DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 2, 2018, pp. 445–462, http://dx.doi.org/10.2478/JOS-2018-0020

Adaptive Design Strategies for Nonresponse Follow-Up in Economic Surveys

Stephen J. Kaputa[1] *and Katherine J. Thompson*[1]

The U.S. Census Bureau is investigating nonrespondent subsampling strategies for use in the 2017 Economic Census. In previous research, we developed an optimized allocation procedure for subsampling nonrespondents that selects larger systematic samples in domains with lower initial response. This article expands on our previous research by exploring improvements to the optimal allocation method; we investigate refinements to the previous procedure that incorporate measures of respondent balance with respect to the original sample. The revised allocation procedures have simultaneous objectives of allocating high proportions of sample in domains that indicate potential nonresponse bias and of equalizing response rates across domains. We examine the effects of the alternative allocation approaches on Horvitz-Thompson estimates via a simulation study using data from the 2014 Annual Survey of Manufactures.

Key words: Adaptive design; nonresponse subsampling; optimal allocation.

1. Introduction

The U.S. Census Bureau is investigating adaptive nonresponse follow-up (NRFU) strategies for small businesses in the 2017 Economic Census. One considered approach is the selection of a probability subsample of small businesses for NRFU. For this, the nonrespondent subsampling design must be robust to the sample design and estimators to the largest extent possible; the Economic Census is a multi-purpose collection with a key set of general statistics items collected from *all* surveyed units in a sector (e.g., annual payroll, total receipts) as well as industry-specific items (e.g., product sales). Missing data treatments and estimation procedures differ by type of item. The term *census* is a bit misleading, as many sectors include a probability sample of small establishments. The sample design can likewise differ by sector. Consequently, we consider a systematic sample of small nonresponding establishments sorted by a measure of size (MOS), a sampling design known to be as efficient as stratified simple random sample without replacement if the list is in random order and more efficient if the list is monotonic increasing or decreasing (Lohr 2010). Within sectors/industries, the largest establishments

[1] U.S. Census Bureau, Economic Statistical Methods Division, 4600 Silver Hill Road, Washington, DC 20233, U.S.A. Emails: Stephen.Kaputa@census.gov and Katherine.J.Thompson@census.gov
Acknowledgments: The authors thank Cha-Chi Fan, Carma Hogue, Eddie Salyers, Edward Watkins III, three referees, and the Associate Editor for their useful comments on earlier versions of this manuscript. This report is released to inform interested parties of research and to encourage discussion. Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau.

are deemed ineligible for nonrespondent subsampling, because of their large contribution to industry totals.

Often, the primary motivation for selecting a subsample of nonrespondents for follow-up is to save cost without inducing additional nonresponse bias. In a business survey setting, the contention that the realized respondent set of small businesses remains a probability sample is debatable given the (traditional) emphasis on obtaining responses from the larger sampled units. In this setting, one could argue that sampled smaller units "opt in" to respond. In fact, several discussions of the summary report of the AAPOR Task Force on non-probability sampling (Baker et al. 2013) specifically question whether "a probability sample with less than full coverage and high nonresponse should still be considered a probability sample." Selecting a probability subsample of nonrespondents may limit this phenomenon, especially if the subsampling is combined with contact strategies known to be effective (especially for "hard-to-reach" populations). In addition, with a probability subsample, one can use accepted quality measures, such as sampling error or response rates for evaluation.

In this article, we propose a subsampling design that can be implemented at any stage of the data collection process and with any sample design, making it quite flexible although not necessarily optimal for selected sample designs and estimators. For this, we introduce an optimized allocation procedure for nonrespondent subsampling that has simultaneous objectives of allocating high proportions of sample in domains that indicate potential nonresponse bias and of equalizing response rates across domains. This procedure builds on earlier work presented in Kaputa et al. (2014). The earlier allocation strategies focused on equalizing response rates in specified domains. If the nonrespondents comprise a random subsample, this strategy should work well when combined with a systematic sample. However, if the nonrespondents differ systematically from the respondents in selected domains, then the earlier allocation strategies will be unlikely to reduce nonresponse bias. Särndal (2011) discusses the concept of a "balanced" sample where the realized respondent set "has the same or almost the same characteristics as the whole population" for selected items. Of course, attaining such a balanced sample would be a primary goal of any adaptive collection design and nonrespondent subsampling by extension. If the auxiliary variable is positively correlated with several survey outcomes, then the modified allocation procedure should be effective in reducing the nonresponse bias in the survey estimates, especially when combined with a robust sampling procedure.

This article expands upon previous research listed above by exploring improvements to the optimized allocation method that minimizes deviations between domain response rates incorporating the balance indicator and distance measure presented in Särndal and Lundquist (2014). This proposed refinement attempts to remediate nonresponse bias effects while maintaining the simple computational structure of the earlier allocation procedures. The nonrespondent subsampling strategy discussed here is based on the deterministic model of nonresponse bias, which partitions the population into a respondent stratum and a nonrespondent stratum (and the two strata are static). These are necessary conditions for specifying a probability subsample. As one referee noted, adaptive design implicitly relies on the stochastic model of nonresponse bias and aims to affect the correlation between the response propensity and the outcome variable. At best, the proposed subsample design will select a balanced subsample of nonrespondents. However, determining the appropriate

contact strategies to obtain responses from the selected units requires additional research (e.g., Thompson and Kaputa 2017). In Section 2, we present the subsample design and introduce the revised allocation procedures. The new allocation procedures have simultaneous objectives of allocating high proportions of sample in domains that indicate potential nonresponse bias and of equalizing response rates across domains. In Section 3, we examine the effects of the alternative allocations on Horvitz-Thompson estimates via a simulation study using data from the 2014 Annual Survey of Manufactures. We conclude in Section 4 with general comments and ideas for future research.

