methods, data, analyses | 2024, pp. 1-22 DOI: 10.12758/mda.2024.03

Combining Information from Multiple Data Sources to Improve Sampling Efficiency

Paul Burton, Sunghee Lee, Trivellore Raghunathan & Brady T. West

University of Michigan, Survey Research Center (SRC)

Abstract

Many surveys target population subgroups that may not be readily identified in sampling frames. In the case study that motivated this study, the target population was households with children between the ages of 3 and 10 from two areas surrounding Cleveland, Ohio and Dallas, Texas. A standard approach is to sample households from these two areas and then screen for the presence of age-eligible children. Based on the estimated number of age-eligible households in these two areas, this approach would have required completing screening interviews with 5.4 to 5.7 households to find one eligible household. We developed a model-assisted sample design strategy to improve screening efficiency by attaching a measure of eligibility propensity to each household in the population. For this, we used a modeling and imputation strategy that combined information from several data sources: (1) the population of addresses for these two areas with demographic covariates from a commercial vendor, (2) external population data (from the American Community Survey and Census Planning Data) for these two areas, and (3) screening data from a large nationally representative survey. We first tested this sampling strategy in a pilot study and then implemented it in the main study. This strategy required 4.2 to 4.3 completed screeners to identify one eligible household. The proposed approach therefore improved the sampling efficiency by about 25% relative to the standard approach.

Keywords: address-based sampling, imputation, rare populations, commercial data, census data, address frame

The Housing and Children Study (H&C) is an evaluation study of housing voucher programs provided by the United States Department of Housing and Urban Development (HUD), specifically about their effect on the environment and experiences of children ages 3 to 10 years old in Dallas, TX and Cleveland, OH. These voucher programs assist low-income families and are operationalized in municipalities (e.g., Dallas) by local HUD branches known as Public Housing Authorities (PHA). H&C was designed to be an in-person survey with two sample components: one with individuals that have applied for housing vouchers ("the voucher sample") and the other with members of the general population ("the population sample"). The voucher sample was drawn from a well-specified sampling frame by PHAs. This paper focuses exclusively on the methodologies used for designing the population sample. Appendix 1 includes a list of acronyms used in this paper along with their definitions.

The population sample was designed as an area-probability sample from the two areas and used income level as a stratification factor. The main goal was to develop strategies to increase the sampling efficiency by reducing the number of households to be screened to identify eligible families and, thereby, reducing the field cost. Eligible families had at least one child ages 3 to 10 years old. This age eligibility rate was estimated nationally at 18.4% based on the American Community Survey (ACS) 2010-2014 5-year public use microdata sample (PUMS) and 17.5% based on the National Survey of Family Growth (NSFG) Cycle 8. Based on these rates, we would be required to screen roughly 5.4 to 5.7 households to identify one age-eligible household.

Due to declining response rates and increasing costs for population-based surveys, survey researchers have started examining the utility of auxiliary data to mitigate such difficulties (Smith, 2011). Commercial databases, typically purchased from sample vendors, are an example of this type of auxiliary data. Developed for commercial purposes, these databases provide a rich set of information at the individual address level, which may allow survey researchers to consider these databases as a means for improving sampling efficiency and nonresponse bias adjustment (e.g., Buskirk et al., 2014; English et al., 2019; Harter et al., 2016; Pasek et al., 2014; West, 2013; West, Wagner, Hubbard, & Gu, 2015).

Direct correspondence to
Sunghee Lee, University of Michigan, Survey Research Center (SRC),
426 Thompson St. Ann Arbor, MI 48104, USA
E-mail: sungheel@umich.edu

Sampling Rare Population Subgroups Using Commercial Databases

When surveys target specific population subgroups that are rare or small in number, a non-trivial amount of resources is required for screening eligible cases. Under this type of sampling scenario, if commercial databases include information relevant to the characteristics of target subgroups, it can be appended to the sampling frame and used for stratification (Kalton, Kali, & Sigman, 2014; Valliant, Hubbard, Lee, & Chang, 2014). A wide range of information is available from commercial databases, from socio-demographics to product purchase behaviors, donations, and voting records, and the amount of information varies by vendors (see Tables A1 and A2 in West et al., 2015). Valliant et al. (2014) also demonstrated a stratified sampling approach for the Health and Retirement Study (HRS), which is a longitudinal survey that targets a specific age cohort every six years using area-probability sampling. In 2016, HRS targeted households whose oldest member was born between 1960 and 1965 with an additional goal of oversampling ethnic and racial minorities. The sampling combined stratification at two levels: (1) stratification of geographic segments using aggregate level information from the decennial Census and ACS; and (2) stratification of addresses based on age and race/ethnicity information about people at the address obtained from commercial databases. With a disproportionate allocation, their design achieved cost savings under a variety of constraints. Similar gains in sampling efficiency were also demonstrated for a telephone survey, the National Immunization Survey (Barron et al., 2015)we assume that information is available at the sampling stage to stratify the general-population sampling frame into high-and low-density strata. Under a fixed constraint on the variance of the estimator of the domain mean, we make the optimum allocation of sample size to the several strata and show that, in comparison to proportional allocation, the optimum allocation requires (a, where landline telephone numbers were stratified by matched commercial data, enabling the targeting of households with a minor member.

