

Review of Research Supporting the Use of Administrative Record Enumeration in the 2020 Census

Robert E. Fay

August 12, 2021

Introduction

The Census Bureau has acknowledged that the 2020 Census faced unprecedented challenges (Reichert and Kelly, 2021) and consequently accelerated the schedule for releasing quality metrics of field operations (Ortman and Chapin, 2021). Following public release in the summer and fall of 2020 of self-response rates, total completion rates, and nonresponse follow-up (NRFU) workload completion rates, the Census Bureau released additional indicators on April 26, 2021 (U.S. Census Bureau, 2021), including a link to a downloadable table in a file named “2020-data-quality-metrics-release_1.xlsx.” Wherever appropriate, the table provided comparable operational metrics for the U.S., the 50 states and DC, and Puerto Rico for the 2010 and 2020 Censuses. In many respects, the results for the two censuses appear broadly similar, as Bentley (2021) observed. But, among the few 2020 metrics for which 2010 comparisons are not possible, the use of administrative records (AR) in 2020 arguably merits the most scrutiny. As part of the NRFU operations, administrative records were used to classify some addresses as vacant, not housing units (deletes), or occupied, generally after a single follow-up attempt. Characteristics of the AR-determined occupied units were also provided by administrative records or completed by imputation.

This review is intended to address two questions:

1. To the extent AR enumerations were incorporated into the 2020 Census, did they maintain the accuracy that would have been obtained by using another approach?
2. Is there currently enough information to answer the first question?

This review is restricted to publicly available research reported by Census Bureau staff relevant to these two questions. Although of interest, the review does not consider commentary by other observers on the Census Bureau’s AR research during the previous decade.

Review of AR Research on using ARs

In their timely Census Bureau report, Mary H. Mulry, Tom Mule, Andrew Keller, and Scott Konicki (2021) summarized the foundations for the models and the evidence supporting the use of Administrative Record (AR) enumeration in the 2020 Census. Their report outlines the history of AR use for other statistical purposes up through 2010, the initial testing of AR enumeration, the development of statistical models of AR quality, testing of the proposed approach leading up to 2020, and the adaptations necessitated by the Covid-19 epidemic during a critical window of time for the 2020 Census.

This review follows the outline of the report by Mulry et al. (2021). The review summarizes parts of Mulry et al. (2021) without further elaboration, leaving the original to provide sufficient detail. For other parts, particularly those dealing with the research and tests leading to the 2020 implementation, the review examines both the summary given by Mulry et al. (2021) and other available sources, including several that Mulry et al. (2021) did not explicitly cite but presumably drew from. Like the report, which

cites Stempowski and Christy (2021) and U.S. Census Bureau (2017), the review does not consider the role of AR in the enumeration of group quarters.

After reviewing the supporting research in detail, the concluding section finds that the evidence on whether AR enumeration was of equal quality to be inconclusive. Early tests showed AR enumeration to be less accurate than achieved by enumerators for the same households, but the strategy of restricting its application to those households where its performance was expected to be best provides some protection against a substantial loss of accuracy. Because of multiple enhancements added closer to its application, the possibility remains that AR enumeration improved the average accuracy of the enumeration compared to the likely result without its use. Further comments are included in the final summary.

The goal was to use AR in the 2020 Census to classify some addresses as Occupied, Vacant, or Nonresidential and to create a roster for the AR Occupied units, steps that had not been implemented in any previous census. Mulry et al. (2021) began by reviewing the use of AR before 2020. Initially AR aggregates, including birth records, were used to evaluate census coverage in 1940 and in subsequent censuses through Demographic Analysis. Studies matching AR at the individual level to census enumerations followed. The Census Bureau has made increasing use of AR for other purposes, and Mulry et al. (2021) noted several uses of AR in the 2010 census and other Census Bureau programs.

Within the overarching goals of census accuracy for the U.S. population as a whole and within each state, Mulry et al. (2021, p. 6) remarked

The AR enumeration is designed in a manner that requires it to assure that the designation of addresses as Occupied, Vacant, or Nonresidential has a high probability of being accurate, and in doing so, AR enumeration contributes to the accuracy of coverage of the population. In addition, when a household is enumerated using ARs, the operation is required to assure that there is a high probability that the AR records reflect the number of household members and their characteristics.

In this document, the use of the term “high quality” ARs is used to mean that there is a high probability that the AR status assigned to an address is accurate. The assignment of the Occupied status means that there is a high probability that the address is occupied, and the household size and composition and characteristics are accurate. The assignment of the Vacant status means that there is a high probability that the address has living quarters, but no one resides there, while the Nonresidential status means that there is a high probability that [the] address does not have living quarters.

Their report does not translate “high probability” into a precise numeric value. One goal of this review is to determine whether this notion is clarified in the context of the research.

Mulry et al. (2021, p. 7) describe the Census Bureau’s Personal Validation System (PVS) designed to validate information on name and address in various government and commercial AR sources. When possible, the person is identified by a Protected Identification Key (PIK), effectively an anonymized SSN. Addresses are assigned a MAFID, an identifier in the Census Bureau’s Master Address File (MAF). This system is the underpinning of the Census Bureau’s ability to link multiple AR files and to associate them to people at a specific location.

Two early tests of AR framed the initial research questions. A 2013 test drew a sample of addresses in Philadelphia and attempted to construct the household based on IRS 1040 forms, Medicare records, and the commercial Targus Federal Consumer File. The results were viewed as promising but requiring further refinement of the methods. A study in the 2014 Census Test in parts of Montgomery County, MD, and the District of Columbia used IRS 1040, Medicare, records from the Social Security Numident file, and the Undeliverable-As-Addressed (UAA) information from the United States Postal Service (USPS). The study was again interpreted as supporting feasibility but requiring further refinement.

This review will intersperse a summary of Mulry et al. (2021) with summaries of other documentation of the progression of the research.

Mulry and Keller (2017). Published in 2017 based on an earlier proceedings paper (Mulry and Keller, 2015), the study of Mulry and Keller framed their research problems in the context of decisions made around the time of the analysis of the 2014 Test Census and the design of the 2015 Test Census. Rather than analyze the data from either test, they simulated replacement of 2010 census NRFU enumerations by AR with methods similar to the 2014 test. They chose the title “Comparison of 2010 Census Nonresponse Follow-Up Proxy Responses with Administrative Records Using Census Coverage Measurement Results,” and, after reviewing the past evidence on the limited accuracy of proxy interviews, stated their research question as, “Are proxy responses for NRFU addresses more accurate than the administrative records available for the housing unit?” (Mulry and Keller, 2017, p. 455). Mulry et al. (2021, p. 9) devoted two paragraphs to a description of the findings from the Mulry and Keller (2017) study.

