Statistical Journal of the IAOS 32 (2016) 177–188
DOI 10.3233/SJI-161002
IOS Press

An approach for using administrative records to reduce contacts in the 2020 Decennial Census

Darcy Steeg Morris[a,*,1], Andrew Keller[b] and Brian Clark[b]
[a]*Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, USA*
[b]*Decennial Statistical Studies Division, U.S. Census Bureau, Washington, DC, USA*

Abstract. For the 2020 Decennial Census, the U.S. Census Bureau is researching how to use administrative record information from government and other sources in place of field visits during the Nonresponse Followup (NRFU) operation. This paper describes an approach for identifying vacant and occupied housing units to be enumerated using administrative records, removing them from the NRFU workload. While the approach allows flexibility in balancing cost and quality, we evaluate one possible scenario via a retrospective study of the 2010 Census.

Keywords: Administrative records, Decennial Census enumeration, nonresponse, statistical modeling

1. Introduction

A primary cost driver of the U.S. Decennial Census is the collection of data from households for which a self-response is not obtained. In the 2010 Census, the Nonresponse Followup (NRFU) operation included about fifty million addresses requiring up to six contacts each, totaling about $1.6 billion [1]. For purposes of planning the 2020 Decennial Census, the U.S. Census Bureau is researching ways to reduce costs of NRFU operations, without sacrificing data quality. One solution may be to use administrative records (AR) in lieu of personal visits. Examples of government and commercial administrative records include sources from agencies and companies such as the Internal Revenue Service (IRS), Center for Medicare and Medicaid Services (CMS), U.S. Postal Service (USPS), and TARGUSinfo (Targus). Whether these administrative records are collected as part of administering a government program (e.g. Medicare) or for business purposes (e.g. Targus), they have at least one thing in common: each have a data point associating a person with an address at some point in time. Pooling this person-place information across sources, U.S. Census Bureau researchers have proposed statistical techniques to extract the best and most relevant information for Census enumeration [2–4]. To that end, this paper presents a model-based approach to determine which housing units to enumerate using such data sources (removing them from NRFU fieldwork) and which to enumerate using NRFU field operations.

The use of administrative records in population censuses is not a new concept either internationally or domestically [5,6]. In fact, the idea of administrative record use in the U.S. Decennial Census dates back to the 1980s [7], while the world's first fully register-based census was implemented in Denmark in 1981 [8]. Due to increasing operational costs and decreasing response rates, alternative sources of population information are of interest to national statistical organizations all over the world [9,10]. As modernizations of traditional censuses continue to be researched, the use of administrative records currently play a pri-

*Corresponding author: Darcy Steeg Morris, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, USA. E-mail: darcy.steeg.morris@census.gov.

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

mary role in fully register-based censuses in countries such as Denmark and the Netherlands [11,12], as well as in partial register-based censuses in countries such as Switzerland, Germany and Poland [13].

We present an approach for the U.S. Decennial Census that seeks to maintain a traditional census that incorporates administrative records in lieu of traditional census fieldwork where appropriate. The administrative record sources available for Decennial Census research and implementation do not fully cover the population by their very nature [14]. For this reason, it is crucial to identify the subpopulations for which administrative records are a viable option. The proposed approach for administrative record use in the U.S. Decennial Census uses predictive models as a screening tool to identify housing units for which we can elicit a sufficiently reliable household roster or a vacancy determination using administrative records. We aim to address the following research questions:

- How can we identify housing units as vacant from administrative records (and remove from the NRFU workload)?
- How can we identify occupied housing units for which we can reasonably rely on administrative records for household count and composition (and remove from the NRFU workload)?
- How would the results from the proposed use of administrative records applied to 2010 Census data compare to the observed 2010 Census data?

2. Data

This research uses two general types of data: administrative records data and Census data. The 2010 Census data is used as a benchmark to evaluate the proposed administrative records methodology, while the administrative records data is a compilation of available federal and commercial data sources to be evaluated.

We define the administrative records composite data as all person-place pairs present in any of the selected administrative records sources at a 2010 Census NRFU address. Four sources are used to define the administrative records composite data: Internal Revenue Service (IRS) Individual Tax Returns (1040) filed for tax year 2009 between weeks 4 and 17 of 2010, IRS Informational Returns (1099) for tax year 2009, the Indian Health Service Patient Database, and the Center for Medicare and Medicaid Services (CMS) Medicare Enrollment Database. The administrative records composite data forms the universe of persons and addresses eligible for administrative record enumeration. This yields a person-level dataset which includes information about the presence/absence of administrative records sources for each of the person-place pairs. This data is used to fit one of our models.

We create an aggregated housing unit-level administrative records dataset that defines a household roster, and consequently an administrative records household count and composition (number of adults and children), as all person records associated with a given address in the administrative records composite data. The administrative records roster is defined as the union of all persons present in any of the administrative records files whose identifying information can be validated to create a unique person identifier. This yields a housing unit-level administrative records dataset which includes information on administrative records housing unit count, household composition and the presence/absence of administrative records sources in a housing unit. This data is used to fit two of our models.

Information from the commercial entity TARGUSinfo – a company that provides services such as person verification – is used to inform models but not used to determine the administrative records household roster (i.e. person-place combinations that occur only in Targus Federal Consumer file do not contribute to defining the administrative records composite universe). In addition, we incorporate data from the United States Postal Service (USPS) Delivery Sequence File, namely USPS undeliverable-as-addressed (UAA) reason codes in our models.

