August 11, 2021

Administrative Data Used in the 2020 Census

Karen D. Deaver

Decennial Census Programs Directorate

United States™ Census Bureau

Intentionally blank

Table of Contents

I. INTRODUCTION 1
II. BACKGROUND 1
III. DATA SOURCES 3
A. TITLE 26 COMPOSITE 3
B. TITLE 13 COMPOSITE 4
C. SOURCES NOT IN A COMPOSITE 5
D. ADDITIONAL SOURCES ONLY USED FOR CITIZENSHIP 5
IV. FRAME DEVELOPMENT 5
V. RESPONDENT MOTIVATION 6
A. INITIAL CONTACT STRATEGY 6
B. ADVERTISING CAMPAIGN 6
VI. SELF-RESPONSE PHASE 7
A. NON-ID ADDRESS ENHANCED MATCHING 7
B. PAPER DATA CAPTURE QUALITY ASSURANCE 8
C. SPECIAL ENUMERATIONS 9
VII. NONRESPONSE FOLLOWUP 9
A. VACANT AND DELETE IDENTIFICATION 10
B. ADMINISTRATIVE RECORDS ENUMERATION 10
C. BEST TIME TO CONTACT MODELING 12
D. PHONE NUMBER AVAILABILITY 12
VIII. RESPONSE DATA VERIFICATION 12
A. SELF-RESPONSE QUALITY ASSURANCE 12
B. ENUMERATOR QUALITY CONTROL 13
IX. POST-RESPONSE PROCESSING 13
A. ADDITIONAL UNDUPLICATION OF PEOPLE IN HOUSING UNITS 13
B. COUNT IMPUTATION 13
C. CHARACTERISTIC IMPUTATION 14
X. PUBLISHING DATA 15
A. CITIZEN VOTING AGE POPULATION (CVAP) TABLES 15
B. COUNT QUESTION RESOLUTION 15

XI. COVERAGE EVALUATION 16
A. POST-ENUMERATION SURVEY 16
B. DEMOGRAPHIC ANALYSIS 16
XII. REFERENCES 16
APPENDIX A –DATA SOURCES USED IN THE 2020 CENSUS 1
APPENDIX B - OVERWRITING EXAMPLE 1

ADMINISTRATIVE DATA USED IN THE 2020 CENSUS

I. INTRODUCTION

Based on the results of the 2018 End-to-End Census Test, prior census tests, and other research, the Decennial Census Programs Directorate determined its approach to using administrative data in many operations of the 2020 Census. Events that occurred during the fielding of the 2020 Census, including the COVID-19 pandemic and natural disasters, required some adjustments to these plans.

In this context, we are using "administrative data" broadly to include:

- Microdata records contained in files collected and maintained by federal, state, and local government agencies (traditionally referred to as administrative records).
- Microdata records contained in files collected and maintained by commercial entities (often referred to as third-party data).

As well as:

- Macro- and microdata from Census Bureau data collections for statistical purposes and address enhancement operations (internal data).
- Macrodata from publicly available sources (public data).

The intended uses of administrative data and the expected sources of data evolved throughout the research and planning for the 2020 Census and their actual use was expanded during 2020 Census fielding. This report documents the final use of administrate data during the 2020 Census. It updates and can replace the similar "intended use" memo made public in 2020 (Deaver, May 2020).

II. BACKGROUND

To meet the strategic goals and objectives of the 2020 Census, the Census Bureau made fundamental changes to the design, implementation, and management of the decennial census. These changes built upon the successes and addressed the challenges of previous censuses, while also balancing objectives of cost containment, quality, flexibility, innovation, and disciplined and transparent acquisition decisions and processes. The 2020 Census Operational Plan included using administrative records, third-party data, internal data, and public sources (collectively referred to in this document as "administrative data") to avoid cost, maintain quality, and improve efficiency of operations (U.S. Census Bureau, December 2018).

Several core sources were used to support this effort, including data collected from Census Bureau operations, which is protected under 13 U.S.C. § 9, and data from the Internal Revenue Service (IRS), which is referred to as Federal Tax Information (FTI), and is protected by 26 U.S.C. § 6103(b)(8). In accordance with Title 26 and a 2013 Memorandum of Agreement between the Department of Treasury, Internal Revenue Service, and the Department of Commerce, U.S. Census Bureau, the scope of work for which FTI may be used includes frame building, enumeration, imputation, and evaluation. Other agreements that the Census Bureau maintains with other administrative data providers similarly delineate allowable uses of their data.

Below are the instances for which the Census Bureau used administrative data in the 2020 Census, excluding evaluations and experiments in support of early planning and research to inform the transition and design of the 2030 Census. Several instances used extracts from either a "Title 26 composite" or a

"Title 13 composite" rather than directly accessing original administrative datasets. The process to create these composites is described in the next section. More detailed information on each instance is provided in sections IV- XI. Appendix A summarizes the sources for each instance. For more information about the 2020 Census operations, see the detailed operational plans (U.S. Census Bureau, various). Final assessments of each operation will be available on www.census.gov in the coming years.

- Frame Development: Develop and update the address frame, including group quarters, and spatial data that serve as the universe for 2020 Census enumeration activities.
- Respondent Motivation:
  - Support the initial contact strategy (appropriate delivery of census invitations and questionnaires).
  - Develop and execute the microtargeted advertising campaign.
- Self-Response:
  - Augment respondent-provided address data to enhance matching of non-ID responses to the updated enumeration address list: the addresses for responses that are returned without the preassigned census identification number and do not easily match to the Master Address File (MAF) will be compared with administrative record information in order to obtain missing address information, or correct errors, such as misspelling. By improving the address, another attempt can be made to associate the response data to a household in the enumeration universe so that no further effort is required to obtain a response, such as follow-up mailings or fieldwork.
  - Quality assurance in the Paper Data Capture operation: Optical Character Recognition (OCR) capture of written-in names is compared to existing administrative data to ensure quality of the OCR process.
  - Enumerate or supplement field enumeration of nontraditional or unique living arrangements, such as group quarters, military installations, and federally affiliated people overseas.
- Nonresponse Followup (NRFU):
  - Reduce contacts for cases in the NRFU workload through identification of vacant housing units and deletes (units that do not meet the Census Bureau's definition of a housing unit).
  - Enumerate nonresponding, occupied housing units with quality, reliable information.
  - Model the "best time to contact" occupied, nonresponding housing units (modeled as intended, but feature disabled during fielding).
  - Provide additional phone contact information to cases in the NRFU workload.
- Response Data Verification:
  - Self-Response Quality Assurance: Corroborate respondent-provided information to detect potentially suspicious responses.
  - Enumerator Quality Control: Validate enumerator-provided information to ensure reporting accuracy and more efficiently target reinterview efforts.
- Post-Response Processing:
  - Additional Unduplication of People in Housing Units: Determine in which household to keep people with resolved duplicate links.

- Count Imputation
    - Housing Unit: Determine final occupied/vacant/nonexistent status for unresolved cases and impute the household count for those addresses determined to be occupied but without a specified number of occupants.
    - Group Quarters: Impute final count for unresolved, occupied facilities.
  - Characteristic Imputation: Impute household and person characteristics where they are missing from the response.
- Publishing Data:
  - Create the Citizen Voting Age Population special tabulation (intended but subsequently cancelled).
  - Resolve Count Question Resolution challenges.
- Coverage Evaluation:
  - Improve matching and characteristic imputation in the Post-Enumeration Survey (PES).
  - Create independent estimates through Demographic Analysis.

