DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 599–624, http://dx.doi.org/10.2478/JOS-2018-0029

A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 U.S. Census

Andrew Keller[1], *Vincent T. Mule*[1], *Darcy Steeg Morris*[1], *and Scott Konicki*[1]

The U.S. Census Bureau is conducting research on using administrative records to reduce the cost while maintaining the quality of the 2020 Census Nonresponse Followup (NRFU). Previous census tests have implemented approaches that use predictive models and optimization procedures to identify vacant and occupied housing units using administrative records. This article details a modification to previous approaches, introducing a simple distance metric to define a quality ranking of housing units to enumerate using administrative records. The distance approach is illustrated, assessed, and compared to a previous approach via a retrospective study of the 2010 U.S. Census.

Key words: 2020 Census; administrative records; nonresponse followup.

1. Introduction

Sample surveys and censuses are historically the primary source for producing official statistics. In order to deal with increasing operational costs and decreasing response rates, national statistical organizations are researching how and when to use administrative records in the census and survey life cycle (Bakker et al. 2015; Fienberg 2015; Wallgren and Wallgren 2007; Brackstone 1987; Federal Committee of Statistical Methodology 1980). Administrative records are data "generated for a different purpose" that "arise organically through administrative processes" (Japec et al. 2015), whether collected through administering a program of a federal government agency or a service of a commercial business. The U.S. Office of Management and Budget has defined administrative records as data held by agencies and offices of the government that has been collected for other than statistical purposes to carry out basic administration of a program (U.S. Office of Management and Budget 2014). This article also considers nonpublic, commercial data similar to administrative records, which is consistent with the wider definition proposed by the United Nations Economic Commission for Europe (UNECE 2011). With respect to surveys, Groves and Harris-Kojetin (2017) outline potential beneficial ways to use administrative records in various stages of the survey life cycle. These include being used as a survey frame, as a replacement for survey data collection, for editing and imputation of missing responses, or for survey evaluation. With

[1] U.S. Census Bureau, Washington, DC 20233, U.S.A. Emails: andrew.d.keller@census.gov, vincent.t.mule.jr@census.gov, darcy.steeg.morris@census.gov, and scott.m.konicki@census.gov
Disclaimer: This article is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

respect to censuses, Steffey and Bradburn (1994) note possible uses of administrative records including for coverage improvement, census evaluation, operational efficiency improvement, or to replace traditional census-taking wholly or partially with an administrative records (i.e., register-based) census. Many countries have indeed adopted full register-based (Thygesen 2015; van Zeijl 2014) or partial register-based censuses (Maris et al. 2012). Using administrative records in such a way offers a cost-saving opportunity in a changing census environment of escalating costs; however, it is equally important to consider the quality implications to guide when the use of administrative records is appropriate.

The goal of the 2020 U.S. Census is to count each person once in their correct location at a lower cost per household (adjusted for inflation) than the 2010 Census while maintaining data quality. To meet this goal, the Census Bureau is researching fundamental changes to the design, implementation, and management of the 2020 Census. One major innovation research area noted in the 2020 Operational Plan (U.S. Census Bureau 2017a) is the development of methodologies to incorporate administrative records (AR) into the census design. The U.S. Census Bureau proposes using administrative records in various parts of the operation including to update the address frame, for effective advertising, and to validate respondent addresses for Internet responses to prevent fraud. The 2020 Operational Plan (U.S. Census Bureau 2017a) also specifically recognizes using administrative records to reduce contacts in the Nonresponse Followup (NRFU) operation.

In the 2010 Census, the NRFU operation sent enumerators to about 50 million addresses in all areas of the country to verify the status for every non-self-responding address. Each NRFU address was allowed up to six enumerator contacts. After over 90 million personal visit attempts across the country with field costs of about USD 1.6 billion (Walker et al. 2012), each address was determined to be occupied, vacant, or nonexistent. The occupied units were assigned a person count and person roster including basic demographic characteristics such as name, age, date of birth, race, Hispanic origin, and relationship to householder.

Modernizing the U.S. decennial census using administrative records to supplement or replace traditional census-taking has been a topic of interest since the 1980s (Alvey and Scheuren 1982; Scheuren 1999). However, unlike other countries that implement full or partial register-based censuses, the U.S. has not had a single administrative records system with a high coverage of the entire population (Mulry 2014). For example, the Census Bureau is provided conditional access to data from organizations such as the Internal Revenue Service (IRS), Social Security Administration (SSA), Center for Medicare and Medicaid Services (CMS), and commercial data vendors. Even though each of these data sources covers just a segment of the entire U.S. population, they provide information relevant to census enumeration such as a person's tax-filing address from IRS and birth date from SSA. Previous research has developed methods to combine and use several administrative sources to identify occupied and vacant units prior to or after minimal NRFU fieldwork, thus reducing the number of enumerator visits (Mule and Keller 2014). The administrative sources are used as an input to decision rules about mode switching in NRFU. In this article, we describe an approach to classify units as vacant or occupied at the beginning of NRFU to enable census field operations to reduce costs, thereby allowing resources to focus on units where administrative data are unreliable or unavailable. Our

approach is developed as a way to classify administrative record data as high quality or not in order to selectively substitute for field responses early in NRFU activities. However, the approach can also be used adaptively throughout the data collection phase or in other census operations, such as for imputation in the data-processing phase (see Subsection 5.2). Or, more generally, the approach can be applied to sample surveys with little alteration when appropriate data are available (see Subsection 5.3).

Through a series of census field tests, various approaches for determining vacant and occupied housing units via administrative records have been tested and refined with increasing levels of complexity and integration with other census operations. In the 2013 and 2014 Census Tests, rules-based approaches were implemented (Walejko et al. 2014; Keller et al. 2016), followed by a predictive modeling approach used in the 2015 Census Test based on linear optimization of logistic regression model predictions (Morris et al. 2016). Most recently, the 2016 Census Test used an adaptation of the modeling approach that is based on a Euclidean distance function (Chapin and Keller 2017). In this article, we present the distance function approach to determine high-quality administrative records in a way that simplifies implementation, while maintaining similar quality to the procedure used in the 2015 Census Test. We retain the same underlying predictive modeling structure, as it naturally incorporates information from multiple administrative records sources and other auxiliary data, but offers a new way to synthesize the model information. We illustrate the utility of this advancement of the predictive modeling methodology in a retrospective study of the 2010 Census. The distance function approach is a direct alternative to the optimization approach in Morris et al. (2016). Comparing the two methods, we find that our method has a high overlap with the linear optimization approach in identifying cases of sufficient quality that can be enumerated using administrative records. At the same time, the distance function yields similar quality metrics (measured through a retrospective study of the 2010 Census) while being easier to implement. Furthermore, the classification mechanism of the distance function approach selects housing units based on their own merit rather than relative to a predetermined set of housing units.

2. Administrative Records Data for the Study of the Decennial Census

The Census Bureau receives separate administrative record data files from various government agencies and private companies for statistical use. To enable the linking of these diverse data sets, an anonymized identifier is assigned to each person record in each administrative record file. The Census Bureau's Person Identification Validation System (PVS) determines the protected identification key (PIK) via a probabilistic matching algorithm between the administrative record source data and a series of reference files. See Wagner and Layne (2014) for details on the PVS algorithm. For simplicity we assume that the PIK assignment is correct and match the files accordingly, however, we acknowledge the importance of linkage error associated with the PVS methodology. See, for example, Layne et al. (2014) for discussion of error associated with PIK assignment given the use of the various reference files.

In our study of the 2010 Census, we use these linkable and anonymized administrative record files to compile a household roster composed of administrative record persons for all housing units in the 2010 NRFU universe in the United States. The 2010 vintage

administrative record sources used to create 2010-level administrative record household rosters are:

- IRS Individual Tax Returns (Form 1040)
- IRS Informational Returns (Form 1099)
- Indian Health Service (IHS) Patient Database
- CMS Medicare Enrollment Database

The resulting administrative record household roster – the collection of PIKs found in any of the selected administrative record files at a given address – is unique by person and address. That is, no persons are duplicated within a housing unit. We use person-level administrative record data, as well as an aggregated housing unit-level administrative record data set that includes characteristics such as administrative record household count and general characteristics of the people in the household. The Social Security Numerical Identification (Numident) File is used to obtain age and sex information for each person in the administrative record household roster.

