Published in final edited form as:

J Surv Stat Methodol. 2014 June ; 2(2): 182–209. doi:10.1093/jssam/smu006.

Efficient Use of Commercial Lists in U.S. Household Sampling

Richard Valliant, Frost Hubbard, Sunghee Lee, and Chiungwen Chang

Institute for Social Research, University of Michigan

Abstract

Sampling households using commercial lists has the potential to reduce costs and to efficiently identify some subgroups for which target sample sizes are desired. However, the information on the lists for demographics like age is usually incomplete and inaccurate. We demonstrate that this inexact information can still be used to improve the efficiency with which some, but not all, demographic subgroups can be located during sampling. The paper also illustrates the use of nonlinear programming as a means for finding sample allocations that are subject to a variety of practical constraints. A commercial address list and data from the National Survey of Family Growth and the Health and Retirement Study are used to illustrate the calculation of allocations to strata of housing units defined by information on the list.

Keywords

commercial lists; household sampling; constrained optimization; domain sampling; nonlinear programming; sample allocation

1. Introduction

Residential address lists sold by various commercial vendors are being used extensively in the U.S. for sampling households. Use of these lists is known as *address-based sampling* (ABS) and is seen as a potential remedy for the rising costs associated with manually listing residential housing units (HUs) by field personnel. The development of an HU frame through field listing may take months. On the other hand, if a commercial list of addresses is of high enough quality, it can be used to directly sample households. Iannacchione, Staab, & Redden (2003) describe one of the first uses of ABS in the U.S. They used a residential mailing list frame as a cost effective way to randomly select and survey approximately 15,000 households in 2000 for The Dallas Heart Study (Victor, et al., 2004). A random subset of the respondent pool was asked to provide blood and urine samples at the end of the interview. This second-phase sample required in-person data collection. The research on ABS, including coverage of this type of frame, is synthesized in a paper by Iannacchione (2011). All ABS sampling frames in the U.S. generally derive from a single source—the U.S. Postal Service (USPS) Address Management System (AMS). Information contained in this system includes: street name and number, mailbox number (if appropriate), city, state, nine-digit ZIP Code, the delivery-sequence number (order that a USPS carrier delivers mail),

and address vacancy indicators. Only commercial vendors that apply and qualify for a license may access the USPS-AMS data through an electronic file called the Computerized Delivery Sequence (CDS) file. According to Iannacchione (2011), the CDS in combination with another USPS file, called the CDS No-Stat file, which contains mainly rural addresses (see Shook-Sa, Currivan, McMichael, & Iannachione, 2013), has nearly complete coverage of the U.S. household population for mail surveys.

In addition to the USPS-AMS characteristics, commercial vendors sell “enhanced” versions of the CDS. The enhancements can include landline telephone numbers, a name associated with the address, Spanish surname indicator, estimated age and/or race/ethnicity of the head of household, as well as some geocoded (i.e., latitude, longitude, and Census tract) information. Latitude and longitude can be used to map ABS geography to Census blocks (U.S. Census Bureau, 2011), groups of which are often used in household sampling. The quality of this information depends on factors such as the address-telephone number match rate, recency of the household contact information, geocoding errors attributed to the approximate mapping of the mailing address to a physical location, and the number of HUs linked to a particular mailing address. At least four advantages of the CDS are noted in the literature:

1. Data collection costs may be reduced either by using the list frame information directly for sampling or by providing listing personnel with starting information. A list of sampled addresses can be used in combination with a technique called half-open interval (HOI) (Eckman & O’Muircheartaigh, 2011; Deming, 1960) to pick up addresses not on the list. This approach is not without flaws. For example, articles cited in Iannacchione (2011) point to varying levels of undercoverage in ABS frames for rural areas of the U.S. Additionally, Eckman (2010) found that field staff tend to ratify that the starter list is correct rather than checking for omissions. In fact, Eckman & O’Muircheartaigh (2011) recommend that HOI be used only in cases where the frame has a good deal of undercoverage.
2. Unclustered samples can be selected within primary sampling units (PSUs), which will reduce the standard error of many estimates. The cost implications can be minimal if the size of the PSUs, measured in terms of travel distance for field staff, is not large.
3. If accurate, the household-level items, like race/ethnicity and age, can improve the efficiency of sample identification by enabling the targeting of different groups.
4. If the contact information is correct, then several modes may be used to contact sample members with the ultimate goal of increasing response rates.