Practical Limitations in Using Commercial Databases for Sampling

There are three issues with utilizing commercial databases for sampling rare population subgroups. First, not all sample addresses (or telephone numbers) may be matched to commercial data (Valliant et al., 2014), with matching rates potentially varying by vendors (West et al., 2015). Second, for the addresses successfully matched with commercial data, variables in the commercial data vary in terms of their missing rates, and this also varies by vendor (West et al., 2015). The third problem is the quality of the information in the commercial databases. The agree-

ment rates between self-reported survey data and commercial data examined. For example, in a study that matched the 2010 U.S. decennial Census with commercial data, Rastogi and O'Hara (2012, Tables 23 and 24) showed varying agreement rates not only by vendors but also by characteristics. For example, on race/ethnicity, the agreement rates between the Census and commercial data was higher for Whites than for minority groups. The rate was around or above 95% for Whites but was around or below 10% for American Indian or Alaska Natives. Moreover, there are no standardized racial/ethnic categories across the commercial data vendors.

In sum, the third issue above deals with data accuracy, and the first two with data availability or completeness. Information incompleteness is directly a missing data issue, which has been discussed as a major limitation of using commercial data for sampling (Kalton et al., 2014; Roth, Han, & Montaquila, 2013), although a recent study reports some improvement (Roth et al. 2018). In addition to the varying missing rates across variables within a database, the missingness in the commercial databases itself appears not necessarily at random. For example, home ownership in the commercial databases is less likely to be missing among home owners than non-owners (Pasek et al., 2014, Table 3).

Imputing Missing Data in Commercial Databases

To maximize potential benefits of the existing commercial data while mitigating the practical limitations of missing data and poor accuracy, this study proposes a new method of using commercial data for sampling rare population subgroups by imputing missing data and using eligibility prediction models. These methods are then demonstrated via application to a case study. In the next section, we present the sampling design used for H&C, the imputation approaches applied to the commercial databases, and the sample design using predicted eligibility assisted by the imputed commercial data at the address level as well as external data aggregated at the geographic segment level. We then examine the accuracy and efficiency of the proposed method as observed in real fieldwork.

To meet the goal of improving screening efficiency on H&C, we used three data sources: (1) the population of addresses enhanced with commercial data for the sampled areas purchased from a vendor, (2) external population data (from ACS and Census Planning Data) for these two areas, and (3) screening data from a large nationally representative survey that includes information relevant to the eligibility in H&C. Using information from these three sources, we developed a two-stage sample design. The first stage involved sampling Census block groups (BGs), and the second stage then sampled addresses within the selected BGs using enhanced address lists. In both stages, we modelled and predicted eligibility using external data. For the first stage, we developed a model to estimate the number of households with at least one child between the ages of 3 and 10 years for each BG and

used this as the measure of size (MoS) in the selection of the BGs. For the second stage, we predicted the probability for having an age eligible child for each address in the selected BGs and used this predicted eligibility as the MoS.

We first implemented this design in a pilot study before refining the strategy for the main study. The next two sections describe the H&C pilot and main study. Within each section, sampling methods and results are presented.

Pilot Study

Sampling Frame

The pilot study was conducted in Dallas, TX, using a sampling frame that included a total of 998 BGs, covering 70.5% of the ZIP codes where potential voucher applicants resided.

Sample Design

The sample design leveraged multiple external data sources: (1) the ACS 2010-2014 5-year summary file (SF); (2) the 2016 Census planning data; (3) a commercial database purchased from Marketing Systems Group (MSG: http://www.m-s-g.com/); and (4) household roster data from the 2011-2015 National Survey of Family Growth (NSFG), an area-probability national sample survey conducted by the Centers for Disease Control and Prevention. It should be noted that NSFG and MSG data are available at the address/household level, while ACS SF and Census planning data are aggregated at various levels of geography as fine as BGs. The availability of NSFG roster data was crucial for the H&C design, because it provided precise data on H&C age eligibility used in both stages of sampling.

Two-stage sampling as illustrated in Table 1 was used to select the sample. In the primary stage, BGs were sampled using a stratified probability proportionate to estimated size (PPeS) design. In the secondary stage, addresses/households were selected from sampled BGs also using a stratified PPeS design. The detail for each stage is described below.

Table 1 Description of Overall Sample Design, Housing and Children Study

	Primary Stage	Secondary Stage
Sampling unit	Census block groups (BG)	Addresses/Households
Measure of Size	Number of households with at least one child aged between 3 and 10 years old	Probability of having at least one child aged between 3 and 10 years old
Estimation Method	Model-based prediction	Model-based prediction
Prediction Model	Grouped logit model with - DV: Household-level age eligibility indicator from the NSFG roster data aggregated to the BG level - IVs: BG-level auxiliary data (ACS SF and Census planning data with the dimensions reduced through principal components analysis)	Individual logit model with - DV: Household-level age eligibility indicator from the NSFG roster data - IVs: Address-level commercial data (with missing data treated through sequential multiple imputation) + BG-level auxiliary data (ACS SF data with the dimensions reduced through principal components analysis)
Prediction	Multiply the proportion of eligible households for each BG in the H&C frame, predicted by fitting the grouped logit model, with the number of households for each BG	Predict the probability of being age eligible for each address in BGs sampled from the primary stage by fitting the individual logit model
Stratification Variable and Method	Proportion of households with an annual income less than $35,000 directly available from ACS SF	Household income from commercial data - If not missing, exact income values from commercial data - If missing, imputed income from sequential multiple imputation

Note. DV: Dependent variable; IV: Independent variable; H&C: Housing and Children Study; NSFG: National Survey of Family Growth; ACS SF: American Community Survey Summary File