Rastogi and O'Hara (2012) compared several administrative record and third-party sources to the 2010 Census. For federal files, IRS 1040 individual tax returns had the highest match rate to the 2010 Census. This is due to the magnitude of persons and the fact that tax filings start in February with a deadline of April 15, close to the April 1 Census Day. The analysis showed that CMS' Medicare Enrollment Database had a high match rate for the elderly population. The IHS Patient Database is chosen to address potential undercoverage of the American Indian population. The Social Security Numident file has been shown to have very high coverage and reliable data for age and sex.

It should be noted that not all housing units have information in the selected administrative record files. Conversely, there are people in the administrative records files that are not enumerated in the census. Hence, undercoverage and overcoverage exists when comparing between a census roster and an administrative record roster for the same unit. Because we are not assuming that the administrative records files have sufficient coverage of the entire population, our approach is to eliminate NRFU visits to addresses for which we are confident in the administrative record data. That is, we are trying to limit the use of administrative records to cases where coverage differences between administrative records and fieldwork are minimized, provided that fieldwork would generate the correct Census Day roster.

In addition to the administrative sources, information from commercial files, is used to inform the models. Variables derived from these data are used as independent variables in the models. We also incorporate data from the United States Postal Service (USPS) Delivery Sequence File (DSF), the American Community Survey (ACS), the Master Address File (MAF), census operational information, and USPS Undeliverable as Addressed (UAA) reason codes obtained from census mailings delivered around Census Day.

3. Models and Methodology

The administrative records data described in Section 2 contains a wealth of timely information about the characteristics of addresses. We employ a modeling approach to extract predictive information from the administrative records to identify housing units with

sufficiently reliable vacancy and roster information. The predictive models described in Subsections 3.1 and 3.2 to follow are the same as those used in Morris et al. (2016). A cursory description of the models is provided here; see Morris et al. (2016) for further details. These models estimate various measures of administrative record quality that are subsequently used to rank housing units based on their likelihood of vacancy or their likelihood of correct enumeration for occupied housing units. In Subsection 3.3, we present the distance function approach as a way to use the predicted probabilities from the models to define a quality ranking and identify high-quality housing units that can be removed from the NRFU workload and enumerated using existing administrative records. We refer to units identified as having sufficiently good information from administrative records to accurately predict a vacant housing unit as *AR Vacant*; we define *AR Occupied* units analogously.

3.1. Model for Determining Vacant Housing Units

To identify vacant units via administrative record information, we rely on a statistical model to estimate predicted probabilities of Census Day housing unit status. We fit a multinomial logit model on the housing unit-level administrative record data to predict the three possible values of housing unit status: occupied $\left(y_h^{unocc} = 1\right)$, vacant $\left(y_h^{unocc} = 2\right)$, or nonexistent $\left(y_h^{unocc} = 3\right)$, where the *unocc* superscript denotes the model used for administrative records removal of unoccupied or vacant housing units, and the h subscript indexes the housing unit. From this model, we estimate the probability of each unit status type in the 2010 Census data (i.e., the training data):

$$\hat{p}_{h,occ}^{unocc} = P\left(y_h^{unocc} = 1\right), \quad \hat{p}_{h,vac}^{unocc} = P\left(y_h^{unocc} = 2\right), \quad \hat{p}_{h,del}^{unocc} = P\left(y_h^{unocc} = 3\right).$$

The predicted probabilities, $\hat{p}_{h,occ}^{unocc}$ and $\hat{p}_{h,vac}^{unocc}$, are passed to the distance function to determine which cases are identified as AR Vacant.

The use of a statistical model naturally allows the incorporation of information from multiple sources. For example, vacancy information from a USPS mailing around Census Day is strongly associated with Census Day vacancy (Keller et al. 2016), however it is not a perfect proxy and is not the only strong predictor. This model combines information from USPS mailing data and persons associated with a housing unit present in, for example, tax returns or the Medicare enrollment database. Specifically, housing unit status – as determined by the training data (2010 Census data in our application) – is modeled as a function of independent variables from administrative records, field collection paradata, and survey information. Such covariate information includes the UAA data from the USPS for each of the census mailings, persons from the administrative record sources listed in Section 2, characteristics associated with the block group as determined by the ACS, and other address-level information. The appendix contains a complete list of independent variables for the vacant model.

3.2. Models for Enumerating Occupied Housing Units

To identify and enumerate occupied units via administrative record information, we rely on two statistical models to measure the quality of the administrative records information for enumerating households accurately.

3.2.1. Person-Place Model

The person-place model estimates the probability of enumerating a person on the administrative records at the same address as the 2010 Census data (i.e., the training data). We fit a logistic regression model on the person-level administrative record data to predict the outcome:

$$y_{ih}^{occ1} = \begin{cases} 1 & \text{if person } i \text{ is found in AR and 2010 Census at the same address } h \\ 0 & \text{otherwise} \end{cases}$$

where the *occ1* superscript denotes the person-place model for determining occupied units, the h subscript indexes the housing unit, and the i subscript indexes the administrative record person. Morris (2014) and Morris (2017) study a version of the person-place model comparing alternative estimation approaches (logistic regression, classification trees, and random forests). The choice of estimation procedure has little impact on the findings, thus logistic regression is used here for consistency with the other models used in this research. This model assigns to all person-place pairs in administrative record files a predicted probability, $\hat{p}_{ih}^{occ1} = P\left(y_{ih}^{occ1} = 1\right)$, that the 2010 Census and the administrative record roster data place the person at the same address. The person-place model includes all administrative record person records associated with the address from the sources in Section 2. The 2010 Census person records are assigned PIKs with the methodology discussed in Section 2. Note that a person in administrative record and not the Census is coded as $y_{ih}^{occ1} = 0$. This category could include possible census omissions. Conversely, a person not in administrative records and in the census is excluded from the modeling universe.

Person-place match is modeled as a function of independent variables from person-level administrative record information (e.g., indicators of the presence of the administrative records person in each source at the address, indicators of presence of the administrative records person at a different address within the same administrative records source), address-level administrative record information (e.g., number of administrative records people associated with an address), field operations information (e.g., USPS mailing information, number of NRFU neighbors), and information from other survey sources (e.g., characteristics of the local geography – such as poverty rate, renter rate, Hispanic rate, vacancy rate – from the ACS). The person- and address-level administrative record information is of particular importance. For example, Morris (2014) finds that the presence of an IRS 1040 record at given address, and conversely, the presence of an IRS 1040 at a different address, are strong predictors in the person-place model. The former is associated with an increased probability of the administrative records placing the person at the census address, whereas the latter is associated with a decreased probability. The appendix contains a complete list of independent variables for person-place model.

The person-place model is fit at the person-level, but decisions are made at the housing unit-level. Therefore, the person-level predicted probabilities, $\hat{p}_{ih}^{occ1}$, are summarized for each address such that the housing unit-level predicted probability for address h was defined as:

$$\hat{p}_{h}^{occ1} = \min\left(\hat{p}_{1h}^{occ1}, \ldots, \hat{p}_{n_h h}^{occ1}\right)$$

where n_h is 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 administrative record household count is defined as the sum of all individuals associated with the administrative record address, and each address has the associated predicted probability of having an administrative record/census address match. These predicted probabilities, $\hat{p}_h^{occ1}$, are passed to the distance function to determine which cases are identified as AR Occupied.

3.2.2. Household Composition Model

The household composition model is used to estimate the probability that the sample address has the same household composition (number of adults and children) determined by NRFU fieldwork as its pre-identified administrative record household composition. We fit a multinomial logistic model on the housing unit-level administrative record data to predict the outcome from the 2010 Census (i.e., the training data):

$$y_h^{occ2} = \begin{cases} 0 & \text{if unit } h \text{ is vacant in 2010 Census} \\ 1 & \text{if unit } h \text{ has 1 adult and 0 children in 2010 Census} \\ 2 & \text{if unit } h \text{ has 1 adult and } \geq 1 \text{ children in 2010 Census} \\ 3 & \text{if unit } h \text{ has 2 adults and 0 children in 2010 Census} \\ 4 & \text{if unit } h \text{ has 2 adults and } \geq 1 \text{ children in 2010 Census} \\ 5 & \text{if unit } h \text{ has 3 adults and 0 children in 2010 Census} \\ 6 & \text{if unit } h \text{ has 3 adults and } \geq 1 \text{ children in 2010 Census} \\ 7 & \text{if unit } h \text{ has } \geq 4 \text{ adults in 2010 Census} \end{cases}$$

where the *occ2* superscript denotes the household composition model for determining occupied units, and the h subscript indexes the housing unit. For every address, this model assigns a predicted probability of each household composition type, $\hat{p}_{h,k}^{occ2} = P\left(y_h^{occ2} = k\right)$ for $k = 0,1,2,3,4,5,6,7$. Note that the construction of the dependent variable assumes that age is nonmissing for all housing units. This assumption is satisfied in our application because we use an edited file that includes imputed age for any nonresponse.

