Statistical Journal of the IAOS 32 (2016) 189–198
DOI 10.3233/SJI-161009
IOS Press

Imputation research for the 2020 Census[1]

Andrew Keller
Decennial Statistical Studies Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, USA
Tel.: +1 301 763 9308; E-mail: andrew.d.keller@census.gov

Abstract. For the 2010 Census, the count imputation procedure filled in housing unit status and size for the small proportion of addresses (less than one-half percent) where this information was unknown. The small proportion was due in part to an extensive nonresponse followup (NRFU) field operation geared towards resolving addresses so that a status and count were known.
For 2020, the Census Bureau is researching two changes to the NRFU field operation to reduce cost. The first is the possible use of administrative records (AR) to provide a status and count for some nonresponding addresses. The second is potentially reducing the number of visits made to nonresponding addresses. Although using AR will help resolve some of the remaining unresolved cases, the proportion of addresses in need of count imputation may be higher in 2020 due to the reduction in NRFU fieldwork.
The 2010 count imputation model was developed assuming a small amount of missing data. This research looks at potential count imputation models to handle increased missingness. The paper also articulates the downstream characteristic imputation ramifications from the same missing data challenge.

Keywords: Count imputation, characteristic imputation, administrative records, nonresponse

1. Introduction

To meet the strategic goals and objectives for the 2020 Census, the Census Bureau must make fundamental changes to the design, implementation, and management of the decennial census. These changes must build upon the successes and address the challenges of the previous censuses while also balancing challenges of cost containment, quality, flexibility, innovation, and disciplined and transparent acquisition decisions and processes. Over the course of the decade, the Census Bureau is completing a series of field tests to understand the implications of possible design changes. In this paper, I specifically focus on the 2015 Census Test design. One goal of this test was to implement new methods during the nonresponse followup (NRFU) operation as a means of reducing overall census cost. These included

- modifying contact strategies by reducing the number of contacts and applying adaptive design methods to manage the work in the field
- using administrative records (AR) to remove cases from the NRFU workload by assigning an unoccupied status or by enumerating housing units.

2. Imputation in the 2010 Census

The enumeration portion of the 2010 Census was essentially completed in three stages. To begin, most of the country received a mail form as part of the self-response stage. Then, nonrespondents from the self-response stage were part of the NRFU operation. For 2010 NRFU, a maximum of six contact attempts was permitted with a proxy response permitted only after the maximum attempts to interview a household member had failed. The contact strategy was fixed for all households and AR was not used.

The third stage was imputation. Historically, imputation constitutes a necessary step occurring at the end of each census in order to produce population totals for both persons and housing units. In the 2010 Census,

[1]This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

count imputation was a one-time process performed following the completion of the NRFU operation. This ensured that each census address was provided a final status of occupied, vacant, or non-existent (also known as delete). Following count imputation, characteristic imputation was also a one-time process, completed to ensure that each person was provided characteristics including age, sex, race, Hispanic origin, and relationship to householder.

Fortunately, due to the extensive visit protocol for the 2010 NRFU operation, the 2010 Census had a very small rate of count imputation. Specifically, less than one-half percent of addresses were imputed a status and, if necessary, population count during count imputation. The count imputation model was a general model in the sense that it imputed from a distribution of addresses with similar characteristics in the same tract. See [1] for details. For person characteristics, the reported response rates were: age −90%, sex −97%, race −94%, Hispanic origin −93%, and relationship −96%.

As noted above, the 2015 Census Test represented a shift in how a census was completed. Relevant to this paper, a reduced NRFU operation while incorporating AR provided a unique challenge in terms of increased missing data for addresses as well as persons. If the Census Bureau plans to reduce the number of contact attempts, it is necessary to fully exploit the AR information so that the unresolved rate can be reduced. This research discusses how we could utilize other information from our AR models in order to end up with an imputation rate closer to that of 2010. This paper demonstrates two ways of doing this:

- We incorporate additional AR data as the census progresses. The AR data used to apply models will be augmented over the course of the year. Hence, it is possible to identify more cases for removal by re-processing the models over additional data.
- We revisit the initial AR model results. We applied thresholds to make an initial decision to identify a unit as occupied or vacant. Can we soften those thresholds to complete unresolved cases and remove them from the imputation universe?

To set the framework for this research, we first explain the modeling methodology and design options used for the 2015 NRFU operation. We then apply them back to the 2010 Census data to get a sense of the count and characteristic missingness that can be expected in a decennial census using this approach.