Primary Stage Design

The primary stage focused on selecting 15 BGs from 998 BGs on the H&C pilot frame using the BG-level number of households with at least one child aged between 3 and 10 years old as the MoS. Note that this MoS is not readily available from any of the external data. We estimated the MoS as follows using NSFG and ACS SF data at the BG level. First, we created a dataset by aggregating the household-level H&C age eligibility in the NSFG roster data to the BG level and appending 160 variables from ACS SF relevant for this age eligibility (see Supplementary Table 1 at https://goo.gl/co4SuZ). Second, for the goal of estimating the proportion of H&C eligible households at the BG level, we fitted a grouped logit model of aggregated eligibility by regressing the aggregated BG-level eligibility rates from NSFG on ACS SF variables. Instead of selecting individual variables from ACS for this model, we used principal component analysis (PCA) to reduce the dimensionality from 160 ACS variables while retaining a similar amount of information. With the PCA suggesting 63 components that explained 95% of the variance in the original 160 variables, we modelled the aggregated BG-level eligibility from NSFG on these 63 components as well as 155 two-way interactions identified from a stepwise variable selection process. This model included a total of 1,909 BGs in NSFG and provided fair fit with an area under the ROC curve of 0.66 and a non-significant Hosmer–Lemeshow goodness-of-fit test (χ^2=7.90, df=8; p=.443).

The estimated model was applied to the 988 BGs on the H&C pilot frame, from which the BG-level proportion of H&C eligible households was predicted. With the counts of total households available from ACS SF, the predicted proportions were multiplied by the household counts, yielding the MoS at the BG level. The minimum MoS was set at 10 eligible households. BGs smaller than the minimum MoS were combined within income strata as described shortly.

BGs on the frame were stratified using the proportion of "low income" households from ACS SF defined as those with annual income less than $35,000. Specifically, we used the tertiles of this distribution as cutoff points, designating BGs into three strata: low (>37.4%; i.e., more than 37.4% of the households in BG with income less than $35,000), middle (19.3-37.4%) and high income (<19.3%). With the overall project goal being to select BGs at the ratio of 3:2:1 from low-, middle- and high-income strata, the pilot study selected 8, 5, and 2 low-, middle- and high-income BGs with PPeS within strata.

Secondary Stage Design

The secondary stage dealt with selecting addresses from the 15 sampled BGs using the predicted probability of a given address being H&C eligible as the MoS, which allowed us to improve our ability to target likely eligible households. With this information not directly available, we leveraged four external datasets through a prediction model, similar to the primary stage design. First, we concatenated all

61,085 addresses in the NSFG roster data with their H&C eligibility indicator and all 10,304 addresses in the 15 BGs sampled for the pilot study from the USPS delivery sequence file. For H&C addresses, the eligibility indicator was naturally missing. To these data, we merged address-level MSG data (15 variables in Table 2) and BG-level ACS SF and Census planning data (483 variables in Supplementary Table 2 at https://goo.gl/ERGWvy). The idea was to model the household-level eligibility as a function of the MSG variables and ACS/Census variables. This required treatments of the missingness in the MSG data and the large dimensionality of the ACS/Census data.

The large dimensionality was handled with PCA, similar to the procedure used for the primary stage. A total of 483 ACS/Census variables was reduced to 113 components that retained 95% of the total variance. The missing rates of MSG variables considered in the pilot study were as low as 17.6% and as high as 83.9% as reported in Table 2. To mitigate this issue, we applied sequential imputation using multivariable regression models through the software package IVEware (Raghunathan, Berglund, & Solenberger, 2018). For numeric variables, ordinary least squares regression models were used; for binary variables, logit models; and for categorical variables, multinomial logit models. The baseline imputation model included the 113 components from the PCA as predictors. We used multiple imputation in order to assess model fit and ascertain uncertainty associated with the random error in the imputation models, which single imputation does not allow. Repeating the imputation 10 times offers sufficient information about this uncertainty (Raghunathan et al., 2018). Because imputed values for the missing cases varied only minimally across imputations, we used the average of 10 imputed values.

Logistic regression was used to model the eligibility of 61,085 addresses in the NSFG roster data with the ACS/Census principal components and imputed commercial data. Across 10 imputations, the model fit was comparable with an area under the ROC curve ranging around 0.71-0.72. The estimated model was applied to the 10,304 H&C addresses in order to predict their probability of being age eligible. The predicted eligibility was around 24-25%, and this result was similar across the 10 imputations as shown in Table 3. The average of the 10 predicted eligibility probabilities was used as the MoS.

H&C addresses were stratified by the income variable in the MSG data whose missingness was treated with imputation as described above. The income strata were formed based on the tertiles of this income distribution. Addresses with income <$30,000 were assigned to the low-income stratum, $30,000-62,500 to the middle-income stratum and >$62,500 to the high-income stratum. Considering the target ratio of 3:2:1 for these income strata, 684 addresses were selected using PPeS with predicted eligibility as the MoS within income stratum for the screening interviews conducted from October to December 2016.

Table 2 Missing Rates of Variables in MSG Data Used for Address-Level Eligibility Prediction, Housing and Children Study

Variable Description	Missing Rate	
	Pilot Study *(n=71,389)**	Main Study *(n=135,716)**
Age of Person 1 in household	48.4%	47.1%
Education of Person 1 in household	28.6%	26.7%
Ethnicity of Person 1 in household	28.6%	26.7%
Gender of Person 1 in household	17.6%	15.1%
Total household Income	17.6%	15.1%
Marital Status of Person 1 in Household	26.4%	27.2%
Flag for Asian Surname of Person 1 in Household	17.6%	15.1%
Flag for Hispanic Surname of Person 1 in Household	17.6%	15.1%
Flag for Name provided for Person 1 in Household	17.6%	15.1%
Number of Adults (18 years and older) in Household	83.9%	85.3%
Number of Children (Under Age 18) in Household	21.8%	24.6%
Does Householder Rent or Own the Household	21.8%	24.6%
Age of Person 2 in Household	75.0%	74.8%
Flag for Phone Number provided of Household	70.7%	83.8%
Flag for Presence of Any Person Age 18 to 24 in Household	17.6%	15.1%
Flag for Presence of Any Person Age 25 to 34 in Household	82.6%	15.1%
Flag for Presence of Any Person Age 35 to 64 in Household	82.6%	15.1%
Flag for Presence of Any Person Age ≥65 in Household	17.6%	15.1%

* Sample sizes indicate counts of addresses in the block groups sampled for the Housing and Children Study and addresses in the National Survey of Family Growth roster data considered in the address-level eligibility prediction model.