III. DATA SOURCES

A. TITLE 26 COMPOSITE

Initially, administrative data were used to create a robust repository of data and information that was used for activities that had been approved by data providers (e.g., the Internal Revenue Service and the Social Security Administration). Data from several federal, state, internal, and commercial primary sources were combined, standardized, and corroborated to create the Title 26 composite, which contained variables such as address, householder name, household roster and relationships, and demographic data about the household inhabitants. FTI such as taxpayer ID and return type were included in this repository. These data were combined with extracts from the Master Address File (MAF)/Topologically Integrated Geographic Encoding and Referencing (TIGER) database (MTdb). The MTdb reflects the latest geographic information for each address. Extracts from the Title 26 composite were provided for additional processing as described in subsequent sections as appropriate.
The following sources were used to create the Title 26 composite and include historic and current vintages as available to the Census Bureau:

- Administrative Records
    - Centers for Medicare and Medicaid Services (CMS) Medicare Enrollment Database (MEDB)
    - Housing and Urban Development (HUD) Public and Indian Housing Information Center (PIC) and Tenant Rental Assistance Certification System (TRACS), now known as the combined “Longitudinal” File
    - HUD Federal Housing Administration (FHA) Integrated Database (IDB), which includes data from Computerized Homes Underwriting Management System (CHUMS)
    - Indian Health Service (IHS) Patient Registration
    - Internal Revenue Service (IRS)
        - 1040 Individual Tax Returns
        - IRS 1099 Information
    - Selective Service System (SSS) Registration

    - Social Security Administration
        - Census Numident (a processed version of the SSA Numeric Identification File, Numident, which does not contain a Social Security number) as well as the Census Numident Alternate Name File
    - State or local program datasets, examples include:
        - Homeless Management Information System (HMIS)
        - Alaska Permanent Fund Dividend (PFD)
        - Supplemental Nutrition Assistance Program (SNAP)
        - Temporary Assistance for Needy Families (TANF)
        - Women, Infants, and Children (WIC)
    - U.S. Postal Service (USPS) National Change of Address (NCOA) File
- Third-Party Datasets[1]
    - Black Knight
    - DAR Partners (Data Advisory Research)
    - Targus (Wireless and Federal Consumer)
    - Veteran Service Group of Illinois (VSGI)
- Census Bureau Data
    - American Community Survey (ACS) data
    - Census 2000 and 2010 Census edited and unedited datasets
    - Census Household Composition Key File (produced using Social Security Administration data and previous census information linking children 18 years and younger with their parents)
    - Contact Frame file (list of phone numbers) compiled from some sources listed here (SNAP, WIC, Alaska PFD, ACS Data, DAR Partners, Targus Wireless and Federal Consumer, and VSGI) as well as the ones below before it was delivered to 2020 Census systems:
        - Experian InSource and Experian End-Dated Records
        - InfoUSA
        - Melissa Data
        - National Sample Survey of Registered Nurses
    - "Best Race and Ethnicity" information compiled from some sources listed here (MEDB, HUD Longitudinal, IHS, Census Numident, ACS data, prior census data, and Targus Federal Consumer) as well as:
        - CMS Medicaid and Statistical Information System (MSIS)
        - National-Level Adult and Child TANF Recipient Files
        - Experian InSource and Experian End-Dated Records
        - InfoUSA

B. TITLE 13 COMPOSITE

Similarly, a robust repository of Title 13 data and information was created from the Title 26 composite and extracts of this composite were provided to various 2020 Census operations that did not need to use FTI. While FTI was used in the creation of the Title 26 composite, in the creation of the Title 13 composite, the FTI was overwritten such that no FTI appeared in the final Title 13 composite and FTI was

[1] Data from other commercial sources (e.g. CoreLogic) were used in census tests but are not considered part of the production datasets.

not the sole source of any information. This composite was therefore considered to be Title 13, not Title 26. An example of how the overwriting worked is in Appendix B.

C. SOURCES NOT IN A COMPOSITE

Some processes used sources, either solely or in part, that were not part of the composites discussed above. Those sources were:

- ACS contact history information
- Department of Defense
    - List of deployed personnel
    - Count of federally affiliated people stationed/assigned overseas, including dependents
- Communications-related
    - Modeled self-response predictions (Predicted Self-Response Score and Internet Proportion of Self-Response)
    - Audience segmentation information
    - Federal Communications Commission (FCC Residential Fixed Internet Access Service Connections per 1,000 Households by Census Tract [publicly available])
- Frame-related
    - USPS Delivery Sequence File (DSF)
    - Additional Geographic Support Program and 2020 Census program data – address and spatial data provided by tribal, federal, state and local governments
    - Lists of group quarters and transitory locations from federal, state, local, and third-party sources as well as ongoing operations
- Planning Database (PDB) (publicly available datasets at the tract and block group levels that assemble a range of housing, demographic, socioeconomic, and census operational data; variables have been extracted from the 2010 Census and ACS databases)
- Group quarters administrator lists
- Lists of residency information for off-campus college students
- Integrated Postsecondary Education Data System (IPEDS)
- USPS Undeliverable As Addressed (UAA) Information

D. ADDITIONAL SOURCES ONLY USED FOR CITIZENSHIP

In addition to some sources mentioned in the sections above, there were several sources that were expected to be used only to research how to produce, and subsequently produce, citizenship information in conjunction with the census. This effort was canceled to be in compliance with Executive Order 13986 issued on January 20, 2021 (Executive Office of the President, January 20, 2021).

IV. FRAME DEVELOPMENT

Throughout the decade, the Census Bureau maintained address and spatial (e.g., roads, boundaries, and geographic areas) data in the Master Address File (MAF)/Topologically Integrated Geographic Encoding and Referencing (TIGER) System. The MAF/TIGER System is regularly updated with data from the United States Postal Service; ongoing geographic partnership efforts with tribal, state, and local governments; and fieldwork. These efforts were used to update the address frame and reflect changes to the housing stock that occurred over time.

For the 2020 Census, additional efforts were employed to finalize the frame. Focused geographic partnership efforts helped improve the address list. State and local governments provided address updates during the Local Update of Census Addresses (LUCA) program and New Construction (NC) efforts. Additional data for group quarters and transitory locations was obtained from multiple sources, including federal partners, state partners, local governments, and a third-party provider.

Frame development did not use either the T26 composite or the T13 composite. Data within the MAF/TIGER System—extracts from the MTdb—were used as inputs to several enumeration operations.

V. RESPONDENT MOTIVATION

Respondent motivation activities did not use T26 or T13 composite data. Most activities used aggregated data at higher geographic or demographic levels; two activities used person-level data. The activities are outlined below.

A. INITIAL CONTACT STRATEGY

Administrative data were used to determine how the Census Bureau would send the initial invitation to respondents. Respondents in areas more likely to respond online received the "Internet First" mailing strategy, where they received invitations to respond online. Those who did not respond online received reminders to respond and a paper questionnaire before NRFU began. Respondents in areas least likely to respond online (as determined by using ACS data and Federal Communications Commission internet connectivity data), received the "Internet Choice" mailing strategy. The Choice strategy consisted of receiving an invitation to respond online, but with a paper questionnaire in the first mailing. Respondents then received reminders to respond either online or via the questionnaire they received earlier. Those who did not respond received another paper questionnaire before NRFU began. The "Internet First" or "Internet Choice" delineation was determined by geographic area (not by individual household/address) and used area-based response likelihoods, not person-level data, to identify which contact strategy the area received.

Addresses in certain geographic areas received bilingual English/Spanish invitations and questionnaires. These areas were designated where 20% or more of households in a census tract (areas with about 4,000 households) need Spanish language assistance — defined as households that have at least one person age 15 or older who speaks Spanish and doesn't speak English "very well" based on ACS data.[1]

B. ADVERTISING CAMPAIGN

To increase the effectiveness of advertising and contact strategies, the Census Bureau, through the communications contract, used demographic and geographic information from various sources to help target the advertising to specific populations.

