Update on Disclosure Avoidance and Administrative Data

John M. Abowd and Victoria Velkoff
Associate Director R&M, Chief Scientist and Associate Director Demographic Programs
U.S. Census Bureau
Census Scientific Advisory Committee
September 13, 2019

Disclosure Avoidance Update

2018 End-to-End Census Test

- Version of the 2020 Disclosure Avoidance System used for 2018 E2E Test code base and draft technical documents released https://github.com/uscensusbureau/census2020-das-e2e
- Test products released from the 2018 E2E Test used a privacy-loss budget of 0.25 for reasons documented here: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019_13.html

Formal Privacy for the American Community Survey

- Formal privacy methods for the American Community Survey *will not be implemented before 2025* (Deputy Director's blog: https://www.census.gov/newsroom/blogs/random-samplings/2019/07/boost-safeguards.html)
- The scientific and user communities will be fully engaged as part of that process
- All current efforts are focused on formal privacy methods for the 2020 Census

2020 Disclosure Avoidance System (Persons)

- Expanded from 2,012 cell national histogram (used for E2E Test) to 370,994 cell national histogram for Demographic and Housing Characteristics-Persons (DHC-P)
- Person-level workload optimization completed for fully interacted:
  - HHGQ (In household or in GQ (7 major types))
  - Race (63 OMB categories)
  - Hispanic/Latino
  - Sex
  - Age (single years to age 84, then (85-89), (90-94), (95-99), (100-104), (105-109), (110-114), (115-max))
  - Citizenship (see administrative record update)

2020 Disclosure Avoidance System (Housing and Households)

- 387,072 cell national histogram for Demographic and Housing Characteristics-Housing (DHC-H)
- Household-level workload optimization completed for fully interacted:
  - Householder attributes: [race, Hispanic/Latino, sex, age]
  - Multigenerational
  - Household size (Top-coded at 7)
  - P60, P65, P75 (Presence of people 60 years and over (respectively, 65 or 75))
  - Household type attributes (see forthcoming technical document for details)

Demonstration Products

- Based on the national 2010 Census Edited File
- For approximately 70% of tables in DHC-P and DHC-H
- Based on each of the workload-optimized approx. 400,000 cell histograms at each geography using new version of TopDown
- Details of the optimization and privacy-loss budget management will be presented at CSAC (pending decisions from the August 29, 2019 DSEP meeting)
- Soft target release date mid-October
- CNSTAT workshop to discuss demonstration products December 11-12
- New code base release in October

Administrative Citizenship Data Update

Apportionment

- The Census Bureau's primary data product from the 2020 Census is apportionment counts that will be delivered to the President by December 31, 2020.
- The counts are used to reapportion the U.S. House of Representatives.
- The apportionment population count for each of the 50 states includes the state's total resident population (citizens and non-citizens), plus a count of overseas federal employees (and their dependents living with them) who are allocated to their home states.
- This is identical to the method used for the 2010 Census, but based on the 2020 Census residence criteria.
- The apportionment counts are calculated using the Census Unedited File (CUF), which is produced by November 30, 2020.
- The CUF does not contain any citizenship data.

Redistricting (PL94-171)

- The Census Bureau is required, under Public Law 94-171, to make data available to the states to assist in redistricting.
- These data are produced from the Census Edited File (CEF), which is produced from the CUF by imputing item missing data using administrative records and statistical models.
- The CEF is produced by January 25, 2021.
- The CEF is sent to the 2020 Census Disclosure Avoidance System, which releases the Micro-data Detail File (MDF) to the tabulation system.
- Redistricting data at the block level are produced from the MDF, and will be released by state from February 18 through March 31, 2021.

Redistricting (PL94-171) Format

- Total population by the 63 detailed race categories – Table P1;
- Total population by Hispanic origin (across all races) and for the non-Hispanic origin population by the 63 detailed race categories – Table P2;
- Total voting-age population by the 63 detailed race categories – Table P3;
- Total voting age population by Hispanic origin (across all races) and for the non-Hispanic origin population by the 63 detailed race categories – Table P4.
- Total Population only - Group Quarters Population by Group Quarters Type – Table P5.
- Housing Unit Counts - Occupancy Status – Table H1.