Initially, the Census Bureau planned to conduct NRFU and then use ARs to enumerate addresses for which enumerators did not obtain [a] response. However, Mulry and Keller (2017) were able to assess the quality of the 2010 Census NRFU roster and the AR roster for a housing unit by comparing both to the roster collected by the 2010 Census Coverage Measurement (CCM). The data for the 2010 CCM included 2010 Census data as well as data collected in an independent listing of addresses in the CCM sample blocks and subsequent interviews conducted at all the addresses on the listing. The 2010 CCM used the collected data in processing that determined whether each person on the 2010 Census rosters and the CCM sample rosters were enumerated correctly, incorrectly, missed in the other survey, or had an unresolved status. The CCM results were used to create a “gold standard” roster, justified by its extensive fieldwork, processing, and clerical matching. Linking the “gold standard” roster for an address to its corresponding 2010 NRFU and AR rosters provided a determination of whether each of the rosters had the correct household members.

Using weighted data, the analysis of 2010 NRFU addresses in the CCM sample found that 51 percent of the addresses with proxy respondents and 61.3 percent of the addresses with household member respondents could be found in ARs. For people, 56.6 percent of the proxy NRFU enumerations and 88.0 percent of the household member NRFU enumerations were at the correct residence. For the people on the AR rosters, 49.1 percent of the AR enumerations at addresses enumerated by proxy respondents and 72.5 percent of enumerations by household member respondents were at the correct address (Mulry and Keller 2017). The low percentage of correct enumerations on the AR rosters at the addresses enumerated by proxy respondents led to narrowing the focus of future research. The attention turned to the identification of the NRFU

addresses with high quality ARs that could be used for enumeration when one contact attempt by a NRFU enumerator did not result in an interview.

To emphasize the points in the second paragraph above, AR could provide a usable account for only 51 percent of proxy NRFU households compared to 61.3 percent of NRFU households with a household respondent. Their analysis was then restricted to only those households where the census and AR rosters could both be compared to the CCM gold standard, 5,310 proxy households and 16,876 NRFU households with a household member. For proxy households, the accuracy of the AR account, 49.1 percent, trailed that for census enumerations at 56.6 percent, seen as a reason to turn away from the goal of improving proxy enumeration with AR. But for NRFU households enumerated with a household member, the AR results at 72.5 percent trailed census enumerations at 88.0 percent by an even larger margin (Figure 1). These results are the only ones cited by Mulry et al. (2021) that directly compare the relative accuracy of NRFU and AR.

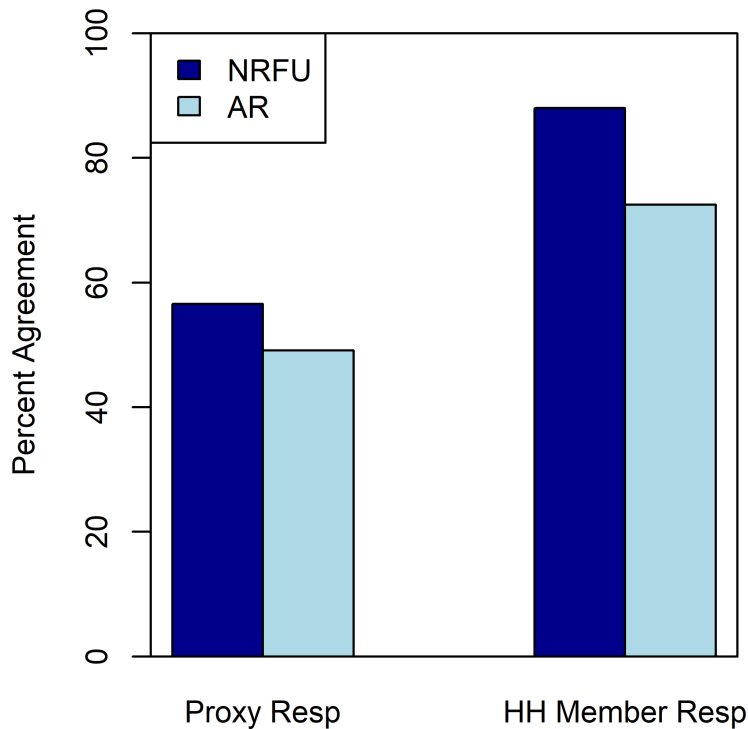


Figure 1. Percent Agreement to combined 2010 CCM for NRFU vs. AR for the 51% of proxy households and 61.3% of NRFU households enumerated with a household member respondent. Source: Mulry and Keller (2017).

To simulate AR performance, Mulry and Keller (2017) reported using data from only two sources: (1) the Internal Revenue Service (IRS) 1040 forms filed in all months of 2010, and (2) the Medicare records for all months of 2010. Although they stated “In addition, the 2014 Census Test operations used only these two sources,” (p. 462), Mulry et al. (2021) reported that the 2014 test also incorporated USPS UAA (undeliverable as addressed) returns in determining the AR status. The UAA operation with follow-up mailings would not have been possible to simulate retrospectively with the 2010 census data. Possibly,

then, AR performance in this study was handicapped by this restriction. (Note that other studies, including Brown, Childs, and O’Hara, 2015, used the National Change of Address (NCOA) files instead.)

Mulry and Keller analyzed the subset of P- and E-sample housing unit addresses in the 2010 NRFU where both the E-sample and P-sample results were available. (The P-sample was a sample of housing units sampled independently of the census to measure census omissions, and the E-sample was a sample from the census to measure erroneous enumerations.) They referred to these cases as the *combined CCM*, noting that this subset excluded E-sample NRFU housing units that could not be linked to any P-sample records. It similarly excluded P-sample housing units that could not be linked to any completed E-sample cases. They did not reweight this data set, but presented weighted data using E-sample base weights for some analyses and unweighted counts otherwise.

The Mulry and Keller (2017) analysis is not straightforward, because when the combined CCM and the AR households are available for comparison, 2.8 percent (weighted estimate) of the combined CCM had insufficient information to be processed, and 20.7 percent were whole person imputations and therefore unmatchable to AR persons. On the AR side, 43.1 percent could not be matched to a PIK at the address, some fraction of which may have corresponded to the combined CCM cases with insufficient information or were whole person imputations. The situation was less ambiguous for NRFU addresses with household respondents, where 2.6 percent had insufficient information and 1.4 percent were whole person imputations, and 21.9 percent of the AR persons did not match a PIK at the same address. In spite of these complexities, census enumeration appeared to outperform AR accuracy according to the 2010 CCM, particularly for household respondents in NRFU, as suggested by the summary presented by Mulry et al. (2021) above.

In hindsight, an additional analysis of the combined CCM could have been examined based on a more interpretable subset, namely, households where each of the P- and E-sample persons had resolved status and had been assigned a PIK. For this subset, the CCM would have defined omissions and erroneous inclusions for both the census household and the associated AR roster.