- The Census Bureau developed predictive models and used these models to estimate tract-level self-response propensity for various mode(s). These predictions were then used to develop the media plan and to aid in campaign optimization. Tract-level self-response rate predictions were aggregated to larger geographic areas and helped determine the media and messaging strategies for the campaign.

[1] This use was inadvertently omitted from the previously released memo.

- Predictions were combined with geographic, demographic, housing, and sentiment information to create audience segments. The segmentation information was used to design and execute targeted advertising and communication strategies for geographic segments and audience groups.
- As households completed the census questionnaire, we used campaign analytics to identify the best messages and modes to reach various segments. Headquarters staff, field teams, and partnership specialists coordinated to prioritize audiences and align messaging. We also used internet address and browsing history to customize user experience with the web platform.

In addition to the above general use, two targeted mailings also made use of administrative data.

- Native Hawaiian and Pacific Islanders (NHPI) in the continental United States received a direct mailing through a partner organization. The partner organization used an address list supplied to them by NHPI-related organizations.
- Residential addresses in targeted USPS carrier routes determined to be at risk for undercounting young children also received a direct mailer. Tract-level response and demographic data from the most recent ACS, the 2010 Census, and the planning database, along with address information from the MAF and route information from the USPS were used to determine the targeted routes.

An additional effort developed to adapt the communication strategy to a COVID-19 environment, was engaging the general public in an email marketing campaign that sought to educate recipients about the 2020 Census, motivate them to self-respond, and encourage them to share information with their family members and friends. Emails for this effort were obtained both from the public subscribing to receive emails on census.gov, 2020census.gov or through the GovDelivery Network, and from the Census Bureau Contact Frame. Between July and August, targeted emails were deployed to low responding regions.

VI. SELF-RESPONSE PHASE

A. NON-ID ADDRESS ENHANCED MATCHING

A goal in the 2020 Census was to make it easy for people to respond anytime and anywhere to increase self-response rates. We did this by providing response options that did not require a unique census identifier (ID).

A “non-ID response” or a “non-ID case” refers to address information and associated person data provided by respondents without the preassigned, unique identification number. The process of comparing these non-ID cases to existing address information in the MTdb is referred to as “Non-ID Processing.” The final MTdb matching and geocoding result for each non-ID case was the outcome of the process. The Census Bureau used administrative, third-party, and census data to augment respondent-provided address data (referred to as “address enhancement”) to support matching to the MTdb[1].

Throughout the entire self-response period, Non-ID Processing matched respondent-provided addresses from non-ID cases to the MTdb. Census Bureau research suggested that initial address matching efforts

[1] The initial census enumeration universe had been updated through Address Canvassing and incorporated into the MAF that was used during Non-ID processing.

would pair about 87 percent of all non-ID cases with an existing record in the MTdb. Research also suggested that an additional 2 to 3 percent could be matched to an existing record in the MTdb as a result of address enhancement, which focused on the non-ID cases that could not be initially matched to the MTdb. Information derived from administrative records and third-party data sources was used to augment respondent-provided address information, creating the most accurate addresses possible. These "enhanced" addresses then were used in a second attempt to match non-ID addresses to the MTdb. The remaining non-ID cases that failed the second matching attempt or were not "enhanced" were validated and geocoded manually.

Only data from the Title 13 composite were used in this operation.

B. PAPER DATA CAPTURE QUALITY ASSURANCE

Administrative data were used to increase the confidence in the information that was captured from paper questionnaires. Manual keying or Optical Character Recognition (OCR) was used to capture the information from returned paper questionnaires, then the information was compared with existing information as a quality check (i.e., to keep error rates low). Administrative data were not used to replace any manual or OCR captured data, just to provide a confidence level about it. Greater confidence increased OCR rates and decreased keying costs.

More specifically, the 2020 Census Paper Data Capture system, Integrated Computer Assisted Data Entry (iCADE), used OCR to electronically capture response data provided by respondents, thereby eliminating most data needing to be captured manually. Each character within a response field was assigned a confidence score by the OCR software for processing.

Results from OCR were also made available to the iCADE application, where low-confidence write-in fields were manually captured to enter a verification process. This Quality Assurance (QA) process performed an independent verification process manually of both OCR and keyed entries to ensure the outgoing error rate was acceptable and met the decennial requirements. Verification was done by comparing the two data sets looking for a match or not. A sample stratum was defined and used to determine the sample.

If the error rate within any sample stratum was determined to be unacceptable, those fields were then sent to a process called Remainder Verification. Remainder Verification manually reprocessed all fields within the targeted stratum that were not sampled and verified. This process looked to correct any other errors in the batch.

Having access to supportive demographic information (e.g., age, race, sex, date of birth) from administrative data allowed iCADE to produce additional matches within a household. This enhanced the quality related to the confidence levels for "First Names, Last Names, Middle Initial, Race, Age, and Date of Birth" assisting the OCR and keying data validation processes.

Names and other external data elements integrated into this process were used as a direct replacement for response data. The data sources, when linked to an MTdb address, were used to increase or decrease the confidence levels of the data captured. External data elements were integrated to validate existing OCR and manual keying results.

Only data from the Title 13 composite were used in this operation.

C. SPECIAL ENUMERATIONS

The Census Bureau focused additional efforts on nontraditional living quarters, which are traditionally harder to data capture because of the unique living arrangements they present.

These enumerations did not use T26 composite or T13 composite data.

A group quarters (GQ) is a place where people live or stay in a group living arrangement that is owned or managed by an entity or organization providing housing and/or services for the residents. This is not a typical household-type living arrangement. These services may include custodial or medical care as well as other types of assistance, and residency is commonly restricted to those receiving these services. People living in GQs are usually not related to each other. GQs include such places as college residence halls, residential treatment centers, skilled nursing facilities, group homes, correctional facilities, workers' dormitories, military barracks, and domestic violence shelters. The Census Bureau worked with GQ administrators to enumerate the people residing at the GQ. GQ administrators provided 2020 Census response data either electronically to a Census Bureau secured portal or provided a census worker with a paper listing of response data. These data may have been used as the sole source of obtaining information or as a supplemental tool to ensure data collection of an entire facility when other enumeration methods were used. These data are not considered administrative records or third-party data, but internally collected data.

The Census Bureau worked with Department of Defense (DOD) to acquire lists of deployed U.S. military and civilian personnel to ensure a complete enumeration of this population. DOD used its administrative records to provide the Census Bureau a data file of military and civilian employees who were deployed outside of the United States (while stationed/assigned in the United States). The Census Bureau used this file to enumerate these employees at their stateside address that matched to an existing address in the MTdb.

DOD also provided counts by home state of U.S. military and federal civilian employees stationed or assigned overseas and their dependent living with them at their overseas duty station as part of the Federally Affiliated Count Overseas (FACO) operation. In addition, the Census Bureau worked with other federal departments and agencies to collect counts of their federally employed individuals stationed or assigned overseas and their dependents living with them.

VII. NONRESPONSE FOLLOWUP

After giving the population multiple opportunities to self-respond to the 2020 Census, addresses for which the Census Bureau did not receive a self-response formed the initial universe of addresses for the NRFU operation. The NRFU operation served two purposes: 1) to determine housing unit status for nonresponding addresses, and 2) to enumerate housing units for which a 2020 Census response was not achieved.

