Administrative Record Use in the 2020 Census For Modeling and Processing

Andrew Keller and Thomas Mule
Decennial Statistical Studies Division
U.S. Census Bureau

Ukraine Presentation
September 23, 2022

Any views expressed are those of the author and not those of the U.S. Census Bureau.
The Census Bureau's Disclosure Review Board has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release.
CBDRB-FY21-DSSD007-0022

Outline

- Administrative Record (AR) Data and Modeling
- AR Enumeration
- AR Use for Count Imputation
- AR Use for Item Missing Data and Whole Person Imputation
- Discussion

AR Data Sources

- **United States Postal Service Information**
  - USPS Undeliverable-as-Addressed (UAA) reasons for census mailings made around April 1, 2020
  - Delivery Sequence File

- **AR Sources**
  - Internal Revenue Service (IRS) 1040 filings
  - IRS 1099 information returns
  - Centers for Medicare and Medicaid Services Medicare Enrollment database
  - Indian Health Service Patient Database
  - Census Household Composition Key
  - Third-party Veterans Service Group of Illinois (VSGI) files

Identify AR Unoccupied Units

- **Unoccupied Model Estimates Address-Level Probability of Occupied, Vacant, Delete**
  - 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 and other addresses
  - American Community Survey area-level estimates: % renters, % poverty, % black, etc.
- **Occupied and Vacant probabilities were ingested to define a vacant distance function.**
- **Pre-defined cutoff threshold to cull out AR Vacant addresses**
- **Similar process to define AR Delete addresses**

Identify AR Occupied Units

Can we reduce contacts for 101 Main Street?

1. Build a roster from administrative record sources – Take union from the following sources:

IRS 1040 & 1099, Medicare, Indian Health Service, Census Household Composition Key

2. Use statistical models to evaluate the roster

A. Person-Place Model - How likely is it that we are counting each person on the AR roster in the right place?

B. Household Composition - How likely is it that the household composition of the AR roster matches the Census

3. Person-Place and Household Composition probabilities were ingested to define an occupied distance function.

4. Decision for 101 Main Street based on pre-defined cutoff threshold to cull out AR Occupied addresses

Methodological References

Statistical Methodology: "A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 U.S. Census"
2020 Census Application: "Administrative Record Modeling in the 2020 Census"

AR Modeling and Contact Strategy

- Models produced AR Occupied, AR Vacant, AR Delete, and AR No Determination outcomes
- AR Occupied, AR Vacant, AR Delete indicated that the AR could be used for enumeration if no response was received.
- AR No Determination indicated that the AR had insufficient quality for enumeration use.
- AR Occupied, AR Vacant, AR Delete outcomes informed the contact strategy during the field operation.

AR Enumeration

- **If an address that met the quality threshold to be AR Occupied, AR Vacant, or AR Delete did not respond by the end of data collection, we made response records using AR data.**
- **AR Enumeration cases were only used if no other response was received.**
- **AR Enumeration cases were ingested in the same manner as self-response or enumerator-based responses.**
- **Of the 151,800,000 addresses in the 2020 Census[1]**
  - 3.20% were enumerated as AR Occupied
  - 1.15% were enumerated as AR Vacant
  - 0.24% were enumerated as AR Delete

References

[1] 2020 Census Data Quality: census-operational-quality-metrics-release_1.xlsx
CBDRB-FY21-DSSD007-0022

AR Enumeration (2)

- **A hierarchy (more later) was built so that people in AR Occupied units then have characteristics directly substituted from their own past reports to the Census Bureau (which include the decennial census or ACS) or AR data.**

- **We assigned characteristics for people in AR Occupied units if we had multiple sources confirming the family lives there. If not, we only assigned a count.**

- **Characteristics include:**
  - Person-Level (Name, Age and Date of Birth, Sex, Race, Hispanic Origin, Relationship to Householder)
  - Household-Level (Tenure)

- **AR Enumeration cases without Census or AR characteristics had remaining characteristics imputed using hot deck procedure.**

AR Enumeration Characteristic Hierarchy

- **Age**

Assign age/DOB from 2010 Census response or information from the Social Security Administration (i.e. the Census Numident).