2. Allocation Procedure

2.1. Two-Phase Subsample Design

Given a stratified probability sample of n units divided into disjoint domains ($h = 1, \ldots, H$), we are interested in selecting a probability subsample of nonrespondents of a fixed size at predetermined NRFU phase (t). Domains can be sampling strata, groups of sampling strata, or any characteristics related to nonresponse and known for all units. All sampled units receive an initial contact, and one or more nonresponse follow-ups may be attempted before subsampling.

In this framework, some units are not included on the nonrespondent subsampling frame; instead, these units are treated as the certainty component of the nonrespondent subsample. For example, the largest units will always receive NRFU because of their expected influence on an industry total. Hereafter, we refer to the units that are excluded from consideration for nonrespondent subsampling as *ineligible* units. Because the optimal allocation procedure attempts to equalize domain response rates, *all* units – including the ineligible units – are included in the allocation procedure, although the subsampling is only performed on the frame constructed from eligible nonresponding units. This division of units within a domain into eligible and ineligible categories complicates the optimization procedures described below, but is an unfortunate reality. If all units in a given domain are eligible for subsampling (none excluded/ineligible) then the optimization problem is considerably simplified.

To summarize, the original sample comprises H disjoint domains. Immediately prior to nonrespondent subsampling, each domain h contains r_{1h} responding units and m_{1h} nonresponding units. A frame of m_1^e eligible nonrespondents is created from the m_{1h} nonresponding units in each domain h, and a subsample of the eligible nonrespondents in each domain is selected with sampling rate K_h. All m_{1h}^i ineligible nonresponding units are subjected to NRFU, and only the subsampled eligible units are subjected to NRFU $\left(m_{1h}^e + m_{1h}^i = m_{1h}\right)$. In addition, the NRFU procedures may differ by unit type ($\text{NRFU}^{\mathbf{i}}$ versus $\text{NRFU}^{\mathbf{e}}$).

Figure 1 illustrates this two-phase design. Notice that each domain contains two distinct sets of units: units that are not considered for subsampling (ineligible units) and units that are (eligible units).

Sample estimates are constructed by combining appropriately adjusted r_{1h} and r_{2h} units, where $r_{2h}^i + r_{2h}^e = r_{2h}$. Adjustment cell weighting or imputation may be used to account for any remaining nonresponse at the end of the collection period.

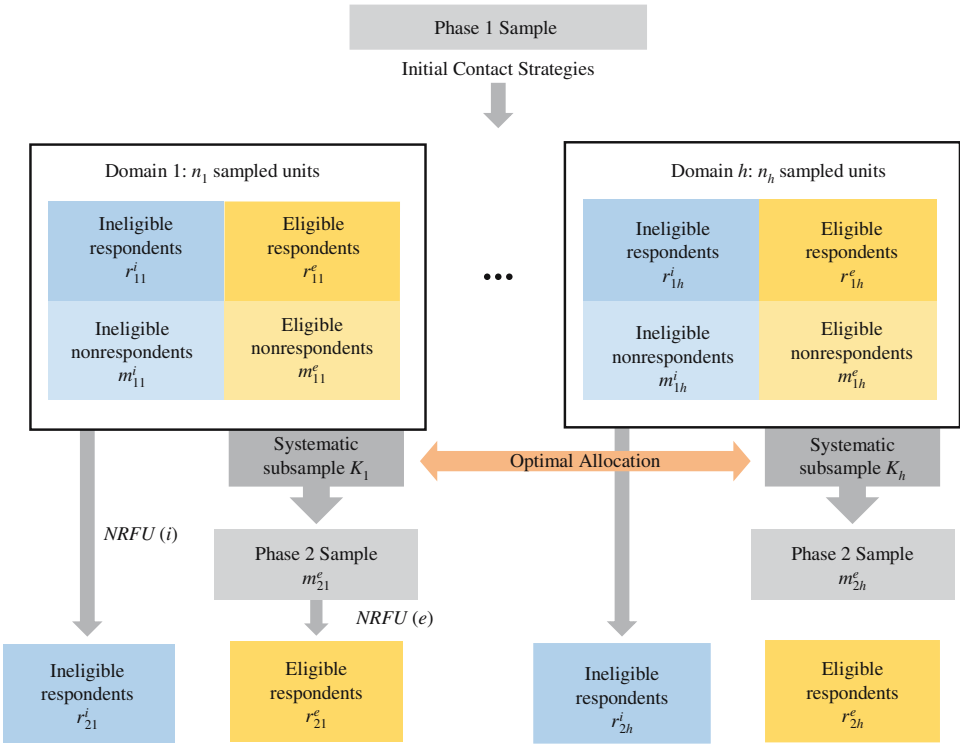

Fig. 1. Illustration of nonrespondent subsampling procedure.

2.2. *Allocation Strategies*

The minURR allocation method introduced in Kaputa et al. (2014) equalizes response rates across domains. If the units within a domain have the same mean and these means differ by domain, then the similar response rates can be indicative of a representative sample (Wagner 2012). Similarly, if the variables used to define the allocation domains are predictive of response, this allocation strategy will reduce nonresponse bias.