Table 3 Distribution of Predicted Address-Level Eligibility Probability from Each Imputation for Addresses in Sampled Block Groups, Housing and Children Study

	Pilot Study					Main Study									
						Dallas					Cleveland				
Imputation	N	Mean	SD	Min	Max	n	Mean	SD	Min	Max	n	Mean	SD	Min	Max
1	10,304	0.246	0.148	0.013	0.774	41,536	0.417	0.268	0.004	0.987	26,000	0.259	0.209	0.004	0.984
2	10,304	0.242	0.144	0.013	0.736	41,536	0.423	0.272	0.004	0.992	26,000	0.258	0.204	0.004	0.985
3	10,303	0.246	0.147	0.013	1.000	41,526	0.420	0.271	0.003	0.985	25,999	0.258	0.207	0.003	0.984
4	10,304	0.244	0.143	0.008	0.757	41,534	0.420	0.270	0.002	0.985	25,501	0.253	0.204	0.004	0.987
5	10,304	0.242	0.143	0.011	0.727	41,536	0.424	0.271	0.004	0.986	25,999	0.261	0.208	0.005	0.988
6	10,303	0.247	0.143	0.009	1.000	41,536	0.419	0.267	0.002	0.987	26,000	0.257	0.206	0.004	0.987
7	10,304	0.247	0.143	0.000	0.760	41,536	0.419	0.269	0.002	0.987	26,000	0.260	0.211	0.003	0.985
8	10,304	0.248	0.144	0.013	0.771	41,536	0.419	0.271	0.003	0.987	25,999	0.255	0.205	0.003	0.987
9	10,303	0.246	0.144	0.000	0.766	41,536	0.415	0.268	0.003	0.984	26,000	0.258	0.207	0.004	0.986
10	10,303	0.247	0.147	0.012	0.766	41,536	0.419	0.270	0.004	0.987	26,000	0.260	0.209	0.000	0.988
Average	*10,304*	*0.245*	*0.145*	*0.009*	*0.806*	*41,536*	*0.420*	*0.265*	*0.004*	*0.983*	26,000	0.258	0.202	0.004	0.986

Note. The sample size may differ across imputation. This occurred when the imputation produced fewer categories of the MSG ethnicity variable for the National Survey of Family Growth addresses than for the Housing and Children Study addresses.

Results

Accuracy

Table 4.A provides results of screening interviews by BG and income strata along with the observed and predicted eligibility of sample addresses. The comparison between predicted and observed eligibility provides information about the accuracy of our predictions. Overall, out of 684 sample addresses, 284 completed the screener; and among them, 78 were eligible for H&C. This resulted in a 27.5% eligibility rate, which is 10 percentage points higher than the national eligibility rates of 17-18% estimated from NSFG and ACS. The predicted eligibility rate of 25.8% mapped onto the eligibility observed in the field, 27.5%. When examining the eligibility by BG, there was a substantial variation in its prediction accuracy across BGs measured by the difference between observed and predicted eligibility rates. Although the small number of BGs considered in the pilot study limited a thorough investigation, BGs in the high-income stratum appeared to be subject to a lower level of variability in prediction accuracy than BGs in the lower-income stratum.

Table 4 Block Group Level Screener Results by Income Strata, Housing and Children Study

A. Pilot Study

Block Group		Counts of Addresses			Eligibility		
Number	Income Strata	Sampled	Interviewed	Eligible	Observed	Predicted*	Pred−Obs
1	Low	53	27	9	33.3%	31.2%	-2.1%
2	Low	41	12	3	25.0%	16.4%	-8.6%
3	Low	42	19	9	47.4%	28.9%	-18.5%
4	Low	60	26	9	34.6%	29.8%	-4.8%
5	Low	56	25	7	28.0%	34.5%	6.5%
6	Low	51	18	6	33.3%	29.5%	-3.8%
7	Low	52	22	2	9.1%	22.1%	13.0%
8	Low	33	19	5	26.3%	43.4%	17.1%
Subtotal: Low-Income		*388*	*168*	*50*	*29.8%*	*29.3%*	*-0.5%*
9	Middle	42	23	7	30.4%	32.6%	2.2%
10	Middle	32	11	4	36.4%	29.0%	-7.3%
11	Middle	42	17	6	35.3%	26.1%	-9.2%
12	Middle	39	13	4	30.8%	27.9%	-2.8%
13	Middle	45	14	1	7.1%	12.3%	5.1%
Subtotal: Middle-Income		*200*	*78*	*22*	*28.2%*	*25.2%*	*-3.0%*
14	High	64	23	3	13.0%	9.2%	-3.8%
15	High	32	15	3	20.0%	20.6%	0.6%
Subtotal: High-Income		*96*	*38*	*6*	*15.8%*	*13.0%*	*-2.8%*
Grand Total		*684*	*284*	*78*	*27.5%*	*25.8%*	*-1.7%*

*Average eligibility predicted for sample addresses in a given block group in the secondary sampling stage.