Brown, Childs, and O’Hara (2015). A few other studies also incorporated the CCM data. Brown, Childs, and O’Hara (2015) showed how both the quality of census enumerations and AR rosters could be assessed by analyzing statistical associations between them. Unlike Mulry and Keller (2017), the analysis did not focus on supporting decisions at the start of NRFU, since key census characteristics were not available when needed to guide NRFU. It is also not clear that all of the AR files considered in this paper would be available in time for use in NRFU. The analyses in the study were complex—arguably more so than the models implemented in 2020. The report also examined a large number of AR sources: Some of them continue to appear in subsequent research, including IRS 1040 and information returns, Medicare, USPS information, and the commercial source Targus. But many of them were dropped subsequently, including state sources from New York, Illinois, and Texas, and a number of commercial sources.

Quality scores for census enumerations were based on 16 potential observable errors (POEs), for example, whether an occupied housing unit was enumerated by proxy and whether a person was duplicated elsewhere in the census. (These are both examples of variables not available at the start of NRFU.) In a number of cases, such as the use of national change of address (NCOA) information, the POEs included information from administrative records, but not from the CCM. The study used the CCM P-sample, but had to exclude part of the sample as out of scope for the analysis, such as P-sample cases not linked to a census address. The POEs were then evaluated based on their ability to predict census-

CCM differences. A logistic regression model for whether the census and CCM housing unit count agreed was fitted including the POEs as predictors, and the predicted values from the model used as quality scores.

The quality of the AR roster was evaluated using two sets of logistic regressions. The first set was based on the set of unduplicated PIKs alive on Census Day across the AR sources and their ability to predict correctly whether the person was in the same place for NRFU households without any POEs, that is, for the most reliable set of NRFU enumerations. Each regression used variables specific to the AR source. The authors described the second set of predictions with the following (p. 9):

A second-stage regression predicts the person-place match propensity for each person-address pair found in at least one of the sources used in the first-stage regressions. The regression incorporates information from the first-stage regressions by including variables indicating whether the person record is in each particular administrative record source at this address or a different one, plus interactions between these dummy variables and the individual match propensities obtained from the first-stage regression corresponding to the variable source for the particular person-place pair. In addition, the regression contains variables regarding the housing structure and decennial census paradata.

The predicted propensities at a person-place level were used to form AR households, scoring each household with the minimum value of the predicted propensities for the AR persons. The authors analyzed the count agreement between the census, AR, and CCM, for the subset of households with high quality administrative records, where *high quality administrative records* were defined in an extended footnote. The footnote stated different cutoffs for the predicted probability that the AR household count will match the census, with a probability of 90 percent or more for AR occupied units and cutoffs depending on UAA (undeliverable as addressed) information from USPS for units likely to be vacant. (Although their paper provided limited detail on this point, their work also modeled the probability that a housing unit was vacant.)

Unlike Mulry and Keller (2017), Brown, Childs, and O'Hara analyzed only whether the count of persons matched the P-sample CCM rather than the accuracy of person-level matching. They presented results (Table 6, p. 13) showing overall agreement between CCM and AR of 93.6%, between Census and AR of 94.9%, and CCM and Census of 94.9%. For the subset of census enumerations without POEs, the corresponding results were 95.6%, 97.7%, and 97.2%. In both comparisons, the census results more closely matched the CCM household counts than did AR, although not by a large margin. They also displayed (Figure 2, p. 15, reproduced below) agreement rates graphically for varying levels of the predicted agreement between AR and the census.

The researchers also divided the results by whether the census enumeration had any POEs. They remarked:

Predicted census enumeration quality and especially the CCM-census agreement rate are much lower when the census enumeration has at least one POE (Figure 3) than it is for those with none (Figure 4). The actual administrative record agreement rates are less strongly associated with predicted administrative record-census agreement when the census enumeration has at least one POE. At the 90 percent predicted administrative record-census agreement level, the CCM-administrative record agreement rate is 96 percent without POEs in the census enumeration, but

it is only 80 percent when there is at least one potential error in the Census. *This again suggests that the census and the CCM tend to have enumeration difficulties in the same housing units.* (Emphasis added.)

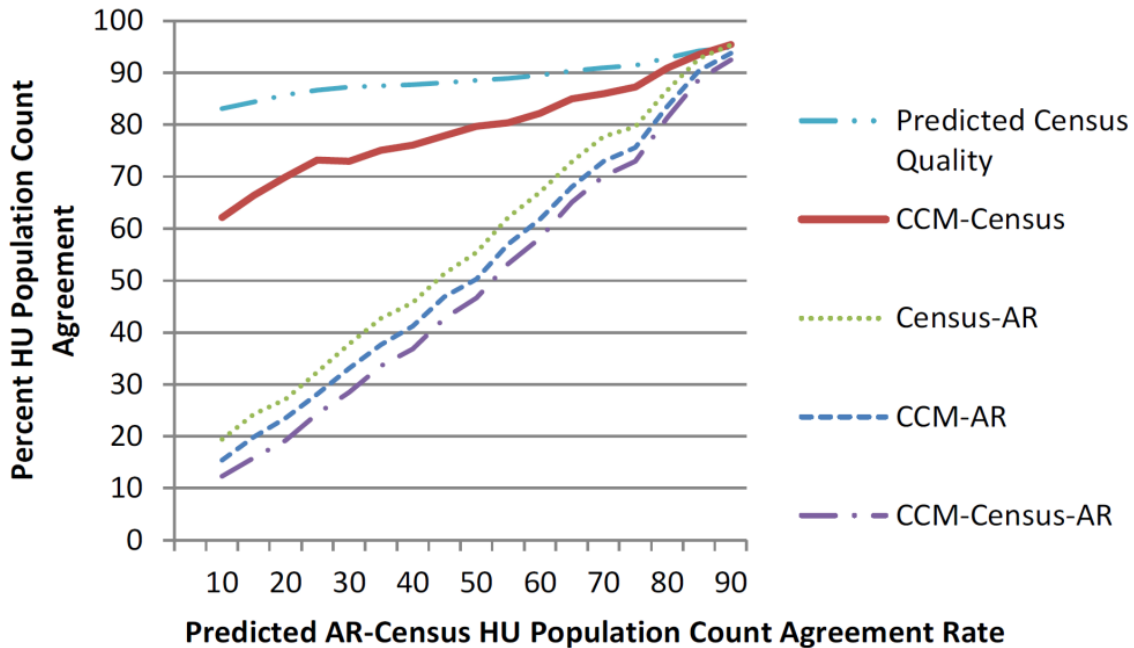


Figure 2. Actual agreement rates for the counts between the 2010 Census and AR track the predicted values closely, while CCM-AR agreement rates fall just below. Both predicted census quality and CCM-Census agreement vary with the predicted AR-Census agreement rate, but stay above the CCM-AR agreement rate. From Figure 2 in Brown, Childs, and O’Hara (2015).