We formulate the allocation as a quadratic program that minimizes the squared deviation in predicted domain response rate from a target unit response rate subject to linear constraints on the size of the nonrespondent sample. First, we introduce the following notation:

q^e_h = Conversion rate for eligible nonresponders in domain h
q^i_h = Conversion rate for ineligible nonresponders in domain h
K = Overall subsampling rate (fixed, usually by budget)
K_h = Domain h specific subsampling rate (solving for this)

The minURR objective function is

$$\min \sum_h \left(URR^P_h - URR^T\right)^2 \tag{2.1}$$

where the predicted domain level unit response rate (URR) and the target response rate are

$$URR_h^P = \frac{\left(r_{1h} + \left(m_{1h}^i\, q_h^i\right)\right) + \left(m_{1h}^e\, q_h^e\, K_h\right)}{n_h} \tag{2.2}$$

$$URR^T = \frac{\sum_h \left(\left(r_{1h} + \left(m_{1h}^i\, q_h^i\right)\right) + \left(m_{1h}^e\, q_h^e\, K_h\right)\right)}{\sum_h n_h} \tag{2.3}$$

with the constraints that all K_h are bounded between zero and one (zero = no subsampling, and one = 100 percent NRFU) and the subsample size is equal to $K \sum_h m_{1h}^e$. We use the SAS® PROC NLP procedure to minimize the objective function, obtaining the set of K_h used for subsampling (Note: The data analysis for this article was generated using SAS software. Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA).

The predicted domain level and target response rates incorporate assumed or estimated conditional response propensities. These estimates can differ for the eligible and ineligible groups within domains. If the same – or similar – NRFU procedures are used for adjacent collections, then conversion rates can be estimated from historic data and possibly adjusted for anticipated changes in response patterns or collection strategies. Otherwise, we recommend using constant values in the allocation, but performing a sensitivity analysis by repeating the procedure with different assumed rates.

Although response rates are easy to compute, they are not necessarily predictors of nonresponse bias; for example, see Peytcheva and Groves (2009). Nevertheless, the response rate is a component in the degree of nonresponse bias on an outcome variable (Andridge and Little 2011; Andridge and Thompson 2015). However, the degree of difference on outcome means of respondents and nonrespondents is an equally important component in the assessment of nonresponse bias. Of course, such measures are not available for collected survey data. Instead, Särndal and Lundquist (2014) present measures that assess the degree to which the response set is similar to the full sample and the degree of difference between respondents and nonrespondents with respect to auxiliary variables or paradata available to all units on the frame. Let

y_{hi} = characteristic of interest, subject to nonresponse
x_{hi} = auxiliary variable available for all sampled units

P_h = weighted response rate = $\sum_{i\in h} w_{hi} I_{hi} / \sum_{hi\in s} w_{hi}$, where I_{hi} is a unit response indicator and w_{hi} is the design weight. Assume that $y_{hi} \approx \beta_h x_{hi} + \varepsilon_{hi}$, where ε_{hi} is a random error term.

Imbalance Measured as $IB_{xh} = (\bar{x}_{rh} - \bar{x}_{sh})' \Sigma_{sh}^{-1} (\bar{x}_{rh} - \bar{x}_{sh})$, $\Sigma_{sh} = \sum_{i\in h} w_{hi} x_{hi} x_{hi}' / \sum_{i\in sh} w_{hi}$. A *balance indicator for variable x* in domain h is given as $BI_{xh} = 1 - 2P_h\sqrt{IB_{xh}}$. This measure is bounded between 0 and 1, with values close to 1 indicating balance on the respondent sample for the studied variable.

Distance The difference between mean value for respondents and mean value for nonrespondents on variable x in domain h. This is measured as $D_{xh} = (\bar{x}_{rh} - \bar{x}_{nr,h})' \Sigma_{sh}^{-1} (\bar{x}_{rh} - \bar{x}_{nr,h})$.

Examined together, these indicators can assess the "representativeness" of the realized respondent sample on selected characteristics. For a given domain, a balance indicator BI_{xh} near one combined with a distance measure D_{xh} near zero are indicative of low nonresponse bias as measured on the auxiliary variable (serving as a proxy for the characteristics of interest). Examined separately, each has its limitation: the balance indicator tends to overestimate the degree of balance between the original sample and respondent sample, and the distance measure is highly sensitive to differences in respondent and nonrespondent sample sizes. This sensitivity can be very apparent in highly skewed business populations.

A domain with a high response rate can have a low balance indicator or a large distance measure, especially if the respondent set comprises primarily the larger or smaller units. In this case, the minURR allocation would be very small (if not zero), and the domain estimates would be subject to high nonresponse bias. We address this deficiency in the minURR allocation procedure by incorporating the balance indicators or distance measures as domain weights into the objective function, proposing the two new objective functions below (the constraints are unchanged).

$$\text{BminURR}: \quad \min \sum_h (1 - BI_{xh}) \left(URR_h^P - URR_{BD}^T \right)^2 \tag{2.4}$$

$$\text{DminURR}: \quad \min \sum_h (D_{xh}) \left(URR_h^P - URR_{BD}^T \right)^2 \tag{2.5}$$

$$URR_{BD}^T = \frac{\sum_h \left(\left(r_{1h} + \left(m_{1h}^i \, q_h^i \right) \right) + \left(m_{1h}^e \, q_h^e \right) \right)}{\sum_h n_h} \tag{2.6}$$

Notice the modification to the target response rate (2.6) in these quadratic programs. The minURR method uses a target response rate given a fixed subsample size and response propensity model. For a given domain, if the predicted response rate is equal to the target response rate, then the squared difference is zero. This is a common occurrence and can mute the effect of the domain weights in the BminURR and DminURR allocations. Of course, the objective of the minURR allocation procedure is to equalize response rates in all domains. Consequently, here we need to create slack in the objective function to allow for the possibility of subsampling in "unbalanced" domains that satisfy the original response rate target. Therefore, we set the target for the BminURR and DminURR allocation methods as the final response rate, given full NRFU (for eligible and ineligible units) under the assumed response propensity models.