The 2010 Census unedited file of NRFU addresses is used as the comparison on which to evaluate the administrative records data. Field responses are treated as "truth" for purposes of comparison. Using 2010 Census is the natural choice as a benchmark, but not the only choice, and future research will involve comparing administrative records to other definitions of "truth". In addition to 2010 Census data, we use information from the American Community Survey (ACS), the Master Address File and Census operational data in our models.

3. Methodology

3.1. Identifying vacant units using administrative records

3.1.1. Model

We have developed a model of housing unit status (occupied, vacant or delete) for purposes of identifying

vacant housing units to be removed from the NRFU workload. The dependent variable of interest is:

$$y_h^{unocc} = \begin{cases} 1 & \text{if housing unit } h \text{ is occupied in the 2010 Census} \\ 2 & \text{if housing unit } h \text{ is vacant in the 2010 Census} \\ 3 & \text{if housing unit } h \text{ is not a housing unit (delete) in the 2010 Census} \end{cases} \quad (1)$$

where the *unocc* superscript denotes the model for determining vacant housing units (i.e. the unoccupied model). We are interested in a predictive model for estimating the probability of each housing status type in the 2010 Census: $\hat{p}_h^{occ} = P(y_h^{unocc} = 1)$, $\hat{p}_h^{vac} = P(y_h^{unocc} = 2)$, and $\hat{p}_h^{del} = P(y_h^{unocc} = 3)$. These probabilities are estimated via a multinomial logistic regression model.

The association between administrative record information and 2010 Census housing unit status is captured by including administrative record information as predictors in this model. We do this via three types of predictors: indicators of the presence of each of the administrative record sources at the housing unit (*here* variables), indicators of the presence of each of the administrative record sources at a different housing unit for a person in the housing unit (*elsewhere* variables), and characteristics of persons in the housing unit as determined by administrative records. For example, in a two-person administrative records housing unit, if at least one of the people had an IRS 1040 filed at that address then the IRS 1040 *here* indicator variable is equal to one. On the other hand, if at least one person had an IRS 1040 form filed at a different address, then the IRS 1040 *elsewhere* indicator variable is equal one. These types of variables are created for five administrative records sources: IRS 1040, IRS 1099, IHS, Medicare, and Targus. Another important administrative record source incorporated into this unoccupied model is USPS undeliverable as addressed (UAA) information. UAAs are flagged by the USPS and assigned a reason code for failed delivery which include reasons such as vacant, no such number, unable to forward, etc. The model also includes as predictors block group characteristics from the ACS planning database and housing unit characteristics from the Master Address File (MAF). See Table 4 in the Appendix for the full set of predictors used in the unoccupied model.

Each housing unit has an associated predicted probability for each of the three housing status categories. These predicted probabilities are passed to the optimization procedure as an input for determining which units are likely to be vacant.

3.1.2. Optimization procedure

We employ a constrained optimization procedure to incorporate information from the housing unit status model in a way that controls misclassification error. The objective is to maximize the removal of vacant units from the NRFU workload subject to some tolerance on the level of the error of misclassification. This linear programming approach allows us to incorporate multiple constraints on the predicted probabilities estimated from the unoccupied model. Specifically, we simultaneously apply two constraints that require that the average predicted probability of vacant status to exceed some threshold and that the sum of the occupied predicted probabilities of identified units not exceed a certain percentage of the estimate of occupied housing units from the ACS. These constraints are used to control the amount of misclassification of occupied or delete units as vacant by requiring high predicted probabilities of vacant status and low predicted probabilities of occupied status.

Let N be the number of NRFU housing units, X_j be an indicator equal to one if housing unit j is identified as vacant and removed from workload, N_a be the number of housing units in area a, N_a^{occ} be the number of occupied units in area a, and T^{unocc1} and T^{unocc2} be chosen thresholds. The optimization procedure seeks to:

$$\begin{aligned} &\text{maximize } U = \sum_{j=1}^{N} X_j \\ &\text{subject to } \frac{1}{U} \sum_{j=1}^{N} X_j \, \hat{p}_j^{vac} \geqslant T^{unocc1}, \\ &\frac{1}{N_a^{occ}} \sum_{j=1}^{N_a} X_{j,a} \, \hat{p}_{j,a}^{occ} \leqslant T^{unocc2} \; \forall \; a. \end{aligned} \quad (2)$$

The utility of incorporating multiple constraints on predicted probabilities from the unoccupied model can be best illustrated via a hypothetical example. Suppose two NRFU housing units with the following predicted probabilities estimated from the unoccupied model:

$$\hat{p}_1^{vac} = 0.80, \hat{p}_1^{occ} = 0.01, \hat{p}_1^{del} = 0.19$$
$$\hat{p}_2^{vac} = 0.80, \hat{p}_2^{occ} = 0.19, \hat{p}_2^{del} = 0.01$$

Both housing units have a $\hat{p}^{vac} = 0.80$, however housing unit one is more likely to be a delete than occupied ($\hat{p}_1^{del} = 0.19 > 0.01 = \hat{p}_1^{occ}$) and housing unit two is more likely to be occupied than a delete ($\hat{p}_2^{del} = 0.01 < 0.19 = \hat{p}_2^{occ}$). While both may be removed as vacant via the linear programming approach,

housing unit two is less likely to be removed because its error is assumed the be more costly. That is, we would rather assign a delete housing unit as vacant than an occupied housing unit as vacant. In short, we want to discriminate between those that are more likely to be delete if they are not vacant and those that are more likely to be occupied if they are not vacant.