For the 2020 Census, we used administrative data to reduce the NRFU workload. Data was used to reduce contacts for vacant and deleted units. The NRFU workload was reduced further by using administrative data, where feasible, to enumerate occupied households that failed to respond after several contact attempts. Additionally, administrative data was used to develop a "Best Time to Contact" model that was intended to reduce NRFU contacts by increasing the likelihood of finding

respondents at home, but this feature was disabled during fielding. Given the unique circumstances in 2020, phone numbers were made available to enumerators to help them complete their work.

A. VACANT AND DELETE IDENTIFICATION

Prior to any fieldwork, an initial set of vacant and nonexistent addresses were identified using administrative data. Pending further information obtained for these addresses, these cases received either one or multiple contact attempts.

Data from the Title 26 composite were used in this operation.

To determine housing unit status, the Census Bureau used data from internal sources, such as the ACS, the MTdb (including USPS DSF status), and the 2010 Census, as well as external sources, such as the U.S. Postal Service's Undeliverable-As-Addressed (UAA) file and IRS records.

More specifically, a multinomial logistic regression model was used to assign predicted probabilities that a given housing unit was vacant or nonexistent. FTI from both IRS 1040 Individual Tax Payer Returns and IRS 1099 Information Returns was used as inputs to the model. The model also included as predictors administrative record household roster information (i.e., count and composition), U.S. Postal Service UAA codes, block group characteristics from the ACS 5-year estimates, and housing unit characteristics from the MTdb. These predicted probabilities were compared with predefined thresholds designed to maximize the identification of vacant and nonexistent units while minimizing misclassification error. The household status probabilities and final status designation was determined in a Title 26 environment and the status designation was sent to a Title 13 environment for use in production. More information on the details of the modeling and its application during 2020 Census fielding can be found in Memorandum 2021.10: Administrative Record Modeling in the 2020 Census (Mulry, Mule, Keller, & Konicki, April 2021).

B. ADMINISTRATIVE RECORDS ENUMERATION

Models were used to identify units with high quality administrative data indicating they were occupied. These models used the same predictors as those used in the vacant and delete model. Occupied households that failed to respond after several contact attempts were eligible for enumeration using their own administrative data in lieu of multiple personal visits. These addresses were matched to administrative data that had been combined to provide the "best" person and housing unit information found throughout the sources.

There were four additional uses of administrative data that were not originally planned in the 2020 Census.

- Original administrative records enumeration did not include American Indian reservations because, as historically hard-to-count areas, we did not want to initially reduce contacts. Given the large number of unresolved cases, we planned to identify Administrative Record (AR) Occupied and AR Closeout Occupied addresses with AR data that could be used if NRFU fieldwork was unable to resolve the status of the address. While not originally planned, this AR processing did result in the AR Occupied and AR Closeout Occupied determinations being sent to the field control system. Thus, these addresses were eligible to receive a reduction in contact attempts during the remainder of the NRFU operation.

- The COVID-19 pandemic caused many college students living in off-campus housing to relocate. To mitigate the risk of undercoverage of off-campus college students, the Census Bureau contacted universities and colleges and requested that they share residency information for their off-campus students. This information was used, if necessary, to enumerate students at the off-campus address.
- Four parishes in Louisiana affected by hurricanes that limited NRFU visits were enumerated with administrative data in lieu of imputation where possible.
- The Census Bureau received permission from the IRS to use a single source of administrative data to determine household size where necessary.

A summary of administrative records enumeration for count and characteristics is below. More information on the details of the modeling and its application during 2020 Census fielding and the enumeration of off-campus addresses can be found in Memorandum 2021.10: Administrative Record Modeling in the 2020 Census (Mulry, Mule, Keller, & Konicki, April 2021). Response records sent to the National Archives and Records Administration (NARA) will not include enumerations with administrative records.

1) Count

Data from the Title 26 composite were used in this operation.

FTI from both IRS 1040 Individual Tax Payer Returns and IRS 1099 Information Returns, as well as non-FTI from sources (including many of those listed in Section IV), was used to determine household count. Where possible, other data sources were used to corroborate information from FTI and unique FTI was overwritten as described in Appendix B so that the final enumeration data were Title 13. The IRS approved the use of its data to determine the number of household members at a NRFU address when the household could not be corroborated with other data sources. This IRS approval applied only to the household size. IRS did not approve the Title 13 use of the names or characteristics of the household members when its data were the sole source. The use of only the household size information for sole source corroboration also was applied to MEDB and IHS data.

2) Characteristics

The non-FTI sources were also used for characteristic enumeration. When possible, we used the 2010 Census and ACS responses before using information from other sources. No FTI was used in characteristic enumeration. Sole-sourced household size cases were not assigned characteristics during administrative records enumeration.

Six characteristics were attempted to be assigned for the 2020 Census, five at the person level and one at the housing unit (HU) level. Only federal sources were used to assign person-level characteristics.

Characteristics Expected to be Enumerated from Administrative Data for the 2020 Census

Person-Level Characteristics	Housing Unit Characteristic
Name	Tenure (own or rent the housing unit)
Age/Date of Birth	
Sex	
Race and Ethnicity	
Relationship to Householder	

C. BEST TIME TO CONTACT MODELING

For the 2020 Census, the goal of enumerator route optimization was to increase the productivity of the enumerators in the NRFU operation by decreasing the miles traveled and total hours spent per case. Assignments were based on factors including enumerator availability, workload location, and the probability of successfully contacting cases at various hours of the day. The enumerators received an optimal routing of attempts to minimize travel.

Data from the Title 26 composite were used in this operation.

Logistic regression models were used to develop predicted probabilities associated with contacting occupied housing units during each hour of the workday (weekdays and weekends). The Census Bureau used data from internal sources, such as the ACS and the MTdb, as well as external sources, such as IRS records, to develop input variables for the model. The predicted contact probabilities were determined in a Title 26 environment and then sent to a Title 13 environment for use in production to optimize case assignments.

While the modeling was done as planned, the best time to contact feature was disabled early in the operation. The program expected that respondents may be home more often given the national response to the COVID-19 pandemic. Disabling the best time to contact component of case assignment allowed more cases to be assigned early in the operation.

D. PHONE NUMBER AVAILABILITY

In areas where in-person follow-up was not permitted due to COVID-19 restrictions, wildfires, or hurricanes, phone numbers associated with an enumerator's workload for the day were displayed. The source of the phone numbers was the Contact Frame. This was an unplanned effort to increase response given the unique circumstances arising during 2020.

VIII. RESPONSE DATA VERIFICATION

The Census Bureau used administrative data to examine the consistency between the respondent-provided or enumerator-collected information and the administrative data sources to determine if further analysis and/or field investigation was warranted.

Name information, in combination with address and other information, can confirm a person’s association with an address as well as the demographic data associated with their household and whether the person can be found at another address. A level of consistency (i.e., matched, nonmatched) between the respondent-provided information or the enumerator-provided data and information available from administrative data was determined.

A. SELF-RESPONSE QUALITY ASSURANCE

Data from the Title 26 composite were used in this operation.

FTI and non-FTI sources were used in matching algorithms to develop a set of Title 13 match results. The match results were then used as inputs to scoring models that help identify suspicious cases for further investigation. The algorithms were run in a Title 26 environment and the match results sent to a Title 13 environment for use in production.

B. ENUMERATOR QUALITY CONTROL

Only data from the Title 13 composite were used in this operation.

Like many of the other processes, quality control efforts used the Title 13 composite described in Section II above as the comparison data set to verify that enumerators conducted interviews appropriately and collected accurate data. The data in the composite was never used in lieu of additional field verification for those housing units with consistency measures below an acceptable level. Furthermore, the administrative data matching results were never used as a sole indicator of whether a response is suspicious and required further analyst and/or field follow-up. Instead, the results from administrative data were used in conjunction with the results of statistical modeling on response data, paradata, and other data elements to flag responses as suspicious.

