

Discussion of Use of Administrative Data and Modeling Efforts for NRFU During the 2020 Census

Robert A. Hummer

University of North Carolina at Chapel Hill

Census Scientific Advisory Committee Meeting

March 30-31, 2017

Critical Importance of Administrative Data and Modeling for NRFU Cases in 2020 Census

- Some 50+ million addresses will fall into NRFU category; obviously (but so important), the fewer the better!
- Use of administrative records (AR) major new strategy that will be utilized to significantly reduce number of field staff and door knocks
- One of 4 key strategies to be used in Census 2020 to simultaneously reduce per household (adjusted for inflation) costs relative to 2010 while producing a high quality Census
- This needs to work well!

Key Uses of Administrative Data in the 2020 NRFU Operation

- 1) Effectively determine as many vacant, nonexistent, and occupied NRFU addresses as possible, while minimizing error in doing so
- 2) Enumerate the occupied NRFU addresses as effectively as possible... obtain an accurate roster and best determine the characteristics of individuals on that roster

... Administrative data & modeling will be used in both of these steps...

Determining Vacant & Non-Existent NRFU Addresses in 2016 Test

- After 2 Census mailings, 1+ “Undeliverable as Addressed” with a reason – e.g., Vacant or No Such Number – were eligible to be AR Vacant or AR Non-Existent
- Core AR’s used: USPS, IRS, CMS, IHS; other government and private sources also used
- Build person records from the AR; assign them PIK. Do people seem to be living at the NRFU addresses?
- 2+ AR sources needed before in-person contacts were reduced (a conservative approach)
- Multinomial logistic regression model: Vacant/non-existent/occupied as outcome variable; USPS info, individual info from AR, MAF info, and neighborhood info from ACS used as predictor variables
- Beta coefficients used to obtain predicted probabilities of which addresses are vacant or non-existent, each relative to occupied

Calculation of Distance Function for Vacant and Non-Existent

- Want to maximize: occupied = 0; vacant = 1
- Euclidian vacant distance function; continuous measure for each address
- Must establish a cut-point for when an address is determined to be vacant
- -----
- Also want to maximize: occupied = 0; non-existent = 1
- Euclidian vacant distance function; continuous measure for each address
- Must establish a cut-point for when an address is determined to be non-existent

Determining AR Occupied Addresses

Use AR data to develop household rosters

- IRS 1040, IRS 1099 Information Returns, CMS Medicare Database, and Indian Health Service
- Other potential data: HUD, Selective Service, NCOA, SNAP, CARRA Kidlink, Tax and Deed Info (to be finalized in Sept 2018)
- *** This will be really important... which I'll touch on below ***

Determining AR Occupied Addresses

1) Person-Place Model: Do AR sources place individuals at the same address as Census enumeration would? Gauged on 2010 data... decisions are made at the housing unit level. Assign lowest person-place probability when there are multiple people within a household.

2) Household Composition Model: Do AR sources predict household composition similarly to 2010 NRFU fieldwork? Smaller, simpler household compositions have higher agreement...

... The “occupied decision” is based on the shortest Euclidean distance between predicted probabilities of these two inputs. In other words, does AR modeling correctly effectively predict: 1) all people who live at this address, **and** 2) household composition? A continuous value of “occupied,” based on the distance between these two probabilities, is assigned to each address.

... Must establish a cut-point for when an address is determined to be occupied.

Assigning Characteristics to AR-Enumerated NRFU Cases

- Some characteristics for people (age, sex) taken from 2010 Census data or the Social Security NUMIDENT file... PIK
- Race and Hispanic Origin assigned based on research by CARRA
- Imputation procedures to be used for other characteristics
- *** This work, while some of it is more straightforward (e.g., age, sex) than the determination of occupied/vacant/non-existent and arguably of “2nd order”, needs to be prioritized in the 2018 test. I think CSAC would benefit from a table/description of results in late 2018. ***

Quality Assessment of AR Work Using 2015 Census Pre-Test in Maricopa County, Arizona

