Administrative Record Use in the 2020 Census For Modeling and Processing

Andrew Keller[1]
U.S. Census Bureau, 4600 Silver Hill Rd., Washington, DC 20233

Abstract

The paper discusses two aspects of administrative records use in the 2020 Census. We divide the discussion into two parts: First, we briefly examine the use of administrative records in the nonresponse followup (NRFU) operation to inform the field contact strategy. The NRFU operation conducted in-person visits to addresses which failed to self-respond. The goal was to reduce the number of contacts for addresses with high-quality administrative records. To accomplish this, statistical models were fit using the 2010 Census NRFU as training data. Predictions of address status and count were then applied using administrative records available during the 2020 Census.

The second aspect of administrative records use occurred in the processing of response data. Processing refers to the post-data collection actions generating the final census data files (US Census Bureau, 2017). After identifying addresses with sufficient quality administrative records to justify the assigned status, administrative records were then used to enumerate occupied, vacant, and delete addresses with no response. That process employed past census and administrative data to assign characteristics to occupied housing units.

Key Words: Census, Administrative Records

1. Introduction

The Census Bureau researched fundamental changes to the design and implementation of the 2020 Census. One major innovation research area noted in the 2020 Operational Plan (U.S. Census Bureau 2018) incorporated administrative records (AR) into the census design. This paper divides the use of AR in the 2020 Census into two major areas. In Section 2, we discuss the use of statistical models informed by AR to reduce the number of contacts for addresses with high-quality AR. During the 2020 Census, predictions of address status and count were applied using the available AR.

The other major area concerning AR use in post-processing is divided into three sections. Section 3 documents AR Enumeration - the process by which an address was assigned a status and subsequently enumerated using administrative data. Section 4 discusses how AR were used for count imputation of housing unit (HU) addresses. This was the

[1] The views expressed on statistical, methodological, technical, or operational issues are those of the author and not those of the U.S. Census Bureau. Results are from 2020 Census Data Quality Metrics: Release 1 (DRB Clearance CBDRB-FY21-DSSD007-0012) 2020 Census Data Quality (Accessed August 2021) and 2020 Census Data Quality Metrics: Release 3 (DRB Clearance CBDRB-FY21-295) 2020 Census Data Quality (Accessed August 2021)

concluding step to produce the final population counts. Section 5 discusses supplementing missing person and HU characteristics on the 2020 Census with 2010 Census and American Community Survey (ACS) responses, as well as administrative data. This information aided in production of the final characteristics. All these uses of AR for post-processing were implemented in the 2020 Census. Section 6 has an introductory set of results and Section 7 summarizes the document. At this time, more detailed results are still being tabulated for later dissemination.

2. Administrative Records Modeling

Intercensal research starting in 2012 developed methods to combine and use several administrative sources to identify occupied, vacant, and delete addresses prior to or after minimal NRFU fieldwork, thus reducing the number of enumerator visits (Keller et al., 2018). A delete status meant that the address did not contain a structure which met the Census Bureau's definition of a HU. During the 2020 Census, the administrative sources were used to create decision rules about field contact strategies in NRFU. This allowed resources to focus on units where administrative data were less reliable or unavailable. Sections 2.1 and 2.2 provide information about the distance function methodology used in the 2020 Census. Mulry et al. (2021) provide more information about the 2020 Census processing and the necessary adjustments to account for COVID-19 delays.

2.1 Administrative Records Enumeration Distance Thresholds

To identify vacant units with AR, we developed a multinomial logit model which predicted the probability that an AR address would have been enumerated as vacant during the 2010 Census. Keller et al. (2018) provide more discussion of the vacancy model. Independent variables in the model included variables indicating whether the census mailings could be delivered to the address and whether the AR sources indicated anyone lived at the address. The dependent variable had three possible values for each AR address in the NRFU universe:

- occupied
- vacant, or
- delete (i.e., not a HU).