The household composition dependent variable y_h^{occ2} is modeled as a function of independent variables from housing unit-level administrative record information (e.g., count of all administrative records person records associated with the address from each of the administrative records sources), person-level administrative record information (e.g., indicators of whether any administrative records person was found at a different address within the same administrative records source), and housing unit-level information from other survey sources (e.g., flags indicating that young children, elderly, Black or White persons from administrative records were associated with the household). The appendix contains a complete list of independent variables for household composition model.

We are solely interested in the predicted probability associated with the household composition observed in the administrative records. That is, for each housing unit we extract the household composition predicted probability associated with the administrative record household composition, defining $\hat{p}_h^{occ2} = \hat{p}_{h,k^*}^{occ2}$ where k^* is the administrative

record household composition. For example, $\hat{p}_h^{occ2} = \hat{p}_{h,3}^{occ2}$ for a housing unit with an administrative record household composition type of two adults and zero children. These predicted probabilities, $\hat{p}_h^{occ2}$, are passed to the distance function to determine which cases are identified as AR Occupied.

3.3. *Identifying Administrative Record Vacant and Occupied Housing Units Using a Distance Function*

We study a direct alternative for the approach described in Morris et al. (2016) that was implemented in the 2015 Census Test. Morris et al. (2016) use linear programming techniques to combine information from the previously described models to determine AR Vacant and AR Occupied housing units. The optimization approach requires setting multiple threshold parameters that are not straightforward to select and interpret. Furthermore, the constraints in the optimization routine involve averages of probabilities over select workloads, where a workload is a set of housing units that requires enumeration.

Specifically, for identifying AR Vacant units, Morris et al. (2016) set constraints that (1) the average vacant predicted probability must exceed a prespecified threshold and (2) the sum of the occupied predicted probability did not exceed a certain percentage of the estimate of occupied housing units from the American Community Survey. With respect to identifying AR Occupied units, the authors set constraints that (1) the average person-place predicted probability must exceed a prespecified threshold and (2) the average household composition predicted probability must also exceed a different prespecified threshold. This is potentially problematic for two reasons: (1) it allows housing units other than the housing unit of interest to contribute to the identification of that unit as AR Vacant or AR Occupied, and (2) the workload over which to take the average must be predefined and has an effect on each housing unit's identification.

Consider a simple example of determining AR Vacant units in two NRFU workloads, each of four addresses with the following vacant probabilities:

$$\text{Workload 1: } \hat{p}_{1,vac}^{unocc} = 0.81,\ \ \hat{p}_{2,vac}^{unocc} = 0.81,\ \ \hat{p}_{3,vac}^{unocc} = 0.75,\ \ \hat{p}_{4,vac}^{unocc} = 0.50$$

$$\text{Workload 2: } \hat{p}_{1,vac}^{unocc} = 0.90,\ \ \hat{p}_{2,vac}^{unocc} = 0.90,\ \ \hat{p}_{3,vac}^{unocc} = 0.72,\ \ \hat{p}_{4,vac}^{unocc} = 0.50$$

Focusing solely on the average predicted probability constraint for illustrative purposes, the optimization approach of Morris et al. (2016) identifies AR Vacant addresses as those contained in the subset of housing unit-level predicted probabilities that maintains an average that exceeds a specified cutoff. Using the cutoff of 0.8 used in Morris et al. (2016), in this example averaging would identify housing units $h = 1$ and $h = 2$ as AR Vacant in Workload 1, and housing units $h = 1$, $h = 2$, and $h = 3$ as AR Vacant in Workload 2. Due to the nature of averaging, the third household ($h = 3$) is identified as AR Vacant in Workload 2 despite that it has a lower predicted probability of vacancy in Workload 2 as compared to Workload 1. This simplistic example illustrates how the averaging of predicted probabilities allows other cases to contribute to identification of AR Vacant units. In the same vein, the AR Vacant determination depends crucially on the set of predicted probabilities included in the average. Average predicted probabilities are

computed over a predefined area; therefore a decision has to be made about over what areas the averaging is done. One possibility would be to run the linear optimization over the entire nation. This could cause a disproportionate amount of cases to be removed in one area, resulting in unbalanced workloads. Another alternative could be to run the linear optimization for each state or county. Doing this would require running the optimization 50 or 3,000 times, which could increase the computational time and complexity. In an environment where field operations are waiting on results from the administrative records models, the days it would take to run the optimization routine would make timing more challenging.

We study a simpler approach using a distance function that avoids the concerns of the optimization approach – in particular, the distance method evaluates each housing unit on its own merit – and relies on a more transparent and interpretable threshold parameter. Furthermore, the distance method is easier to implement in that real-time workload adjustments can be determined by simply changing the threshold parameter rather than rerunning the optimization procedure. This alternative is partially motivated by the use of a decision criterion for identifying cases to enumerate using administrative records based on distances measured via Receiver Operator Characteristic (ROC) graphs (Morris 2014, 2017). We define distance functions that take multiple measures of the quality of the administrative records, with respect to determining vacancy and for enumeration of occupied housing units, as inputs to output a single measure. This scalar distance measure combines multiple predicted probabilities – which are themselves based on the combination of multiple sources of information via the statistical models – to allow (1) a ranking of the housing units by quality and (2) a definition of a subset of the highest quality housing units by choosing a threshold.

With regard to vacancy determination, we define the housing unit-level *vacant distance* based on the vacant probability, $\hat{p}_{h,vac}^{unocc}$, and occupied probability, $\hat{p}_{h,occ}^{unocc}$, estimated via the housing unit status model discussed in Subsection 3.1. These predicted probabilities can be thought of as a two-dimensional plane with each probability on one dimension with values between 0 and 1. Based on the two probabilities, each address would have a point in this two-dimensional space. The most likely vacant cases would be those that have shortest distance to the point where the occupied probability equals 0 and the vacant probability equals 1 (i.e., the (0,1) point). As a result, we define the Euclidean vacant distance, d_h^{vac}, for each unit h, as

$$d_h^{vac} = \sqrt{\left(1 - \hat{p}_{h,vac}^{unocc}\right)^2 + \left(\hat{p}_{h,occ}^{unocc}\right)^2}.$$

With regard to identifying occupied housing units for administrative record enumeration, we define the housing unit-level *occupied distance* based on predicted probabilities from the two occupied models: the minimum person-place probability for the address, $\hat{p}_h^{occ1}$, and the household composition probability associated with the observed administrative record household composition, $\hat{p}_h^{occ2}$. Both of these probabilities are measures of quality (count match and household composition match, respectively) such that the housing units with higher quality administrative records are associated with higher estimated probabilities. Even though the predictions from these two models are correlated, Morris et al. (2016) show higher agreement in population count and household composition when both models

are used together as compared to using one or the other. Accordingly, we use results from both the person-place and household composition model as inputs for the distance function. Similar to the construction of the vacant distance, the most likely occupied and correct enumeration cases would be those that have shortest distance to the point where the predicted probability from both models equals 1 (i.e., the (1,1) point). Based on this idea, we use the Euclidean distance to define the occupied distance, d_h^{occ}, for each unit h as

$$d_h^{occ} = \sqrt{\left(1 - \hat{p}_h^{occ1}\right)^2 + \left(1 - \hat{p}_h^{occ2}\right)^2}.$$

The distances d_h^{vac} and d_h^{occ} are used to determine AR Vacant and AR Occupied housing units, respectively. That is, we define a given distance cutoff targeting a certain rate of removal of cases from the face-to-face follow-up. We then treat those administrative records as a reasonably correct representation of the true status for those addresses.