3.2. Identifying occupied housing units for AR enumeration

3.2.1. Models

We have developed two separate models for capturing the "quality" of administrative records for purposes of identifying addresses for which we can elicit a sufficiently reliable enumeration from administrative records. The intent of these models is to estimate the confidence in the accuracy of each housing unit's administrative records household roster, not simply the presence/absence of people at the address. The models and methods for determining administrative records quality for enumeration purposes is used to inform the direct use of administrative records in lieu of personal visits. For this reason, we seek to identify occupied housing units for which we can reasonably rely on the administrative records to determine household count and household composition. Note that this is a different research question and requires a different approach than that of developing a traditional count imputation model. Models are fit to a 1% sample and applied to all NRFU housing units.

Person-place model

Census Bureau researchers have proposed creating and analyzing a composite dataset for identifying housing units with administrative records of suitable quality for enumeration purposes. The compilation of person-place pairs in administrative record files from federal sources are matched to 2010 Census person-place pairs to define the dependent variable of interest in the person-place model:

$$y_{ih}^{occ1} = \begin{cases} 1 \text{ if (person } i\text{, place } h\text{) pair from} \\ \quad \text{AR found in 2010 Census} \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

where the $occ1$ superscript denotes the person-place model for determining occupied units for AR enumeration. We are interested in a predictive model for estimating the probability, $p_{ih}^{occ1} = P(y_{ih}^{occ1} = 1)$, that the 2010 Census and the administrative records composite data place the person at the same address. These probabilities are estimated via a logistic regression model[1].

In addition to the collection of person-place combinations, the administrative records composite data contains information as to the characteristics of those combinations, such as, which sources had those person-place combinations and how many sources had those combinations. The model includes as predictors this information about the source of the person-place combination: for each person-place pair, indicators of the presence of the person-place pair in each of the administrative record sources (*here* variables), and the presence of the person in the person-place pair at a different place in each of the administrative records (*elsewhere* variables) are included as predictors. For example, if the person filed an IRS 1040 at that address then the IRS 1040 *here* indicator variable is equal to one. However, if the person also filed an IRS 1040 form at a different address, then the IRS 1040 *elsewhere* indicator variable is equal to one. These types of variables are created for five administrative records sources: IRS 1040, IRS 1099, IHS, Medicare, and Targus. Administrative record household roster information is also incorporated as predictors in the model, specifically, count and household composition. The model also includes as predictors block group characteristics from the ACS planning database, housing unit characteristics from the Master Address File (MAF), and USPS UAA information. See Table 4 in the Appendix for the full set of predictors used in the 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 aggregated within an address so that the housing unit-level predicted probability for address h is defined as:

$$\hat{p}_{h}^{occ1} = \min(\hat{p}_{1h}^{occ1}, \ldots, \hat{p}_{n_h h}^{occ1}) \tag{4}$$

where n_h is the number of people at address h. This minimum criteria assigns to the housing unit the predicted probability for the person in the housing unit for which we have the lowest confidence – a relatively conservative approach. Because the administrative records household roster for enumeration is defined as the union of all individuals associated with the address in the administrative records composite data, taking the minimum provides some protection against erroneously enumerating people in otherwise high

[1] Morris [3] finds that logistic regression and machine learning techniques (classification trees and random forests) exhibit similar predictive power for this person-place model.

match probability households. The potential impact of these assumptions suggest future research about alternative ways to define the administrative household roster, including approaches that remove persons – for which we have low confidence – from an address.

Each address has an associated predicted probability of having good "quality" person-place matches between the administrative records composite data and the 2010 Census data. These are the predicted probabilities that are passed to the optimization procedure as an input for determining occupied units for which we can reasonably rely on the administrative records data for household count and composition.

Household composition model

The results from the 2014 Census Test motivated the development of the household composition model described here. Specifically, we observed that units identified as occupied via a rules-based administrative records enumeration approach [15] were more likely to be occupied in NRFU if the household composition of the administrative records unit was a single adult, a two-person adult unit without children, or a two-person adult unit with children. This association can be captured via a housing unit level model with 2010 Census household composition as the dependent variable:

$$y_h^{occ2} = \begin{cases} 0 & \text{if housing unit } h \text{ has } 0 \text{ occupants in the 2010 Census} \\ 1 & \text{if housing unit } h \text{ has 1 adult, 0 children in the 2010 Census} \\ 2 & \text{if housing unit } h \text{ has 1 adult, } >0 \text{ children in the 2010 Census} \\ 3 & \text{if housing unit } h \text{ has 2 adults, 0 children in the 2010 Census} \\ 4 & \text{if housing unit } h \text{ has 2 adults, } >0 \text{ children in the 2010 Census} \\ 5 & \text{if housing unit } h \text{ has 3 adults, 0 children in the 2010 Census} \\ 6 & \text{if housing unit } h \text{ has 3 adults, } >0 \text{ children in the 2010 Census} \\ 10 & \text{otherwise} \end{cases} \quad (5)$$

where the $occ2$ superscript denotes the household composition model for determining occupied housing units for AR enumeration. We are interested in a predictive model for estimating the probability of each household composition type in the 2010 Census, $p_{h,k}^{occ2} = P(y_h^{occ2} = k)$ for $k = 0, \ldots, 6, 10$. These probabilities are estimated via a multinomial logistic regression model.