3. Administrative records modeling for NRFU

The 2015 Census Test occurred in parts of Maricopa County, Arizona (including Phoenix). Prior to conducting the NRFU operation, a self-response operation was conducted. To identify NRFU units that were occupied or vacant using AR data, models were fit on the 2010 Census Maricopa County NRFU universe and then applied to the 2015 Census Test sample area. These were binomial or multinomial logistic regression models. In this paper, we describe a national-level application of the same models that we applied during the 2015 Census Test. For the simulation in Section 7, we use the 2015 methodology to fit our AR models on a sample of the 2010 Census NRFU universe. We then apply the fit to the entire 2010 Census NRFU universe. See [2] for more methodological detail on administrative records modeling research for the 2020 Census.

3.1. Data and methodology

To begin, we compiled a household roster composed of AR persons for all housing units in the 2010 NRFU universe for the United States. We ensured that no persons were duplicated within a housing unit. For the 2010 data compiled above, we created separate person and address-level data sets for modeling.

The 2010 vintage AR sources used to create household rosters are:

- Internal Revenue Service (IRS) Individual Tax Returns (1040)
- IRS Informational Returns (1099)
- Indian Health Service Patient Database
- Center for Medicare and Medicaid Services (CMS) Medicare Enrollment Database
- Social Security Numident File

Information from the Targus Federal Consumer file, a third-party file, was used to inform the models but not used when compiling the household roster. In addition, we incorporated 2010 vintage data from the United States Postal Service (USPS) Delivery Sequence File, the American Community Survey (ACS), the Master Address File, and 2010 Census operational information. We also used the USPS Undeliverable-As-Addressed (UAA) reasons obtained from the second mailing that was delivered around April 1, 2010. Sections 3.2 and 3.3 provide more information about the vacant and occupied models.

3.2. Identifying vacant units

To identify vacant units, we developed a multinomial logit model, which estimated the unit status probability as of Census Day. The dependent variable had three possible values for each NRFU address record: occupied, vacant, or delete. We then used a linear program to maximize the NRFU workload identification subject to constraints on the predicted probabilities resulting from the vacancy model. These constraints were determined based on analysis of 2010 Census NRFU data.

3.2.1. Vacant model optimization

The motivation of this research is to maximize the reduction of NRFU workload through the identification of vacant units prior to NRFU operations. To accomplish this, we simultaneously applied two constraints when maximizing our NRFU workload reduction. The first constraint was that the average vacant probability of identified units be above some threshold. The second constraint was that the sum of the occupied probabilities did not exceed a certain percentage of the estimate of occupied units from the ACS.

The first constraint attempted to reduce the amount of misclassification of occupied or delete units as vacant units. Those units for which the model was most confident of vacancy status had a high vacancy probability. We specified an average vacant probability of no less than 0.8.

The second constraint tried to reduce the amount of misclassification of occupied units as vacant. It imposed a restriction on the occupied probability. As discussed earlier, each NRFU address had a three-vector probability space. The second constraint helped to distinguish between households with comparable vacant probabilities yet different probabilities of being occupied by allowing only a desired tolerance of the amount of occupied units possibly being misclassified. We identified a threshold of 0.5 percent. Therefore, the sum of the occupied probabilities was no greater than one-half percent of the number occupied units from the ACS over the relevant geography.

3.3. Identifying occupied units

Two models were developed to identify occupied units, a person-place model and a household (HH) composition model. The person-place model predicted the probability that an AR person would be enumerated at the sample address if fieldwork was conducted. 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.

To integrate information from the two models, we used linear programming to identify occupied units. The predicted probabilities from the two models were passed to the linear program to identify cases determined to be occupied. For the linear program, we maximized the occupied identification subject to constraints on the predicted probabilities resulting from both models. In addition, we added constraints that the size of the AR unit must not be greater than six people and the AR HH composition had between one and three adults (with or without children).

3.3.1. Person-place model

We compiled person-place pairs in AR files mentioned above and the 2010 Census person-place pairs to define the dependent variable of interest in the person-place model:

$$y_{ih} = \begin{cases} 1 \text{ if person } i \text{ is found in AR and 2010} \\ \quad \text{Census at the same address} \\ 0 \text{ otherwise} \end{cases}$$

We are interested in a predictive model for estimating the probability $p_{ih} = P(y_{ih} = 1)$, that the 2010 Census and the AR roster data place the person at the same address. These probabilities were estimated via a logistic regression model. The research in [3] shows that logistic regression and machine learning techniques (classification trees and random forests) exhibit similar predictive power for this person-place model. Logistic regression was used for the 2015 Census Test.

