Any other reason once	11	7.3%
Total	151	100.0%

### UAA Reasons for AR Delete/Fieldwork Occupied Units

Reasons	Number	Percent
No Such Number in both	26	28.6%
No Such Number in one	29	31.9%
Any other reason twice	27	30.8%
Any other reason once	9	9.9%
Total	91	100.0%

# NRFU Mailing for Administrative Record Cases

	AR Occupied (Phase 1)		AR Vacant		AR Delete	
	Count	%	Count	%	Count	%
<b>Total</b>	9,353		2,856		1,252	
<b>AR Mailing Sent</b>	8,418		2,848		1,252	
<b>UAA on AR Mailing Sent</b>	125	1.5	1,631	57.3	944	75.4

For 6 weeks after Census Day

AR Vacant were UAA 57.3 percent of the time.

People could have moved in since Census Day

AR Delete were only UAA 75.4 percent of the time

Seems less likely to change from not a housing unit to a housing unit

# Results

## 2010 Census Analysis

# AR Determination by Percent Hispanic Population in Block Group

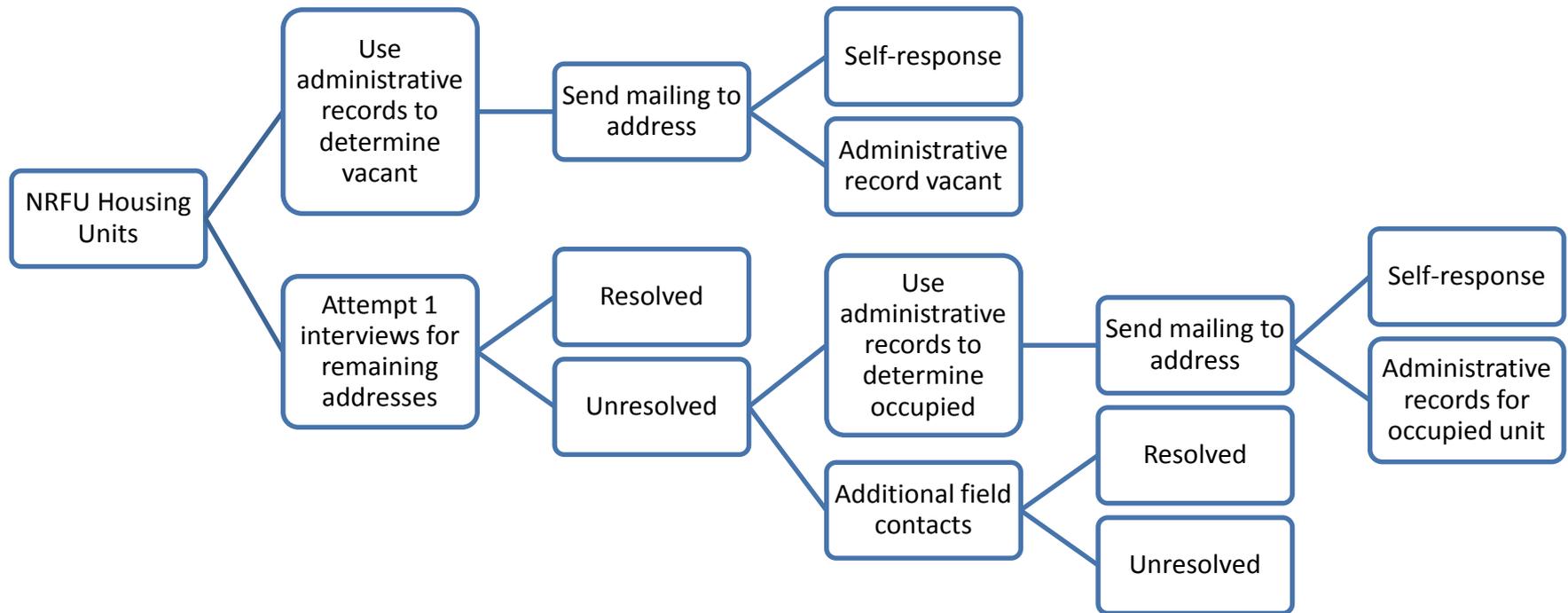
ACS 5-Year Estimate of Percent of Block Group that is Hispanic	2010 NRFU Addresses (millions)	AR Determination (row percent)		
		AR Vacant (%)	AR Occupied (%)	Full Contacts (%)
0 to 10 percent	31.3	11.6	15.9	72.5
10 to 20 percent	6.8	9.4	15.7	74.9
20 to 30 percent	3.6	8.5	14.4	77.1
30 to 40 percent	2.2	7.6	13.3	79.1
40 to 50 percent	1.6	7.5	12.0	80.5
50+ percent	4.2	4.3	9.6	86.1
Total	49.8	10.1	15.0	74.9