In the quadratic programs specified by (2.4) and (2.5), domains with larger weights will have a larger impact on the summed squared difference of the objective function. Since this is a minimization problem, the intent is to allocate the majority of the subsample in domains with lower than desired response rates or poorer than desired representativeness

on the measures for the studied auxiliary variable. One major limitation of this approach is that the balance indicator and distance measure use a single auxiliary variable that may not be positively correlated with all of the characteristics of interest. Consequently, it may not be possible to improve the representativeness of the respondent subsample for all survey variables. Additional auxiliary variables can easily be incorporated into these measures. However, as is often the case with business surveys, we have a limited number of possible covariates, and they tend to be highly collinear (e.g., frame measure size based on census payroll and administrative payroll for the current year).

These optimized allocation procedures (minURR, BminURR, and DminURR) have obtained feasible solutions on all of our studied datasets. Although equal domain size (in terms of total number of units) is not a requirement, domains should be relatively similar in size for maximum benefit. When domains vary widely in size, these optimizations often designate the smaller domains for the largest subsamples and may not require subsampling in the largest domains. Timing is also important. Early in the collection stage, when few units have provided responses, there will be little difference between the three realized allocations, and any advantages of incorporating the respondent sample measures might be lost, especially if the earlier responders differ consistently with later responders on distinct characteristics. If performed late in the collection stage, then weighted allocation procedures may do little to correct nonresponse bias; theoretically, the DminURR allocation could overcompensate for artificially high distances if (1) the URR is high in selected domains and (2) the data are highly skewed.

3. Simulation Study

3.1. ASM Background

We examine the effects of the alternative allocation approaches on Horvitz-Thompson estimates via simulation study using microdata from the 2014 Annual Survey of Manufactures (ASM). The ASM is an establishment survey designed to produce "sample estimates of statistics for all manufacturing establishments with one or more paid employee(s)" (http://www.census.gov/manufacturing/asm/). Sampled ASM establishments are surveyed for the four years between censuses. Strata are defined by a six-digit industry code using the North American Industry Classification System (NAICS). Each industry stratum is subdivided into certainty and noncertainty substrata. Approximately 20,000 establishments are included with certainty, and the remaining 30,000 establishments are selected with probability proportional to a composite frame measure of size (MOS) primarily based on annual payroll. The ASM uses composite imputation to account for unit and item nonresponse, imputing complete records. The ASM publishes totals for a variety of characteristics, including annual payroll, total shipments, cost of materials, and inventories, and uses a difference estimator for totals (Särndal et al. 1992, 221–225).

Because the items collected by the ASM questionnaire are a subset of the Economic Census' manufacturing sector items, the ASM is often used to pretest new Economic Census processing or data collection procedures. The ASM NRFU procedures are very similar to the Economic Census procedures. Initial contact and all four rounds of NRFU are conducted via mail, but the largest establishments in each industry have the highest

priority for phone follow-up. Furthermore, because a large company can comprise several establishments, sets of multi-unit (MU) establishments corresponding to the company can be designated for phone follow-up. The remaining nonresponding cases receive some reminders, but the noncertainty single unit (SU) establishments are very unlikely to receive a personal phone call.

3.2. Simulation Study Design

Our simulation study uses frame data from the 2014 ASM. Consistent with the planned procedures for upcoming embedded experiments and the 2017 Economic Census, multi-unit establishments are ineligible for nonrespondent sampling; eligible units are restricted to single unit establishments. We use three-digit NAICS industry (NAICS3) as subsampling domains: these larger classifications include several six-digit NAICS industries, each containing both certainty and noncertainty strata. The ASM publishes at the NAICS3 and an aggregated industry level, and all NAICS3 industries are considered equally important for publication.

With the ASM, the first NRFU attempt is historically very effective. The planned nonrespondent subsample selection occurs immediately before the second NRFU attempt. The second NRFU attempt for the single unit establishments is generally the most expensive (certified mail compared to standard mail). The ASM implements four rounds of NRFU after an initial collection period that lasts twelve days. After that, the length of the NRFU periods differ: the first round of NRFU lasts 48 days; the second and third NRFU rounds each last 35 days; the last round lasts 88 days. We incorporate these differing time-lengths into the response propensity modeling described in Subsection 3.3. Note that we do not attempt to incorporate any logistic regression model variability into the simulation. Consequently, our results are conditional on our assumed response mechanism.

Our simulation mimics the ASM NRFU. The simulation independently repeated the following procedure 3,000 times for each allocation procedure including full NRFU (no nonrespondent subsampling):

1. Randomly induce initial response and a single round of NRFU response in the complete dataset using the predicted response propensities described in Subsection 3.3., generating an average response rate of 43 percent.
2. Construct a frame of eligible nonrespondents.
    a. Compute balance and distance indicators and allocation response rates (response rates used in the optimized allocation procedures) within domain using all sampled units (eligible and ineligible).
    b. Perform optimized allocation procedures (minURR, BminURR, DminURR).
3. Sort eligible units within domain by frame measure of size (MOS) and select a systematic subsample of eligible nonresponding units using the nonrespondent subsampling rate determined above (minURR, BminURR, and DminURR).
4. Simulate unit response for the remaining three NRFU rounds in each sample. After assigning response status, compute cumulative cost, URR, and sample estimates obtained using each allocation method, along with 100 percent NRFU.