IX. POST-RESPONSE PROCESSING

A. ADDITIONAL UNDUPLICATION OF PEOPLE IN HOUSING UNITS

During post-processing, various steps were planned to remove duplicates (a person counted more than once). As we tallied census results, we compared these results to other benchmark data and determined that additional duplicates remained in the 2020 Census, even after taking all the steps we had planned. Consequently, after data collection was complete, we took more steps to identify and remove additional duplicated individuals from the census. We did a broad search for duplicates within each state and the District of Columbia. If we found individuals who were initially counted at two different housing units, we first checked if there was any indication from census responses that one of the units was vacant or nonexistent. If so, we removed the duplicated people from the vacant or nonexistent unit and kept them at the other address.

If neither housing unit appeared vacant or nonexistent, we checked whether the people were only associated with one of the addresses in administrative records. If so, we kept them at that address. Data from the Title 26 composite were used in this operation. Review of the administrative data was done in a T26 environment and only a list of the duplicated addresses to drop from the person records was sent to a T13 environment for additional processing. Administrative sources corroborated a person's 2020 Census enumeration at the location. We only used Protected Identification Keys (PIKs) and MAF Identification Numbers (MAFIDs) from the administrative sources. For more information on unduplicating responses, see the Census Bureau blog "How We Unduplicated Responses in the 2020 Census (Keller & King, 2021).

B. COUNT IMPUTATION

After data collection, count imputation 1) assigned final statuses (i.e., occupied, vacant, and nonexistent) to those cases that were unresolved and 2) assigned population counts to records with an occupied status, but without a specified number of occupants. Response records sent to NARA will not include imputations from administrative records.

1) Housing Units

Data from the Title 26 composite was used in this operation.

We used the USPS Undeliverable As Addressed (UAA) file as well as information from IRS 1040, IRS 1099, Census Master Address File, 2020 Census paradata, American Community Survey, and other sources including Medicare and Indian Health Service to place all addresses into groups with similar status and household size (if occupied)[1]. Within each group, a hot-deck imputation process was used to fill in the missing status and count (if occupied) from a census address with a complete status to the address without a complete status. FTI was never used as a direct source for final status assignment.

- The classification of administrative records information for each address was completed in a Title 26 environment using the overwriting process described in Appendix B. These variables were transferred to a Title 13 environment to model the final status when it was unknown and the household roster count for occupied units where the count was unknown.
- UAA information was assigned to each CensusID. The National Processing Center downloaded this information from USPS for the Census Bureau.
- We only used PIKs and MAFIDs from the IRS data.

2) Group Quarters

For some group quarters, we did not have response data to determine precisely how many people should be assigned to an occupied facility. We developed, tested, and applied imputation procedures to cases that could not be resolved through other means. These imputations did not use T26 composite or T13 composite data. Instead, we used information already available on the group quarters under consideration—such as the expected count or the maximum capacity the group quarters reported during advance contact or from data collected in our current surveys. We used the IPEDS database to help determine the GQ count for some colleges and universities if the information from advance contact or current surveys was not available.

The Census Bureau had never before conducted count imputation on unresolved group quarters. This change resulted from the unique difficulties presented by the COVID-19 pandemic, including higher numbers of unresolved cases and difficulty obtaining information through the intended avenues, like personal contact with GQ representatives who may not have been at the GQ location. For more information, see the Census Bureau blogs on this topic (Cantwell, 2021; Jarmin, 2021).

C. CHARACTERISTIC IMPUTATION

After all responses are processed and the final population count is established, there are sometimes missing, inconsistent, or nonvalid characteristic data. Administrative records sources were used as part of the process to impute missing characteristics. Imputation methods included direct substitution from prior Title 13 responses or from other administrative record sources. They also included procedures

[1] The 2010 Census was incorrectly listed as an intended source in the previously published version of this memo. We did use 2010 Census counts for the 2018 End-to-End Census Test model, but further research showed that other sources were more accurate for the 2020 model so we did not use 2010 Census data in the 2020 production. We also used the Census Numident, which was inadvertently omitted in the previous memo.

where missing data were imputed from donors with similar characteristics at the person- or housing-unit level. Response records sent to NARA will not include imputations from administrative records.

Only data from the Title 13 composite above were used in this operation. Sources of data included 2010 Census data, ACS responses, the Census Numident, the Best Race and Ethnicity file, Census Household Composition Key File, HUD PIC/TRACS, and Black Knight. When possible, we tried to impute from the 2010 Census and ACS data before imputing from other sources. No FTI was used in characteristic imputation.

Five characteristics were imputed for the 2020 Census, four at the person level and one at the housing unit (HU) level. Only federal sources were used to impute person-level characteristics.

Characteristics Expected to be Imputed for the 2020 Census

Person-Level Characteristics	Housing Unit Characteristics
Age/Date of Birth	Tenure (own or rent the housing unit)
Sex	
Race and Ethnicity	
Relationship to Householder	

X. PUBLISHING DATA

A. CITIZEN VOTING AGE POPULATION (CVAP) TABLES

The Census Bureau suspended all work on the Post-2020 Census CVAP Special Tabulation on January 12, 2021, following the Executive Order on Ensuring a Lawful and Accurate Enumeration and Apportionment Pursuant to the Decennial Census. The Census Bureau had planned to produce this tabulation using 2020 Census and administrative records data and publish it to the census block level. On January 21, 2021, the Census Bureau advised that it would reengage with the U.S. Department of Justice (DOJ) to confirm that the Citizen Voting Age Population (CVAP) data produced from the American Community Survey (ACS) continue to meet its statistical needs. On February 16, 2021 the Census Bureau received a letter from DOJ stating that the CVAP data from the ACS, on which it has traditionally relied, are adequate for its enforcement of Section 2 of the Voting Rights Act. DOJ did not request compilation or release of additional citizenship or CVAP data beyond this ACS data.

B. COUNT QUESTION RESOLUTION

The 2020 Count Question Resolution program (CQR) allows tribal, state, and local government elected officials to request review for corrections to their jurisdiction's 2020 Census counts. There are historically a small percentage of cases where an incorrect geographic boundary or coding of a housing unit was used to produce the official census population and housing count for a local area. There may also be cases where, because of processing errors, the Census Bureau mistakenly duplicated or deleted living quarters that were identified during the census. The Census Bureau does not collect any additional data during the CQR process but will access existing information to research cases. The CQR process will not research whether respondents or census enumerators incorrectly determined an address's occupancy status or household size during field operations, but only resolve boundary disputes, address geocoding errors, and address coverage errors. Census Bureau records used in the research phase include the MTdb, records collected or used during frame development activities (see section IV), and

2020 Census return metadata. Case submission for the CQR operation will take place from January 3, 2022, through June 30, 2023. Case research and final disposition will complete by September 30, 2023.

XI. COVERAGE EVALUATION

A. POST-ENUMERATION SURVEY

The Post-Enumeration Survey (PES) is one of two primary evaluation tools used to produce estimates of census coverage. For the PES, addresses in the selected sample blocks were listed and enumerated in operations that were independent of the 2020 Census. The Census Bureau identifies matches and nonmatches and discrepancies between the 2020 Census and the PES, for both housing units and people in the sample area. Both computer and clerical components of matching are conducted. The system that conducted computer matching used telephone numbers from administrative data (that is, the Contact Frame) for census records in the sample areas when no telephone number was reported in the census. As a result, the use of the updated telephone numbers could improve computer match rates, thereby improving the efficiency of clerical matching and potential followup operations. Additionally, during PES data processing, the same characteristic imputation methodology that was used for the 2020 Census will be used for the PES. Final PES work will be completed in the summer of 2022.