4. Application: 2010 Decennial Census Data

We apply the distance function methodology for determining AR Vacant and AR Occupied housing units in a retrospective study of the NRFU operation of the 2010 Census. In this analysis, the vacant model and two occupied models are fit to a sample of the NRFU housing units in the 2010 Census. The fitted coefficients are then applied to all NRFU housing units to obtain the predicted probabilities ($\hat{p}_{h,vac}^{unocc}$ and $\hat{p}_{h,occ}^{unocc}$ for the vacant model, $\hat{p}_h^{occ1}$ and $\hat{p}_h^{occ2}$ for the occupied models) and the associated distances (d_h^{vac} and d_h^{occ}) for each housing unit h.

4.1. Identifying Administrative Record Vacant Housing Units

Figure 1 plots the estimated vacant probability, $\hat{p}_{h,vac}^{unocc}$, and occupied probability, $\hat{p}_{h,occ}^{unocc}$, for the 50 million NRFU housing units in the 2010 Census. The vacant distance measure, d_h^{vac}, is used to create percentile bands generated by assuming varying cutoffs. The upper

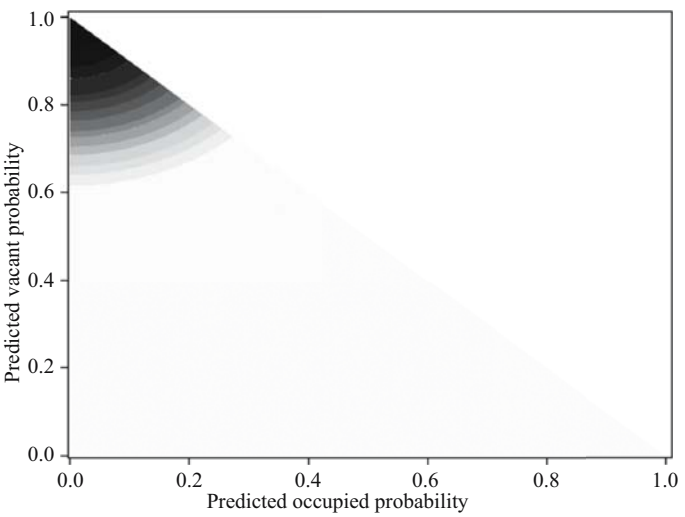

Fig. 1. AR vacant predicted probabilities by vacant distance percentile (source: 2010 Census simulation).

leftmost area, denoted by the black shading, represents the top one percent of NRFU cases with the smallest vacant distance. If we were to restrict our AR Vacant identification total to 500,000 cases, removing these cases from the NRFU workload would reduce the number of visits during NRFU at the smallest predicted expense of quality. The band just below the upper leftmost area, denoted by the darkest gray, are those housing units between the top one percent and two percent of NRFU housing units with the smallest vacant distance. Dividing the data by percentile bands yields the partial concentric circles in Figure 1 depicting various scenarios of target NRFU workload reduction.

To assess accuracy for varying vacant distance cutoffs, we treat the 2010 NRFU housing unit status as the gold standard and compare field vacancy determination to administrative record vacancy determination. Figure 2 shows the true positive rate – the percent of AR Vacant cases that were resolved as vacant during the 2010 NRFU – for each mutually exclusive percentile band up to the 15th percentile, with the lowest vacant distance starting at the first percentile. We see in Figure 2 that, for the top one percent of cases (500,000 NRFU cases) with the shortest vacant distance between the (0,1) point and $\left(\hat{p}_{h,occ}^{unocc}, \hat{p}_{h,vac}^{unocc}\right)$, the true positive rate is 90.8 percent – indicating that among the 500,000 NRFU cases identified as AR Vacant using the distance function approach, 90.8 percent were resolved as vacant through NRFU fieldwork. For the second best one percent of cases (i.e., cases of rank 500,001 to 1,000,000), the true positive rate is 84.9 percent. There is a gradual decrease in the true positive rate as the percentiles increase, depicting the decrease in the quality of administrative records information for cases with a vacant distance that is further from the optimal (0,1) point.

Based on the analysis and the tradeoff between cost reduction and quality, a decision can be made about how many bands to designate as being AR Vacant. The tradeoff exists because by identifying more AR Vacant cases, thereby reducing costs incurred by NRFU followup, we see a larger percentage of cases return as occupied.

Morris et al. (2016) use linear optimization processing of the same predicted probabilities – $\hat{p}_{h,occ}^{unocc}$ and $\hat{p}_{h,vac}^{unocc}$ – to determine about ten percent of the NRFU universe (5,132,613 addresses) as AR Vacant. We are interested in comparing the performance of the linear optimization approach with the simpler distance function approach presented in

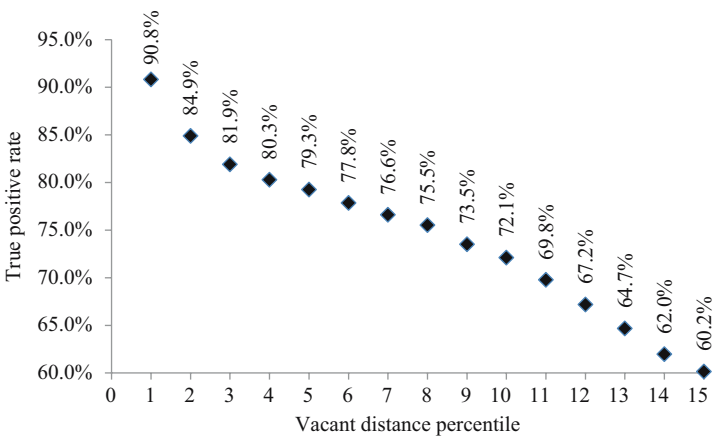

Fig. 2. AR vacant true positive rate by vacant distance percentile (source: 2010 Census simulation).

this article. To do this, we sort the housing units from smallest to largest vacant distance and identify the 5,132,613 addresses with the smallest vacant distance to be AR Vacant. Among these AR Vacant cases, the smallest vacant distance value is $d_h^{vac} = 0.0078$ and the largest vacant distance value is $d_h^{vac} = 0.3559$. We find that 91 percent of the addresses determined to be AR Vacant using the distance function are also identified as AR Vacant by the linear optimization approach. The two methods largely identify the same AR Vacant cases. However the distance function is easier to operationalize.

We further evaluate the distance function and linear optimization AR Vacant cases compared to their 2010 NRFU results. Table 1 shows the results from contrasting the optimization approach versus the distance approach for the same workload. We find similar observed 2010 distributions between the two identification approaches. The distance approach does slightly better in terms of agreement with the NRFU result – the percentage of AR Vacant cases with a vacant NRFU status is higher for the distance approach versus the optimization approach (79.0% vs. 78.1%). Regardless of the approach, not all cases identified as AR Vacant were vacant in the 2010 NRFU. Some of the misclassification between administrative records and census may be due to errors in the 2010 Census. Keller and Konicki (2016) show that approximately ten percent of persons enumerated in these AR Vacant and field occupied units are erroneous enumerations and 20 percent are imputed.

To further assess quality implications, we can look to other 2010 coverage results. Cresce (2012) showed that the 2010 Census continued the trend from the 1990 Census and 2000 Census of underestimating the vacancy rate as compared to other estimates like the American Housing Survey and the Current Population Survey. The Census Coverage Measurement program found that vacant housing units were undercounted by 4.8 percent in 2010 (Mule and Konicki 2012). These evaluation results suggest that by conducting interviews between March and August to assess the population on April 1, the decennial census may have enumerated people in units that were vacant on Census Day.

4.2. *Identifying Administrative Record Occupied Housing Units*

Our assessment of the identification of AR Occupied units is analogous to that of identifying AR Vacant units in the previous section; however, the distance function for identifying AR Occupied units depends on predicted probabilities from two separate models rather than one model. Figure 3 plots the predicted probability from the person-place model, $\hat{p}_h^{occ1}$, and for the household composition model, $\hat{p}_h^{occ2}$, for the eligible NRFU housing units. Only those NRFU addresses with an associated administrative record person are eligible to be AR Occupied. The occupied distance measure, d_h^{occ}, is used to create percentile bands generated by assuming varying cutoffs. The upper rightmost

Table 1. AR vacant versus NRFU status assigned – optimization approach versus distance approach (source: 2010 Census simulation).