Household samples that require personal interviews generally select samples in several stages. First, PSUs, that in the U.S. are usually counties or groups of counties, are selected. At the second stage, secondary sampling units (SSUs) are selected that may be block groups or aggregations of individual blocks defined by the Census Bureau. Depending on interests of the survey, the frame of SSUs may be stratified by characteristics like percent of Black or

Hispanic persons collected in the decennial census or, more recently, the American Community Survey. A list of housing units is compiled within each sample SSU. A sample of HUs is selected, households are identified and data are collected. Commercial lists can be used in several ways in designing these samples and collecting data. Given a list of counties or census blocks, vendors can provide the addresses (that correspond to HUs) that fall into those areas. The lists can then be used as described in (1) or (2) above.

The addition of demographic data by some vendors of commercial lists has opened up the possibility of stratifying by characteristics like age or race/ethnicity of head of household as suggested in (3) above. The difficulties with using the information are that many addresses have no data and, in cases where data are present, it may be incorrect. Whether this incomplete and inaccurate information can still be useful for sampling is the main subject of this paper. Use of inaccurate information has also been studied by Waksberg, Judkins, & Massey (1997) who examined the effects of out-of-date census information for designing household surveys. Amaya, LeClere, Fiorio, & English (2014) studied the accuracy of some fields on the Postal Service file, including the vacancy and drop-point indicators, and concluded that even though the information was sometimes inaccurate, it could still be useful for sampling. Brock-Roth, Montaquila, & Han (2013) previously examined whether educational attainment, Hispanic ethnicity, household income, and home tenure from a list could be effectively used for stratification. They concluded that the list information was not particularly useful because of high rates of missing values. We use a different approach to analysis here that is described in the following sections. Our general approach is to allocate a sample in a way that minimizes the cost of field data collection subject to constraints on design effects due to weighting. The target size of the sample should be determined to achieve precision goals for full population or domain estimates—a step that we do not cover here. Sections 2 and 3 present two examples of how samples can be allocated using stratifying information from commercial lists. In the section 2 example, the goal is to achieve target sample sizes of different types of persons. In section 3 we determine an allocation to obtain target numbers of housing units that contain persons of specified age and race/ethnicity. The technique used here is to minimize survey costs subject to constraints on domain sample sizes, variation in survey weights, and other survey parameters. Section 4 is a conclusion.

2. An Example Using National Survey of Family Growth Data

Commercial lists can be used in a household survey to help reach a target number of persons, as we illustrate in this section using the National Survey of Family Growth (NSFG). Using a national probability sample of households, NSFG gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health (Groves, Mosher, Lepkowski, & Kirgis, 2009). The NSFG is an area probability sample with 197 PSUs; approximately 5,300 persons are interviewed per year. In the NSFG screener, a roster of all persons is collected from each responding HU along with limited demographic data. In particular, the age of each person is obtained and is the variable used to define domains in this section. Although the target age range in NSFG is 15–44, the age of each person in the housing unit (HU) is obtained during screening.

The particular commercial vendor used for the analyses in this paper was Marketing Systems Group (MSG), which was also used by Brock-Roth et al. (2012). However, our analyses are not limited to a single vendor because MSG aggregates information from other companies. At the time of this analysis, MSG aggregated information from four other companies to produce its proprietary address list. To be on the list, at least one piece of commercial data must be present for an address. The four companies used by MSG were Targus, Experian, InfoUSA, and Acxiom. Targus provides daily updates to MSG; the other three companies provide updates every other month¹. In addition, MSG purchases the CDS file from the USPS periodically and merges the commercial data from its proprietary address list to the CDS file. When MSG can match an address from its proprietary list to the CDS file, MSG can provide information on, at most, two persons for each address. At the time of this study, MSG did not provide addresses from the No-Stat file. As noted in Shook-Sa et al. (2013) almost 90% of the addresses on that file were for vacant HUs.