The person-place model was fit at the person-level, but decisions were made at the housing unit-level. Therefore, the person-level predicted probabilities, $\hat{p}_{ih}$, were summarized for each address such that the housing unit-level predicted probability for address h was defined as:

$$\hat{p}_{ih} = \min(\hat{p}_{1h}, \ldots, \hat{p}_{n_h h})$$

where n_h was the number of people at address h. This minimum criterion assigned to the housing unit the predicted probability for the person in the housing unit for which we had the lowest confidence – a relatively conservative approach. The AR HH count was defined as the sum of all individuals associated with the AR address, and each address had the associated predicted probability of having an AR/Census address match. These were the predicted probabilities that were passed to the linear programming portion to decide which cases were determined to be occupied.

3.3.2. Household composition model

The results from the 2014 Census Test motivated the development of the HH composition model. During that test, we observed that units that we identified as occupied with AR were more likely to be occupied in NRFU if the HH composition of the AR unit was a single adult, a two-person adult unit without children, or a two-person adult unit with children.

We began by categorizing each AR HH roster in this manner:

No AR persons
1 Adult, 0 Child
1 Adult, $>$ 0 Child
2 Adult, 0 Child
2 Adult, $>$ 0 Child
3 Adult, 0 Child
3 Adult, $>$ 0 Child
Someone with undetermined age in HH
Other

We then created a dependent variable from the HH composition on the 2010 Census. The categorization was similar except that, in all units, all persons have an age. The reason was that we were using the Census Edited File as the basis for forming the census HH composition. Since this data had age imputed, there were no missing values for age. We fit a multinomial logistic model with the 2010 Census HH composition as the dependent variable over a sample of the data. The predicted probability for the housing unit was the multinomial probability associated with the AR HH type. These predicted probabilities were then passed to the linear programming portion to determine which cases were identified as occupied.

3.3.3. Occupied model optimization

The motivation underlying using linear programming to identify occupied units was that we could integrate information from multiple occupied models. To achieve this, we simultaneously applied two constraints when maximizing our occupied identification. For example; suppose the HH composition model indicated that a unit with two adult and children had a high-predicted probability of also having been a two adult with children HH in the census. Furthermore, suppose that the person-place model showed that one of the children had a low predicted probability of being in the census. Using only the information from the HH composition model would probably have led to removing the unit from the NRFU workload. However, also including the information from the person-place model may have caused the unit to remain in the workload since the status of one person is in question. In short, incorporating information via a linear program allowed for a type of consistency check across both models.

The first constraint was that the average probability of identified units for the person-place model is above some threshold. We identified a threshold of 0.68 for this research. The second constraint was that the average probability of identified units for the HH composition model is above some threshold. We identified a threshold of 0.57 for this research. We will continue to research how to identify initial thresholds as we continue with the AR research.

4. NRFU design options

The NRFU operation in the 2015 Census Test had three panels (control, hybrid, full) that employed different contact strategies and different ways of using AR data, including a control panel that used no AR.

4.1. Control panel

The control panel mimicked the 2010 Census NRFU contact strategy as closely as possible. A maximum of six (6) contact attempts was permitted with a proxy response permitted only after the maximum attempts to interview a household member had failed. The contact strategy was fixed for all households and the panel did not use AR in any way.

4.2. Hybrid panel

In the hybrid panel, housing units identified as vacant via the AR modeling in Section 3 did not receive any NRFU visits. For the remaining housing units not identified as vacant and for which we had AR indicating an occupied status, enumerators made only one personal visit attempt. No proxies were allowed for these cases. Cases unresolved after one personal visit were enumerated using AR data. Units without any determination were allowed a pre-specified number of visits according to the adaptive design procedure described in Section 5.

4.3. Full panel

In the full panel, housing units identified as vacant or occupied did not receive any NRFU visits. Units without any determination were allowed a pre-specified number of visits according to the adaptive design procedure.