AR vacant approach	Workload removal	Occupied (%)	Vacant (%)	Nonexistent (%)	Unresolved (%)
Optimization	5,132,613	9.1	78.1	11.9	0.9
Distance	5,132,613	8.8	79.0	11.3	0.9

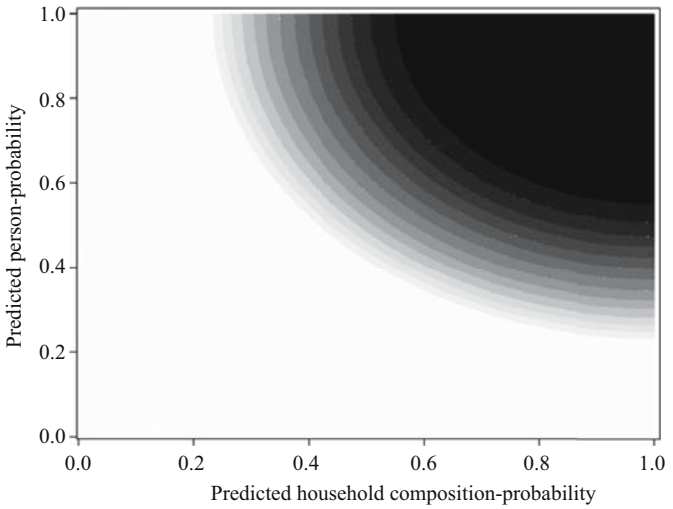

Fig. 3. AR occupied predicted probabilities by occupied distance percentile (source: 2010 Census simulation).

area, denoted by the black shading, represents the top one percent of NRFU cases with the smallest occupied distance. Dividing the data by percentile bands yields the concentric circles in Figure 3 depicting various scenarios of target NRFU workload reduction.

To assess accuracy for varying occupied distance cutoffs, we again treat the 2010 NRFU housing unit status as the gold standard and compare field occupancy determination to administrative record occupancy determination. Figure 4 shows the true positive rate – the percent of AR Occupied cases that were resolved as occupied during the 2010 NRFU – for each mutually exclusive percentile band up to the 15th percentile, with the lowest occupied distance starting at the first percentile. We see in Figure 4 that, for the top one percent (500,000 NRFU cases) with the shortest occupied distance between the (1,1) point and $\left(\hat{p}_h^{occ1}, \hat{p}_h^{occ2}\right)$, the true positive rate is 94.6 percent – indicating that among the 500,000 NRFU cases identified as AR Occupied using the distance function approach, 94.6 percent were resolved as occupied through NRFU fieldwork. There is a gradual decrease in

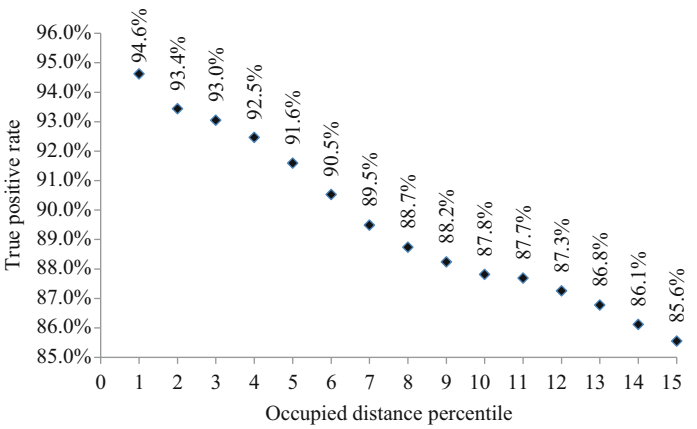

Fig. 4. AR occupied true positive rate by occupied distance percentile (source: 2010 Census simulation).

the true positive rate as percentile increases, similar to administrative records vacant identification.

In addition to determining occupancy and the household count, the decennial census collects information on the characteristics of the people in occupied housing units. It is important to recognize the ramifications on characteristics for cases that are enumerated via administrative records rather than fieldwork. In the previous example of implementing administrative record enumeration for the top one percent based on the occupied models, 500,000 housing units are assigned persons from administrative records. However, because no interviews are completed, characteristics for people in these housing units must be obtained from the administrative records or imputed.

Some characteristics are readily available from the administrative records sources: age is a necessary requirement to be AR Occupied as the household composition model depends on age by definition. Obtained from the Numident file, sex is also usually a nonmissing characteristic. Other characteristics are less straightforward, namely race and Hispanic origin. We use administrative record data from various sources to identify race and Hispanic origin for persons enumerated in AR Occupied units. See Ennis et al. (2015) for a full explanation of how race and Hispanic origin are assigned to persons in the administrative record data. Figure 5 shows the housing unit-level missing data rate for race and Hispanic origin for housing units identified as AR Occupied by each percentile of the occupied distance (starting at the first percentile). For example, of the 500,000 NRFU units identified as AR Occupied in the second percentile, about 0.50 percent of housing units are missing Hispanic for all persons. This would necessitate assigning Hispanic origin for all persons in these housing units via an imputation procedure.

Similar to the vacant cases, we are interested in comparing the performance of the linear optimization approach with the simpler distance function approach presented in this article. We sort the housing units from smallest to largest occupied distance and identify the 7,292,195 addresses with the smallest occupied distance to be AR Occupied. In this case, about 15 percent of the NRFU universe is identified as AR Occupied corresponding to a occupied distance threshold of $d_h^{occ} = 0.7140$, where the smallest observed occupied distance value is $d_h^{occ} = 0.1907$. We find that 93 percent of the addresses determined to be

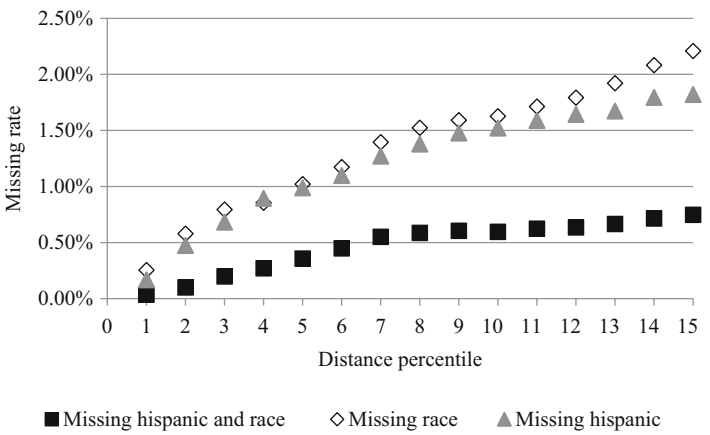

Fig. 5. Housing unit missing data rates by occupied distance percentile (source: 2010 Census simulation).

Table 2. AR occupied versus NRFU status assigned – optimization approach versus distance approach (source: 2010 Census simulation).

AR occupied approach	Workload removal	Occupied (%)	Vacant (%)	Nonexistent (%)	Unresolved (%)
Optimization	7,292,195	90.2	7.9	1.6	0.3
Distance	7,292,195	89.7	8.4	1.7	0.3

AR Occupied using the distance approach are also identified as AR Occupied by the linear optimization approach. Hence, the two methods largely identify the same cases to be removed from the operation and enumerated as occupied via administrative records.

Table 2 shows similar observed distributions of 2010 housing status when contrasting the optimization approach and the distance approach for the same workload amount. The optimization approach does slightly better in terms of agreement with the NRFU result – the percentage of AR Occupied cases with an occupied NRFU status is higher for the optimization approach versus the distance approach (90.2% vs. 89.7%). Note that not all cases identified as AR Occupied were occupied in the 2010 NRFU. Some of the misclassification between administrative records and census may be due to omissions in the census.

5. Conclusion and Discussion

To prepare for the 2020 Census, the Census Bureau is researching cost-saving changes to NRFU. The use of administrative records to reduce field contacts in NRFU is one cost-saving measure specifically noted in the 2020 Operational Plan (U.S. Census Bureau 2017a). We propose a modeling approach for assessing the quality of administrative records for enumerating housing units in conjunction with a distance function to identify AR Vacant and AR Occupied units. The results from the retrospective study of the 2010 Census provide evidence of the internal validation of the model and methodology as the distance function approach accurately recognizes vacancy and occupancy in the vast majority of AR Vacant and AR Occupied cases, respectively. Similarly, the 2016 Census Test provided external validation of the distance approach (Chapin and Keller 2017). We contrast the distance function approach with the optimization approach discussed in Morris et al. (2016) and implemented in the 2015 Census Test. Even though we find that the two methods perform similarly on the 2010 Census data, we favor the distance function approach for its simplicity and operational ease to document in a production environment. This new approach provides a more objective way to define thresholds that dictate the cost and quality tradeoff. The choice of the distance measure cutoff implies a cost reduction in that the addresses identified would receive fewer visits during NRFU, but quality metrics such as true positive rates must be factored in as well.