B. Main Study -- Dallas

Block Group		Counts of Addresses			Eligibility		
Number	Income Strata	Sample	Interviewed	Eligible	Observed	Predicted*	Pred−Obs
Quarter 1							
1	Low	111	63	18	28.6%	42.9%	14.3%
2	Low	110	50	7	14.0%	9.9%	-4.1%
3	Low	103	45	9	20.0%	20.7%	0.7%
4	Low	81	31	10	32.3%	22.8%	-9.5%
5	Low	166	80	27	33.8%	30.9%	-2.9%
6	Low	78	47	16	34.0%	23.6%	-10.5%
7	Middle	81	36	4	11.1%	28.5%	17.4%
8	Middle	94	35	9	25.7%	68.3%	42.6%
9	Middle	87	49	9	18.4%	22.4%	4.0%
10	Middle	97	22	7	31.8%	92.1%	60.3%
11	High	90	47	3	6.4%	61.7%	55.3%
12	High	98	45	14	31.1%	39.5%	8.4%
Subtotal: Quarter 1		*1,196*	*550*	*133*	*24.2%*	*38.3%*	*14.1%*
Quarter 2							
1	Low	122	43	6	14.0%	25.8%	11.8%
2	Low	161	72	13	18.1%	42.6%	24.6%
3	Low	135	73	24	32.9%	21.0%	-11.9%
4	Low	140	51	15	29.4%	67.8%	38.4%
5	Low	70	22	12	54.5%	29.8%	-24.8%
6	Low	132	58	15	25.9%	12.5%	-13.3%
7	Middle	105	14	0	0.0%	80.2%	80.2%
8	Middle	156	66	17	25.8%	10.4%	-15.4%
9	Middle	106	79	14	17.7%	30.8%	13.1%
10	Middle	122	55	13	23.6%	11.4%	-12.2%
11	High	99	45	5	11.1%	11.4%	0.3%
12	High	99	45	5	11.1%	26.6%	15.5%
Subtotal: Quarter 2		*1,447*	*623*	*139*	*22.3%*	*30.8%*	*8.5%*
Quarter 3							
1	Low	87	54	15	27.8%	36.0%	8.2%
2	Low	93	29	6	20.7%	31.3%	10.6%
3	Low	99	64	32	50.0%	82.0%	32.0%
4	Low	93	49	4	8.2%	24.6%	16.4%
5	Low	95	58	20	34.5%	57.7%	23.2%
6	Low	102	53	12	22.6%	27.2%	4.5%
7	Middle	168	115	37	32.2%	35.2%	3.1%
8	Middle	100	40	11	27.5%	40.8%	13.3%

Block Group		Counts of Addresses			Eligibility		
Number	Income Strata	Sample	Interviewed	Eligible	Observed	Predicted*	Pred−Obs
9	Middle	96	68	19	27.9%	52.7%	24.8%
10	Middle	111	50	7	14.0%	79.8%	65.8%
11	High	93	55	13	23.6%	16.5%	-7.2%
12	High	82	24	2	8.3%	6.5%	-1.8%
Subtotal: Quarter 3		*1,219*	*659*	*178*	*27.0%*	*41.6%*	*14.6%*
Quarter 4							
1	Low	143	69	7	10.1%	24.7%	14.6%
2	Low	243	145	31	21.4%	30.4%	9.1%
3	Low	136	77	16	20.8%	16.9%	-3.9%
4	Low	115	52	16	30.8%	21.1%	-9.7%
5	Low	122	60	16	26.7%	81.8%	55.1%
6	Low	151	76	15	19.7%	32.1%	12.3%
7	Middle	87	58	7	12.1%	11.9%	-0.2%
8	Middle	87	41	2	4.9%	13.0%	8.1%
9	Middle	116	64	19	29.7%	38.7%	9.0%
10	Middle	130	61	17	27.9%	49.0%	21.1%
11	High	124	66	15	22.7%	28.8%	6.1%
12	High	92	53	9	17.0%	50.6%	33.6%
Subtotal: Quarter 4		*1,546*	*822*	*170*	*20.7%*	*33.4%*	*12.8%*
Reserve							
1	Low	65	32	8	25.0%	38.3%	13.3%
2	Low	81	33	17	51.5%	81.5%	29.9%
3	Low	59	38	9	23.7%	75.5%	51.9%
4	Middle	104	37	8	21.6%	35.4%	13.7%
5	Middle	70	29	11	37.9%	28.1%	-9.9%
6	High	57	19	4	21.1%	47.0%	25.9%
Subtotal: Reserve		*436*	*188*	*57*	*30.3%*	*50.2%*	*19.8%*
Subtotal: Low-Income		*3,093*	*1,524*	*396*	*26.0%*	*36.1%*	*10.1%*
Subtotal: Middle-Income		*1,917*	*919*	*211*	*23.0%*	*40.1%*	*17.1%*
Subtotal: High-Income		*834*	*399*	*70*	*17.5%*	*31.4%*	*13.8%*
Grand Total		*5,844*	*2,842*	*677*	*23.8%*	*36.7%*	*12.9%*

*Average eligibility predicted for sample addresses in a given block group in the secondary sampling stage.