Overall, the results provide mixed encouragement for AR use in 2020. On the one hand, the models failed to identify a subset of AR records exceeding or exactly matching the quality of census enumerations. On the other, at the upper end of their predicted accuracy, the AR results were nonetheless close in accuracy to the census, to a degree more encouraging than the results reported by Mulry and Keller (2017). The use of NCOA (national change of address) data illustrated the value of updates from USPS, which would take the form of targeted mailings to check for UAA returns.

Keller and Konicki (2016). Another study also incorporated the 2010 CCM in the analysis. The study was set in the context of the 2015 Census Test in Maricopa County, which incorporated the strategy of modeling the accuracy of AR records to select only those with high predicted probabilities for use. Keller and Konicki cite Morris, Keller, and Clark (2016) for details of the modeling, a paper discussed below. Keller and Konicki simulated both the models determining the status as occupied or vacant. (They simplified the modeling to combine vacant and deleted units.) Their simulation indicated that about 25% of the NRFU universe would be enumerated by AR. In addition to simulating the modeling on the 2010 Census, Keller and Konicki observed (p. 705) “At the core of this paper is the idea that solely comparing possible AR modeling methods to previous 2010 Census results is insufficient because census results

have errors.” By matching to the CCM E-sample, including its imputed values, they estimated the classification errors for AR. In their Table 3 (p. 706) the 2010 Census counted 987 thousand persons in AR Vacants but CCM estimated that about only 698 thousand were correct enumerations based on CCM. Balancing that finding, the simulation indicated a somewhat higher AR population count in AR Occupied than the 2010 Census obtained. Their findings indicated small shifts in the estimated percent undercount nationally and by age and sex. Their analysis was more limited, however, than Mulry and Keller (2017), because it derived CCM results only in aggregate form rather than as a comparison of the AR rosters and census rosters by CCM status, a comparison possible with the combined CCM file used by Mulry and Keller.

The 2015 Census Test. Mulry et al. (2021) summarized the design of the 2015 Census Test in Maricopa County, AZ, and its findings. A key milestone of the test was the development and deployment of models to differentiate Occupied, Vacant, or Nonresidential Addresses, and models to assess the accuracy of the AR rosters for occupied addresses. The test added IRS Informational Returns, such as W-2 statements and interest and dividend 1099s, and the Indian Health Service Patient Database (IHS) to the two sources used in 2014: IRS 1040s and Medicare. A linear programming method determined which AR rosters were to be considered high quality, but the method was replaced by another in the 2016 Census Test and will not be further reviewed here.

Morris (2014) and Morris, Keller, and Clark (2016) provide further information on the development of the modeling approach to predict high quality AR determinations. Mulry, Mule, and Clark (2016) analyzed the experimental comparison of AR enumeration with standard NRFU operations.

Morris (2014). Morris (2014) reported on preliminary work underlying the methods used in the 2015 Test Census. The paper primarily focused on modeling the person count in occupied units. The author cited an unpublished report of Brown (2013) (not readily available on the web) for the idea of modeling person-place combinations found in the available AR sources, that is where an administrative record placed a person at a specific address. The 2010 census was assessed using person-place pairs from 19 AR sources. For each MAFID, indicator variables were created identifying whether an AR record placed the person at the address. Separate indicator variables signaled whether an AR source placed the person elsewhere. According to Morris, Brown (2013) had proposed a 2-stage model (quite possibly the one in Brown et al. (2015)), but Morris implemented the simpler version based on a single stage (also proposed by Brown, 2013) in a simulated application to the 2010 Census NRFU.

Morris compared logistic regression to random forests and decision trees. All three methods gave relatively similar findings, but the results from random forests and logistic regression were closest to each other. A possible limitation of the finding is that only a 1% sample of the 2010 NRFU households were used as the training set, leaving open the question of whether machine learning methods, including alternatives to random forests, would have significantly outperformed logistic regression if trained on a much larger subset of the 2010 Census, such as 80%.

Households were formed from the AR person-pairs with high predicted probabilities, scoring the household with the minimum of the probabilities of the persons forming the AR household, an approach used subsequently. The modeling of household formation evolved further, however, so the results from this effort are not summarized here.

The paper also described the optimization approach used in the 2015 Census Test to determine which AR results to incorporate. The ROC curve (Receiver Operating Characteristics) measures in this context the tradeoff between false positives (e.g., classifying an occupied unit as vacant) and false negatives (failing to classify a unit as vacant when the AR classification does so correctly). This tradeoff can be thought of as specificity vs. sensitivity or as type 1 error vs. type 2 error. Morris (2014, 2017) pointed out the option of weighting this tradeoff and showed that the optimum threshold varied considerably with changes in the weighting. For example, it may be more serious to classify an occupied unit as AR Vacant than to fail to accept a correct AR Vacant determination. The same consideration applies to the distance metric described by Keller, Mule, Morris, and Konicki (2018). The specific optimization approach in Morris (2014) and applied in the 2015 Census Test was replaced in the 2016 Census Test.

Morris, Keller, and Clark (2016). Using the 2010 Census NRFU, this collaboration simulated a revised modeling strategy. The paper did not identify its relationship to the 2015 Census Test, but it appears to have some of the same innovations as the test and may have incorporated others. The authors used the primary set of AR sources as the 2015 test to form the AR composite household: Internal Revenue Service (IRS) Individual Tax Returns (1040), IRS Informational Returns (1099), the Indian Health Service Patient Database (IHS), and the Medicare Enrollment Database. They also incorporated information from TARGUS to evaluate the quality, but did not use TARGUS in forming the AR roster. They also used USPS UAA codes that the authors described as obtained from the Delivery Sequence File (DSF). (References to UAA codes elsewhere in this review generally resulted from the Census Bureau mailing postcards or other mail to the address.)

A multinomial logistic regression model was fitted to predict whether a household was vacant, occupied, or not a housing unit (delete). Predictors included the here and elsewhere variables previously described by Morris (2014). They fitted two different models to determine the AR roster for occupied housing units: the person-place model described by Morris (2014) and household composition model in the form of a multinomial logistic regression predicting eight statuses for the housing unit:

0. Unoccupied
1. 1 adult w/o children
2. 1 adult w/child(ren)
3. 2 adults w/o children
4. 2 adults w/child(ren)
5. 3 adults w/o children
6. 3 adults w/child(ren)
10. Other

They report on the optimization procedure used in their study, but because the optimization procedure was subsequently replaced, it is not further summarized here.