After each round of NRFU, we computed estimates using a nonresponse reweighted Horvitz-Thompson estimator. Within eligible domains, original responders' values $\left(r_{1h}^{e}\right)$

were weighted by their sample weights. All subsampled responding units were weighted by the product of their original sampling weight and the subsampling interval (K_h). In addition, subsampled respondents' weights were adjusted to account for remaining nonrespondents at each NRFU round using the unweighted inverse response rate within domain; see Little and Vartivarian (2005). If an eligible domain was not subsampled or less than two units in the subsample responded, then the original responders' weights were adjusted to account for nonresponse at each round of NRFU. Ineligible responders were always adjusted separately by their sample weight and inverse response rate within ineligible domain.

This estimation method does not match the ASM methodology, nor is it necessarily our recommended approach. Instead, we simplified the ASM estimation and imputation procedures to highlight differences in estimate quality obtained under the different allocations.

We compute the relative bias and the mean squared error of the final estimates for payroll (Y_1) and receipts (Y_2) to evaluate the statistical properties of the estimates obtained via each allocation over repeated samples. The relative bias of an estimate ($\hat{Y}$) at NRFU round t for a given allocation method a (Full NRFU, minURR, BminURR, DminURR) is given by

$$RBE(\hat{Y})_{at} = 100\left[\frac{\sum_s^{3000}\hat{Y}_{ats} - Y}{Y}\right]\frac{1}{3000},$$

The mean squared error at NRFU round t for a given allocation method a is

$$MSE(\hat{Y})_{at} = \frac{\sum_s^{3000}(\hat{Y}_{ats} - Y)^2}{3000}$$

where $\hat{Y}_{ats}$ is the estimated total for sample s and Y is the population total for the studied variable.

3.3. Response Propensity Modeling

To obtain unit-level response propensities, we fit logistic regression models on ASM survey data by NAIC3 industry, creating separate models by eligibility subdomain (i.e., for single and multi-units). Each model uses frame MOS as the independent variable and response status of unit i at time t as dependent variable. Frame MOS is available for all sampled units and is highly correlated with response because the ASM subject matter experts emphasize larger sampled units in the NRFU efforts. We did not include survey weights in the propensity modeling, using analogous reasoning to Phipps and Toth (2012), who emphasized that "the main objective is to identify and understand the characteristics of an establishment that are most strongly associated with the propensity to respond to the . . . survey and not to adjust the estimator for nonresponse bias."

We fit five response propensity models per NAICS3 industry/subdomain using the multi-unit data (the ineligible units in the simulation) and the single unit data (the eligible units in the simulation). The initial model defines a respondent as a unit that provided a

completed questionnaire during the first twelve days of collection. To fit the next model, we removed the units that responded during the initial collection period and defined a respondent as a unit that provided a completed questionnaire between the first and second round of NRFU. We continued this iterative process until we obtained five separate models. Ultimately, five different response propensity probabilities were assigned to each unit on the frame, each conditional on having *not* responded at a prior cycle.

We validated our models by comparing the predicted response propensities within domain to the observed response propensities from the ASM sample. For this, we created four size-category cells within domain using MOS quartiles and compared the average predicted response propensities to their observed counterparts from the ASM sample. Table 1 provides the final averaged conditional response propensities by NAICS3 code and subdomain (eligible and ineligible). The conditional averaged response propensities show a clear decline in overall probability of responding at each stage of NRFU, except for the last round, which contains the longest collection period.

These response propensity estimates should not be used for inference about the ASM. They are simply valuable inputs to the simulation.

3.4. Results

3.4.1. Allocations

As mentioned above, the ASM NRFU procedures are changing, beginning with the 2015 collection, and the effects of these changes on conversion rates are difficult to predict. Instead, we have fixed the nonrespondent conversion rates used in allocation (Equations 2.2 and 2.3) at 0.50 for the eligible and ineligible units (q_h^e and q_h^i, respectively). These rates are not inconsistent with the historic conversion rates used in a prior study (Kaputa et al. 2014). Moreover, using constant values in the allocations removes a source of confounding in the analyses; differences in allocation are entirely due to domain sample sizes, allocation response rates, and balance/distance measures (both obtained with frame MOS).

Table 2 presents the average response rates, allocation weights, and subsampling rates for each allocation method. The second column (URR before subsampling) provides the ratio of the count of units that provided a questionnaire prior to nonrespondent subsampling to the total count of ASM sampled units (may include out-of-scope units). Technically, these are check-in rates, not unit response rates, as the quality of the completed response data is not validated at this point in the ASM survey data collection. See Thompson and Oliver (2012). For simplicity, hereafter the term URR refers to the proxy rates.

The allocation URR in Table 2 is the predicted final response rate for all units, assuming the eligible units receive no further NRFU. The target URR for the minURR allocation procedure method is 0.68 and the target URR for the weighted allocation methods (BminURR and DminURR) is 0.72.

Subsampling rates range from zero to one, where zero indicates all eligible units that will receive no further NRFU and one indicates all eligible units that will receive further NRFU (full NRFU). In general, the ASM is well-balanced on MOS (Thompson and

Table 1. Average predicted conditional response propensities, with sample size provided as number of establishments and response propensities in percentages.