C. Main Study -- Cleveland

Block Group		Counts of Addresses			Eligibility		
Number	Income Strata	Sample	Interviewed	Eligible	Observed	Predicted*	Pred−Obs
Quarter 1							
1	Low	171	79	27	34.2%	23.8%	-10.3%
2	Low	151	56	11	19.6%	19.7%	0.1%
3	Low	186	48	8	16.7%	14.3%	-2.3%
4	Low	179	66	38	57.6%	61.4%	3.8%
5	Low	226	73	19	26.0%	13.0%	-13.1%
6	Low	160	28	5	17.9%	16.5%	-1.4%
7	Middle	171	74	16	21.6%	17.4%	-4.2%
8	Middle	168	33	8	24.2%	15.3%	-8.9%
9	Middle	168	50	4	8.0%	14.5%	6.5%
10	Middle	171	72	20	27.8%	16.0%	-11.8%
11	High	166	91	14	15.4%	15.2%	-0.2%
12	High	169	90	9	10.0%	13.3%	3.3%
Subtotal: Quarter 1		*2,086*	*760*	*179*	*23.6%*	*20.0%*	*-3.5%*
Quarter 2							
1	Low	180	85	23	27.1%	22.6%	-4.5%
2	Low	130	35	14	40.0%	61.6%	21.6%
3	Low	165	42	9	21.4%	13.9%	-7.6%
4	Low	100	32	12	37.5%	19.1%	-18.4%
5	Low	142	12	0	0.0%	20.3%	20.3%
6	Low	148	29	0	0.0%	38.3%	38.3%
7	Middle	150	50	12	24.0%	20.2%	-3.8%
8	Middle	137	58	13	22.4%	17.7%	-4.8%
9	Middle	152	41	8	19.5%	21.4%	1.9%
10	Middle	132	34	9	26.5%	17.4%	-9.1%
11	High	142	56	10	17.9%	13.4%	-4.5%
12	High	132	10	1	10.0%	19.1%	9.1%
Subtotal: Quarter 2		*1,710*	*484*	*111*	*22.9%*	*23.5%*	*0.6%*
Quarter 3							
1	Low	147	46	17	37.0%	20.8%	-16.2%
2	Low	135	17	4	23.5%	52.7%	29.2%
3	Low	147	47	12	25.5%	21.7%	-3.9%
4	Low	243	74	14	18.9%	27.0%	8.1%
5	Low	140	67	32	47.8%	36.4%	-11.4%
6	Low	135	23	2	8.7%	12.2%	3.5%
7	Middle	135	43	8	18.6%	17.3%	-1.3%
8	Middle	166	41	10	24.4%	15.8%	-8.6%

Block Group		Counts of Addresses			Eligibility		
Number	Income Strata	Sample	Interviewed	Eligible	Observed	Predicted*	Pred−Obs
9	Middle	212	35	6	17.1%	11.6%	-5.6%
10	Middle	143	28	13	46.4%	28.4%	-18.1%
11	High	135	41	5	12.2%	16.9%	4.7%
12	High	132	58	9	15.5%	10.8%	-4.7%
Subtotal: Quarter 3		*1,870*	*520*	*132*	*25.4%*	*22.4%*	*-3.0%*
Quarter 4							
1	Low	141	43	6	14.0%	16.4%	2.4%
2	Low	156	58	11	19.0%	18.1%	-0.9%
3	Low	132	62	30	48.4%	57.5%	9.1%
4	Low	104	52	30	57.7%	91.4%	33.7%
5	Low	131	54	13	24.1%	15.0%	-9.1%
6	Low	184	66	17	25.8%	16.2%	-9.6%
7	Middle	120	51	7	13.7%	20.0%	6.2%
8	Middle	172	51	7	13.7%	13.9%	0.2%
9	Middle	141	30	1	3.3%	8.8%	5.5%
10	Middle	145	77	15	19.5%	15.1%	-4.4%
11	High	128	47	7	14.9%	18.6%	3.7%
12	High	130	43	5	11.6%	17.2%	5.6%
Subtotal: Quarter 4		*1,684*	*634*	*149*	*23.5%*	*23.8%*	*0.3%*
Reserve							
1	Low	167	64	11	17.2%	19.0%	1.8%
2	Low	150	47	9	19.1%	10.1%	-9.0%
3	Low	147	23	6	26.1%	7.3%	-18.8%
4	Middle	167	105	21	20.0%	49.3%	29.3%
5	Middle	131	35	3	8.6%	13.5%	5.0%
6	High	141	30	4	13.3%	46.6%	33.2%
Subtotal: Reserve		*903*	*304*	*54*	*17.8%*	*24.7%*	*7.0%*
Subtotal: Low Income		*4,197*	*1,328*	*380*	*28.6%*	*26.4%*	*-2.2%*
Subtotal: Middle Income		*2,781*	*908*	*181*	*19.9%*	*18.5%*	*-1.4%*
Subtotal: High Income		*1,275*	*466*	*64*	*13.7%*	*18.9%*	*5.2%*
Grand Total		*8,253*	*2,702*	*625*	*23.1%*	*22.6%*	*-0.6%*

* Average eligibility predicted for sample addresses in a given block group in the secondary sampling stage.

Sampling Efficiency

Sampling efficiency was examined by comparing sample sizes under the current design and under simple random sampling (SRS) of addresses within BGs. The current design yielded 78 eligible cases from 684 addresses with a screener cooperation rate of 41.5% and an eligibility rate of 27.5%. Under SRS, the eligibility rate would be similar to the national rate (around 17.5%.) To yield 78 cases, SRS would have required 1,074 addresses (=78 / (41.5% cooperation rate x 17.5% eligibility rate)), which is an increase of almost 400 sample addresses.

Main Study

The main study targeted households with at least one child aged between 3 and 10 years old in Dallas, TX and Cleveland, OH. Given the results from the pilot study, an identical sample design was employed in the main survey with more streamlined and updated external data.

Frame

The frame included 998 BGs from the city of Dallas proper as done in the pilot study and 850 BGs from within the city of Cleveland proper, covering 70.5% and 85.3% of the ZIP codes where the voucher applicants resided in the respective locations.