They simulated their approach on the 2010 Census NRFU. Among their findings was that combining the person-place model and the composition model increased the accuracy of prediction over either model alone. By choosing parameters, they were able to vary the proportion of NRFU workload that would be removed, and presented results for 10%, 15%, and 20%. To illustrate, they targeted a 15% removal rate and determined parameters that actually reduced the NRFU case load for vacants by 10.8% and occupied housing units by 14.6%. Over 90% of the AR vacants were either vacant or deletes, while about 90% of the AR occupied were occupied. Of those, about 65% have a housing unit count match and 67%

have a household composition match. They noted that removing NRFU proxy households would raise the previous two percentages to 70% and 74%, respectively. But it is difficult to translate these results into a comparison of the accuracy of the AR results and the census results for the same households. In their conclusion section, the authors stated:

The caveats of this research dictate an interesting and important future research agenda. The 2010 Census is a natural comparison for evaluating the “quality” of administrative records, but it is necessary to evaluate any approach for using administrative records on other versions of “truth” – for example, American Community Survey (ACS) and Census Coverage Measurement (CCM) data. We present results from one possible scenario for using administrative records, but a complete cost-benefit analysis of the effect of this alternative data collection strategy is warranted.

Mulry, Mule, and Clark (2016). This proceedings paper presented the analysis of the AR test incorporated in the 2015 Census Test. The paper cites the statistical models of Morris, Clark, and Keller (2016) and the earlier work of Brown, Childs, and O’Hara (2015) and of Morris (2014). The 2015 Census Test was split into three panels, two of which tested the operational implementation of AR enumeration, and the third was kept as a control, where standard NRFU procedures were deployed. In the control panel, the results of NRFU could be compared to what AR enumeration would have produced. An Evaluation Follow Up (EFU) was then conducted on a sample of 4,098 NRFU housing units, using specially trained enumerators and questionnaires. The EFU was intended to establish a gold standard for the sampled households by which to evaluate both the NRFU and AR enumeration results. The EFU was effectively the equivalent for the 2015 Census Test of the 2010 CCM for AR simulations using the 2010 Census. The sample was restricted to housing units where there was a discrepancy between the NRFU and AR counts, with the further restriction that all AR records had PIKs.

The analysis of the EFU was divided into nine categories depending on characteristics of the NRFU, AR, and EFU outcomes. Mulry et al. (2016) reported on just three of them in the paper:

- AR Occupied and NRFU household respondent occupied but counts differ
- AR Occupied and NRFU proxy respondent occupied but counts differ
- AR Vacant and NRFU occupied.

The remaining categories were analyzed in an internal document. The EFU included 1,961 households classified as AR Occupied with a NRFU household respondent, where 839 of them had different AR and NRFU counts. Excluding 42 housing units from the comparison because of noninterview, unresolved, or EFU vacant, the EFU count matched the AR counts for 147 housing units, the NRFU counts for 468 housing units, and neither for 182. Although EFU did not investigate the 1,122 housing units where the AR and NRFU counts agreed, assuming (optimistically) that the EFU would also agree in all cases would give agreement rates between the EFU and NRFU of 82.9% and between the EFU and AR of 66.1%. The EFU was another opportunity to test that AR quality would match NRFU quality for household respondents, but the results fell short. (Note, however, that AR enumeration here did not include the single followup visit implemented in later tests and in 2020, or other enhancements.)

Similarly, the EFU included 765 households classified as AR Occupied with a NRFU proxy respondent, where 314 of them had different AR and NRFU counts. Excluding 37 housing units for noninterview, unresolved, or EFU vacant, the EFU count matched 105 NRFU counts, 102 proxy counts, and 70 neither. Assuming that 451 housing units without EFU data would have agreed with the AR and NRFU counts, the agreement rates between the EFU and NRFU would be 76.3% and between the EFU and AR of 75.8%. These hypothetical calculations are again upper bounds, however, because it is likely that the EFU would have disagreed with the NRFU and AR counts for some of the housing units where NRFU and AR agreed.

The 2016 Census Test. Mulry et al. (2021) summarized the findings of the 2016 Census Test:

The 2016 Census Test sites were in Los Angeles County, CA, and Harris County, TX. The search for a new modeling approach began with preliminary studies prior to the 2016 Census Test that examined using multinomial models where the dependent variable had three levels that represented the address statuses Occupied, Vacant, and Nonresidential. The multinomial models produced a probability for each of the three address statuses for each address. Several types of multinomial models were evaluated, including multinomial logistic regression and random forest. The studies found that none of the multinomial models were reliably able to distinguish among the three address categories of Occupied, Vacant, and Nonresidential when compared to field results. That is, assigning the status of the highest predicted probability without consideration of the other outcomes did not provide a high enough level of accuracy. Subsequently, the focus of the research on AR enumeration shifted to examining the application of model-based Euclidean distance programming to aid in identifying addresses with high quality ARs. The approach focused on using a Euclidean distance function and identifying threshold values that the highest of the three estimated probabilities had to exceed for the assignment of an Occupied, Vacant, or Nonresidential status to an address (U.S. Census Bureau Administrative Records Modeling Team 2017).

Comparisons of classifications of addresses based on the AR modeling were compared with field classifications since the test did not include a control panel. Other analyses investigated whether using Undeliverable-As-Addressed (UAA) categories, which U.S. Postal Service mail carriers assign to addresses when their mail cannot be delivered were helpful as independent variables in models for determining the AR Vacant and AR Nonresidential statuses. The results of the investigations of Euclidean distance programming and the UAAs were presented to the Census Bureau Scientific Advisory Committee at their meeting in March of 2017 (U.S. Census Bureau Administrative Records Modeling Team 2017). The committee agreed that the methods tested in the 2016 Census Test showed promise (Census Scientific Advisory Committee 2017).

Note that the first two of the cited paragraphs states that the performance in 2016 was less than adequate, but it suggests that use of “Euclidean distance programming” would improve the outcome.

Administrative Records Modeling Team (2017). This overview was prepared for a 2017 meeting of the Census Scientific Advisory Committee and presented results not found elsewhere in this review. As stated in Mulry et al. (2021), one of its purposes was to summarize the results of the 2016 Census Test. After detailing the AR sources and justifying the use of predictive modeling to identify the most reliable AR housing unit classifications and rosters, the report briefly commented on the similarity of the results from logistic models and random forests. It then described how individual model predictions were

evaluated with distance functions, described in more detail in Mule et al. (2018) below, rather than through optimization. The paper summarized the person-place and household composition models, adding a category for “Someone with undetermined age in household.” The distance function for occupied units creates a score from these two probabilities. The report also discussed the earlier optimization approach implemented with linear programming, but again this alternative will not be reviewed.