	Ineligible						Eligible					
NAICS3	Sample size	Initial	NRFU 1	NRFU 2	NRFU 3	NRFU 4	Sample size	Initial	NRFU 1	NRFU 2	NRFU 3	NRFU 4
311	3,766	11	31	22	14	51	801	27	32	23	21	27
312	508	7	15	6	13	68	162	30	37	32	15	23
313	358	16	26	22	12	51	133	31	30	27	10	22
314	256	14	20	20	20	42	250	27	34	23	15	22
315	166	16	28	23	13	37	456	20	19	22	20	17
316	78	15	23	33	21	22	105	30	23	11	20	11
321	1,380	17	37	16	17	56	933	35	33	29	22	28
322	1,207	8	44	27	16	56	186	33	42	34	17	19
323	1,005	16	21	13	11	34	1,020	33	33	29	15	24
324	823	15	30	34	27	63	43	16	24	21	14	23
325	2,898	9	22	26	25	52	388	35	34	21	17	25
326	2,423	15	25	20	21	45	543	33	32	26	16	27
327	2,947	13	24	27	15	50	745	32	28	25	16	22
331	1,214	14	30	17	23	39	263	25	37	27	15	20
332	4,012	16	29	21	19	46	4,969	37	36	28	19	29
333	2,723	16	26	18	22	47	1,570	34	33	34	16	37
334	1,520	12	25	18	23	38	653	32	33	26	16	32
335	894	14	22	23	15	48	279	26	34	22	19	31
336	2,007	13	29	22	22	44	413	28	27	31	17	27
337	646	13	29	16	15	42	965	36	34	27	20	30
339	899	15	26	22	19	38	987	32	30	25	14	21

Table 2. Subsampling rates for each allocation, averaged over 3,000 repeated samples.

			Allocation Weights		Eligible Unit Subsampling Rates		
NAICS3	URR before subsampling %	Allocation URR %	Balance	Distance	minURR Target = 0.68	BminURR Target = 0.72	DminURR Target = 0.72
311	41	66	0.05	0.01	0.44	0.46	0.50
312	30	59	0.05	0.01	1.00	0.98	0.92
313	41	64	0.04	0.01	0.60	0.93	0.85
314	42	59	0.06	0.02	0.75	0.97	0.94
315	36	44	0.04	0.01	0.98	0.99	0.94
316	41	55	0.06	0.03	0.80	0.96	0.91
321	51	67	0.01	0.00	0.12	0.05	0.02
322	51	73	0.03	0.01	0.00	0.00	0.00
323	44	61	0.08	0.03	0.61	0.86	0.89
324	41	69	0.02	0.00	0.03	0.50	0.32
325	32	64	0.04	0.01	1.00	0.98	0.91
326	39	65	0.01	0.00	0.57	0.11	0.05
327	38	64	0.03	0.00	0.80	0.47	0.29
331	42	67	0.03	0.00	0.22	0.49	0.35
332	51	64	0.07	0.02	0.33	0.25	0.39
333	45	64	0.05	0.01	0.43	0.52	0.54
334	40	63	0.04	0.01	0.70	0.82	0.69
335	37	63	0.06	0.02	0.90	0.99	0.97
336	40	65	0.03	0.00	0.56	0.60	0.43
337	50	62	0.07	0.02	0.44	0.66	0.68
339	45	60	0.03	0.00	0.61	0.61	0.42

Kaputa 2017). The balance and distance allocation weights reflect this. However, there are some interesting patterns:

- If the allocation URR is greater than the target URR, then the domain subsampling rate is zero. For example, see domain 322, whose allocation URR is 73 percent.
- The minURR method always takes larger subsamples in low responding domains and smaller subsamples in high responding domains.
- The weighted allocations methods more often select larger subsamples in unbalanced domains than their minURR counterpart. For example, NAICS3 323 has highest imbalance measure. The minURR subsampling rate is 0.61, whereas the BminURR and DminURR subsampling rates are 0.86 and 0.89, respectively.

3.4.2. Quality Measures

Table 3 contains the survey-level RBE, MSE, average final response rates and an estimate of cost from each allocation method for *all units* (which includes the responses from eligible and ineligible units) and from the *eligible units* (which comprises single unit establishment responses prior to subsampling and responses from subsampled units). Detailed industry-level results can be provided upon request. We use the full NRFU results as our quality baseline. Note that this baseline is not the "gold standard", but provides a frame of reference. The same NRFU procedures are used in all allocation scenarios, and the estimates that include subsamples will have higher sampling variance by design.

Obviously, full NRFU is more costly. Unfortunately, we cannot provide an accurate estimate of cost, so we provide a relative cost. Due to the bulk mailing procedures, combined follow-up with other surveys, and ongoing changes in collection procedures, the subject matter experts could not provide detailed cost-per-unit estimates. Instead, we used an unrealistic cost model: cost of USD 1/mailed letter and cost of USD 2/certified letter. Therefore, the presented average cost does not reflect real survey cost. It simply illustrates the effect of nonrespondent subsampling on total survey cost in our scenario. For the eligible units, the URRs decrease by approximately 13 percent. Still, including the ineligible subdomain units in the URR greatly mitigates this effect overall.

For payroll, the RBEs and MSEs are similar at the survey (all-unit) level for the three allocation methods. At first glance, it appears that the full NRFU estimates are clearly superior. Technically, this is true. However, it is important to note the differences in respondent sample composition for eligible units and all units (survey level). Notice that the subsampling reduces the *magnitude* of the nonresponse bias for eligible units. Interestingly in the eligible subdomains, the full NRFU estimates are more negatively biased than their subsampled counterparts when aggregated across all domains. This is an outcome of the allocation procedure, which uses *all* of the units (eligible and ineligible) to compute the allocation response rates, balance measures, and distance measures. Here, these survey level measures tend to indicate positive bias. The BminURR and DminURR procedures attempt to reduce the bias at the survey level by allocating units to imbalanced domains as well as to domains with low response rates. In an imbalanced domain, the respondents are systematically different from the nonrespondents. In our application, the positive bias and the non-zero distance in these domains indicate that their respondents are larger than the nonrespondents (in terms of MOS). These domains are identified in the

Table 3. Quality measures for complete sample and subsampled units.