The association between administrative record and 2010 Census household composition type is captured by including administrative records household type (count and composition) as a predictor in this model. The model also includes as predictors information about the presence/absence of each particular administrative record source for the housing unit – these are generally the same administrative record source variables that are included in the unoccupied model, which is also fit at the housing unit-level. The model also includes as predictors block group characteristics from the ACS planning database, housing unit characteristics from the Master Address File (MAF), and USPS UAA information. See Table 4 in the Appendix for the full set of predictors used in the household composition model.

Each housing unit has an associated predicted probability for each category of household composition, $\hat{p}_{h,k}^{occ2}$. The predicted probability corresponding to the observed administrative record household composition type, $\hat{p}_h^{occ2} = \hat{p}_{h,k^*}^{occ2}$ where $k^* =$ AR household composition, is passed to the optimization procedure as an input for determining occupied units for which we can reasonably rely on the administrative records data for household count and composition.

To summarize, an overview of the characteristics of the two models are presented in Table 1.

Optimization procedure

We employ a constrained optimization procedure to jointly incorporate information from both models as well as "lessons learned" from previous Census tests. The objective is to maximize the use of administrative records for identifying occupied units subject to some tolerance on the level of "quality" of the administrative records used in lieu of personal visits. The linear programming approach allows us to incorporate constraints on the average predicted probability of identified units for both occupied models as well as other housing unit level constraints. Specifically, we use housing unit-level constraints that require: a maximum household size allocated from administrative records of six people (the person limit in the IRS 1040 data) and a pre-specified set of types of households determined by administrative records (1–3 adults with or without children: types found to be associated with good match rates in the 2014 Census test). The optimization procedure is run over the universe of housing units for which we have an associated administrative record.

Table 1
Summary of occupied model definitions

Level of analysis	Universe	Dependent variable	Probability model
Person	All person-place combinations in AR	Address match between AR and 2010 Census for the person	Logit
Housing unit	All 2010 NRFU addresses	2010 Census household composition	Multinomial logit

Let N^* be the number of housing units for which we have an associated administrative record, X_j^* be an indicator equal to one if administrative records enumeration is used for housing unit j, T^{occ1} be a chosen threshold for the person-place model, and T^{occ2} be a chosen threshold for the household composition model. The optimization procedure seeks to:

$$\begin{aligned}
&\text{maximize} \quad U^* = \sum_{j=1}^{N^*} X_j^* \\
&\text{subject to} \quad \frac{1}{U^*} \sum_{j=1}^{N^*} X_j^* \, \hat{p}_j^{occ1} \geqslant T^{occ1}, \\
&\qquad\qquad \frac{1}{U^*} \sum_{j=1}^{N^*} X_j^* \, \hat{p}_j^{occ2} \geqslant T^{occ2}, \\
&\qquad\qquad (\text{AR household size})_j \leqslant 6 \; \forall \; j, \\
&\qquad\qquad (\text{AR household composition})_j \\
&\qquad\qquad \in (1,2,3,4,5,6) \; \forall \; j.
\end{aligned} \tag{6}$$

The utility of incorporating constraints on predicted probabilities from both models can be best illustrated via a hypothetical example. Suppose Census field operations enumerate a two adult/one child housing unit; while administrative records place two adults and two children at this address. Such a housing unit satisfies the constraints on administrative record household size and administrative record household composition type. Next, suppose that the household composition model indicates that this unit has a high predicted probability of being a two adult with children household in the Census. Furthermore, suppose that the person-place model shows that one of the children has a low predicted probability of being at that particular address in the Census. Using only the information from the household composition model would likely lead to removing this case from NRFU workload and assigning a household count of four. However, including the information from the person-place model may cause the case not to be used for administrative records enumeration since the quality of one of the person records is in question. In short, incorporating information via a linear programming procedure allows for a type of consistency check across various models.

4. Evaluation: Retrospective study of 2010 Census

The proposed approach for using administrative records to reduce contacts in the 2020 Census is evaluated via a retrospective study of the 2010 Census. We apply the proposed approach by fitting the models and running the optimization procedure described in Section 3, by state, over all 50 million 2010 NRFU housing units. All NRFU housing units are eligible to be identified as vacant units via administrative records – and thus removed from the NRFU workload; but only about 54% of the NRFU housing units – those with associated administrative records – are eligible to be enumerated via administrative records.

We present two categories of evaluation criteria: housing unit level comparisons and aggregated comparisons. Housing unit level evaluation criteria can be used to assess accuracy on a case-by-case basis, while aggregated criteria can be used to assess accuracy at various geographic and/or demographic level. The housing unit-level metric for the unoccupied model is 2010 Census housing unit status distribution for vacant removals. The housing unit-level metrics for the occupied models are: household composition match between administrative records and 2010 Census (including and excluding proxy responses), household count match between administrative records and 2010 Census (including and excluding proxy responses), household count match within one person (including and excluding proxy responses), and 2010 Census housing unit status. At the national level we report the workload reduction (the percent of units with housing status and enumeration determined via administrative records) and the population coverage ratio (the count of total administrative records persons divided by the count of 2010 Census persons). In this paper, we only provide results for the aggregated metrics at the national level. Future research will look at these metrics at lower levels of geography and by demographic groups.