We defined a Euclidian vacant distance function for AR Vacant identification as:

$$d_{AR_{Vac}} = \sqrt{(1-\hat{p}_{vacant})^2 + \left(0-\hat{p}_{occupied}\right)^2}$$

where:

$\hat{p}_{vacant}$: probability that address was a vacant unit via vacancy model

$\hat{p}_{occupied}$: probability that address was an occupied unit via vacancy model

The formula shows that cases with the smallest distance were those with the highest vacant probability and lowest occupied probability. Starting with the smallest vacant distance, AR Vacant cases were identified by allowing for increased vacant distance values up to a threshold. This threshold was based on analysis of 2010 Census NRFU data.

We defined a Euclidian delete distance function $d_{AR_{Del}}$ for AR Delete identification as:

$$d_{AR_{Del}} = \sqrt{(1-\hat{p}_{delete})^2 + \left(0-\hat{p}_{occupied}\right)^2}$$

where:

$\hat{p}_{delete}$: probability that address did not contain a structure which met the Census Bureau's definition of a HU via vacancy model

$\hat{p}_{occupied}$: probability that address was an occupied unit via vacancy model

The formula shows that cases with the smallest distance were those with the highest delete probability and lowest occupied probability. Starting with the smallest delete distance, AR Delete cases were identified by allowing for increased delete distance values up to a threshold. This threshold was based on analysis of 2010 Census NRFU data.

Two models were developed to identify AR Occupied units: a person-place model and a household (HH) composition model. Independent variables in the occupied models included variables indicating which AR sources placed people at the address and whether these people were found at a different address in the AR sources. The person-place model predicted the probability that an AR person would be enumerated at the sample address if fieldwork was conducted. The dependent variable was whether the AR person was at the address in the 2010 Census. The HH composition model predicted the probability that the sample address would have the same HH composition determined by NRFU fieldwork as its pre-identified AR HH composition. HH composition is defined by the number of adults in the unit and the absence or presence of children. The dependent variable was the 2010 Census HH composition. Keller et al. (2018) provide more discussion of the person-place and HH composition models.

Similar to AR Vacant and AR Delete, we defined a Euclidian occupied distance function $d_{AR_{Occ}}$ for AR Occupied identification as:

$$d_{AR_{Occ}} = \sqrt{\left(1 - \hat{p}_{person-place}\right)^2 + \left(1 - \hat{p}_{HH\ composition}\right)^2}$$

where:

$\hat{p}_{person-place}$: probability that an AR person would be enumerated at the sample address if fieldwork was conducted via person-place model

$\hat{p}_{HH\ composition}$: probability that the sample address would have the same HH composition determined by NRFU fieldwork as its pre-identified AR HH composition via HH composition model

The formula shows that cases with the smallest occupied distance were those where the person-place probability was closest to one and the HH composition probability was closest to one (i.e. the (1,1) point). Starting with the smallest occupied distance, AR Occupied cases were identified by allowing for increased occupied distance values up to a threshold. This threshold was based on analysis of 2010 Census NRFU data.

2.2 Administrative Records Distance Threshold Application

For each address in the 2020 Census, the vacant, delete, occupied thresholds were established. The distances were then categorized based on the quality of their AR data. The classification assignment enabled the NRFU field operation to operationalize their treatment for cases that had not previously responded by the time fieldwork began.

For the NRFU operation, nonresponding addresses were classified into four categories according to the quality of AR data for that address. They were AR Occupied, AR

Vacant, AR Delete, and AR No Determination. These categories informed the NRFU contact strategy. Generally, units identified as AR Occupied, AR Vacant, or AR Delete received fewer address visits. If those cases were not resolved by enumerators in the field or by a self-response, then the administrative data were used as response data. All cases in the NRFU workload received at least one visit regardless of the quality of the AR data. There were seven statuses assigned.