Sample Design

A stratified two-stage PPeS design, identical to the pilot study, was implemented with more up-to-date auxiliary data. Specifically, the ACS 2011-2015 5-Year SF, the NSFG roster data from 2011 to 2017 and the MSG data purchased in 2017 were used in the main study. In particular, the NSFG data included 68,180 addresses from 2,007 BGs. Note that Census planning data was not considered in the main study design, due to a large overlap in its information with ACS SF. Data collection was planned for a year with the fieldwork implemented via 4 replicates. Hence, the sample drawn at the beginning of the project was released sequentially by replicate.

Primary Stage Sampling

The eligibility rate of addresses aggregated from all 2,007 BGs from NSFG was regressed on BG-level variables in ACS SF. The grouped logit model included these 84 components extracted from PCA of 236 variables in ACS SF (see Supplementary Table 3 at https://goo.gl/KtRcfD) and 188 two-way interactions of some com-

ponents as predictors. This model showed an improved fit compared to the pilot study (area under the ROC curve: 0.67; Hosmer–Lemeshow goodness-of-fit test: χ^2=2.65, *df*=8, *p*=.955).

With the updated ACS data, the income stratification changed. For Dallas, BGs with >51.0% households with annual income less than $35,000 were classified as the low-income stratum; those with 27.0%-51.0% into the middle-income stratum; and those with <27.0% into the high-income stratum. For Cleveland, 62.1% and 38.7% were the respective income cut-off points. Overall, 54 BGs were selected for each site using PPeS for a 3:2:1 ratio of low-, middle- and high-income strata BGs, where 48 BGs were randomly split into 4 replicates and the remaining 6 BGs were set aside as reserve sample.

Secondary Stage Sampling

The address-level eligibility model included 68,180 addresses from the NSFG roster data (41,536 in Dallas and 26,000 in Cleveland). Address-level eligibility was modelled using address-level MSG variables as well as BG-level ACS SF data, where the missingness of the MSG data was handled through sequential multiple imputation and the dimensionality of the ACS data was reduced through PCA. The distribution of predicted eligibility across the 10 imputations is shown in Table 3. The predicted eligibility was similar across imputations and, on average, higher for Dallas (approximately 0.42) than Cleveland (approximately 0.26). The average predicted eligibility from the 10 imputations was used as the MoS.

Income-based stratification used the household income variable in MSG. Unlike the pilot study, the income tertiles calculated *within each BG* were used. This means that the stratification did not use "hard boundaries" but varied by BG. Within each BG, one third of addresses were assigned to low-, middle- and high-income strata. Considering the target ratio of 3:2:1 for these income strata as well as predicted eligibility rates of addresses, 5,844 addresses from Dallas and 8,258 addresses from Cleveland were sampled for data collection, which ran from May 2017 to September 2018.

Results

Accuracy

The results of the screener fieldwork are in Tables 4.B and 4.C. Among the 2,842 households in Dallas that completed the screener, 677 were eligible. This overall eligibility rate of 23.8% was lower by 12.9 percentage points than the predicted eligibility rate of 36.7%. Although the over-prediction of eligibility was persistent across all replicates and across income strata, the observed eligibility rate was still higher than the national average of 17-18%. For Cleveland, the overall eligibility

rate was 23.1%, closely matching the predicted eligibility of 22.6% and higher than the national average eligibility. With the exception of BG 11 of Dallas in Quarter 1, the variability in accuracy was smaller for the addresses in the high-income stratum.

Sampling Efficiency and Cost Considerations

Our design yielded 677 eligible cases with a screener cooperation rate of 48.6% and an eligibility rate of 23.8% for Dallas. To yield the same number of eligible households under SRS, the design would have required screening 7,954 addresses (= 677 / (48.6% cooperation rate x 17.5% eligibility rate)), an increase of a little over 2,100 sample addresses. For Cleveland, with a yield of 625 eligible cases, a screener cooperation rate of 32.7%, and an eligibility rate of 23.1% under the current design, SRS would have required 10,909 addresses (=625 / (32.7% cooperation rate x 17.5% eligibility rate)), an increase of over 2,600 sampled addresses. Our design offered a net reduction in required sample size of 27% (5,844 under our design vs. 7,954 under SRS) for Dallas and 24% (8,253 under our design vs. 10,909 under SRS) for Cleveland.

In order to assess the cost savings through improvement in screening efficiency, we fitted a simple cost model with interviewer as the unit of analysis as follows:

$$T_i = \beta_0 + \beta_1 S_i + \beta_2 I_i + \varepsilon_i$$

Where T_i is the total billed hours by interviewer i; S_i is the number of completed screeners by interviewer i; and I_i is the number of completed interviews by interviewer i. Therefore, coefficients β_1 and β_2 are, respectively, the interviewer hours per completed screener and per completed main interview. Using the data from 60 interviewers for the main study, the estimated model ($R^2 = 0{,}913$) suggested about 1.9 hours (SE: 0.4) per completed screener and about 10.8 hours (SE: 1.2) per completed interview.

To estimate the cost savings, we consider a counter factual that assumes the same cooperation rate for the screening interview and yields the same number of eligible households (677 in Dallas and 625 in Cleveland) with the national eligibility rate of 17.5%. The standard approach would have required completing screener interviews with 3,869 households (=677/17.5%) in Dallas and 3,571 households (=625/17.5%) in Cleveland, as opposed to 2,842 and 2,702 completed screeners in the respective areas under our design given in Tables 4.B and 4.C. This equates to a 25% reduction in required screener completion. This also means that, with 1.9 interviewer hours estimated per completed screener, our design saved nearly 3,600 interviewer hours. This ignores the additional costs of sampling a larger number of households to reach the required eligible households using the standard approach.