4.1. Thresholds for optimization procedures

The optimization procedures used for identifying vacant and occupied NRFU housing units to be removed from the NRFU workload are crucially dependent on the choice of threshold parameters. Setting up the problem in this manner allows flexibility in the approach. For illustrative purposes, we present results for one scenario where the thresholds are determined as discussed in this section.

4.1.1. Thresholds for identifying vacant housing units

The thresholds for the unoccupied model directly influence the cost-quality balance of the usage of administrative records for removing vacant units from the NRFU workload. As we decrease T^{unocc1} we increasingly remove housing units that are less likely to be vacant according to the model (i.e. those we are less confident in). As we increase T^{unocc2} we are loosening the restriction that keeps the associated predicted probability of being occupied low. The constraints work together to impose both an average vacant probability restriction in addition to requiring that the ratio of the sum of occupied probabilities with total occupied units be below some threshold. We propose $T^{unocc2} \in (0.002, 0.005, 0.01)$ as possible threshold for the second constraint and have studied the effects of varying T^{unocc1} between 0.75 and 0.95 [16].

4.1.2. Thresholds for identifying occupied housing units for AR enumeration

The thresholds for the two occupied models directly influence the cost-quality balance of the administrative records usage approach. Figure 1 depicts this trade-off. Figure 1 plots the quality metric on the y-axis (count match, household composition match and coverage ratio) against workload removal on the x-axis. Each line represents a different setting of the thresholds. The dark solid lines trace out the effect of changing the threshold on the person-place model (T^{occ1}) with no restriction on the predicted probabilities from the household composition model ($T^{occ2} = 0$). Conversely, the dark dashed lines trace out the effect of changing the threshold on the household composition model (T^{occ1}) with no restriction on the predicted probabilities from the person-place model ($T^{occ1} = 0$). The remaining lines trace out the effect of changing the threshold on the person-place model (T^{occ1}) with different restrictions on the predicted probabilities from the household composition model ($T^{occ2} = 0.45, 0.50, 0.55, 0.60, 0.65$).

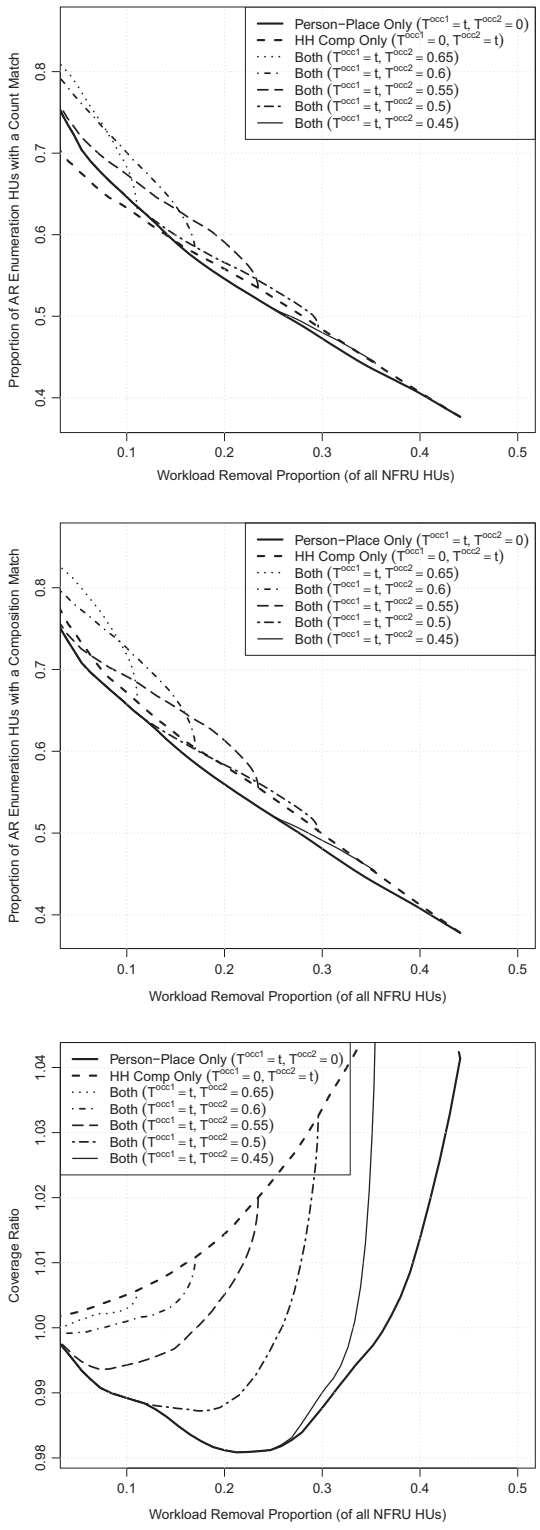

Fig. 1. Cost-quality curves for occupied model threshold determination.