		Relative bias %		Mean squared error			
	Method	Payroll	Receipts	Payroll	Receipts	Response rate %	Modelled cost
All units	Full NRFU	1.53	2.34	7.56E + 15	1.80E + 18	79.45	196526
	minURR	1.65	2.64	8.72E + 15	2.26E + 18	75.14	183715
	BminURR	1.61	2.48	8.43E + 15	2.01E + 18	75.14	183709
	DminURR	1.60	2.54	8.29E + 15	2.12E + 18	75.18	183701
Eligible units	Full NRFU	− 1.05	0.18	1.41E + 14	1.73E + 15	80.37	56536
	minURR	− 0.48	2.70	1.17E + 14	3.40E + 16	67.45	43724
	BminURR	− 0.64	1.09	1.81E + 14	1.79E + 16	67.46	43720
	DminURR	− 0.64	1.75	1.57E + 14	2.44E + 16	67.57	43713

weighted allocation procedures – they may not be by the minURR procedure – and the resultant subsampled domains will tend to contain smaller cases. An example of this occurs with NAICS 324. Recall from Table 2 that the minURR method subsamples at a rate of 0.03, nearly ending follow-up, whereas the BminURR and DminURR subsample at a rate of 0.50 and 0.32 respectively. These larger subsamples reduced the relative bias in this domain substantially for payroll and receipts when compared to those obtained using the minURR method; in fact the BminURR method nearly produces results equivalent to full NRFU. On the other hand, using the weighted allocation does not detrimentally affect the relative bias in NAICS 327, where there was no strong evidence of a sample imbalance prior to subsampling. Although the realized subsampling rates are different with the three allocations – 0.80 (minURR method), 0.47 (BminURR), and 0.29 (DminURR) – the ultimate result was relative bias estimates below 1 percent for all methods and variables.

Recall that the allocation procedures are all designed to reduce nonresponse bias. A consequence of the quadratic program is that the domain subsampling rates can be quite variable, in turn increasing the sampling variance. The subsampling rates obtained with the weighted allocations tend be more homogeneous, since the procedures use two different criteria for allocation. In our application, all estimates include a constant sampling variance term (for the original sample). With payroll, the full NRFU MSE can be viewed as the sum of the sampling variance plus a squared bias term (nonresponse). The minURR payroll MSE includes the sampling variance from both stages, but has reduced bias squared component. Here, the subsampling variance component is large due to the variable subsampling rates. In contrast, the weighted allocations exhibit a better balance between reduced survey level bias and increased variance due to subsampling.

With receipts, the results are slightly different. First, the full NRFU estimates in the eligible unit subdomains are nearly unbiased (0.18%), whereas the survey level estimate is biased. Again, it appears that the ineligible unit estimates are positively biased. Intuitively, this makes sense given the emphasis on obtaining responses from larger units in the contact strategies. Receipts are positively correlated with payroll (Thompson and Kaputa (2017) report average correlations of approximately 85% for single unit strata), and one would generally expect units with large payroll to have large receipts. The majority of the nonresponse bias in the receipts estimate comes from these larger ineligible unit subdomains. In contrast, the units in the eligible subdomains are more homogeneous with respect to receipts.

Interestingly, the subsamples from weighted allocations are more balanced than those obtained from the unweighted allocation, accordingly reducing the RBE and MSE values. This is probably a function of both the allocation procedures and the systematic sampling. We suspect but cannot prove that the measure of sample imbalance on payroll *underestimates* the corresponding measures for receipts. In other words, an apparently small imbalance for payroll could be much larger for receipts. The BminURR allocation selects slightly larger subsamples in domains with relatively large imbalances. Recall that the units within sampling domain are sorted by MOS (essentially payroll) before systematic subsampling. Of course, receipts are not always a strictly increasing function of payroll: large businesses can operate at loss; small businesses could have unexpectedly good revenue. A realized systematic subsample for receipts could contain an atypically high proportion of large or small units. If this happens, then the unweighted (minURR)

allocation would yield subsamples with equal response rates but unrepresentative samples on receipts. In contrast, the BminURR allocation would be more robust, as it appears to be here. Finally, we hypothesize that improvements shown by the BminURR allocation over the DminURR allocation may be a consequence of the distance measure's sensitivity to differences in sample size between respondents and nonrespondents within domain.

4. Conclusion

Nonrespondent subsampling was originally proposed over sixty years ago in Hansen and Hurwitz (1946). From the beginning, the driving motivation for the subsampling was cost savings, but the allocation strategies were designed to minimize the expected increases in sampling error caused by the additional stage of subsampling, and these increases were estimated under the assumption of 100 percent response from the subsampled units. This latter assumption is generally not true. In reality, additional or different outreach efforts are used to convert nonrespondents, especially in hard-to-reach populations. This is the central tenet of the responsive design approach introduced in Groves and Heeringa (2006).