5.1. Contact Strategy

The proposed distance approach for identifying AR Vacant and AR Occupied cases can be used operationally in the context of a broader contact strategy. Here we provide an overview of a NRFU field visit strategy related to units identified as AR Vacant or AR

Occupied. This contact strategy – which was implemented in the 2016 Census Test (Chapin and Keller 2017) – illustrates how and when administrative records may substitute for face-to-face interviews, thus reducing costs of field operations. Research and testing programs continue to adapt and refine this contact strategy leading up to the 2020 Census, but generally suggest using administrative records in a reasonably similar manner (U.S. Census Bureau 2017b).

Prior to the start of the 2016 Census Test NRFU operation, each address was eligible to receive up to four mailings before and after Census Day. If the address did not respond to these mailings, the Census Bureau decided how many times to visit the address during the NRFU operation in accord with the quality of administrative record data. Figure 6 shows the flowchart of the visit strategy for NRFU housing units in the 2016 Census Test. The distance function methodology was used to identify the AR Vacant and AR Occupied housing units shaded in the flowchart in Figure 6 (Chapin and Keller 2017).

Housing units identified as AR Vacant did not receive any visits during NRFU. In general, the AR Vacant units were those units with Undeliverable as Addressed reason codes returned from the initial census mailings and an absence of administrative record presence (i.e., no sign of life in the administrative records). As part of the NRFU contact strategy, a postcard was mailed to the AR Vacant units to allow an additional opportunity for self-response.

The cases not identified as AR Vacant received one field visit. This visit allowed cases to be resolved in several ways: completion of an interview with the household member, field determination of vacancy, or field determination that the address was not a housing unit. If the enumerator did not make contact with anybody at the housing unit, the enumerator left a notice of visit regardless of whether the unit was AR Occupied or not. This notice of visit included self-response information to encourage the household to respond by going online, dialing the questionnaire assistance number, or returning the paper questionnaire sent earlier. Units determined to be AR Occupied received only this one visit in the 2016 Census Test. After one visit, if the housing unit remained unresolved then AR Occupied housing units received an additional postcard mailing with self-response information. All other unresolved housing units (those not identified as AR Occupied) were contacted via the usual protocol (i.e., additional contacts). As shown, there were several ways before and during NRFU that the Census Bureau attempted to obtain and use self-responses before enumerating cases via administrative record information.

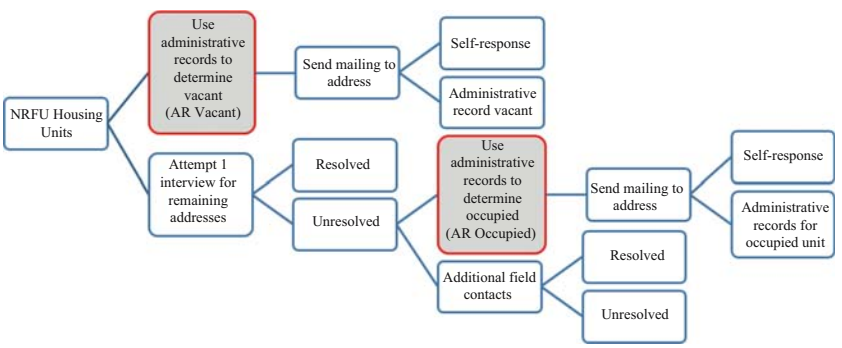

Fig. 6. Nonresponse followup visit strategy (2016 Census test).

5.2. Adaptive Uses of the Distance Function Method

The contact strategy described in Subsection 5.1 shifts AR Vacant and AR Occupied units from an approach solely reliant on enumerator visits. In practice, the AR Occupied units are only allowed a maximum of one enumerator visit before having to respond via another mode. This tailored contact strategy results from identifying cases for removal based on administrative record information available at the start of NRFU. The distance function approach assumes a fixed set of data on which the underlying models are fit. However, the approach can be implemented adaptively as new administrative record information is obtained during the NRFU operation. In the context of the 2015 Census Test, Keller (2016) describes multi-phased integration of administrative record modeling as an adaptive component throughout NRFU. For this test, after initial AR Occupied and AR Vacant cases were identified, the Census Bureau received additional IRS 1040 and IRS 1099 information. After processing these data, the administrative record models were refit. Additional units were identified as AR Occupied and subsequently enumerated via the new administrative record data. Although this was not preplanned, this adaptation enabled the resolution of cases using administrative record data that had not been available at the start of NRFU. Doing so in real-time allowed the Census Bureau to shift resources to units that had proven to be more difficult to enumerate.

The distance function methodology can also be used after data collection is complete, as an alternative to unit imputation of status and population size for unresolved housing units. In the context of the 2015 Census Test, Keller (2016) documents a modification of the optimization approach: refitting and determining AR Vacant and AR Occupied cases by lowering the average constraint values in the optimization approach – thus identifying more AR Vacant and AR Occupied cases. The new cases that remained unresolved addresses after the full visit strategy are assigned occupancy status and enumerated using administrative records rather that via an imputation. In the same fashion, rather than relying on the optimization approach, the new distance function approach can be extended to allow additional unresolved addresses to be assigned an AR Vacant or AR Occupied status by lowering the threshold.

To elaborate on this scenario, Figure 7 shows a hypothetical example for the AR Occupied determination. A distance threshold can be specified to identify the dark gray area in the upper right corner of the figure. Addresses with predicted probabilities in this area will receive no more than one visit. A second distance threshold can be specified to identify the medium gray area. These cases would receive the full visit strategy during NRFU; however, if they are unresolved after fieldwork is completed, then administrative records information would be used to determine occupancy status and a roster, if occupied, instead of using count imputation for these cases. The administrative records for the remaining housing units in the light gray area would not be utilized, as they are not of sufficient quality. This figure and hypothetical scenario exemplify the clarity and ease of communicating the distance function approach for reducing visits or avoiding imputation.

5.3. Implementing on Surveys

We have focused exclusively on using administrative records to replace household responses specifically for the decennial census. Using administrative records to curtail

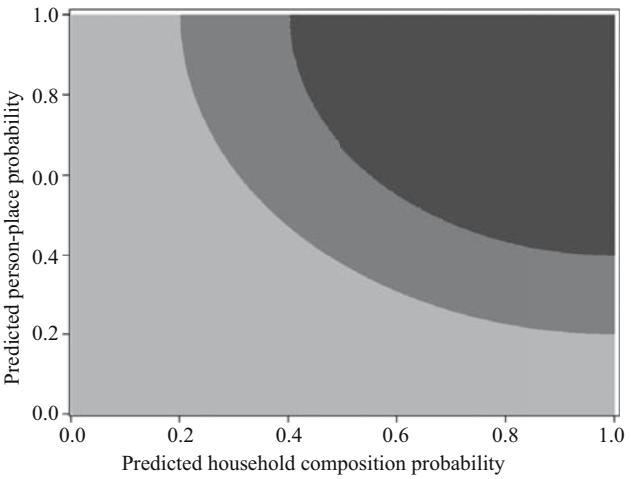

Fig. 7. Hypothetical example of using different occupied distance thresholds (source: 2010 Census simulation).

contacts or reduce respondent burden can be generally useful in surveys. However, admittedly, the use of administrative records in the decennial census is a less arduous problem due to the limited number of interview questions. The decennial census is only charged with forming a Census Day household roster of persons, to include minimal demographic data such as age, sex, Hispanic origin, race, and relationship for persons in occupied units. However, surveys such as the ACS have more data items with more complex topics. Nevertheless, provided that the administrative data relevant to the subject of measurement is available to the survey administrator, the methodology presented in this article can potentially be adapted for survey use.

