2020 Census: Stages and Challenges in Post-Data Collection Processing

Presentation to the National Academy of Sciences, Engineering, and Medicine's Panel to Evaluate the Quality of the 2020 Census

Michael Thieme, Assistant Associate Director for Systems and Contracts
Barbara LoPresti, Chief Decennial Information Technology Division

July 21, 2021

Shape your future START HERE >

United States Census 2020

2020 Census
Post Data Collection Processing Flow

Geographic Processing

The 2020 Census Geographic Processing is based on the decades of work by the Census Bureau and our Partners to produce the most accurate and complete geographic foundation ever – Master Address File/Topologically Integrated Geographic Coding and Referencing (MAF/TIGER) System.

Over 152 million addresses were included in the 2020 Census Final enumeration universe.

DRF1 (Decennial Response File 1)

The DRF1 is the first file produced after data is collected.

It contains all response data—including duplicate responses.

The complete inventory of every residential address in the nation is linked to every response we received during data collection.

DRF2 (Decennial Response File 2)

The Primary Selection Algorithm selects which data from the DRF1 should represent a housing unit on the CUF.

CUF (Census Unedited File)

Count imputation is applied to unresolved cases in DRF2.

The CUF contains the final universe of addresses, enumeration status, and population count.

COMPLETE

APPORTIONMENT COUNTS

CEF (Census Edited File)

Finalizes characteristic data. Editing and imputation are applied to missing and erroneous values for all items.

Disclosure Avoidance/ Microdata Detail File

Confidentiality protection applied using the Disclosure Avoidance System (DAS)

The DAS creates the Microdata Detail File (MDF) which is used for tabulation.

TAB Tabulation

Turns our massive data store from the 2020 census into the easily understand-able and usable data tables

REDISTRICTING RELEASE

Shape your future START HERE >

United States Census 2020

2020 Census
Apportionment and Redistricting Key Milestones

Processing Step	Start	Finish
Decennial Response File 1 (DRF1)	10/29/2020	12/26/2020
Decennial Response File 2 (DRF2)	12/26/2020	2/24/2021
Census Unedited File (CUF)	2/25/2021	3/10/2021
Apportionment Preparation and Release*	3/12/2021	4/26/2021
Census Edited File (CEF)	4/20/2021	6/24/2021
Disclosure Avoidance Application/Microdata Detail File	6/25/2021	7/18/2021
Tabulation File Processing	7/19/2021	8/16/2021
Redistricting Preparation and Release	8/17/2021	9/30/2021

COMPLETE

2020 Census
Processing Anomalies

- An anomaly is “something different, abnormal, peculiar, or not easily classified” that can be identified at any stage in processing
- Every census has processing anomalies – 2020 is no different
- Anomalies found in processing are not errors in the census, but they can turn into errors if we don’t review and resolve them
- Far from raising concerns, the fact that the Census Bureau's planned process expects and identifies anomalies demonstrates that it is *working*
- Identifying and resolving anomalies is not only expected, it is evidence that quality assurance processes continue to validate our commitment to accuracy

2020 Census
Processing Anomalies

Anomaly Categories

1. **Standard problems that arise in processing any large survey**
   - Routine coding anomalies
   - Basic errors in processing code
   - Processing errors in data handoffs between systems
   - Errors in business rules

Examples:

- Processing software was miscalculating the age of respondents if those respondents did not include the month and day of their birthday in their response. A simple code correction fixed this, but if it had not been fixed, it may have reduced our ability to match and remove duplicate responses (for example when we get responses from two people in the same household).
- Some responses collected via internet self-response were duplicated in responses from Group Quarters (GQ) addresses. The processing software specified invalidating the Housing Unit responses in favor of the GQ response but did not execute that invalidation properly. A code correction fixed this, but if it had not been fixed the result could have produced an overcount of persons. This is also a good example of addressing responses from multiple modes.

Anomaly Categories (cont.)

2. **Anomalies resulting from unanticipated respondent action - especially with the impact of the pandemic**

Example:

- In a small number of colleges and universities the total student population in all dorms was submitted by those institutions for *each* dorm, potentially inflating the population count on those campuses. A code fix enabled distribution of the correct population among the college dorms where this occurred.

Anomaly Categories (cont.)

3. Anomalies resulting from unanticipated enumerator action

Example:

- During GQ enumeration, there were isolated instances in which some enumerators mistakenly set a GQ rework indicator for one or more persons in the GQ. This invalidated all responses at the GQ. A GQ rework indicator is intended to be set *only* for the entire GQ, not persons within the GQ. A software fix enabled and validated all responses at the GQ**.**

Anomaly Categories (cont.)

4. **Anomalies when characteristic data processing produced unlikely results**

   Examples:

   - A single-sex GQ with a large proportion of generic first names (e.g., "person") are being assigned sex based on first-name, resulting in mixed-sex distribution.
   - A Group Quarters facilities reporting repeated age within all GQ units

Key Message

Our data processing (including the handling of anomalies) has not shown any critical errors caused by data collection mistakes or omissions that we could not fix. We are happy to say we have fixed or are fixing anomalies that our systems and processes have identified thus far. All processing decisions favored accuracy, not expediency in meeting the schedule.

Questions

Shape
your future
START HERE >

United States Census 2020

END