The report noted that it was possible to update the model as additional AR information became available. Although the report did not provide this detail, the use of separate AR/here and AR/elsewhere variables in the model might allow the model to function well while the AR sources were in flux. The pair of variables could both be set to zero at all locations when an AR source for the PIK had not been provided, then change to 1’s (typically 1’s for the elsewhere indicator in all places where the person’s PIK appears in the census except the place associated with the AR record).

The report indicated that alternatives to training the models on the 2010 Census were still under consideration. ACS was specifically mentioned, but a possible role for CCM was not identified.

The report reviewed in detail quality measures from the 2015 Census Test, the 2016 Census Test, and simulations using the 2010 census. The 2015 AR procedures were simulated on a control panel. When the model for AR Vacant produced unexpectedly poor results, changes in the handling of AR Vacant were incorporated into the 2016 test. The revised method required at least one field visit and at least one UAA return to reduce misclassification of occupied and nonexistent housing units. The 2015 AR occupied units were occupied in the test census approximately 91% of the time, which was similar to the 2014 test.

The 2016 Census Test was evaluated on a subset held as an evaluation sample. The report noted that many cases in the evaluation sample were unresolved, particularly among AR vacant and delete cases, reducing the reliability of the assessment.

The report closely examined the role of household moves on census accuracy, checking against the USPS National Change of Address (NCOA) file. An analysis of monthly NCOA entries in 2016, provided evidence on the likely timing of moves. In a number of AR Vacant/census occupied cases, the NCOA records suggested that the unit was vacant on Census Day. The report presented further evidence that from 1990 to 2010 decennial censuses may have understated the proportion of vacant units based on evidence from current surveys, suggesting that the test census may have erroneously enumerated some housing units that were vacant on April 1.

Application of the 2016 models to the 2010 census produced a set of NRFU AR vacants, 80% of which were vacant in the census, but 10% were occupied in the census and 10% non-existent. About 90% of the NRFU AR occupied were occupied in the census, but about 8% were vacant and 2% non-existent. In both cases, the simulation was unable to reflect the benefit in 2016 from the single NRFU visit.

An examination of vacancy rates for block groups classified by the percent Hispanic and by percent Non-Hispanic Black led to a concern over possible overestimation of AR Vacant in blocks with a high concentration of the latter population group. This evidence led to the inclusion of an additional mailing for any address with an AR determination. Whether the quality of AR enumeration will vary in other respects for members of previously undercounted groups remains an important question.

Keller, Mule, Morris, and Konicki (2018). Besides emphasizing the adoption of the distance measure to combine the predicted model probabilities, this paper provided an additional summary of the state of modeling research, in line with the report from the Administrative Records Modeling Team (2017). This paper cites both Morris (2014) above, and Morris (2017) for the observation that the logistic regression model was competitive with the random forest in fitting the 2010 NRFU. As noted above, the 2014 comparison is based on only a 1% training sample, and the same is true for the 2017 comparison, which was also based on a 1% training sample. Again, the machine learning model could have benefited from a larger training sample.

The abstract and title of this paper highlight the distance function approach, but a comparison using the 2010 NRFU results in essentially a tie between the new approach and the optimization approach that it replaced. (This finding differs from the apparent suggestion in Mulry et al. (2021) that the use of the Euclidean distance function would result in an improvement.) The paper notes reasons, however, to prefer the new approach in terms of simplicity and applying a uniform standard across areas.

The 2016 Census Test appears to be the last in the series where a large scale test of AR enumeration was conducted and analyzed. The 2017 test was of internet response without including NRFU. Mulry et al. (2021, pp. 11) report on additional refinements incorporated in the 2018 End-to-End Census Test.

- Adding a requirement for a second AR source to corroborate an AR Occupied status.
- Use of a Household Composition Key File (Deaver 2020, p. 4) to associate children with parents in an effort to improve AR coverage of children.
- Possible use for determining household size for households still unresolved at the end of NRFU was under consideration, but Mulry et al. (2021) is unclear whether this was implemented
- As mentioned earlier, adding an additional mailing to validate AR Vacant and Nonresidential. If a UAA was not returned, the address was added back to the NRFU workload.

Mulry et al. (2021) summarized the final NRFU strategy for AR Vacants and AR Deletes with the diagram shown in Figure 3.

Section 4 of Mulry et al. (2021) usefully summarizes the final models, while this review has provided a more detailed account of their evolution. Section 5 of their report presents the flow of the original processing plan, while Sections 6 and 7 discuss the modifications due to the Covid-19 epidemic. Modifications were also summarized by Mule (2020) in a presentation to the Census Scientific Advisory Committee.

Mulry et al. (2021) describe processing for four Louisiana parishes where field operations were truncated because of Hurricane Laura and Hurricane Delta. Where housing units had not yet been enumerated by mid-October, the Census Bureau added to the housing units that qualified for AR enumeration after either one or six NRFU visits by AR enumeration of additional housing units under more lenient standards. Mulry et al. (2021) detail the criteria, which were judged against the alternative of imputation, given the decision to end field activities in mid-October. For example, to be AR Enumerated with characteristics, one of the criteria was that the multinomial model had to indicate a probability of 50 percent or more that the housing unit was occupied. It is beyond the scope of this review to assess how well the approach worked.

As an interesting parallel, to complete the 2016 Census, Statistics Canada turned to a similar extensive use of administrative records because of the Fort McMurray fire affecting that community and the surrounding Wood Buffalo community in Alberta (Statistics Canada, 2017). To represent the usual residents of the community, the agency moved the reference day to May 1, 2016, instead of Census Day as May 10 for the rest of Canada. The blog posted noted that past research on administrative record enumeration had made this approach possible.

Summary. Introduction of AR enumeration into the 2020 Census was a large conceptual change, and Census Bureau staff published detailed accounts of the evolution of the methods to be used. Nonetheless, the available evidence from the published record does not support a definitive answer to the question of whether use of AR enumerations reduced or improved the quality of the 2020 Census. To be clear, the available evidence does not eliminate the possibility that the AR enumerations were of equal or higher quality than would have been expected from application of NRFU enumeration for those households. But the evidence also does not eliminate the possibility that the quality of the AR enumerations instead fell short. The following summary attributes this situation primarily to two factors, gaps in the research program and difficulty assessing the cumulative benefit of later features added to AR enumeration procedures. This summary will conclude, however, with some suggestions on how the 2020 Post Enumeration Survey (PES, the successor to the 2010 CCM) may be able to provide some of the answers currently out of reach.