B. DEMOGRAPHIC ANALYSIS

Demographic Analysis (DA) is the other approach for measuring census coverage. DA refers to a set of methods that have historically been used to develop national-level estimates of the population for comparison with decennial census counts. 2020 DA estimates were developed from current and historical vital statistics from the National Center for Health Statistics (NCHS), data on international migration from ACS, and Medicare records. The data were independent of the census being evaluated. The results will be compared with the census counts by age, sex, limited race groups, and Hispanic origin to produce estimates of net coverage error in the census. The DA program initially planned to use administrative records on Legal Permanent Residents (LPR) from the Office of Immigration Statistics (OIS), which is part of the Department of Homeland Security. Ultimately, these records were not used because of changes to the method for estimating the foreign-born population. For more information, see the technical documentation on the Census Bureau's website (Jensen, et al., 2020). IRS tax returns and the Census Numident file will also be used to develop state and county DA estimates for young children age 0 to 4; however, this is an experimental set of estimates and not one of the official series of 2020 DA estimates. Final DA work will be completed by the end of 2022.

XII. REFERENCES

Cantwell, P. (2021, April). *How We Complete the Census When Households or Group Quarters Don't Respond*. Retrieved from https://www.census.gov/newsroom/blogs/random-samplings/2021/04/imputation-when-households-or-group-quarters-dont-respond.html

Deaver, K. D. (May 2020). *Memorandum 2020.06: Intended Administrative Data Use in the 2020 Census.* Washington, DC: U.S. Census Bureau. Retrieved from https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/memo-series/2020-memo-2020_06.html

Executive Office of the President. (January 20, 2021). *Ensuring a Lawful and Aaccurate Enumeration and Apportionment Pursuant to the Decennial Census.* Retrieved from https://www.federalregister.gov/documents/2021/01/25/2021-01755/ensuring-a-lawful-and-accurate-enumeration-and-apportionment-pursuant-to-the-decennial-census

Jarmin, D. R. (2021, February). *2020 Census Processing Updates*. Retrieved from https://www.census.gov/newsroom/blogs/director/2021/02/2020-census-processing-updates.html

Jensen, E. B., Knapp, A., King, H., Armstrong, D., Johnson, S. L., Sink, L., & Miller, E. (2020). *Methodology for the 2020 Demographic Analysis Estimates.* U.S. Census Bureau. Retrieved from https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2020da_methodology.pdf

Keller, A., & King, R. (2021, April). *How We Unduplicated Responses in the 2020 Census*. Retrieved from https://www.census.gov/newsroom/blogs/random-samplings/2021/04/how_we_unduplicated.html

Mulry, M. H., Mule, T., Keller, A., & Konicki, S. (April 2021). *Memorandum 2021.10: Administrative Record Modeling in the 2020 Census.* Washington, DC: U.S. Census Bureau. Retrieved from https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/memo-series/2020-memo-2021_10.html

U.S. Census Bureau. (December 2018). *2020 Census Operational Plan v4.0.* Washington, DC: U.S. Census Bureau. Retrieved from https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/planning-docs/operational-plan.html

U.S. Census Bureau. (various). *2020 Census Operational Plan and Detailed Operational Plans*. Retrieved from https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/op-plans.html

APPENDIX A –DATA SOURCES USED IN THE 2020 CENSUS

Use / Source	Frame Development	Resp Motivation	Non-ID Address Validation	Paper Data Capture QA	Special Enumerations.
USPS DSF	XXX				
Ongoing MAF/TIGER System updates from partnerships and fieldwork	XXX				
2020 Census partnership data (LUCA, NC)	XXX				
Lists of GQs and TLs from federal sources (Medicare, BOP, ICE, Bureau of Indian Affairs, U.S. Marshals, Maritime Agencies, Military), state & local partners	XXX				
STR Inc. list of hotels/motels	XXX				
MTdb Extracts			XXX	XXX	XXX[1]
Modeled self-response predictions		XXX			
Audience segmentation information		XXX			
FCC Internet Connectivity		XXX			
GQ administrator lists					XXX
Off-campus student records					
IPEDS					
DOD Deployed data and FACO inputs *					XXX
CMS MEDB			XXX	XXX	
HUD FHA IDB (CHUMS)			XXX	XXX	
HUD Longitudinal (PIC/TRACS)			XXX	XXX	
IHS Patient Registration			XXX	XXX	
IRS 1040					
IRS 1099					
NCHS Vital Statistics (births and deaths)					
SSS Registration			XXX	XXX	
State/Local Program Data			XXX		
USPS NCOA			XXX	XXX	
ACS Data		XXX[1]			
ACS Contact History					
2000 Census Data	XXX		XXX	XXX	
2010 Census Data	XXX		XXX	XXX	
Census Numident			XXX	XXX	
Census Numident Alternate Names			XXX		
Census HH Composition Key					
Contact Frame		XXX	XXX	XXX	
Best Race and Ethnicity					
Black Knight					
DAR Partners			XXX	XXX	
Targus Federal Consumer and Wireless			XXX	XXX	
VSGI			XXX	XXX	
USPS UAA					
Planning Database		XXX			
* Other federal agencies provided their FACO numbers directly into the system, not as a file to 2020 production systems and are therefore not listed here.					

[1] This source was inadvertently omitted from the previously released memo.

Source \ Use	NRFU Vacant/Delete Identification	NRFU AR Enumeration			NRFU Phone Avail.	NRFU Best Time to Contact
		Occupied / AR Enum. Eligible	Count	Characteristics		
USPS DSF						
Ongoing MAF/TIGER System updates from partnerships and fieldwork						
2020 Census partnership data (LUCA, NC)						
Lists of GQs and TLs from federal sources (Medicare, BOP, ICE, Bureau of Indian Affairs, U.S. Marshals, Maritime Agencies, Military), state & local partners						
STR Inc. list of hotels/motels						
MTdb Extracts	XXX	XXX				XXX
Modeled self-response predictions						
Audience segmentation information						
FCC Internet Connectivity						
GQ administrator lists						
Off-campus student records			XXX	XXX		
IPEDS						
DOD Deployed data and FACO inputs *						
CMS MEDB	XXX	XXX	XXX			
HUD FHA IDB (CHUMS)						
HUD Longitudinal (PIC/TRACS)				XXX		
IHS Patient Registration	XXX	XXX	XXX			
IRS 1040	XXX	XXX	XXX			XXX
IRS 1099	XXX	XXX	XXX			XXX
NCHS Vital Statistics (births and deaths)						
SSS Registration	XXX	XXX				
State/Local Program Data						
USPS NCOA	XXX	XXX				
ACS Data	XXX	XXX		XXX		XXX
ACS Contact History						XXX
2000 Census Data						
2010 Census Data	XXX	XXX		XXX		
Census Numident	XXX	XXX	XXX	XXX		XXX
Census Numident Alternate Names						
Census HH Composition Key	XXX	XXX	XXX	XXX		XXX
Contact Frame					XXX	
Best Race and Ethnicity	XXX	XXX		XXX		XXX
Black Knight				XXX		
DAR Partners	XXX	XXX				
Targus Federal Consumer and Wireless						
VSGI	XXX	XXX				XXX
USPS UAA	XXX	XXX				
Planning Database	XXX	XXX				XXX

* Other federal agencies provided their FACO numbers directly into the system, not as a file to 2020 production systems and are therefore not listed here.