Citizen Voting-Age Population Data

- The Paperwork Reduction Act clearance package for the 2020 Census and the President's Executive Order 13880 commit the Census Bureau to releasing Citizen Voting-Age Population (CVAP) data by March 31, 2021.
- These data will be produced by combining administrative data from a number of federal, and possibly state, agencies into a separate micro-data file that will contain a "best citizenship" variable for every person in the 2020 Census.
- The citizenship micro-data file and the CEF will be simultaneously sent through the 2020 Disclosure Avoidance System, which will do the final record linkage and place a confidentiality protected citizenship variable on the same MDF as will be used to produce the redistricting data.
- CVAP data will be produced at the block-level from the MDF and released to the public by March 31, 2021.

CVAP Data Format

- No final decisions have been made regarding the methodology and format of the block-level CVAP data.
- No decisions have been made regarding the future of the American Community Survey-based CVAP data that have been produced annually since 2011.
- The Census Bureau's internal working group has set March 31, 2020 as the final date for determining the viability of each potential administrative data source on citizenship.
- March 31, 2020 is also the final date for releasing the specifications of the CVAP data to be released by March 31, 2021.
- The Census Bureau is considering the release of demonstration products based on historical data using the proposed methodology for the 2020 CVAP data.

Using Administrative Data

- Following the Secretary's March 26, 2018 instructions, modeling efforts focused on using survey responses (to the question on the 2020 Census) and administrative records
- When the Supreme Court upheld the injunction on asking the question, and the President issued Executive Order 13880, modeling efforts focused on using more administrative record sources
- The Director convened the Interagency Working Group, which consists of high-level executives in federal agencies that have person-level data relevant to estimating citizenship
- Primarily two uses of administrative data for estimating citizenship:
- (1) Keeping the names, addresses, and other PII in the record linkage system current
- (2) Determining citizenship status from variables on the files and eligibility conditions

Current Sources of Citizenship Data

- Social Security Administration NUMIDENT
- Contains place of birth and citizenship status for approximately 94% of its universe
- Individual Taxpayer Identification Numbers
- NOTE: the Census Bureau does not receive, and has not requested, application for ITIN data
  ITINs can be identified when they are used in the SSN field of a form the Census Bureau does receive
- IRS 1040 and 1099 forms
- Primarily used to keep the record linkage system current
- CMS Medicare and Medicaid/CHIP
- Contain some citizenship information but are primarily used to keep the record linkage system current
- Housing and Urban Development
- Federal Housing Administration, Public and Indian Housing Information Center, Tenant and Rental Assistance Certification System,
  Low-income Housing Tax Credits, Computerized Homes Underwriting Management System used to keep the linkage system current

Additional Federal Citizenship Data

- Department of Homeland Security USCIS/CBP/ICE
- Lawful permanent residents and naturalization data (CIS), visas (ICE), arrival/departure (CBP)
- Department of State (Passport Services)
- Passport data
- Social Security Administration
- Master beneficiary record
- Indian Health Service
- Patient registration
- Department of Justice
- US Marshals and Citizenship and Immigration Data Collection

Research Program

- Citizenship modeling
  - Develop statistical models that efficiently and accurately combine multiple sources of administrative citizenship data to estimate “best citizenship” for each person known to the Person Identification Validation System (PVS), which is the production record linkage system for the 2020 Census
  - Use these models to prepare a micro-data file outside the 2020 Census production system that can be combined with the 2020 Census Edited File to provide the 2020 Disclosure Avoidance System with the “best citizenship” variable to tabulate block-level CVAP tables
  - This research began in April 2018, final specifications and modeling details are planned for release before March 31, 2020, which is the internal deadline for finalizing the input administrative record sources

Research Program II

- Enhanced record linkage capabilities
  - The production PVS can link persons found in the SSA NUMIDENT and ITIN universes; about 90% of the U.S. resident population
  - Many of the requested files from DHS, State, and others, are expected to provide the PII that enables record linkage for much of the balance of the resident population, provided that the PII on the 2020 Census is as reliable as it was in 2010

Confidentiality Protection

- As with all administrative data ingested by the Census Bureau, the citizenship data will be used only for statistical purposes
- As with all administrative data ingested by the Census Bureau, the confidentiality of the citizenship data will be fully protected by Title 13, Section 9, which prohibits:
  - "… mak[ing] any publication whereby the data furnished by any particular establishment or individual under this title can be identified"
- The CVAP tables will be produced using the 2020 Census Disclosure Avoidance System, which implements differential privacy using the TopDown algorithm
- The CVAP tables will share the privacy-loss budget determined by the Data Stewardship Executive Policy Committee for the 2020 Census publications

Thank you.

John.Maron.Abowd@census.gov and Victoria.A.Velkoff@census.gov