Table 2
Analysis for threshold determination on 1% sample of NRFU housing units

Workload removal target	Metric to optimize	T^{occ1}	T^{occ2}	Count match (no proxy)	HH comp Match (no proxy)	Coverage ratio
20%	Count match	0.62	0.55	0.651	0.688	1.005
15%		0.68	0.57	0.696	0.734	1.001
10%		0.73	0.60	0.751	0.789	1.001
20%	HH comp match	0.62	0.55	0.651	0.688	1.005
15%		0.66	0.59	0.698	0.739	1.003
10%		0.73	0.60	0.751	0.789	1.001
20%	Coverage	0.64	0.53	0.637	0.670	0.998
15%	ratio	0.68	0.56	0.687	0.723	1.000
10%	(all NRFU)	0.73	0.59	0.744	0.781	1.001

Table 3
Evaluation criteria for selected scenario

Universe	Evaluation metric	Result
AR vacant housing units	% Occupied	8.5
	% Vacant	78.8
	% Delete	11.8
	% Unresolved	0.9
	Workload reduction (%)	10.8
AR occupied housing units	% Occupied	90.2
	% Count match	64.6
	% Count match ± 1	89.2
	% HH comp. match	67.0
	% Count match (no proxy)	69.9
	% Count match ± 1 (no proxy)	91.7
	% HH comp. match (no proxy)	73.6
	Population coverage ratio	1.005
	Workload reduction (%)	14.6

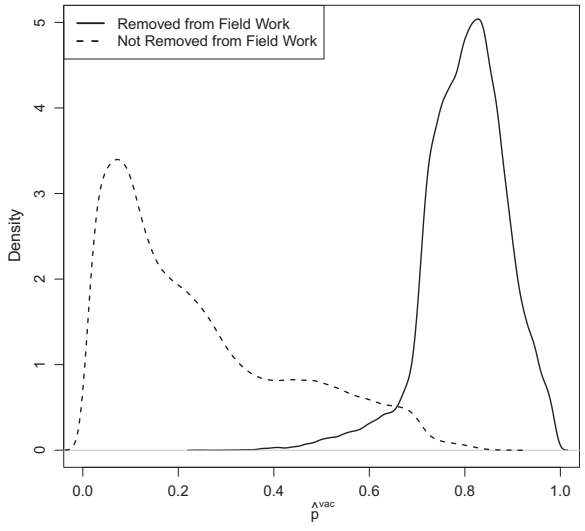

Fig. 2. Distribution of housing unit predicted probabilities by administrative record usage for unoccupied model in selected scenario.

There are three main takeaways from these plots. First, Fig. 1 shows that using the information from both of the models achieves better quality at the same cost. For example, at a 10% workload reduction, the person-place model alone and household composition model alone result in a count match rate of about 0.65 and 0.64, respectively. However, using a combination of the two occupied models results in a count match rate of about 0.70 for the same amount of workload removal (10%). Second, Fig. 1 clearly shows that as more housing units are removed from the NRFU workload (the thresholds are set lower) and enumerated using administrative records, the housing unit count match rate and household composition match rate between the administrative records and the 2010 Census decrease, while the coverage ratio generally diverges from 1.00. Lastly, the cost-quality curves can be used to determine discrete points for possible removal scenarios. Table 2 presents six such scenarios determined by targeting a particular workload removal (i.e. cost) while optimizing particular quality metrics. These scenarios are dictated by choosing T^{occ2} as the highest curve on the cost-quality plots and then finding the corresponding T^{occ1} associated with the workload removal target. For example, with a goal of maximizing the count match rate between administrative records