For the period September 2011 – December 2012, NSFG data on respondents to the screener were compared to information obtained from MSG. The weighted screening response rate (corresponding approximately to AAPOR rate RR4) during this period was 94%, accounting for a nonresponse follow-up subsample. The listings from NSFG constituted a full, national sample. A total of 24,346 sampled HUs were sent to MSG for matching. MSG reported whether the HU was in the MSG files and, if so, data on individual persons in the HU.

Table 1 shows counts of persons by age group estimated from the NSFG screener compared with estimates based on the MSG data for the same addresses (HUs). The weights are base weights for HUs that inflate the sample to the 50 United States. These weights do not include a nonresponse adjustment or any type of calibration to population controls. The commercial list contained 58.7% of the total number of persons actually found by NSFG. About 36.8% of the MSG-persons had no age information. In addition, the commercial list provides no information on persons under 18 years old. Coverage is poorest for 18–24 and 25–29 at 9.1% and 21.3%. Coverage of persons age 40 and over is generally better than for the ages younger than that.

Finding a set of sampling rates for households that will produce a desired number of HUs containing eligible sample persons is a fairly complex problem for several reasons. A commercial address list such as MSG's is helpful because it allows some addresses that are likely to be eligible HUs to be oversampled. Commercial lists, however, will be less than perfect as the information is unavailable, incomplete, or inaccurate. Having several age-ethnicity domains of interest with target sample sizes in each further complicates the problem. An allocation can be made by using equal probability sampling, or given appropriate data, an allocation can be found using nonlinear programming (NLP). NLP is an extremely useful technique for finding allocations in complex problems where it is impossible to obtain a direct, closed-form solution. Methods for finding approximate solutions have been developed in the field of operations research. Winston & Venkataramanan (2003) provide an easy introduction to the subject. Although NLP appears in the literature for sample allocation, it may be under-used in practice. Bethel (1989) gives a

¹As of 1 January 2013, MSG no longer uses Acxiom.

specialized nonlinear algorithm for some stratified allocation problems. Valliant, Dever, & Kreuter (2013) give a number of examples of the use of NLP for sample allocation. One of the more complex applications is the multistage sample allocation for the U.S. Consumer Price Index, which is based on NLP (Leaver, Johnson, Baskin, Scarlett, & Morse, 1996).

To illustrate how a commercial list might be used for a sample targeted at certain age groups, we consider two groups using the NSFG data: 18–44 and 65+. We use the following notation:

D = target age domain (18–44 or 65+)

h = sampling substratum based on commercial list information on individual addresses

1. List has record; 0 persons in age group
2. List has record; 1 or more persons in age group
3. List has record; no age information
4. No List record

(We use the term substratum because in the NSFG, and other similar household surveys, higher-level strata of PSUs are used. Note that the commercial list classification into substrata 1 and 2 will be inaccurate for some addresses.)

$p_h(D)$ = proportion of HUs in sampling substratum h that have at least 1 person in the age domain;

$a_h(D)$ = average number of persons per HU in sampling substratum h that are actually in domain D ; this average is based on all HUs including those with no persons in the domain;

n_h = number of sample HUs allocated to stratum h ;

$n(D) = \sum_h n_h a_h(D)$ = expected number of sample persons found to be eligible by being in domain D .

To approximate costs, suppose that the cost of screening and dropping an ineligible HU is c_S and the average cost of screening a HU and interviewing all persons in an eligible HU is c_{S+I} . The expected cost of a randomly selected sample HU in sampling stratum h , when screening is done to locate a member of domain D , is

$$c_h(D) = p_h(D)c_{S+I} + [1 - p_h(D)]c_S.$$

Assume that a self-weighting sample of HUs will be selected within each substratum, i.e., all sample HUs in a substratum will have the same selection probability. To accomplish this, the sampling rates used within geographic segments would be adjusted to account for the prior stages of sampling. The overall sampling rates will be allowed to vary from one substratum to another.