Source \ Use	SRQA	Enum. QC	Additional Unduplication	Count Imputation		Character. Imputation	CQR	PES	DA
				HU	GQ				
USPS DSF									
Ongoing MAF/TIGER System updates from partnerships and fieldwork							XXX		
2020 Census partnership data (LUCA, NC)							XXX		
Lists of GQs and TLs from federal sources (Medicare, BOP, ICE, Bureau of Indian Affairs, U.S. Marshals, Maritime Agencies, Military), state & local partners							XXX		
STR Inc. list of hotels/motels							XXX		
MTdb Extracts	XXX	XXX	XXX	XXX	XXX**		XXX		
Modeled self-response predictions									
Audience segmentation information									
FCC Internet Connectivity									
GQ administrator lists									
Off-campus student records									
IPEDS					XXX				
DOD Deployed data and FACO inputs *									
CMS MEDB	XXX	XXX	XXX	XXX					XXX
HUD FHA IDB (CHUMS)		XXX							
HUD Longitudinal (PIC/TRACS)	XXX	XXX				XXX		XXX	
IHS Patient Registration	XXX	XXX	XXX	XXX					
IRS 1040	XXX		XXX	XXX					XXX
IRS 1099	XXX		XXX	XXX					
NCHS Vital Statistics (births and deaths)									XXX
SSS Registration	XXX	XXX							
State/Local Program Data									
USPS NCOA	XXX	XXX							
ACS Data				XXX	XXX	XXX		XXX	XXX
ACS Contact History									
2000 Census Data	XXX	XXX					XXX		
2010 Census Data	XXX	XXX				XXX	XXX	XXX	
Census Numident	XXX	XXX	XXX	XXX		XXX		XXX	XXX
Census Numident Alternate Names	XXX	XXX							
Census HH Composition Key			XXX	XXX		XXX		XXX	
Contact Frame	XXX	XXX						XXX	
Best Race and Ethnicity						XXX		XXX	
Black Knight						XXX		XXX	
DAR Partners	XXX	XXX							
Targus Federal Consumer and Wireless	XXX	XXX							
VSGI	XXX	XXX							
USPS UAA				XXX					
Planning Database									

* Other federal agencies provided their FACO numbers directly into the system, not as a file to 2020 production systems and are therefore not listed here

** The count at the GQ from the last ACS or current surveys visit is contained in this source.

APPENDIX B - OVERWRITING EXAMPLE

This is an example of how the U.S. Census Bureau can use FTI and information from multiple sources in the 2020 Census to inform census fieldwork decisions and conduct administrative record enumeration. This approach adheres to procedures provided by IRS Office of Safeguards regarding validation and overwriting FTI. This example was presented to IRS Statistics of Income staff in July 2016 during the review and approval process for the 2020 Census Production Predominant Purpose Statement.

This example presents a hypothetical Mars family who filed an IRS 1040 return at 101 Main Street. The source, validation, and overwriting involving different variables and values are presented. Data variables and values that are FTI are shown in red shaded background with white letters. Data variables and values that have been overwritten in accordance with the IRS regulations are shown in white background with green letters.

1. IRS 1040 and 1099 Information

Each month, the IRS delivers the IRS 1040 Individual tax returns. Table 1 shows the two hypothetical families at 101 and 102 Main Street. The Mars family has four people and lives at 101 Main Street. John Smith lives by himself at 102 Main Street. Since this is the 1040 information delivered to the Census Bureau, all of the variables are FTI and thus colored in red. Among the variables, the IRS provides the Census Bureau with the address, tax identification number (TIN) of filers and dependents, filing status, and other FTI variables. For 101 Main Street, primary filer Michael Mars had a full name, SSN present but the other people have only last name and SSNs provided. For 102 Main Street, single primary filer John Smith did not provide an SSN when filing. While not shown in the table, the IRS also delivers a similar file containing the 1099 Information return file as well.

Table 1: IRS 1040 FTI Record of the Mars Family at 101 Main St.

Address	Primary Filer TIN	Primary Name	Secondary Filer TIN	Secondary Name	Filing Status	Dependent 1 TIN	Dependent 1 Name	Dependent 2 TIN	Dependent 2 Name	Dependent 3 TIN	Dependent 3 Name	Other FTI Variables
101 Main St.	111-11-1111	Michael Mars	111-11-1112	Mars	Joint Married	111-11-1113	Mars	111-11-1114	Mars	None	None	…
102 Main St.	blank	John Smith	None	None	Single	None	None	None	None	None	None	…

2. Census Person Validation System (PVS) Processing of IRS 1040 and 1099 Delivered File

After receipt of the IRS 1040 file, the Census Bureau processes the IRS 1040 file through the Person Validation System. This processing uses the address provided by the IRS and links this address to an address already to the Census Master Address File (MAF). This assigns the Census Bureau's MAF Identification Number (MAFID) to the record. The PVS processing was able to assign the Census Bureau's Protected Identification Key (PIK) to replace the IRS TIN based on combinations of SSN, Name and Address fields. For John Smith, the PVS processing was unable to assign a PIK. This is an example of a file that has a combination of FTI and non-FTI variables. The address and PIKs are colored in green since they have been overwritten in accordance with IRS regulations. IRS provided name information has been removed. The Census Bureau does similar processing of the 1099 Information file as well.

Table 2: IRS 1040 Record for Mars Family after Census PVS Processing

Census MAFID	Address	Census PIK 1	Census PIK 2	Filing Status	Census PIK 3	Census PIK 4	Other FTI Variables …
123412341	101 Main St.	999-99-9991	999-99-9992	Joint Married	999-99-9993	999-99-9994	…
123412342	102 Main St.	Blank	Blank	Single	Blank	Blank	

Since 102 Main St. has no PIKs, we are not able to build a roster using IRS 1040 data for this address. The remainder of this document will focus on using the IRS data to see if we can reduce contacts for 101 Main St. For the 2020 Census, the Census Bureau will be building rosters for households using **multiple sources** determined by September 2018**. The remainder of this document will assume that an additional source besides IRS information can corroborate that the Mars family lives at 101 Main St.**

3. Processing of Administrative Records to Determine Census Fieldwork

The Administrative Records Modeling team has been researching under the DMS 863 project the possibility of using administrative records and third-party data to determine if the number of contacts can be reduced. The 2020 Production DMS 987 project will implement this usage in the 2020 Census. In this example, we will show parts of the processing. Step 2 showed that the Census Bureau had a file that had a combination of FTI and non-FTI variables. Our research has shown that other variables created using FTI are powerful predictors in our determination. This example highlights two of them. Table 3 shows an example of where we create additional variables based on FTI information to be used in the

determination about how many contacts to conduct during fieldwork. There is a history of using IRS information combined with other sources to determine fieldwork for the decennial census. For the 2010 Census, part of the workload for the Coverage Followup operation was based on the comparison of census rosters to administrative records households that included IRS 1040, IRS 1099, and other non-FTI sources.

An example of two variables created using FTI information is:

- Was any person in the Mars household found on another 1040 return filed this year? Our research has found that if any person in the household is associated with another 1040 return, then there are possible questions about the administrative record roster assembled.

- Was any person in the Mars household found on the 1040 return filed at this address last year? Our research has seen that finding at least one person on 1040 returns at the same address in consecutive years is more highly correlated with count and household composition agreement.

Table 3: Example of Creating Two Additional FTI Variables to be used in Determination

Census MAFID	Census PIK	Filing Status	Processing Week	Found on another 1040 return	Found on last year 1040 at same address
123412341	999-99-9991	Joint Married	12	No	Yes
123412341	999-99-9992	Joint Married	12	No	Yes
123412341	999-99-9993	Joint Married	12	No	Yes
123412341	999-99-9994	Joint Married	12	No	Yes

The Administrative Record determination processing uses predictive models. These predictive models are based or “trained” on comparison of the rosters built for 2010 NRFU addresses using people from 2010 versions of the administrative record sources to the person enumerated at those addresses in the 2010 Census. Based on building a roster using the most recent versions of the administrative record files, we can use predictive models can allow us to make determinations if the roster from administrative record sources is similar enough to a census fieldwork enumeration that we can reduce the number of contacts. In our example, we have built a roster of the Mars family.