5. Adaptive design

The 2010 Census employed a fixed contact strategy for NRFU. Regardless of location, each housing unit was allowed six contacts. The Census Bureau has been researching an adaptive design procedure that allows the maximum number of contact attempts to vary across areas. The goal of this approach is to contain costs while equalizing a measure of area-level data quality.

The 2010 Census showed that generally proxy respondents provided less complete information than household members did. Hence, for our current research, the data quality measure we have been working with is the proxy rate. Currently, the Census Bureau is researching an approach that identifies a maximum number of visits across block groups while minimizing the variance in proxy reporting across those areas. See [4] for more details.

6. Applying AR modeling and design options

Once the NRFU universe is determined, we immediately apply the results from the AR modeling to identify vacant and occupied units. Supposing we use the hybrid design, we first remove the units we identified as vacant. Next, we allow at least one contact for all remaining units. For the units we identified as occupied via AR, we take the interview result from the single personal visit. If the personal visit is unable to resolve the unit, we call it occupied and enumerate the housing unit via AR. For the remaining housing units for which we were unable to identify a vacant or occupied status via AR modeling, we allow the number of NRFU visits as specified by the adaptive design procedure in Section 5. Note that different permutations can be identified with respect to AR modeling, NRFU designs, and adaptive design options that would affect the magnitude of count and characteristic imputation. A single simulation is demonstrated in Section 7 to provide the reader with a qualitative understanding of how AR can be used to decrease the amount of imputation.

7. Simulation

The following simulation shows an example how we could incorporate additional AR data as well as revisit the initial AR model results to reduce the unresolved rate. This simulation assumes that we use AR models described in Section 3, the hybrid design described in Section 4, and the adaptive design procedure from Section 5. The results shown are on the national NRFU data from the 2010 Census. We show a step-wise progression to reduce the unresolved rate. We consider the ramifications on misclassification error and characteristic missing data.

7.1. Running the initial AR models and applying hybrid and adaptive design

To begin, we run the AR models for our example. The NRFU universe is 49,817,252 cases. Table 1 shows the distribution of cases identified as vacant and occupied by AR models and those for which no determination was made.

Table 1 shows that about 14.6% of the NRFU cases are identified as AR Occupied and 10.3% are identified as AR Vacant. This is a similar percentage as seen in [2]. Next, we apply the hybrid and adaptive designs. Under the hybrid design, the 5,132,613 units identified as vacant are assigned a vacant status. The 7,292,195 units identified as occupied are allowed one personal visit. Cases resolved by the one visit are given the status determined through the NRFU interview. Cases unresolved after one personal visit are assigned as occupied and enumerated using AR data.

Finally, the No Determination cases are allowed the maximum number of visits as determined by the adaptive design procedure. For this simulation, if the number of visits from the 2010 NRFU operation is less than or equal to the maximum number visits allowed by the adaptive design procedure, we assign the status and population count determined by the 2010 NRFU visit. Conversely, if the number of visits from the 2010 NRFU operation exceeds the maximum number visits allowed by the adaptive design procedure, we assign an unresolved status. Table 2 shows the distribution of assigned NRFU status for each AR Model category. There are four categories: occupied (Occ), vacant (Vac), delete (Dele), or unresolved (Unres).

Recall that, under the hybrid design, all cases identified as AR Vacant are assigned a NRFU vacant status. However, even though 7,292,195 units are identified as AR Occupied, about 2.5% are assigned a vacant status and 0.8% are assigned a delete status. This is because we allowed one personal visit which obtained the non-occupied NRFU interview result.

Table 1
NRFU Universe by AR model category

AR model category	Total	No determination	AR occupied	AR vacant
N	49,817,252	37,392,444	7,292,195	5,132,613
Percent	100.0%	75.1%	14.6%	10.3%

Table 2
AR model category NRFU status assigned via simulation

AR model category	Total	NRFU status assigned via simulation				%			
		Occ	Vac	Dele	Unres	Occ	Vac	Dele	Unres
No Determination	37,392,444	18,401,622	9,060,944	4,047,126	**5,882,752**	49.2%	24.2%	10.8%	**15.7%**
AR Occupied	7,292,195	7,047,201	183,370	61,624	0	96.6%	2.5%	0.8%	0.0%
AR Vacant	5,132,613	0	5,132,613	0	0	0.0%	100.0%	0.0%	0.0%
Total	49,817,252	25,448,823	14,376,927	4,108,750	5,882,752	51.1%	28.9%	8.2%	11.8%

7.1.1. Assignment ramifications

After applying the hybrid and adaptive designs, we investigate the misclassification error and characteristic missing data ramifications. To begin, we look at the AR Occupied and AR Vacant cases. For the AR Occupied cases, the hybrid approach allows an initial NRFU interview. If the unit is not resolved in the first contact, the unit is assigned an occupied status. Among the 7,292,195 AR Occupied cases, 2,622,845 were resolved on the first contact. The other 4,669,350 were resolved after the first contact.