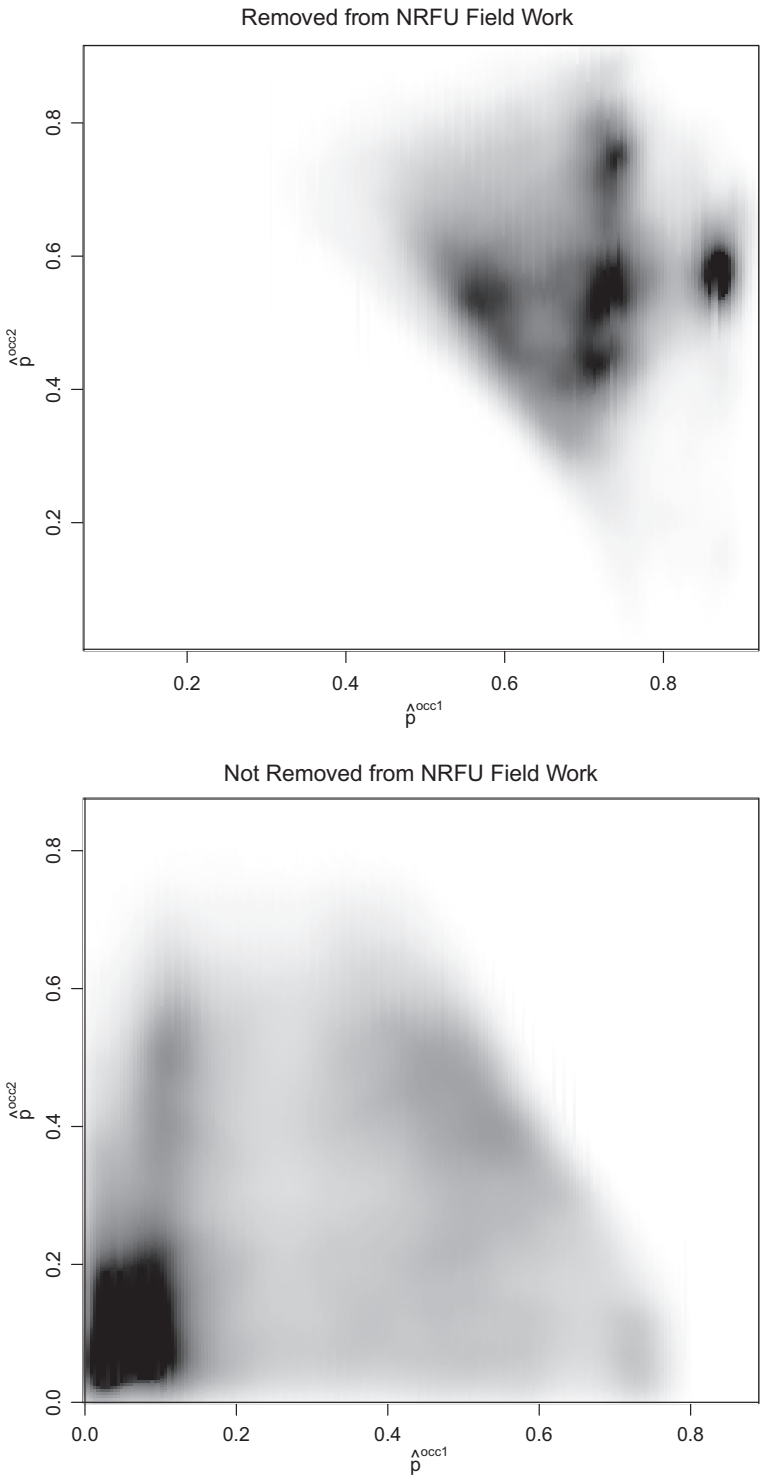

Fig. 3. Bivariate distributions of housing unit predicted probabilities by administrative record usage for occupied models in selected scenario. Darker shading is higher density.

and the 2010 Census and targeting a 20% workload removal, the curve associated with $T^{occ2} = 0.55$ is the highest and the corresponding point for the person-place threshold is $T^{occ1} = 0.62$. This is the first line of Table 2.

4.2. Results

We present an evaluation of one scenario of administrative records usage to reduce contains in the 2020 Census. We set the threshold parameters for the occupied removal in the manner described previously. Specifically, we target a 15% workload removal with a high count match agreement: this corresponds to $T^{occ1} = 0.68$ and $T^{occ2} = 0.57$ (see Table 2, row 2). Threshold parameters for the unoccupied model are set as $T^{unocc1} = 0.80$ and $T^{unocc2} = 0.005$. We find that in this scenario, the procedure for removing vacant housing units via administrative records reduces the NRFU case load by 10.8% and the procedure for removing occupied housing units via administrative records reduces the NRFU case load by 14.6%; see Table 3. While about 79% of the housing units removed as vacant via administrative records are indeed vacant, over 90% are unoccupied (vacant or delete). Similarly, about 90% of the housing units removed to be enumerated via administrative records as occupied are indeed occupied. About 65% have a housing unit count match and about 67% have a household composition match. These numbers increase to about 70% and 74%, respectively, when proxy responses are excluded from the comparison. About 90% of the removed housing units have a count match within one person. On the aggregate, these errors balance out to a population coverage ratio of about 1.005.

To get a sense of the housing units that were removed from NRFU workload in this scenario, we can look at the distribution of the predicted probabilities for those units removed from workload via administrative records usage and those housing units that remain for field work. Figure 2 displays the kernel density of the predicted probability of vacancy, $\hat{p}_h^{vac}$, by administrative record usage (vacant removal). We see that those determined to be vacant via the constrained optimization procedure and thus removed from NRFU fieldwork have a distribution that is clearly shifted higher than the distribution of the housing units that remained in NRFU fieldwork. There is some overlap in the distributions due to the second constraint in the unoccupied procedure which penalizes housing units which are more likely than others to be occupied as opposed to delete despite their potentially large predicted probability of being vacant; however this overlap is not very large. Figure 3 displays the estimated bivariate distribution of the predicted probabilities from the two occupied models (with the person-place model probability $\hat{p}_h^{occ1}$ on the x-axis, and the household composition model probability $\hat{p}_h^{occ2}$ on the y-axis). The constrained optimization procedure for determining occupied units incorporated four constraints – these figures provide insight as to how those constraints work together to determine the housing units to be enumerated via administrative records. For those housing units removed from fieldwork, we see a distinct linear boundary illustrating the interplay between the two sets of predicted probabilities. For example, the procedure allows a housing unit with a low $\hat{p}_h^{occ2}$, say 0.3, to be removed and enumerated as occupied if the corresponding $\hat{p}_h^{occ1}$ is large enough, say 0.7. This is the benefit of incorporating two models of administrative record "quality" into the procedure: even if one model deems the housing unit not quite good enough, the housing unit may be removed for administrative records usage if the other model favors it strongly enough.

5. Conclusion

This paper presents an approach for identifying vacant and occupied housing units to be enumerated using administrative records in the Decennial Census, removing them from the NRFU workload. We use a constrained optimization approach that incorporates predicted probabilities from various models as well as "lessons learned" from Census tests. This approach allows flexibility in two key ways: constraints can easily be incorporated into the procedure as new information is learned, and threshold parameters used in the constraints can be chosen to achieve a desired balance between cost and quality. It is crucial to emphasize that the level of removal and corresponding quality is determined explicitly by setting the parameters in the optimization routines; T^{unocc1} and T^{unocc2} for removal of vacant units, and T^{occ1} and T^{occ2} for removal of occupied units. This flexibility allows a continuous range of possible scenarios that the decision-maker can control. We present results of one scenario of threshold parameter settings to illustrate the performance of this approach for using administrative records as compared to 2010 Census results.