One-visit AR Vacant, **One-visit AR Delete**, and **One-visit AR Occupied** cases had the lowest distance thresholds for their respective status. They signaled that the address would only need to be visited one time provided that other business criteria were met. One-visit AR Occupied cases also indicated that no proxy response was to be sought. **Closeout AR Vacant**, **Closeout AR Delete**, and **Closeout AR Occupied** meant that the address could be taken out of the NRFU operation if it was not resolved by the time NRFU began its closeout operation. The resulting enumeration would be generated from the AR status. **AR No Determination** addresses indicated that no AR status was assigned due to the inability to assign a vacant, delete, or occupied status from the AR data. Because new AR data was received on a monthly basis, generally monthly updates to the statuses were provided to the field between May 2020 and the end of September 2020. Mulry et al. (2021) provide more information regarding the monthly updates and 2020 Census processing with respect to AR modeling.

3. Administrative Records Enumeration

Following the completion of the NRFU operation, AR Enumeration assigned either an occupied, vacant, or delete status based on the pertinent AR data available for the address. AR Enumeration occurred only after two conditions had been satisfied. First, the address had not responded through any other census operation including census mailings and enumerator visits to the address. In short, preference was given to any other response; AR Enumeration was only used for addresses lacking a response. For example, if there was a HH or proxy response from NRFU, AR Enumeration was not used. Second, the address had sufficient quality AR data justifying the assigned status. Keller et al. (2018) document the criteria to determine addresses with appropriate AR data for enumeration. In the scenario of an occupied AR Enumeration, we constructed a list of residents for the HU taken from AR. This AR roster was formed using the union of unduplicated people associated with that HU from Internal Revenue Service data, Center for Medicare and Medicaid Services data, Indian Health Service data, and the Census Household Composition Key file. Note that multiple sources had to confirm presence of at least one person listed as part of the AR roster for an AR Enumeration to occur.

In terms of processing, AR Enumeration cases were added to the set of census responses in the same manner as self-response or enumerator-based responses. The similarity stemmed from the fact that the AR response data were constructed from information associated exclusively with the persons on the AR roster and assumed no matching error. People in occupied AR Enumeration units then had characteristics directly substituted from their own past reports to the Census Bureau (which included the 2010 Census or ACS) or AR data. This stands in contrast to address-level count imputation of HU addresses (discussed in Section 4) where their final status and count were supplied by donor addresses.

3.1 Administrative Records Enumeration Methodology

AR Occupied responses constructed a HH from AR data. If possible, each person in the HH was linked to its corresponding 2010 Census or ACS reported data or administrative sources. Each address was linked to its corresponding administrative housing record. Table 1 shows the assignment hierarchy for each characteristic.

Table 1: AR Enumeration Characteristic Assignment Hierarchy

Characteristic	Hierarchy
Age and Date of Birth	1. 2010 Census Age and Date of Birth Report 2. Census Numident Date of Birth[2]
Sex	1. 2010 Census Sex Report 2. Census Numident Sex
Race	1. 2018 through 2010 American Community Survey (ACS) Race Report 2. 2010 Census Race Report 3. 2018 through 2010 ACS Ancestry Report 4. 2018 through 2010 ACS Place of Birth Report 5. Census Numident Place of Birth 6. Center for Administrative Records Research and Application (CARRA) Best Race and Hispanic Origin File Race Value
Hispanic Origin	1. 2018 through 2010 American Community Survey (ACS) Hispanic Origin Report 2. 2010 Census Hispanic Origin Report 3. 2018 through 2010 ACS Ancestry Report 4. 2018 through 2010 ACS Place of Birth Report 5. Census Numident Place of Birth 6. CARRA Best Race and Hispanic Origin File Hispanic Origin Value
Relationship to Householder	1. Census Household Key file
Tenure	1. Housing and Urban Development Public and Indian Housing Information Center Tenant Rental Assistance Certification System file 2. Third-Party Commercial Data

In general, for AR Enumeration in the 2020 Census, age/date of birth and sex were assigned for each person from either the 2010 Census report or from the Census Numident. This was intentional, as presence of date of birth was a necessary condition to call a case AR Occupied. Race and Hispanic origin were assigned from 2010 Census, ACS, or AR data.