For example, the Census Bureau has been researching the use of administrative records to reduce the difficulty and length of the American Community Survey to address concerns about respondent burden (Stempowski 2015). Ruggles (2015) identified potential administrative record sources for replacing or supplementing field response data. The American Community Survey Office (ACSO) at the Census Bureau has an active research program to further study topics and variables suggested in Ruggles (2015), for example, income (O'Hara et al. 2016), year built (Moore 2015), and housing value (Kingkade 2013). The preliminary work from ACSO has addressed the potential for an all-or-nothing use of administrative records to eliminate ACS questions. However the distance method could serve as an intermediate solution for reducing respondent burden by tailoring the survey questions based on the administrative record data availability and quality for each respondent. Specifically, historic survey data and the relevant administrative record data could be used to model the quality of administrative record data for a given question. The quality measures resulting from applying the model fit to a current round of data collection could then be used as in the distance method to repurpose the administrative record information via item substitution only for those units with quality exceeding some threshold. Such an adaptive strategy would reduce respondent burden, particularly those with consistently high-quality administrative record data. To our knowledge, such an implementation has not been studied or operationalized in any surveys. However,

relatedly, Chesnut (2013) has studied modeling approaches for adaptive mode switching – model-based tailoring of contact strategies – to reduce respondent burden in the ACS.

5.4. Future Work

The definition of the distance measure for determining vacant and occupied units assumes equal weighting on the two corresponding predicted probabilities. We conjecture that this may be an issue for our AR Occupied identification because it uses two predictions with different census quality ramifications. The person-place model concerns counting people in the right place, whereas the household composition model concerns the agreement between administrative record household composition and census household composition. Differential weighting may be desired if it makes practical and empirical sense to emphasize one model over the other. Alternatively, transformations of the predicted probabilities may have an impact on the conclusions. The distance function uses the raw predicted values; however, the two dimensions each have a different dependent variable such that the distribution of predicted probabilities for each are not likely the same. Further work will examine if transformations including standardizations of the two probabilities can be useful in the determination.

An underlying assumption of the models in this research is that the relationships between the administrative records and the 2010 Census will remain consistent for the 2020 Census. The approach assumes the estimated relationships from the training data (e.g., 2010 Census data) can be reasonably applied for predicting the test data (e.g., 2020 Census data). Additionally, the approach assumes the 2010 Census data as "truth," even though there exists inherent error in Census results. Although the 2010 Census data is a reasonable basis for model development, the Census Bureau is actively researching the feasibility of using alternate training data to fit the administrative record models. For example, the use of more current ACS data as training data in conjunction with 2015 Census Test data could be treated as the gold standard (Chow et al. 2017).

Appendix

Table A1. List of independent variables for vacant and occupied models.

			Occupied models	
	Variable	Vacant model (Section 3.1)	Person-place (Section 3.2.1)	HH composition (Section 3.2.2)
	American community survey block group level variables			
% of	persons in block group (BG) between 25 and 44 years old	X	X	X
	persons in BG greater than 64 years old	X	X	X
	persons in BG identifying as Black	X	X	X
	persons in BG identifying as Hispanic	X	X	X
	occupied housing units in BG with at least 2 related HH members	X	X	X
	persons over 4 in BG speaking language other than English at home	X	X	X
	housing units in BG considered as mobile homes	X	X	X
	housing units in BG where householder/spouse are members of HH	X	X	X
	occupied housing units in BG that are not owner occupied	X	X	X
	housing units in BG vacant at time of interview		X	X
	housing units in BG occupied at time of interview	X		
	persons in BG living below poverty level	X	X	X
	Housing unit characteristics			
	# of neighbors in Non Response Followup (NRFU)	X		
	USPS Undeliverable As Addressed (UAA) reason (two mailings)	X	X	X
	USPS UAA reason agreement – Kappa Coefficient	X		
	housing unit type (e.g., multi-family)	X	X	X
	within structure description	X		

Table A1. Continued.

			Occupied models	
	Variable	Vacant model (Section 3.1)	Person-place (Section 3.2.1)	HH composition (Section 3.2.2)
	has Delivery Sequence File "X" status and both neighbors are in NRFU	X		
	on fall Delivery Sequence File of 2009	X	X	X
	apartment with Unable to Forward UAA reason code on 1st mailing	X		
	Housing unit characteristics from administrative records			
> = 1 person in HU is...	White			X
	Black	X		X
	Hispanic	X		
	missing ethnicity	X		
	age <2	X		X
	age <10	X		X
	age 10–17	X		X
	age 18–24			X
	age 25–44			X
	age 65+	X		X
	Housing unit level administrative record source information			
> = 1 person in HU is placed at this HU according to...	Internal Revenue Service (IRS) 1040 Tax Year (TY) 2009	X		X
	IRS 1099 TY 2009	X		X
	Indian Health Service Patient Database (IHS)			X
	Medicare			X
	Commercial data	X		X
	IRS 1040 TY 2008	X		
	Administrative Records (AR) HH count		X	
	AR HH composition	X	X	X
	HH with IRS 1040 TY 2008 persons, no AR persons in current year	X		
	IRS 1040 TY 2009 persons also in IRS 1040 TY 2008 at same unit			X

Table A1. Continued.

	Variable	Vacant model (Section 3.1)	Occupied models: Person-place (Section 3.2.1)	Occupied models: HH composition (Section 3.2.2)
> = *1 person* in HU is placed at another HU according to...	IRS 1040 TY 2009	X		X
	IRS 1099 TY 2009	X		X
	Medicare	X		
	Commercial data	X		
Person level administrative record source information				
Person is placed at this HU according to...	IRS 1040 TY 2009		X	
	IRS 1099 TY 2009		X	
	IHS		X	
	Medicare		X	
	Commercial data		X	
	IRS 1040 TY 2008		X	
Person is placed at another HU according to...	IRS 1040 TY 2009		X	
	IRS 1099 TY 2009		X	
	IHS		X	
	Medicare		X	
	Commercial data		X	

6. References

Alvey, W. and F. Scheuren. 1982. "Background for an Administrative Record Census." in JSM Proceedings, Social Statistics Section, American Statistical Association, Cincinnati, OH, August 16–19, 1982. Washington, DC: American Statistical Association. 137–152.

Bakker, B.F.M., P.G.M. van Heijden, and S. Scholtus. 2015. " "Preface" to a Special Issue on Coverage Problems in Administrative Sources." *Journal of Official Statistics* 31(3): 349–355. Doi: http://dx.doi.org/10.1515/jos-2015-0021.

Brackstone, G.J. 1987. "Issues in the Use of Administrative Records for Statistical Purposes." *Survey Methodology* 13(1): 29–43. Available at: http://www.statcan.gc.ca/pub/12-001-x/1987001/article/14467-eng.pdf (accessed November 2017).

Chapin, M. and A. Keller. 2017. "Administrative Records Research and Planning, 2018 End-to-End Census Test: Nonresponse Followup." from 2020 Census Program Management Review–April 21, 2017. Available at: https://www2.census.gov/programs-surveys/decennial/2020/program-management/pmr-materials/04-21-2017/pmr-update-testing-2017-04-21.pdf (accessed November 2017).

Chesnut, J. 2013. "Model-Based Mode Switching from Internet to Mail in the American Community Survey." DSSD 2013 American Community Survey Memorandum Series

#ACS13-MP-01. Available at: https://census.gov/content/dam/Census/library/working-papers/2013/acs/2013_Chesnut_01.pdf (accessed November 2017).

Chow, M.C., H.P. Janicki, M.J. Kutzbach, L.F. Warren, and M. Yi. 2017. "A Comparison of Training Modules for Administrative Records Use in Nonresponse Followup Operations: The 2010 Census and the American Community Survey." Center for Economic Studies Working Paper #CES 17-47. Washington, DC: U.S. Census Bureau. Available at: https://www2.census.gov/ces/wp/2017/CES-WP-17-47.pdf (accessed November 2017).

Cresce, A. 2012. "Evaluation of Gross Vacancy Rates From the 2010 Census Versus Current Surveys: Early Findings from Comparisons with the 2010 Census and the 2010 ACS 1-Year Estimates." Federal Committee on Statistical Methodology 2012 Research Conference, Washington, DC, January 10–12, 2012. Available at: https://www.census.gov/housing/files/FCSM%20paper.pdf (accessed November 2017).

Ennis, S.R., S.R. Porter, J.M. Noon, and E. Zapata. 2015. "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. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2015/adrm/carra-wp-2015-08.pdf (accessed November 2017).