# AR Determination by Percent Non-Hispanic Black Population in Block Group

ACS 5-Year Estimate of Percent of Block Group that is Non-Hispanic Black	2010 NRFU Addresses (millions)	AR Determination (row percent)		
		AR Vacant (%)	AR Occupied (%)	Full Contacts (%)
0 to 10 percent	33.6	10.5	16.0	73.5
10 to 20 percent	5.4	8.2	15.6	76.2
20 to 30 percent	2.9	8.1	14.1	77.8
30 to 40 percent	1.8	8.3	13.0	78.7
40 to 50 percent	1.2	8.9	11.9	79.2
50+ percent	4.9	12.1	9.6	78.3
Total	49.8	10.1	15.0	74.9

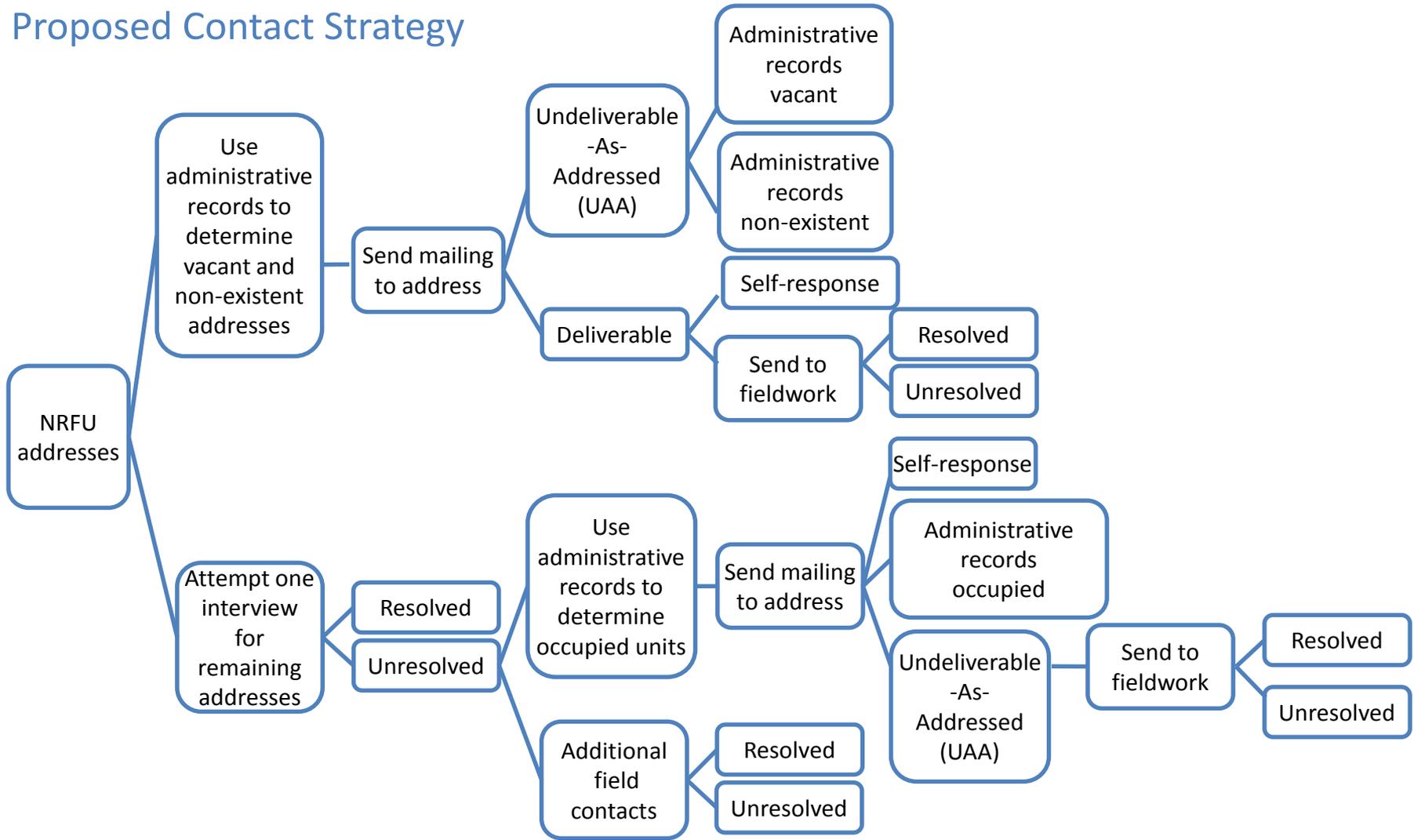
# Changes to 2018 Contact Strategy

# 2016 Contact Strategy



# 2018 End-to-End Census Test

## Proposed Contact Strategy



# Additional Analysis

- American Community Survey instead of 2010 Census for training data
- Distance function: alternatives to Euclidean
- Additional file analysis
  - Census Bureau CARRA Kidlink
  - Supplemental Nutrition Assistance Program and other state participation data

# Discussion

What are your reactions and suggestions for the current algorithm to identify occupied, vacant and non-existent addresses?

What is your reaction to the changes in the contact strategy to be used in the 2018 End-to-End Census Test?