With respect to relationship to householder for AR Enumeration, the householder was assigned to be the oldest person on the roster. Relationship to householder is a relative characteristic in the sense that the correct value can change depending on who is identified as the householder. With that in mind, the Census Bureau developed a Census Household Key file to link children to their parents. The dataset identifies parent-child relationships using Social Security data from the Census Numident. For more on the Census Household Key file, see Luque and Wagner (2015). For AR Enumeration, the

[2] The Census Numident is an internal file developed from the Social Security Administration Numident file, containing one record for each Social Security Number.

Census Household Key file was used to determine whether any people in occupied AR Enumeration units could be assigned a biological child relationship to the householder.

Occupied AR Enumeration units were searched for on the Public and Indian Housing Information Center Tenant Rental Assistance Certification System file from the Department of Housing and Urban Development. If they were found, a rental status was assigned. If not, they were subsequently searched for on third-party commercial data to see if the unit was corporate owned or if the unit had an active mortgage in which the owner was the resident. If the unit was corporate owned, a rental status was assigned. If the unit had an active mortgage in which the mortgage owner was the resident, an owned with a mortgage status was assigned.

As mentioned earlier, AR Enumerations were processed like traditional methods of enumeration such as self-response or enumerator-based responses. Similar to those methods, some AR enumerations had missing characteristics. In cases where AR enumerations were missing items, they were imputed through characteristic imputation during the formation of the final characteristics file. This was the same imputation processing that occurred for people and HUs responding in the traditional manner with missing characteristic data.

AR Vacant cases were assigned a vacant status if the address had not responded through any other census operation and the address had sufficient quality AR data justifying the vacant assignment. In the case of AR Vacant units, the detailed vacancy reason was imputed during the creation of the final characteristics file. Similar to AR Vacant units, AR Delete cases were assigned a delete status if the address had not responded through any other census operation and the address had sufficient quality AR data justifying the delete assignment. In contrast to vacant HUs, deleted addresses are not HUs and were not part of the final count of HUs. Hence, no characteristic imputation was needed.

4. Administrative Records Use for Count Imputation of HU Addresses

To create the final population counts, each address was assigned a final status of occupied, vacant, or delete. If the status was occupied, then the HU record had a population count greater than zero. At the end of the census data collection operations including self response, enumerator response, and AR enumerations, some addresses lacked an address status or population count or both. The population count from the AR roster was used along with other address-level and operational covariates to model the address status and population count for addresses when either or both were unknown. Specifically, the AR count from the unresolved address helped place the address in a model cell. Note that we did not directly assign the AR count as the final population count for addresses in the count imputation universe.

5. Using Administrative Records for Item Missing Data and Substitution

This section describes how 2010 Census and ACS responses, and AR data were used to impute person and HU characteristics. Those data were then used to construct the final characteristics file. Section 5.1 discusses AR use for characteristic imputation when only some characteristics for the census HH roster were missing. Section 5.2 discusses characteristic imputation when the only the population count was provided for the HU.

5.1 Using Administrative Records for Item Missing Data

To begin, the final counts of persons and HUs were created. To use the AR characteristics, people on the final population counts file were linked to the AR reference file. The link resulted from matching person data reported on the census response such as name, date of birth, and sex to the AR reference file that contained those same covariates and the unique identifier.

It was not necessary to have all characteristics reported to link to AR. For example, people with missing sex or date of birth could still link to the AR reference file. Of the 308.7 million people in the 2010 Census, Rastogi and O'Hara (2012) show that about 279.2 million person records on the 2010 Census could be linked to AR. People were matched to their corresponding person record in AR. Similarly, HUs were matched to their corresponding HU record in AR. The associated characteristics were then used for imputation when data was missing.