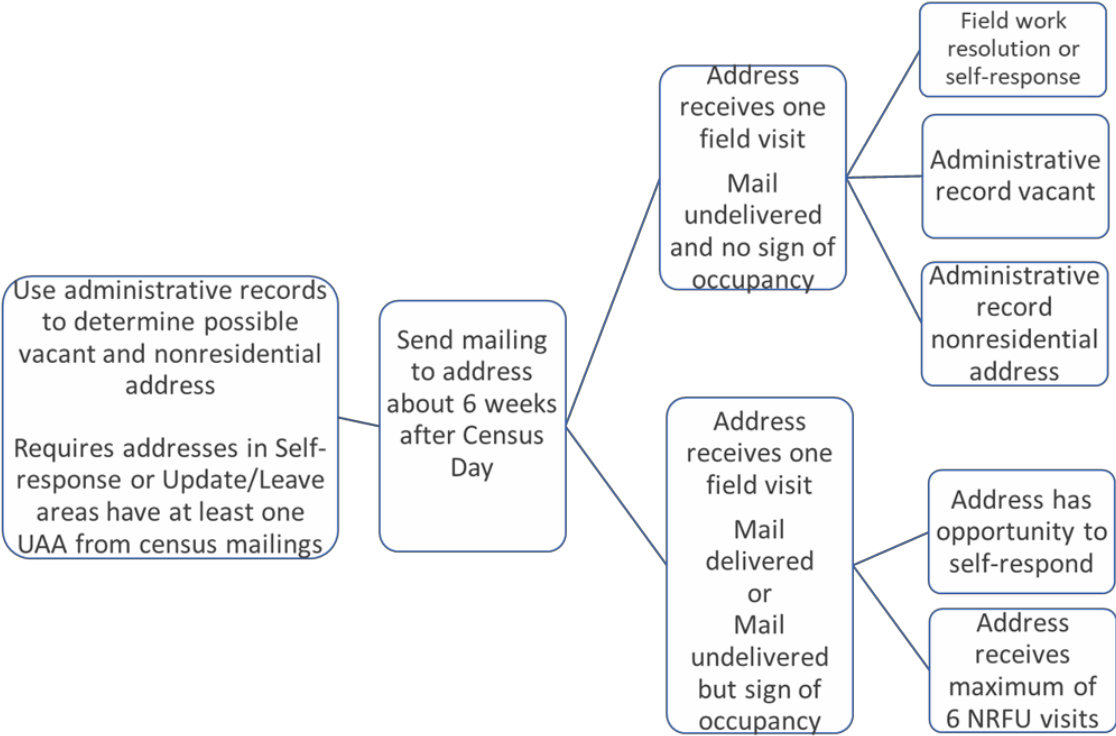


Figure 3. 2020 NRFU contact strategy for Vacant and Nonresidential Addresses. From Figure 1 of Mulry et al. (2021).

The question “Are there methods to identify a group of households where AR enumeration is of higher quality than NRFU enumeration?” sharpens the measurement issue that the Census Bureau’s research program addressed, once the agency decided to shift from improving proxy responses to using AR enumeration where models indicated the AR enumerations would be of high quality. It appears that the only basis for an affirmative answer to that question requires a third source, plausibly able to be used as a “gold standard.” Evidence that AR can produce answers similar to NRFU in most cases is insufficient. One study, Mulry and Keller (2017), stands out in that regard. Although not representing the entire household population, the combined 2010 CCM sample in that study supports the identification of individuals where conflicts between the AR and NRFU rosters can be adjudicated in favor of either. In fact, the study produced disappointing AR results both for NRFU households enumerated by proxy and NRFU households enumerated with a household respondent. The paper provides relatively few details on the construction of AR rosters, but the approach appears minimalist, leaving AR handicapped in the comparison. For example, the paper does not mention use of USPS information, either NCOA or UAA.

Other studies also use the 2010 CCM. Brown, Child, and O’Hara (2015) developed an elaborate AR model for occupied and vacant units and evaluated the results against the CCM on the basis of whether the number of persons agreed. Although agreement in number is a weaker criterion than agreement of individuals as in Mulry and Keller (2017), Brown et al. (2015) also found a correlation between accurate NRFU enumeration and accurate AR enumeration, with AR appearing to be closest to NRFU quality where AR enumerations were predicted to be of highest quality. Keller and Konicki (2016) used a weighted analysis of the CCM to measure net aggregate change. Morris, Keller, and Clark (2016) simulated AR enumeration on 2010 NRFU, but they made a concluding recommendation to use other sources for evaluation, including the CCM. Otherwise, the practice of evaluating AR methods against the CCM appeared to wane towards the latter part of the decade, even as some studies continued simulations using the 2010 NRFU. The 2010 NRFU was the training data set for AR in 2020.

The 2015 EFU was also an attempt to compare AR enumeration to standard NRFU, but the account of Mulry, Mule, and Clark (2016) suggests that the EFU was resource constrained. For example, the study did not sample the entire population of housing units where NRFU could be matched to AR enumeration.

Some studies, especially as reported by the Administrative Record Modeling Team (2017), examined how the timing of a household’s move from or to an address near Census Day could increase the probability of NRFU misclassifying the occupancy status of the address on Census Day. Appropriate administrative records, including NCOA files or UAA recorded in the DSF (as reported by Morris, Keller, and Clark, 2016) could contribute to more accurate AR occupancy status than NRFU for those households. The use of postcard reminders in the 2020 Census sent to any candidate AR address similarly addresses the issue of household moves, although the shift in NRFU schedule distances the UAA response from the status of the housing unit on Census Day.

In hindsight, more could have been learned from the simulations and other studies by applying the general principle of Varying One Thing At a Time (VOTAT) from STEM education. Each implementation of AR enumeration involved complex choices and possible interactions, but more could have been learned by varying each of them individually to discover their individual contributions and possible interactions. For example, the contribution of different approaches to NOCA data could have been better

understood. Possibly, such experiments were conducted in-house, but their results were not reported in the literature reviewed here.

Some arguments support the quality of AR enumerations as implemented in the 2020 census as a replacement for NRFU enumeration:

- In Brown, Childs, and O’Hara (2015) the accuracy of AR enumeration approached that of NRFU at the upper end of the predicted agreement rate of AR with the population count.
- Particularly after the 2016 Test Census, a number of modifications have been introduced to address specific concerns that have been identified and documented. The cumulative effect of these improvements potentially has raised the quality of the AR enumerations implemented in 2020 to their expected NRFU level. Until the 2020 PES becomes available, there was no obvious way to evaluate most of these improvements given time constraints.
- The AR enumerations freed up resources that may have contributed to increasing the completeness of census operations for other NRFU housing units.

A number of concerns can also be mentioned:

- Comparisons of AR enumeration as a substitute for NRFU were not initially favorable.
- Over the span of studies in the decade, the evidence for the quality of AR enumeration relative to NRFU procedures became less quantitative.
- Using 2010 NRFU as training data for the 2020 application may produce suboptimal predictions.
- There were and are opportunities for the Census Bureau to clarify some technical details of the implementation, such as the numerical values of the thresholds.
- Few (perhaps only one) studies examined the impact of AR enumeration on historically undercounted groups.