Under the hybrid approach, all AR Vacant cases are assigned a vacant status before any NRFU visits occur. We compare the vacant assignment versus the 2010 NRFU status to get an understanding of the error. Table 3 compares the assignment status for the AR Occupied – First Contact, AR Occupied – More Contacts, and AR Vacant cases against the 2010 NRFU status to get an understanding of the misclassification error.

Table 3 shows that, of the 2,622,845 AR occupied cases we resolved on the first contact, about 7.0% we would have assigned as occupied when field visits later determined they were vacant. For the 4,669,350 AR occupied cases we did not resolve in the first contact, we would have assigned these as occupied under the hybrid approach. In these cases, about 8.5% were vacant and 1.2% were delete. Last, of the 5,132,613 AR vacant cases, about 9.1% were occupied and 11.9% were delete.

It is important to identify the ramifications on characteristics for cases we assign as occupied via AR. Table 3 shows that we would have enumerated 4,669,350 cases as occupied via AR. In these units, we would have assigned 10,236,982 persons. However, since no interviews were completed, characteristics would have to be taken from AR or imputed. Czajka [5] discusses directly substituting AR for survey data. We investigate the impact on missing data for race and Hispanic origin by using AR data before imputation. To identify race and Hispanic origin for persons enumerated in AR Occupied units, we use AR data from multiple sources. Ennis et al. [6] explain how race and Hispanic origin are assigned to persons in AR data.

With respect to other characteristics, note that, in order to identify a unit as occupied via AR models, it must have all ages filled. In addition, sex is usually a non-missing characteristic because of its presence on the Numident file. Relationship to householder is not considered in this table, but is a subject of ongoing research. Table 4 shows the missing data rate for race and Hispanic origin separated and combined for these persons.

7.2. Incorporating additional AR data – Phase 2 (During NRFU)

Table 2 shows that, by applying the hybrid and adaptive designs, 5,882,752 addresses are unresolved. This unresolved rate is about 12% of the overall NRFU universe. In comparison, the actual unresolved rate in 2010 was about 1% of the total NRFU universe. This high unresolved rate occurs because we are not allowing for the six contact attempts allowed in 2010. In practice, this simulated unresolved rate may be lower due to changes in field procedures or other operations that could be undertaken during a decennial census. As a result, it may be that the 5,882,752 overstates the unresolved universe.

The NRFU operation begins during the middle of May. Hence, the initial AR models are run during that time to identify occupied and vacant units. However, the Census Bureau receives additional AR data throughout the NRFU operation. For example, during the 2015 Census Test, after an initial AR model was run, the Census Bureau received additional IRS

Table 3
NRFU status assigned via simulation versus 2010 NRFU status

AR model category	Total	2010 NRFU status				%			
		Occ	Vac	Dele	Unres	Occ	Vac	Dele	Unres
AR Occupied – First Contact	2,622,845	2,377,851	183,370	61,624	0	90.7%	7.0%	2.3%	0.0%
AR Occupied – More Contacts	4,669,350	4,199,592	395,181	54,929	19,648	89.9%	8.5%	1.2%	0.4%
AR Vacant	5,132,613	466,977	4,009,243	610,490	45,903	9.1%	78.1%	11.9%	0.9%

Table 4
Missing characteristic data from AR occupied assignment

AR model category	Total housing units assigned	Total persons assigned	% Missing Hispanic origin	% Missing race	% Missing combined race & Hispanic origin
AR Occupied – More Contacts	4,669,350	10,236,982	12.3%	13.4%	10.7%

1040 and IRS 1099 information. In 2015, the AR models were then rerun incorporating the new data and additional units were identified as AR occupied. We assigned these units an occupied status. We call this Phase 2 because it entails incorporating new AR data during the NRFU operation.