- Important note: This AR work used an earlier set of methods.
- Those determined to be occupied in AR: 90.8% were occupied based on fieldwork
- Of the occupied, person count agreed exactly 60.2% of the time. It was off by one person 26.9% of the time.
- Those determined to be vacant in AR: Just 46% were found to be vacant in fieldwork. PROBLEMATIC.
 - 19.2% were found to be occupied in fieldwork (very problematic)
 - 34.8% were found to be non-existent (less problematic)

Quality Assessment of AR Work Using 2016 Census Pre-Tests in Harris County (TX) and Los Angeles County (CA)

- Those determined to be occupied in AR: 80.3% were occupied based on fieldwork. This looks lower than the 2015 results, but 13.2% were unresolved in fieldwork.
- Of the occupied, person count agreed exactly 67.7% of the time. It was off by one person 23.0% of the time.
- Those determined to be non-existent in AR: Just 48.6% were found to be non-existent in fieldwork. (And 29.1% were occupied). PROBLEMATIC
- Those determined to be vacant in AR: Just 42.8% were found to be vacant in fieldwork. STILL PROBLEMATIC.
 - 21.1% were found to be occupied in fieldwork (still very problematic)
 - 20.7% were found to be non-existent (less problematic)
 - 15.4% were unresolved in fieldwork

Quality Assessment of AR Work Using 2010 Census

- NRFU Households: ~50,000,000
- 10.1% Determined to be Vacant: ~5,000,000
- 15% Determined to be Occupied: ~7,500,000
- 0.1% Determined to be Non-Existent: ~50,000
- ... With lots of agreement from 2010 NRFU fieldwork results.

- AR didn't effectively determine: ~37,450,000

- Possible to decrease this number without making too many mistakes???
- AR strategy less frequent in determining vacancies or occupancies in neighborhoods with high % Hispanic
- AR strategy less frequent in determining occupancies but more frequent in determining vacancies in neighborhoods with high % Black

Additional Improvement & Considerations for 2018 Test (and 2020?)

- Will send additional mailing in May to initial AR determination cases. Only those with (a 2nd determination of) UAA information will be eligible for AR determination. This is a good change and should yield better results.
- Use additional state-specific data (e.g., SNAP) as AR data? YES
- Possible to obtain & use other data from states as additional sources in your modeling (e.g., birth records, prisons, driver's licenses, state IDs, etc.)?
- Use CARRA Kidlink file to improve under-coverage of children in AR sources? YES
- Potentially use very recent ACS as training data for the AR modeling? YES

Reaction to Current Algorithm to Identify Occupied, Vacant, and Non-Existent Addresses?

- Distance algorithm is well conceptualized; should work well...with the very important caveat being that the input AR data needs to be as rich and detailed as possible, especially for the people/addresses most likely to be in NRFU!
- What are the best cut-points for AR vacant, non-existent, and occupied housing units? Should the cut-points vary across geographic areas? Analysis of sensitivity/specificity of cut-points?
- Good data inputs are going to be essential, especially for reducing NRFU load and maximizing the potential of ARs. NRFU cases are a select group; need data to more effectively predict vacant, non-existent, and occupied addresses for the kinds of addresses that end up in NRFU
 - SNAP; TANF? Birth/Death Records? Medicaid? Driver's Licenses? State ID Cards? Utility Records? Prison/Jail Records? Etc. Don't know what's possible, but need to think creatively and focus on individuals, addresses, and geographic areas most likely to be in NRFU...

Reaction to Changes in Contact Strategy in the 2018 End-to-End Test?

- Increase in contact strategy is an important innovation and seemingly necessary; most importantly, will help remove AR vacant/non-existent cases that are not vacant-vacant or nonexistent-nonexistent and put them back in the field
- Together with improved input data, this change will yield very important results in 2018
- Strongly recommend CSAC see these results as quickly as possible; fall 2018? Spring 2019 too late to make any final adjustments?

Thank You

- To the Administrative Data Team for making the paper and slide deck available to me in plenty of time
- To Tom Mule, Jr. for contacting me a couple of weeks before the CSAC Meeting and offering to help with any questions about the documents; this was much appreciated
- From the CSAC perspective, this was an ideal way to get prepared for the meeting, especially with such a complex and important topic