The caveats of this research dictate an interesting and important future research agenda. The 2010 Cen-

sus is a natural comparison for evaluating the "quality" of administrative records, but it is necessary to evaluate any approach for using administrative records on other versions of "truth" – for example, American Community Survey (ACS) and Census Coverage Measurement (CCM) data. We present results from one possible scenario for using administrative records, but a complete cost-benefit analysis of the effect of this alternative data collection strategy is warranted. All analysis presented in this paper is at the national level. The linear programming procedure was implemented by state using results from models that were fit nationally. It will be important to assess differences when incorporating geographic information in the models and/or running the linear programming procedure by finer geographies and/or setting thresholds differently by geography and/or evaluating the approach using metrics at finer levels of geography.

References

[1] S. Walker, S. Winder, G. Jackson and S. Heimel, 2010 Census Nonresponse Followup Operations Assessment. 2010 Census Planning Memoranda Series, no. 190; 2012. Available from: https://www.census.gov/2010census/pdf/2010_Census_NRFU_Operations_Assessment.pdf.

[2] K.M. Shaw and J. Boies, Nonresponse Followup Modeling and Microsimulation: Examining Cost-Benefit Tradeoffs for 2020, in JSM Proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association, 2013.

[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, pp. 1729–1743.

[4] J.D. Brown, J.H. Childs and A. O'Hara, Using the Census to Evaluate Administrative Records and Vice Versa, in Proceedings of the 2015 Federal Committee on Statistical Methodology (FCSM) Research Conference. Washington, DC: FCSM, 2015.

[5] F. Scheuren, Administrative Records and Census Taking, *Survey Methodology* **25**(2) (1999), 151–160.

[6] D.L. Steffey and N. Bradburn, *Counting People in the Information Age.* Washington, DC: National Academy Press, 1994.

[7] W. Alvey and F. Scheuren, Background for an Administrative Record Census, in JSM Proceedings, Social Statistics Section. Washington, DC: American Statistical Association, 1982, pp. 137–152.

[8] A. Lange, The Population and Housing Census in a Register Based Statistical System, in the Proceedings of the 59th World Statistics Congress of the International Statistical Insitute. The Hague, Netherlands: International Statistical Institute, 2013, pp. 2315–2320.

[9] B.F.M. Bakker, P.G.M. van Heijden and S. Scholtus, "Preface" to a Special Issue on Coverage Problems in Administrative Sources, *Journal of Official Statistics* **31**(3) (2015), 349–355.

[10] S.E. Fienberg, "Discussion" of a Special Issue on Coverage Problems in Administrative Sources, *Journal of Official Statistics* **31**(3) (2015), 527–535.

[11] L. Thygesen, The Use of Administrative Sources for Censuses: Merits and Challenges, *Statistical Journal of the IAOS* **31**(3) (2015), 381–389.

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

[13] M. Maris, E.S. Nordholt and J. van Zeijl, 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; 2012. Available from: (www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2012/use_of_register/WP_3_Netherlands.pdf).

[14] S. Rastogi and A. O'Hara, "2010 Census Match Study," 2010 Census Planning Memoranda Series, No. 247; 2012. Available from: (www.census.gov/2010census/pdf/2010_Census_Match_Study_Report.pdf).

[15] 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 Report, 2015.

[16] B. Clark, Evaluating Optimization Constraints when Maximizing Nonreponse Followup Workload Reduction, U.S. Census Bureau Report, 2015.

Appendix

Table 4
Model covariates for unoccupied and occupied models

			Occupied models	
	Variable	Unoccupied model	Person-place	HH comp.
	ACS block group level variables			
% of block group...	age 25–44, 65+	X	X	X
	black, hispanic	X	X	X
	related family	X	X	X
	other language	X	X	X
	mobile home	X	X	X
	married	X	X	X
	owner-occupied, rental	X	X	X
	vacant		X	X
	in poverty	X	X	X
	Housing unit characteristics			
	# of neighbors in NRFU	X		
	recent delivery sequence file information	X	X	X
	USPS UAA flag	X	X	X
	USPS UAA reason	X	X	
	housing unit type (e.g. multi-family)		X	X
	Housing unit characteristics from administrative records			
$\geqslant 1$ *person in HU is...*	white, black	X		X
	hispanic, missing ethnicity	X		
	age $< 9, 10-17, 65+$	X		X
	Housing unit level administrative record source information			
	AR HH count		X	X
	AR HH composition		X	X
$\geqslant 1$ *person in HU is placed at this HU according to...*	IRS 1040 TY 2009	X		X
	IRS 1099	X		X
	IHS	X		X
	Medicare	X		X
	Targus	X		X
	IRS 1040 TY 2008			X
$\geqslant 1$ *person in HU is placed in another HU according to...*	IRS 1040	X		X
	IRS 1099	X		X
	IHS	X		X
	Medicare	X		X
	Targus	X		X
	Person level administrative record source information			
Person is placed in this HU according to...	IRS 1040 TY 2009		X	
	IRS 1099		X	
	IHS		X	
	Medicare		X	
	Targus		X	
	IRS 1040 TY 2008		X	
Person is placed at another HU according to...	IRS 1040		X	
	IRS 1099		X	
	IHS		X	
	Medicare		X	
	Targus		X	