When the sample is allocated to the sampling strata, the sampling rates may vary quite a lot among the substrata even if a self-weighting sample is selected within each substratum. We can attempt to control this by constraining the “design effect due to weighting” (Kish, 1965), denoted by $deff_w$ here:

$$deff_w = 1 + \frac{1}{n\bar{w}^2} \sum_{i \in s} (w_i - \bar{w})^2 \quad (1)$$

where s denotes the set of n sample HUs across all substrata, w_i is the weight for sample HU i , and $\bar{w} = \sum_s w_i / n$. The w_i 's are inverse selection probabilities that account for all stages of sampling. If the sample is multistage, as are most household samples in the U.S., a design effect that accounts for both differential weighting and homogeneity of units within clusters would be more appropriate (Kish, 1987; Gabler, Haeder, & Lahiri, 1999). This more elaborate $deff$ is equal to $deff_w$ times the factor, $1 + (\bar{m} - 1)\rho$, where $\bar{m}$ is the average number of HUs selected per cluster and ρ is an intraclass correlation for a particular variable collected in the survey. In the following illustration, we deal only with $deff_w$. When the $deff_w$ is computed from just the units in domain D , we denote it by $deff_w(D)$.

There are a variety of ways that an optimization problem can be set up to control the allocation of the sample to substrata. The formulation here is to minimize total cost subject to constraints on the number of sample persons obtained in the age group, a minimum of 250 HUs allocated to each substratum, the number of sample HUs must be less than the population count (which has no effect here since the population is large), and the design effect due to weighting is less than or equal to 1.5, denoted as d_0 . This target value is not particularly large for U.S. household surveys. For example, in the 2007 National Survey on Drug Use and Health the design effects due to weighting for national estimates for age groups ranged from 1.63 to 3.39 (see Table 3.6 in (RTI International, 2012)). The target design effect for prevalence estimates for sex-age domains in the 1999–2006 National Health and Nutrition Examination Surveys was 1.5 (see p. 3 of National Center for Health Statistics, 2012).

To make the effective sample size of eligible persons the same as for an equal probability sample of $n_{eq}(D) = 2,000$, we set the constraint of expected, effective number of age eligibles to be $n_{eff}(D) = 2,000$ where $n_{eff}(D) = n(D)/deff_w(D)$ for each age group. Without this adjustment, the NLP allocation will look unrealistically good compared to an equal probability allocation because its total sample size will be much smaller, but the precision of estimates would be worse than those for the equal probability allocation. The statement of the optimization problem is:

Objective: Find $\{n_h\}_{h=1}^H$ to minimize total cost, $C_D = \sum_{h=1}^4 n_h c_h(D)$;

Subject to the constraints:

$$n_h \geq n_{\min} \text{ for all } h$$

$$n_{eff}(D) = n_{eq}(D)$$

$n_h \leq N_h$ with N_h being the number of HUs in the population in substratum h .

$deff_w(D) = d_0(D)$, a fixed constant

Although the objective function is linear in the substratum sample sizes, the design effect, $deff_w(D)$, has the sample sizes in the denominators of the w_j weights, making this an NLP problem. Solving this type of NLP requires specialized software. There are several choices available—the `alabama` package in R (Varadhan, 2010), the procedures `NLP` and `OPTMODEL` in SAS® (URL: www.sas.com), and the Solver add-on (URL: <http://www.solver.com/excel-solver.htm>) which comes bundled with Microsoft Excel® are three. We used the last of these.

We set $n_{eff}(D) = 2,000$, $n_{min} = 250$, and $d_0(D) = 1.5$ as noted above. We treat the determining of sample allocations for ages 18–44 and 65+ as separate problems to compare solutions for two groups that are different proportions of the population (see Table 1). Table 2 lists the other parameters needed to solve the allocation problem— $p_h(D)$, $c_h(D)$, and $a_h(D)$. When the commercial list shows that an HU has 1 or more persons in the target age group (sampling substratum 2 in Table 2), the actual proportion of HUs is much higher than in the other substrata. For example, the proportion of HUs with 1+ persons who are 18–44 is 0.720 in substratum 2 but 0.207, 0.465, and 0.372 in substrata 1, 3, and 4.

