

Census Scientific Advisory Committee

Administrative Records Working Group Draft Report

**Barbara (babs) Buttenfield, Convener
University of Colorado – Boulder**

30 March 2018

Working Group (WG) Context

- Census goals:
 - reduce costs of Non-Response Follow-Up (NRFU) for 2020 Census
 - maintain quality of data collection / reporting
- Administrative Records (ARs) drawn from federal, state, 3rd party data sources to create and populate roster of NRFU units
- Strategies include statistical modeling, added field contacts (if needed)

Why is AR Use Critically Important?

- Roughly 50 million NRFU addresses
(30M Occupied, 15M Vacant, 5M Non-Existent)
 - ARs can significantly reduce field staff, repeated visits
 - Potential \$1.4 billion reduction in follow-up data collection costs
- WG notes that statistical testing / simulations in recent years confirm the utility of AR use
- NFRUs for upcoming Census data collections may rise given the American public's shifting attitudes about government oversight

Uses for ARs in 2020 NRFU Operation

- Categorize as many NRFU addresses as possible, distinguishing Occupied, Vacant or Nonexistent, and minimizing error.
- Generate accurate roster of Occupied addresses using ARs and statistical modeling.
- Determine characteristics of individuals and household compositions as accurately as possible.

Current Sources for ARs

- **Social Security**
- **IRS 1040 and 1099 Forms**
- **CMS Medicare and Medicaid**
- **Indian Health Service**
- **CARRA Best Race and Hispanic Origin data**
- **US Postal Service** (UAA undeliverable list)
- State-level veterans, health and human service data (SNAP, **KidLink**)
- MLS, tax, deed and parcel data (CoreLogic)
- 2010 and ACS 5 year block group-level estimates

Working Group Members and SMEs

CSAC

Barbara Battenfield

Allison Plyer

Ken Simonson

Jack Levis

Krishna Rao

Barbara Anderson, ex officio

Census Subject Matter Experts (SMEs)

Tom Mule

Quentin Brummet

Andrew Keller

Moises Yi

Nicholas Jones

Scott Konicki

Larry Warren

Mark Leach

Deborah Wagner

Gary Chappel

Darcy Morris

Jennifer Ortman

Tommy Wright (Census Designated Federal Officer)

WG Activities

1. Consider effectiveness of various AR data sources
2. Streamline workflow for handling NRFUs
3. Categorize NRFU units (Occupied, Vacant, Non-existent)
4. Compare national with sub-national models
5. Assign demographic characteristics (e.g., race)
6. Use ARs to improve American Community Survey (ACS) data collection on population and housing (covered in previous presentation)

WG Activities

1. Effectiveness of AR Data Sources

- Accuracy and reliability criteria of state and local level data, guided by Title 13, Title 26, and FRPA
- Lack of (or unreliable) addresses caution against use of drivers licenses, birth certificates
- Highly variable availability of SNAP, TANF, WIC
- CoreLogic MLS data also considered (advantages and concerns discussed earlier)

WG Activities

2. Census Workflow for NRFUs

- Hierarchical strategy to build AR roster of NRFUs
 - Use ARs to distinguish Occupied from Vacant or Non-Existent
 - For unoccupied, follow-up mailing trying to identify Vacant
- Handling remaining unresolved addresses
 - Rely on ARs and mail to Occupied
 - Some physical follow-ups
- Added mailings, CAPI and CATI drive up costs

WG Activities

3. Categorizing NRFUS

- Statistical modeling to categorize Vacant / Non-Existent
 - Multinomial logistic regression assigns probability
 - USPS, IRS, CMS, IHS, ACS, 2010 decennial categorization
- Second regression for Occupied addresses
 - Regression assigns probability of correct address using same AR sources as above, plus SSA, CARRA, VSGI
 - Household composition uses ARs from ACS and 2010 census
- Model reliability varies with ACS MoEs, or covariates of income, mobility, household composition
 - Tests on 2015 ACS show that inclusion of covariates does not improve error rate of Unoccupied address removal.

WG Activities

4. National / Sub-National Models

- Single model to determine Occupied units effective for the entire nation?
 - Compare 4 models('urban / rural', renter / owner occupied, Hispanic ethnicity, and Census geographic regions)
 - Field checks for true and false positive assignments
 - Current results show insufficient differences from national model to warrant subnational models, although owner/renter distinguished Occupied somewhat better.
- Potential concerns (time constraints)
 - Single test study area (Maricopa County AZ) may not fully represent national range of demographic conditions
 - Subnational model parameters (20% renters, 20% Hispanic) don't match national averages

WG Activities

5. Assign Demographic Characteristics

- ARs to characterize age, sex, race / ethnicity, composition.
 - Quality varies with AR source, geographic region, collection time
 - For housing, characterize tenure or vacancy details
- Tested to simulate 2010 Census for 16.7 million persons
 - Assign race, Hispanic origin with ACS, 2000 Census, SSA, SNAP, CARRA. With ARs handled ~95%
 - Further assignment with Within Household, Nearest Neighbor rules improve by up to another ~0.57% (varies by race and origin)
- WG concern about decision to eliminate combined race question / MENA – suggest revisiting these simulations
- Up to 98% reports of child in households reflected in KIDLINK AR, with improvements in high poverty areas

Recommendations for CSAC Consideration

1. Regression Models to Categorize NRFU Units (Vacant or Occupied)

Census should continue to explore strategies for incorporating ACS area-level Margins of Error (MoEs) into the multinomial models for 2020 census, recognizing that MoEs increase for smaller enumeration areas and anticipating that 2020 Census participation may drop relative to previous data collections.

Recommendations for CSAC Consideration

2. Compare National with Sub-National Models

Census should test additional thresholds for owner-renter and for Hispanic concentration sub-national models. The 20% threshold used does not reflect national averages.

Census should examine more than one geographic area in their testing. A single study area (Maricopa County, Arizona) cannot reflect all demographic concentrations (race/ethnicity, urban income extrema, family composition) that may occur across the United States.

Recommendations for CSAC Consideration

3. *Assigning Characteristics to AR Roster Units*

Continue testing to further reduce bias in assigning characteristics of age, sex, race / origin and tenure in documented low income and marginal demographic communities, where inaccurate assignments may be more pronounced.

Pursue statistical simulations using additional ARs to further improve counts of young and minority children (historically undercounted).

Review / revise estimates for race and Hispanic origin categories, paying special attention to simulations applying “Within Household” and “Nearest Neighbor” hot deck assignments, given the recent decisions against a combined question for collecting race / ethnicity, or a separate “Middle Eastern or North African” (MENA) category.

Recommendations for CSAC Consideration

4. *Manage Administrative Records Testing Workflows*

Implement formal protocol for risks, dependencies priorities for AR use in handling NRFU units, for the upcoming decennial census data collection and for future efforts.

Here's why:

1. Testing and analysis take time but are less costly than repeated NRFU follow-ups. Many times, WG heard that testing could not be extended (additional parameters / thresholds, testing more than a single study area) simply because time is too short.
2. Incomplete testing might provide incomplete information. Revised workflow and added statistical staff would permit more comprehensive testing. Given that ARs improve reliability and quality, adjust workflow to maximize return on investment, in the upcoming census and into the future.