In this article, we are concerned with the sample design aspect of a response or adaptive design. For this, we use response propensities modeled from historic data and assumed nonrespondent conversion rates in our optimal allocation. We do not alter the underlying response propensity models for the subsampled units, implicitly assuming unchanged contact strategies. The objective of the allocation is to obtain a balanced sample of respondents in all survey domains for a multi-purpose survey, noting that other optimal allocation strategies may be preferable for a survey with one or two key items or with more auxiliary information. We are encouraged by the promising results from our simulation study. Of course, if the respondent set is balanced on a characteristic before subsampling, then the simpler allocation approach presented in earlier work might be preferable. That said, our example shows the benefits of taking the respondent sample balance and distance into account in the allocation, even when the respondent set appears to be fairly balanced. In our application, guarding against potential imbalance ultimately improved the subsampled estimate quality for both items. Moreover, if the respondent sample is truly balanced before subsampling occurs, then the domain weights will be approximately equal, which effectively yields the unweighted allocation.

We find it difficult to make a general recommendation on the choice of weights in the optimization. The balance indicator tends to be overly optimistic and distance measure is very sensitive to sample size, especially with skewed populations. Based on our results, we are inclined to favor the weighted allocation that makes use of the balance indicator, which tends to favor equalizing response rates but does not ignore deficiencies in very imbalanced domains. In practice, however, we prefer to produce both allocations and investigate the cases where large differences exist.

The increased variability in design weights and reduction in response rates are less than desirable effects caused by subsampling. Moreover, an adaptive or responsive design would apply different contact strategies to the subsampled units (e.g., Schouten et al. 2013), such as personal contact or certified mailings. If these contact strategies are expensive, then any cost savings could be lost. And of course, if cost is not greatly reduced, then the sacrifice in precision may not be justified.

Nonrespondent subsampling designs can be developed and evaluated with little or no field work, especially when historic data are available. Measuring the effects of different contact strategies cannot be accomplished by simulation. Field tests are necessary, gathering data from focus groups or via designed experiments. Several agencies are conducting such experiments: see Marquette et al. (2015), Wilson et al. (2016), and Thompson and Kaputa (2017) for recent examples on establishment surveys. The full benefits for survey quality of implementing our proposed subsampling design cannot be evaluated without field testing that combines the probability subsampling with alternative field procedures.

5. References

Andridge, R.R. and R.J.A. Little. 2011. "Proxy Pattern-Mixture Analysis for Survey Nonresponse." *Journal of Official Statistics* 27: 153–180. Available at: http://www.jos.nu/Articles/abstract.asp?article=272153.

Andridge, R.R. and K.J. Thompson. 2015. "Using the Fraction of Missing Information to Identify Auxiliary Variables for Imputation Procedures via Proxy Pattern-Mixture Models." *International Statistical Review* 83(3): 472–492. Doi: http://dx.doi.org/10.1111/insr.12091.

Baker, R., J.K. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, and R. Tourangeau. 2013. "Summary Report of the AAPOR Task Force on Non-Probability Sampling." *Journal of Survey Statistics and Methodology* 1(2): 90–137.

Groves, R.M. and S.G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society Series A* 169: 439–457. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00423.x.

Hansen, M.H. and W.N. Hurwitz. 1946. "The Problem of Non-Response in Sample Surveys." *Journal of the American Statistical Association* 41: 517–529.

Kaputa, S.J., L. Bechtel, K.J. Thompson, and D. Whitehead. 2014. "Strategies for Subsampling Nonrespondents for Economic Programs." In *Proceedings of the Section on Survey Research Methods*, August 6, 2014. Alexandria, VA: American Statistical Association.

Little, R.J. and S. Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31(2): 161–168.

Lohr, S.L. 2010. *Sampling: Design and Analysis* (*2nd Edition*). Boston, MA: Brooks/Cole.

Marquette, E., M. Kornbau, and J. Toribio. 2015. "Testing Contact Strategies to Improve Response in the 2012 Economic Census." In *Proceedings of the Section on Government Statistics*: American Statistical Association, August 10, 2015. Alexandria, VA: American Statistical Association.

Peytcheva, E. and R.M. Groves. 2009. "Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates." *Journal of Official Statistics* 25: 193–201.

Phipps, P. and D. Toth. 2012. "Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data." *The Annals of Applied Statistics* 6(2): 772–794. Doi: http://dx.doi.org/10.1214/11-AOAS521.

Särndal, C.E. 2011. "The 2010 Morris Hansen Lecture: Dealing with Survey Nonresponse in Data Collection in Estimation." *Journal of Official Statistics* 27: 1–21.

Särndal, C. and P. Lundquist. 2014. "Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation." *Journal of Survey Statistics and Methodology* 2(4): 361–3087.

Särndal, C., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*, New York: Springer Verlag.

Schouten, B., M. Calinescu, and A. Luiten. 2013. "Optimizing Quality of Response through Adaptive Survey Designs." *Survey Methodology* 39(2): 29–58.

Thompson, K.J. and S. Kaputa. 2017. Investigating Adaptive Nonresponse Follow-up Strategies for Small Businesses through Embedded Experiments. *Journal of Official Statistics*. Doi: https://doi.org/10.1515/jos-2017-0038.

Thompson, K.J. and B.E. Oliver. 2012. "Response Rates in Business Surveys: Going Beyond the Usual Performance Measure." *Journal of Official Statistics* 28: 221–237. Available at: http://www.jos.nu/Articles/abstract.asp?article=282221.

Wagner, J. 2012. "A Comparison of Alternative Indicators for the Risk of Nonresponse Bias." *Public Opinion Quarterly* 76(3): 555–575. Doi: http://dx.doi.org/10.1093/poq/nfs032.

Wilson, T., J. McCarthy, and A. Dau. 2016. Adaptive Design in an Establishment Survey: Targeting, Applying and Measuring 'Optimal' Data Collection Procedures in the Agricultural Resource Management Survey. *Proceedings of the Fifth International Conference on Establishment Surveys*.

Received October 2016
Revised August 2017
Accepted October 2017