The preceding considerations are reason to withhold judgment on whether the AR enumerations in the 2020 census were of equal quality to the NRFU interviews they replaced. But the 2020 PES, in a role similar to the 2010 CCM, could clarify the accuracy achieved by AR enumerations, depending on the success of the PES effort. Unlike simulations on the 2010 NRFU and comparisons to the CCM, the 2020 PES will assess the 2020 Census as actually executed. The 2020 PES will also reveal the impact of Covid-19 on the 2020 Census, which data from the 2010 Census cannot reflect. The analysis of the 2020 PES to investigate the quality of AR enumeration will not be straightforward, however, because the question will be the accuracy of the AR enumeration compared to the result standard NRFU would have obtained, a counterfactual.

References

Administrative Records Modeling Team (2017). Administrative Records Modeling Update for the Census Scientific Advisory Committee. Unpublished U.S. Census Bureau document, February 24, 2017, <https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/administrative-record-modeling-in-the-2020-census.pdf>. (accessed May 1, 2021).

Bentley, M. (2021). Examining Operational Quality Metrics. (Census Bureau blog post, April 26, 2021) <https://www.census.gov/newsroom/blogs/random-samplings/2021/04/examining-operational-metrics.html>.

- Brown, J.D., Childs, J.H., and O'Hara, A. (2015). Using the Census to Evaluate Administrative Records and Vice Versa, Proceedings of the 2015 Federal Committee on Statistical Methodology (FCSM) Research Conference, https://nces.ed.gov/fcsm/2015_research.asp. (accessed July 5, 2021).
- Deaver, K.D. (2020). *Intended Administrative Data Use in the 2020 Census*. Washington, DC: U.S. Census Bureau. <https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/administrative-data-use-2020-census.pdf>. (accessed June 21, 2021).
- Keller, A. (2019), "Analyzing Tradeoff Between Administrative Records Enumeration and Count Imputation," Proceedings of the 2019 Joint Statistical Meeting, Government Statistics Section, 2221-2230. <http://www.asarms.org/Proceedings/index.html>.
- Keller, A. and Konicki, S. (2016). Using 2010 Census Coverage Measurement Results to Better Understand Possible Administrative Records Incorporation in the Decennial Census. In JSM Proceedings, Survey Research Methods Section, American Statistical Association, Chicago, IL, July 30–August 4, 2016. Alexandria, VA: American Statistical Association. 701–710. <http://www.asarms.org/Proceedings/index.html>.
- Keller, A., T. Mule, D.S. Morris, and S. Konicki (2018). A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 U.S. Census. *Journal of Official Statistics*, Vol. 34, No. 3, 2018, pp. 599–624. <http://dx.doi.org/10.2478/JOS-2018-0029>.
- Konicki, S. (2012), 2010 Census Coverage Measurement Estimation Report: Adjustment for Correlation Bias, DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-11. Washington, DC: U.S. Census Bureau. <https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/g-series/g11.pdf> (accessed June 20, 2021).
- Morris, D.S. (2014). A Comparison of Methodologies for Classification of Administrative Records Quality for Census Enumeration, Proceedings of the Joint Statistical Meetings, American Statistical Association, pp. 1729-1743. <http://www.asarms.org/Proceedings/index.html>.
- Morris (2017). A Modeling Approach for Administrative Record Enumeration in the Decennial Census, *Public Opinion Quarterly*, 81, 357-384.
- Morris, D.S., Keller, A., and Clark, B. (2016), An Approach for Using Administrative Records to Reduce Contacts in the 2020 Decennial Census, *Statistical Journal of the IAOS* 32, 177-188.
- Mule, T. (2020). Update on Administrative Record Usage. Presentation to the Census Scientific Advisory Meeting, Sept. 17, 2020. U.S. Census Bureau. <https://www2.census.gov/cac/sac/meetings/2020-09/presentation-update-administrative-record-usage.pdf>
- Mulry, M.H. (2007). Summary of Accuracy and Coverage Evaluation for Census 2002. *Journal of Official Statistics*, 23, 345-370.
- Mulry, M.H. and Cantwell, P.J. (2010). Overview of the 2010 Census Coverage Measurement Program and Its Evaluations. *Chance*, 23, 46-51.
- Mulry, M.H. and Keller, A. (2015). Are Proxy Responses Better Than Administrative Records?. Proceedings of the Joint Statistical Meetings, American Statistical Association, Alexandria, VA. pp. 2465-2479. <http://www.asarms.org/Proceedings/index.html>.

Mulry, M.H. and Keller, A. (2017). Comparison of 2010 Census Nonresponse Followup Proxy Responses with Administrative Records Using Census Coverage Measurement Results. *Journal of Official Statistics*. 33(2). 455–475. DOI: <https://doi.org/10.1515/jos-2017-0022>.

Mulry, M.H., Mule, T., and Clark, B. (2016). Using the 2015 Census Test Evaluation Followup to Compare the Nonresponse Followup with Administrative Records. Proceedings of the Joint Statistical Meetings, American Statistical Association, Alexandria, VA. pp. 503-516, <http://www.asasrms.org/Proceedings/index.html>.

Mulry, M.H., Mule, T., Keller, A., and Konicki, S. (2021), Administrative Record Modeling in the 2020 Census, U.S. Census Bureau, April 27, 2021. <https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/planning-docs/administrative-record-modeling.html> (accessed June 21, 2021).

Ortman, J. and Chapin, M. (2021). Introduction to Quality Indicators: Operational Metrics. (Census Bureau blog post, March 18, 2021). https://www.census.gov/newsroom/blogs/random-samplings/2021/03/introduction_to_qual.html.

Reichert, J. and Kelly, D. (2021). Adapting Field Operations to Meet Unprecedented Challenges. (Census Bureau blog post, March 01, 2021). <https://www.census.gov/newsroom/blogs/random-samplings/2021/03/unprecedented-challenges.html>.

Statistics Canada (2017). StatCan and the Albert Wildfire. March 16, 2017 (blog). <https://www.statcan.gc.ca/eng/blog/cs/wildfire>.

Stempowski, D. and J. Christy (2021). *2020 Census Group Quarters*. March 16, 2021. Washington, DC: U.S. Census Bureau. <https://www.census.gov/newsroom/blogs/random-samplings/2021/03/2020-census-group-quarters.html>. (accessed July 24, 2021).

U.S. Census Bureau (2017). 2020 Census Detailed Operational Plan for: 15. Group Quarters Operation (GQ). Washington, DC: U.S. Census Bureau. <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/planning-docs/GQ-detailed-op-plan.html>. (unable to access).

U.S. Census Bureau (2021). Census Bureau Releases Quality Indicators on 2020 Census. April 26, 2021. <https://www.census.gov/newsroom/press-releases/2021/quality-indicators-on-2020-census.html>.