- **Sex**

Assign sex 2010 Census or the Census Numident.

- **Race or Hispanic Origin**

Assign race and Hispanic origin from the 2010 Census, American Community Survey, Census Numident, or other federal sources.

- **Tenure**

Assign tenure based on a set of rules using federal administrative records sources, such as information from Housing and Urban Development (HUD), and commercial tax and deed information.

Availability of characteristics for assignment for 2020 AR Enumerations

Characteristic	Census 2020
Age	96.4%
Sex	96.5%
Race or Hispanic Origin	83.6%

CBDRB-FY22-172

AR Use for Count Imputation

- At the end of the census data collection operations, some addresses lacked a housing unit status or population count.
- We did not directly assign the AR count as the final population count.
- Of the 151,800,000 addresses in the 2020 Census[1], 0.93% underwent count imputation.
- The population count from the AR roster and UAA status was used along with other address-level and operational covariates to model the housing unit status and population count for addresses when either or both were unknown.
- Specifically, the AR count and UAA status from the unresolved address placed the address in a model cell.

[1] CBDRB-FY21-DSSD007-0022

AR Use for Item Missing Data

1) Editing – Run through responses for issue

a. Detect out of range or inconsistent values
b. Remove invalid or duplicate responses
c. Convert to proper values

2) Assignment – Responses are missing or inconsistent and information can be determined from:

a. Other response provided for that person
b. Previous Census or AR data

3) Allocation – Responses are missing or inconsistent and information can be determined from:

a. Response provided for other persons in household
b. Similar nearby housing units

AR Use for Item Missing Data (2)

- **We attempted to link all 2020 Census people to their corresponding person record in AR.**

- **Notes about the person link**
  - Created by matching person data reported on the census response such as name, date of birth, and sex to the AR reference file that contains those same covariates and the unique identifier
  - Some people cannot be linked to AR because not all information provided.
  - Does not require all characteristics in order to link to AR.

- **Some people in the census did not report all characteristics.**

- **Used the same hierarchy of rules to assign characteristics from past Census and AR data as was used for AR enumeration.**

National Usage of Past Census and Administrative Records during Characteristic Imputation Processing

- **Preliminary usage results for 331.4 million census people:**

	Percent of 331.4 million census people assigned from	
Characteristic	Past Census or ACS response	Other Administrative Records
Race	3.2%	1.5%
Hispanic Origin	3.2%	0.9%
Age	2.6%	1.4%
Sex	2.6%	1.2%

CBDRB-FY22-172

AR Use for Whole Person Imputation

- **For some occupied units, only a population count was known.**
  - These whole person imputation units had all people missing all characteristics.
  - Occurred in two situations:
    - Respondent or proxy had only provided a population count for the unit and no person demographic data
    - Address was imputed as occupied
- **For whole person imputation units, we checked that the AR roster count was the same as the census count.**
- **If the counts agreed, the AR roster was copied into the whole person imputation unit.**
- **Whole person imputation units without a matching AR count were imputed characteristics from a responding unit in the census of the same count and enumeration type where all people had all characteristics reported.**

Conclusion

- AR Modeling developed methodology to assess quality of AR data at the address level to inform the contact strategy during the Nonresponse Followup operation.
- We developed methodology to enumerate occupied, vacant, and delete addresses with sufficient quality.
- Occupied units were assigned characteristics in a hierarchical manner, primarily using previous census and ACS data. AR data was then used.
- AR roster counts and UAA status, along with other address-level and operational covariates were used to group responding and unresolved addresses for count imputation.
- Links were made between the current census files and AR data so that past census, ACS, and administrative data aided in characteristic imputation.
- For scenarios where only a census population count was known and that count matched the AR count, the AR roster was copied into the housing unit.

email: andrew.d.keller@census.gov