For every 2020 household roster built from administrative record sources, we can predict:

- How likely is it that all of the Mars family should be counted at the address? From a scale of 0 to 1, we want this to be as close to 1 as possible.

- How likely is it that the household composition observed for the Mars family from IRS 1040 and other sources would match census fieldwork[6]? From a scale of 0 to 1, we want this to be as close as possible to 1 as well.

Figure 1 shows a graphical description of the number of contacts during fieldwork determination. Each axis represents one of the two predictions listed above. The results are shown in three colors of green, yellow, and red.

For the 2020 Census, we are implementing the following three classifications. In our example, the Mars family will have high enough results for the two predictions that we will reduce the contacts so that this address is classified as administrative record occupied and will have reduced contacts during the NRFU operation.

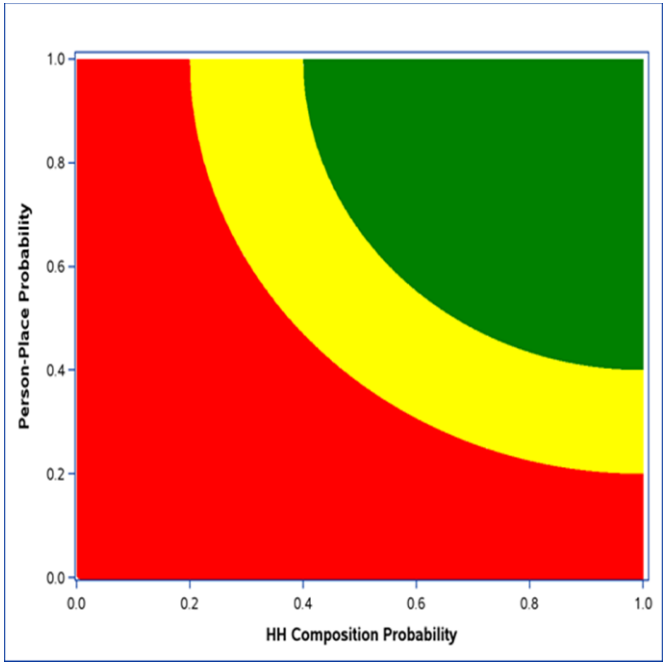
- Administrative Record Occupied in NRFU (Green): If the two probabilities land the address in the green area then the address will be administrative record occupied. We will be reducing contacts for these addresses.

- Administrative Record Occupied for Imputation (Yellow): The Census Bureau has done research on using administrative records after all of the contacts in the NRFU as an alternative to count imputation. For the 2016 Census Test, these addresses will receive six contacts. If the two predicted probabilities land the address in the yellow area, then the address we will use the roster of administrative records as an alternative to imputation

- Not Use (Red) – If the predicted probabilities are in red, then we would implement full fieldwork. We would not want to use these rosters built from administrative records as an alternative to count imputation.

[6] In this example, we would be saying how likely it would be that the census would enumerate a household with two adults and one or more children present.

Figure 1: Graphical Representation of Hypothetical Administrative Record Occupied Determination



4. Overwriting for Enumeration Purposes

Based on the result for the Mars family, we have determined that we will reduce contacts during NRFU. While we used FTI in determining how to conduct census fieldwork, the following will use the Mars family as an example to show how we will not be using FTI for enumeration.

For enumeration, the only information originally sourced from FTI that we use is the address and the PIK. Based on the IRS Office of Safeguards response to the Census Bureau's January 2016 overwriting inquiry, the Census MAFID and PIK have overwritten the FTI. In accordance with the guidance from IRS, we will document the overwriting and the date that it has occurred. This file is no longer FTI. Table 4 provides an example.

Table 4: Example of Overwriting Documentation File for Future IRS Office of Safeguards Review

Census MAFID	Unique Person Number	Census PIK	Overwriting Date
123412341	1	999-99-9991	May 2, 2016
123412341	2	999-99-9992	May 2, 2016
123412341	3	999-99-9993	May 2, 2016
123412341	4	999-99-9994	May 2, 2016

5. Obtain Short Form Characteristics from non-FTI sources

For census enumeration purposes, we need to obtain the name, age, date of birth sex, relationship, race, Hispanic origin, and tenure information for the people at this address. For these characteristics, this information will be obtained from non-IRS Title 26 sources. This includes past Census Bureau responses or other administrative record information sources.

Table 5 shows the administrative record enumeration for 101 Main Street. All of the fields are colored green to indicate that no FTI information was used in these variables. Name was obtained from the Census Numident file or past census responses. To further, emphasize in this example that no FTI information is being used for enumeration purposes, the relationship for Mars family is set to missing. From the IRS 1040 return, we did see that the Mars family filed a Joint Married return, but we do want to point out that we are not using that information. Since IRS is the sole source of the relationship, we are unable to use this information as a direct assignment. Our characteristic imputation will have to assign the relationship status for this household.

Table 5: Example of Administrative Record Enumeration of Mars Family Using non-FTI information

Census Address	Unique Person Number	Name	Age	Date Of Birth	Sex	Relationship	Race	Hispanic Origin
101 Main St.	1	Michael Mars	43	1/1/1973	Male	Missing	White	Non-Hispanic
101 Main St.	2	Mary Mars	40	2/2/1976	Female	Missing	White	Non-Hispanic
101 Main St.	3	Hershey Mars	17	3/3/1999	Male	Missing	White	Non-Hispanic
101 Main St.	4	Lucy Mars	14	3/4/2002	Male	Missing	White	Non-Hispanic

Note: Census Day is April 1, 2016 for this example.

Step 3 showed that for some addresses it may be determined administrative records would only be used as an alternative to imputation. For these addresses, they will receive the full amount of contacts during the NRFU operation. If at the end of the census fieldwork, their occupancy and/or population status are still unresolved then administrative records will be used in a similar way as laid out in this section.

6. Creating a Census Response File

Based on this processing, this information can then be delivered to our Title 13 server and be combined with the other Census Bureau responses. This information can be used for all Title 13 purposes of the 2020 Census including archiving to be made available 72 years after Census Day. Table 6 shows a simple example of variables and values that would be included.

Table 6: Census Response File

Address	Enumeration Operation	Unique Person Number	Name	Age	Sex	Relationship	Race	Hispanic Origin
101 Main St.	Administrative Records	1	Michael Mars	43	Male	Missing	White	Non-Hispanic
101 Main St.	Administrative Records	2	Mary Mars	40	Female	Missing	White	Non-Hispanic
101 Main St.	Administrative Records	3	Hershey Mars	17	Male	Missing	White	Non-Hispanic
101 Main St.	Administrative Records	4	Lucy Mars	14	Female	Missing	White	Non-Hispanic
102 Main St.	Internet Self-Response	1	Thomas Lindt	28	Male	Householder	Black	Non-Hispanic
102 Main St.	Internet Self-Response	2	Sally Lindt	26	Female	Spouse	Black	Non-Hispanic
103 Main St.	NRFU	1	John Haribu	33	Male	Householder	SOR	Hispanic

Summary

This document shows how FTI will and will not be used in relation to administrative record enumeration in the 2020 Census. While FTI will be used during the processing to determine the number of contacts for census fieldwork purposes, the example shows that FTI will not be used in the direct enumeration.