Discussion

Our goal in this study was to improve sampling efficiency and thereby reduce the data collection costs of the H&C study. The screening for eligible members of the target population from the larger sampling population frame contributes greatly to the cost of surveys of uncommon and hard-to-reach populations. For implementing measurements about child development and parent-child interactions, H&C required a face-to-face mode.

Survey research organizations can leverage information from previous studies combined with commercial databases to develop model-assisted sampling designs that may improve sampling efficiency and reduce costs. This case study illustrates a methodology that can be used to leverage information from imperfect sources through imputation and modeling. We note that practical limitations exist for using commercial databases directly for sampling. However, when reflecting on our proposed approach that used imputation and the modeling of study eligibility, it is feasible to address the well-documented availability and accuracy issues of commercial data. It is important to note that, for studies designed to oversample addresses/areas with characteristics associated with lower availability or accuracy of the commercial data (e.g., lower income), the prediction accuracy may be lower as shown in the case of over-prediction of eligibility in Dallas (Table 4.B) than for studies without such oversampling requirements.

Efficiency can also be gained by performing model-based analysis when commercial data is available on all households in the selected geographies and the ACS data is available on all geographies used as sampling units. Alternatively, a Bayesian prediction model can be used to synthesize the entire population through simulations and then construct inferences from the simulated populations, offering a gain in inference efficiency. Whatever the method used, we believe that our case study demonstrates that these methods have great potential for leveraging commercial data to improve efficiency in sampling and inference.

References

Barron, M., Davern, M., Montgomery, R., Tao, X., Wolter, K. M., Zeng, W., … Black, C. (2015). Using Auxiliary Sample Frame Information for Optimum Sampling of Rare Populations. *Journal of Official Statistics*, *31*(4), 545–557. https://doi.org/10.1515/JOS-2015-0034

Buskirk, T. D., Malarek, D., & Bareham, J. S. (2014). From Flagging a Sample to Framing It: Exploring Vendor Data That Can Be Appended to ABS Samples. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 111–124.

English, N., Kennel, T., Buskirk, T., & Harter, R. (2019). The construction, maintenance, and enhancement of address-based sampling frames. *Journal of Survey Statistics and Methodology*, *7*(1), 66–92. https://doi.org/10.1093/jssam/smy003

Harter, R., Battaglia, M. P., Buskirk, T. D., Dillman, D. A., English, N., Fahimi, M., … Zukerberg, A. L. (2016). *Address-based Sampling.* Retrieved from https://www.aapor.org/Education-Resources/Reports/Address-based-Sampling.aspx

Kalton, G., Kali, J., & Sigman, R. (2014). Handling Frame Problems When Address-Based Sampling Is Used for In-Person Household Surveys. *Journal of Survey Statistics and Methodology*, *2*(3), 283–304. https://doi.org/10.1093/jssam/smu013

Pasek, J., Jang, S. M., Cobb, C. L., Dennis, J. M., & Disogra, C. (2014). Can marketing data aid survey research? Examining accuracy and completeness in consumer-file data. *Public Opinion Quarterly*, *78*(4), 889–916. https://doi.org/10.1093/poq/nfu043

Raghunathan, T., Berglund, P., & Solenberger, P. (2018). *Multiple Imputation in Practice: With Examples Using IVEware*. https://doi.org/10.1198/000313001317098266

Rastogi, S., & O'Hara, A. (2012). *2010 Census Match Study Report* (No. 247). Retrieved from https://www.census.gov/content/dam/Census/library/publications/2012/dec/2010_cpex_247.pdf

Roth, S. B., Han, D., & Montaquila, J. M. (2013). The ABS Frame: Quality and Considerations. *Survey Practice*, *6*(4), 1–6. https://doi.org/10.29115/SP-2013-0021

Roth, S., Caporaso, A., & DeMatteis, J. (2018). Variables Appended to ABS Frames: Has Data Quality Improved? *Paper Presented at the Annual Conference of American Association for Public Opinion Research, Denver, CO.*

Smith, T. W. (2011). The Report of the International Workshop on Using Multi-level Data from Sample Frames, Auxiliary Databases, Paradata and Related Sources to Detect and Adjust for Nonresponse Bias in Surveys*. *International Journal of Public Opinion Research*, *23*(3), 389–402. https://doi.org/10.1093/ijpor/edr035

Valliant, R., Hubbard, F., Lee, S., & Chang, C. (2014). Efficient Use of Commercial Lists in U.S. Household Sampling. *Journal of Survey Statistics and Methodology*, *2*(2), 182–209. https://doi.org/10.1093/jssam/smu006

West, B. T. (2013). An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 211–225. https://doi.org/10.1111/j.1467-985X.2012.01038.x

West, B. T., Wagner, J., Hubbard, F., & Gu, H. (2015). The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology*, *3*(2), 240–264. https://doi.org/10.1093/jssam/smv004

Appendix 1

List of Acronyms

Acronym	Definition
ACS	U.S. American Community Survey
BG	U.S. Census Block Group
DV	Dependent Variable
H&C	Housing & Children Study
HRS	Health and Retirement Study
HUD	United States Department of Housing and Urban Development
IV	Independent Variable
MoS	Measure of Size
MSG	Marketing Systems Group
NSFG	National Survey of Family Growth
PCA	Principle Component Analysis
PHA	Public Housing Authority
PPeS	Probability Proportionate to Estimated Size
PUMS	U.S. Census Public Microdata Sample
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SE	Standard Error
SF	Summary File
SRS	Simple Random Sample

Date of online first publication: 2024-04-05