5.2 Using Administrative Records for Population Count-Only Cases

For some occupied units, only a population count was known. These units were called substitution units and all people in them were missing all characteristics. Substitution units occurred primarily in two situations. First, a respondent or proxy only provided a population count for the unit and no person demographic data. Second, the address was imputed as occupied. Since no characteristic data were reported, individual person records did not have the minimum characteristic information to link to AR.

For substitution units, we checked that the AR roster count was the same as the census count. Section 3 gives a high-level summary of how an AR roster was created. If the counts agreed, the AR roster was copied into the substitution HU. Substitution units without a matching AR count were imputed characteristics from a responding unit in the census of the same count where all people had all characteristics reported. Hence, no AR data were used to assign characteristics for those cases.

6. Results

Of the 151.8 million addresses in the 2020 Census, 3.20 percent were enumerated as AR Occupied, 1.15 percent were enumerated as AR Vacant, and 0.24 percent were enumerated as AR Delete. This was 4.59 percent of the total address universe. In other words, on average one of every 22 addresses was enumerated using AR. Among the NRFU universe, 9.51 percent were enumerated as AR Occupied, 3.43 percent were enumerated as AR Vacant, and 0.70 percent were enumerated as AR Delete. This was 13.64 percent of the NRFU address universe[3].

Looking at the county level, the national average of NRFU occupied cases enumerated via AR was 16.47 percent with a standard deviation of 5.99 percent and a median of 17.24 percent. At the tract level, the national average of NRFU occupied cases enumerated via AR was 19.75 percent with a standard deviation of 8.05 percent and a median of 19.43 percent[4]. Subsequent results will show the impact of AR on characteristic imputation.

7. Conclusion

This paper describes how AR data were used to inform the field operation. It then describes how AR were used in post-processing for the 2020 Census to create final population counts and characteristics. First, AR were used to enumerate a select set of occupied, vacant, and delete addresses. However, the AR data were only used for addresses where there was no response and AR data were of sufficient quality to justify the assigned status. For units enumerated as occupied using AR, rosters were developed. Second, AR roster counts along with other address-level and operational covariates were used to group responding and unresolved addresses for count imputation of HU addresses. Last, links were made between the current census files and AR data so that 2010 Census, ACS, and administrative data could aid 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 HU.

8. References

Keller, A., Mule, V.T., Morris, D.S., Konicki, S. (2018). “A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 U.S. Census, Journal of Official Statistics, 34(3), 599-624. DOI: http://dx.doi.org/10.2478/JOS-2018-0029

Luque, A. and Wagner, D. (2015). “Assessing Coverage and Quality of the 2007 Prototype Census Kidlink Database.” Center for Administrative Records Research and Applications Working Paper #2015-07. Washington, DC: U.S. Census Bureau.

Mulry, M.H., Mule, V.T., Keller, A., and Konicki, S. (2021). “Administrative Record Modeling in the 2020 Census,” 2020 Census Planning Documents: #2021-10. U.S. Census Bureau. Administrative Record Modeling in the 2020 Census (accessed August 2021).

Rastogi, S. and O’Hara, A. (2012). “2010 Census Match Study Report,” 2010 Census Planning Memorandum Series. Available at: 2010 Census Match Study Report (accessed August 2021).

[3] Results are from 2020 Census Data Quality Metrics: Release 1 (DRB Clearance CBDRB-FY21-DSSD007-0012) 2020 Census Data Quality (Accessed August 2021)

[4] Results are from 2020 Census Data Quality Metrics: Release 3 (DRB Clearance CBDRB-FY21-295) 2020 Census Data Quality (Accessed August 2021)

United States Census Bureau (2017). 2020 Census Detailed Operational Plan for: 19: Response Processing Operation (RPO)," Version 1.0, May 24, 2017. https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/RPO_detailed_operational_plan.pdf (accessed August 2021).

United States Census Bureau (2018). 2020 Census Operational Plan: Version 4.0. Washington DC: Census Bureau. Available at: http://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-oper-plan4.pdf (accessed August 2021).