We apply the Phase 2 procedure by incorporating additional AR data. We then rerun our models to determine if any previously unresolved units can be assigned an AR occupied status. Table 5 shows that 130,902 of the previously unresolved units could be assigned an AR Occupied – During NRFU status. This is about a 2.3% reduction of unresolved cases. We then compare against the 2010 NRFU status to get an understanding of the misclassification error. Table 6 shows the missing data rate for race and Hispanic origin separated and combined for these persons identified in Phase 2.

Table 5 shows that we identify an additional reduction of 130,902 units from the original 5,882,752 unresolved units incorporating the more recent data. These cases were about 93% occupied in the 2010 census. In comparison, the AR Occupied – More Contacts cases were only 90% occupied in 2010. However, the characteristic missing data rates are a little higher when comparing Table 6 to Table 4. For example, the persons in the AR Occupied – More Contacts units have a 12.3% missing data rate for Hispanic origin while persons in the AR Occupied – During NRFU units have a 17.3% missing data rate for Hispanic origin. After applying Phase 2, 5,751,850 units remain unresolved.

7.3. Revisiting initial AR model result – Phase 3 (Close-out NRFU)

Before NRFU began, we applied initial thresholds to the AR models to make a decision to identify a unit as occupied or vacant. Now, with our remaining unresolved cases at the end of NRFU, we revisit those initial thresholds to identify remaining unresolved units as occupied or vacant. In particular, we seek to identify more unresolved units to call occupied or vacant by softening our initial thresholds for removal. We call this Phase 3.

At one extreme, we could disregard all the remaining units with AR and immediately treat all the unresolved cases with a more general imputation procedure as used during 2010 imputation. Table 7 compares the distribution of the 5,751,850 remaining unresolved cases against a similar count imputation procedure to what was used during the 2010 Census.

Table 7 shows the 2010 NRFU housing units had a higher occupied distribution of about 5.4% as compared to using the count imputation procedure. Naturally, there is a motivation to look at using a different count imputation model due to the higher amount of unresolved cases expected during 2020. However, there are other practical considerations as to why we want to revisit the initial thresholds to identify a unit as occupied or vacant as opposed to using a general imputation procedure immediately. First, by using more AR data, we can use the person-level information within that AR data. This enables us to avoid having to impute every characteristic for cases that we count impute as occupied. Second, by using the remaining unit-level AR data instead of count imputing, we preserve unit-level household compositions seen in the AR data. Third, this is an adaptive extension of the AR modeling as opposed to a count imputation.

Section 3 discusses how we identified the initial AR occupied and AR vacant cases. In doing so, we used the 2010 NRFU data to identify starting cut-off threshold probabilities for the three models by looking at the ramifications on quality versus workload removal. As a result of that research, we specified an average vacant probability of no less than 0.8. For the person-place

Table 5
2010 NRFU status of AR occupied – During NRFU cases

AR model category	Total	2010 NRFU status				%			
		Occ	Vac	Dele	Unres	Occ	Vac	Dele	Unres
AR Occupied – During NRFU	130,902	121,272	7,422	700	1,508	92.6%	5.7%	0.5%	1.2%

Table 6
Missing characteristic data from AR occupied – Phase 2 assignment

AR model category	Total housing units assigned	Total persons assigned	% Missing Hispanic origin	% Missing race	% Missing combined race and Hispanic origin
AR Occupied – During NRFU	130,902	325,062	17.3%	17.8%	15.6%

Table 7
2010 NRFU versus count imputation procedure

	Total	Occ	Vac	Dele
2010 NRFU status		73.0%	23.1%	3.9%
Count imputation procedure	5,751,850	67.6%	27.0%	5.5%

Table 8
New threshold probabilities for Phase 3 scenarios

Scenario	Vacant model	Person-place model	HH composition model
A	0.77	0.65	0.54
B	0.74	0.62	0.51
C	0.71	0.59	0.48
D	0.68	0.56	0.45

model, we identified an average 0.68 threshold. For the HH composition model, we identified an average 0.57 threshold.

In this section, we look at four scenarios where we lower the thresholds. Each scenario has three associated probabilities, one for each of the three models. We look at the ramifications on workload removal, true positive rates, and the downstream effects on status distribution when the same count imputation model used in Table 7 is then reapplied over the smaller universe. Table 8 shows scenarios.