Federal Committee of Statistical Methodology. 1980. "Report on Statistical Uses of Administrative Records." *Working Paper 6*, Washington, DC. Available at: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/spwp6.pdf (accessed November 2017).

Fienberg, S.E. 2015. "Discussion" of a Special Issue on Coverage Problems in Administrative Sources. *Journal of Official Statistics* 31(3): 527–535. Doi: http://dx.doi.org/10.1515/jos-2015-0032.

Groves, R.M. and B.A. Harris-Kojetin (editors), National Academies of Sciences, Engineering, and Medicine (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington DC: The National Academies Press.

Japec, L., F. Kreuter, M. Berg, P. Biemer, P., Decker, C. Lampe, J. Lane, C. O'Neil, and A. Usher. 2015. "AAPOR Report on Big Data: Report of the AAPOR Big Data Task Force." Available at: http://www.aapor.org/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15_b.pdf (accessed November 2017).

Keller, A. 2016. "Imputation Research for the 2020 Census." *Statistical Journal of the International Association of Official Statistics* 32: 189–198. Doi: http://dx.doi.org/10.3233/SJI-161009.

Keller, A., T. Fox, and V.T. Mule. 2016. "2014 Census Test – Analysis of Administrative Record Usage." U.S. Census Bureau. Available at: https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-analysis-2014-census-test-ad-rec.pdf (accessed November 2017).

Keller, A. and S. Konicki. 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. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2016/files/389544.pdf (accessed November 2017).

Kingkade, W. 2013. "Self-assessed Housing Values in the American Community Survey: An Exploratory Evaluation Using Linked Real Estate Records." in JSM Proceedings, Government Statistics Section, American Statistical Association, Montreal, Quebec, Canada, August 3-8, 2013. Alexandria, VA: American Statistical Association. 990–1004. Available at: https://ww2.amstat.org/MembersOnly/proceedings/2013/data/assets/handouts/308063_80215.pdf (accessed November 2017).

Layne, M., D. Wagner, and C. Rothhaas. 2014. "Estimating Record Linkage False Match Rate for the Person Identification Validation System." Center for Administrative Records Research and Applications Working Paper #2014-02. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-02.pdf (accessed November 2017).

Maris, M., E.S. Nordholt, and J. van Zeijl. 2012. "Comparing Approaches of Different (Partly) Register-based Countries." from the United Nations Economic Commission for Europe Conference of European Statisticians: UNECE-Eurostat Expert Group Meeting on Censuses Using Registers. Available at: www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2012/use_of_register/WP_3_Netherlands.pdf (accessed November 2017).

Moore, B. 2015. "Preliminary Research for Replacing or Supplementing the Year Built Question on the American Community Survey with Administrative Records." U.S. Census Bureau Working Paper. Available at: www.census.gov/library/working-papers/2015/acs/2015_Moore_02.html (accessed November 2017).

Morris, D.S. 2014. "A Comparison of Methodologies for Classification of Administrative Records Quality for Census numeration." in JSM Proceedings, Survey Research Methods Section, American Statistical Association, Boston, MA, August 2–7, 2014. Alexandria, VA: American Statistical Association. 1729–1743. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2014/files/311864_88281.pdf (accessed November 2017).

Morris, D.S. 2017. "A Modeling Approach for Administrative Records Enumeration in the Decennial Census." *Public Opinion Quarterly: Special Issue on Survey Research, Today and Tomorrow* 81(S1): 357–384. Doi: http://dx.doi.org/10.1093/poq/nfw059.

Morris, D.S., A. Keller, and B. Clark. 2016. "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census." *Statistical Journal of the International Association of Official Statistics* 32: 177–188. Doi: http://dx.doi.org/10.3233/SJI-161002.

Mule, V.T. and A. Keller. 2014. "Using Administrative Records to Reduce Nonresponse Followup Operations." in JSM Proceedings, Survey Research Methods Section, American Statistical Association, Boston, MA, August 2–7, 2014. Alexandria, VA: American Statistical Association. 3601–3608. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2014/files/313148_90659.pdf (accessed November 2017).

Mule, T. and S. Konicki. 2012. "2010 Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Housing Units." DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-02. https://www.census.gov/coverage_measurement/pdfs/g02.pdf (accessed November 2017).

Mulry, M.H. 2014. "Measuring Undercounts for Hard-to-Survey Groups." In *Hard-to-Survey Populations* (Chapter 3), edited by R. Tourangeau, N. Bates, B. Edwards, T. Johnson, and K. Wolter. Cambridge, England: Cambridge University Press. 37–57.

O'Hara, A., A. Bee, and J. Mitchell. 2016. "Preliminary Research for Replacing or Supplementing the Income Question on the American Community Survey with Administrative Records." U.S. Census Bureau Working Paper. Available at: www.census.gov/content/dam/Census/library/working-papers/2016/acs/2016_Ohara_01.pdf (accessed November 2017).

Rastogi, S. and A. O'Hara. 2012. "2010 Census Match Study Report." 2010 Census Planning Memorandum Series. Available at: https://www.census.gov/2010census/pdf/2010_Census_Match_Study_Report.pdf (accessed November 2017).

Ruggles, P. 2015. "Review of Administrative Record Sources Relevant to the American Community Survey." U.S. Census Bureau Working Paper. Available at: www.census.gov/library/working-papers/2015/acs/2015_Ruggles_01.html (accessed November 2017).

Scheuren, F. 1999. "Administrative Records and Census Taking." *Survey Methodology* 25(2): 151–160. Available at: http://www.statcan.gc.ca/pub/12-001-x/1999002/article/4878-eng.pdf?contentType=application%2Fpdf (accessed November 2017).

Steffey, D.L. and N.M. Bradburn (editors), National Research Council. 1994. *Counting People in the Information Age*. Washington DC: The National Academies Press.

Stempowski, D. 2015. "Agility in Action: A Snapshot of Enhancements to the American Community Survey." ACS Information Series Memorandum Number 2015-05. Available at: https://www.census.gov/content/dam/Census/programs-surveys/acs/operations-and-administration/2015-16-survey-enhancements/Agility%20in%20Action%20v2.0.pdf (accessed November 2017).

Thygesen, L. 2015. "The Use of Administrative Sources for Censuses: Merits and Challenges." *Statistical Journal of the IAOS* 31(3): 381–389. Doi: http://dx.doi.org/10.3233/SJI-150909.

United Nations Economic Commission for Europe (UNECE). 2011. *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices*. United Nations Publication. www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf (accessed November 2017).

United States Census Bureau. 2017a. *2020 Census Operational Plan: Version 3.0*. Washington DC: Census Bureau. Available at: http://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-oper-plan3.pdf (accessed November 2017).

United States Census Bureau. 2017b. "Administrative Records Modeling Update for the Census Scientific Advisory Committee." Presented at the Census Scientific Advisory Committee Meeting–March, 30, 2017. Available at: https://www2.census.gov/cac/sac/meetings/2017-03/admin-records-modeling.pdf (accessed November 2017).

United States Office of Management and Budget. 2014. *M-14-06: Guidance for Providing and Using Administrative Data for Statistical Purposes*. Available at: https://obama-whitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf (accessed November 2017).

van Zeijl, J. 2014. "From Traditional to Register-Based Censuses in the Netherlands." from the National Academies of Science: International Conference on Census Methods. Available at: https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_088800.pdf (accessed November 2017).

Wagner, D. and M. Layne. 2014. "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software." Center for Administrative Records Research and Applications Working Paper #2014-01. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-01.pdf (accessed November 2017).

Walejko, G., A. Keller, G. Dusch, and P.V. Miller. 2014. "2020 Research and Testing: 2013 Census Test Assessment." U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/programs-surveys/decennial/2020-census/2013_Census_Test_Assessment_Final.pdf (accessed November 2017).

Walker, S., S. Winder, G. Jackson, and S. Heimel. 2012. "2010 Census Nonresponse Followup Operations Assessment." 2010 Census Planning Memoranda Series, No. 190, April 30, 2012. Available at: https://www.census.gov/2010census/pdf/2010_Census_NRFU_Operations_Assessment.pdf (accessed November 2017).

Wallgren, A. and B. Wallgren. 2007. *Register-based Statistics: Administrative Data for Statistical Purposes*. New York: John Wiley and Sons.

Received March 2016
Revised November 2017
Accepted March 2018