For example, in Scenario A we specify an average vacant probability of no less than 0.77 (as opposed to 0.8). For the person-place model, we identify an average 0.65 threshold (as opposed to 0.68). For the HH composition model, we identify an average 0.54 threshold (as opposed to 0.57). These methods result in another 100,538 units being identified as AR vacant and another 366,005 units being identified as AR occupied. To look at the true positive rate, Table 9 shows the status distribution of the AR vacant and AR occupied cases for the 2010 NRFU operation. This is shown for each of the four scenarios.

Table 9 shows that for Scenario A, among the AR Vacant units, there is a true positive rate of 56.0%. This is lower than the 78.1% true positive rate seen among the Phase 1 AR Vacant cases in Table 3. This large difference is partially due to the high unresolved rate. Table 9 shows that for Scenario A, among the AR Occupied units, there is a true positive rate of 87.3%. This is lower than the 90.2% true positive rate seen among the Phase 1 AR Occupied cases in Table 3.

The 2013 Census Test [7] and 2014 Census Test [8] showed that cases with a Vacant UAA reason were better indicators of a vacant status in NRFU than cases with Attempted Not Known UAA reason, Unable to Forward UAA reason, or any of the remaining UAA reasons. Figure 1 compares the distribution of UAA reasons across Phase 1 and the four scenarios in Phase 3.

Figure 1 shows that Phase 1 and Scenario A of Phase 3 have the highest percentage of AR Vacant cases from a Vacant UAA reason. In addition, as the Phase 3 scenarios identify more AR Vacant cases, they have a smaller proportion of cases with a Vacant UAA reason. This explains why the true positive rate decreases as we identify more AR Vacant cases.

With respect to occupied units, the 2014 Census Test [8] showed that units with an AR HH composition of single adult with no children, two adults without children, or two adults with children were better indicators of an occupied status than other HH compositions. They also had higher rates of agreement between population counts when comparing the AR HH count

Table 9
True positive rates for Phase 3 scenarios

Scenario	AR vacant	Occ	Vac	Dele	Unres	AR occupied	Occ	Vac	Dele	Unres
A	100,538	22.5%	56.0%	8.8%	12.7%	366,005	87.3%	10.5%	0.7%	1.4%
B	151,838	20.6%	51.8%	11.8%	15.9%	742,385	86.5%	11.3%	0.8%	1.4%
C	171,068	19.6%	48.4%	13.9%	18.1%	1,118,007	85.7%	12.0%	0.8%	1.5%
D	176,727	19.2%	45.5%	15.8%	19.5%	1,470,583	84.8%	12.8%	0.8%	1.5%

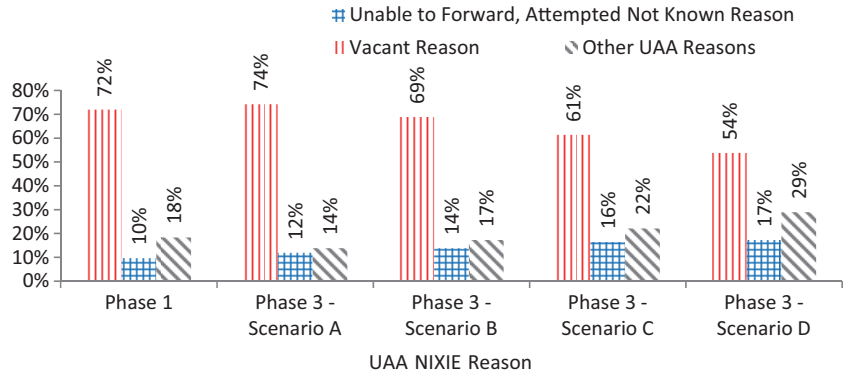

Fig. 1. Distribution of UAA NIXIE reasons for AR vacant units by phase.

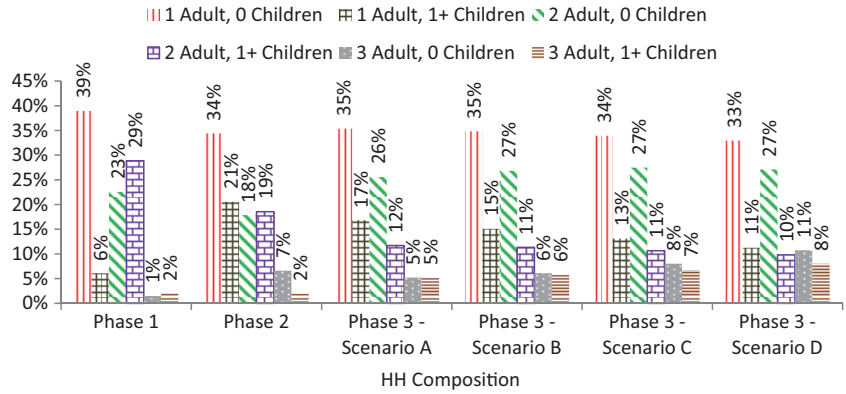

Fig. 2. Distribution of HH composition for AR occupied units by phase.

versus the NRFU HH count. Figure 2 compares the distribution of HH compositions across Phase 1, Phase 2, and the four scenarios in Phase 3.

Figure 2 shows that, across all scenarios, Phase 3 identifies more three-adult units and fewer two-adult units. This is to be expected since Phase 1 is intended to identify the highest quality AR Occupied units.

Table 9 shows that we resolve an additional 466,543 cases by lowering the thresholds in Scenario A. This is 8.1% of the 5,751,850 unresolved cases resulting after Phase 2. After that, 5,254,073 unresolved cases remain. To see the cumulative effect on the additional AR occupied and vacant cases and the count imputation universe, Table 10 shows the resulting status distribution on all the 5,751,850 unresolved cases. To do this, we apply the same count imputation model used in Table 7.

Table 10 shows results by first applying the different scenarios under Phase 3 and then completing count imputation. Recall that, in Table 7, about 73% of the 5,751,850 unresolved cases were occupied in 2010 NRFU distribution. Hence, all of the scenarios appear to be more closely in line with the 2010 NRFU distribution over the same cases as compared to just applying count imputation after Phase 2. In general, identifying more AR resolved cases via instituting lower thresholds from Phase 1 seems to be a more fruitful approach as compared to count imputing directly after Phase 2. In addition, by not applying count imputation immediately after Phase 2, the additional AR occupied

Table 10
Status distribution for Phase 3 scenarios

Scenario	Additional AR resolved	% Reduction of NRFU unresolved universe	Resulting status distribution of 5,751,850		
			% Occupied	% Vacant	% Delete
A	466,543	8.1%	65.5%	25.1%	9.4%
B	894,223	15.5%	67.7%	23.8%	8.5%
C	1,289,075	22.4%	70.2%	22.1%	7.7%
D	1,647,310	28.6%	72.5%	20.4%	7.0%

units will not have all characteristics imputed and the overall count imputation rate will decrease.

8. Conclusions and future work

For the 2010 Census, imputation models had to account for a small amount of missing data. For count imputation, this missingness rate was less than one-half percent. For characteristics, at least 90% of each characteristic (age, race/Hispanic origin, sex, tenure, relationship) was reported. For 2020, changes to census operations including a reduced NRFU operation will necessitate the use AR data beyond initial AR modeling to assign vacant and occupied cases in NRFU. This paper documents the beginning of the imputation research, namely how we plan to use information from AR models and incorporate additional AR data to cut down on imputation.

References

[1] M. Pritts, Census 2010: Overview of Count Imputation, DSSD 2010 Decennial Census Memorandum Series J-08, 2010.

[2] D.S. Morris, A. Keller and B. Clark, An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census, in JSM Proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association, 2015, 3278–3292.

[3] D.S. Morris, A Comparison of Methodologies for Classification of Administrative Records Quality for Census Enumeration, in JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association, 2014, 1729–1743.

[4] S. Konicki and T. Adams, Adaptive Design Research for the 2020 Census, in JSM Proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association, 2015, 1703–1714.

[5] J. Czajka, Can Administrative Records Be Used to Reduce Nonresponse Bias? *The ANNALS of the American Academy of Political and Social Science January* **645** (2013), 171–184.

[6] S.R. Ennis, S.R. Porter, J.M. Noon and E. Zapata, When Race and Hispanic Origin Reporting are Discrepant Across Administrative Records and Third Party Sources: Exploring Methods to Assign Responses. Center for Administrative Records Research and Applications Working Paper #2015-08. Washington, DC: U.S. Census Bureau, 2015.

[7] G. Walejko, A. Keller, G. Dusch and P.V. Miller, 2020 Research and Testing: 2013 Census Test Assessment, U.S. Census Bureau, 2014.

[8] A. Keller, T. Fox and V.T. Mule, Analysis of Administrative Record Usage for Nonresponse Followup in the 2014 Census Test. U.S. Census Bureau, 2015.