As a surrogate for monetary cost, we use hours and set $c_S = 3$, $c_{S+I} = 10$. The expected cost per HU in substratum 2 is higher than for the other substrata because the cost includes both screening and interviewing and a higher proportion of HUs will be eligible for interviewing in substratum 2. The expected number of persons per HU in the desired age ranges is also larger in substratum 2 than in the other substrata as shown in columns 7 and 8. As estimated from the NSFG screener responses, there is an average of 1.249 persons 18–44 per HU and 0.947 persons 65+ per HU in substratum 2 and considerably fewer in the other substrata. However, the commercial list is not completely accurate since the expected numbers per HU are larger than 0 in substrata 1 and 3. In addition, substratum 4 (no list record) does have some HUs with persons in ages 18–44 and 65+, implying that some sample HUs should be allocated to that substratum to ensure complete coverage of the population.

Table 2 also shows the estimated national counts of HUs in the last two columns from the NSFG screener sample. These will be used in computing the substratum sampling rates for the optimal allocations presented in Tables 3 and 4. As estimated from NSFG (“Total” row of Table 2), about 40.9% of HUs contain persons age 18–44 while 17.6% have one or more 65+ year olds. The average number of persons per HU in age group 18–44 from Table 1 and Table 2 is 0.671 (87,369/130,199). For 65+ the average per HU is 0.236 persons (30,674/130,199). Thus, if equal probability samples of HUs were selected for a sample size of 2000 persons, the approximate number of HUs needed to obtain 2,000 persons in each age domain would be 2,980 ($\approx 2000/0.671$ for 18–44 and 8,489 ($\approx 2,000/0.236$ for 65+.

Table 3 shows the allocation solution to the NLP for ages 18–44. The constraints of 3,000 eligible persons and a design effect due to weighting of 1.5 are met exactly. Most of the sample (50%) is allocated to substratum 2; the other substrata receive allocations that range from 466 (14%) to 691 (21%). The largest share of the cost (13,088/22,196) is due to sampling from substratum 2. The bulk of the eligible 18–44 persons will also be picked up

from substratum 2 (2,033 of 3,000). Results for ages 65+ in Table 4 are similar. The design effect due to weighting for 65+ was 1.5, which was the value of the constraint.

Table 5 compares results for the NLP solution to an equal probability allocation. Note that the effective sample size of 2,000 is set by design to be the same for the NLP and equal probability allocations. The numbers of HUs to screen, which are repeated from Tables 3 and 4, are smaller for 65+ than for an equal probability sample but larger for 18–44. Another way of comparing the allocations is the expected cost of each. An equal probability allocation will (in expectation) be distributed proportional to the national counts of HUs in the last two columns of Table 2. The expected number of sample HUs in substratum h in the equal probability sample for domain D is $n_{h,eq}(D) = n_{eq}(D) W_h$ where $W_h = N_h/N$ is the population proportion of HUs in substratum h with substrata defined as appropriate for domain D . The expected cost of the equal probability sample is then

$$C_{eq}(D) = n_{eq}(D) \sum_h c_h(D) W_h.$$

Calculating the expected costs for an equal probability allocation for the 18–44 and 65+ domains gives the results in the last three columns of Table 5. The total cost of the NLP allocation is 27% more expensive for 18–44 year olds but only 81% as expensive for 65+. For 65+, the NLP screening costs are 56% of those of the equal probability allocation. Note, however, that the interviewing costs are considerably higher for the 65+ NLP allocation because more HUs have to be interviewed to obtain the same effective sample size as in the equal probability allocation.

The differences in expected numbers of eligible persons per substratum in Table 2 are more pronounced for 65+ than 18–44, leading to a gain in efficiency for sampling 65+ from the list but a loss for 18–44 year olds. The 18–44 year olds are fairly prevalent, being in 40.9% of HUs, compared to persons 65+, who are in 17.6% of HUs. Consequently, 18–44 years can be located with some regularity in an equal probability sample, implying that the more complicated allocation to substrata is not useful for that age group.

As Tables 3 and 4 show, weights in substrata 1, 3, and 4 differ from those in substratum 2 by factors of about 2.6 to 5.8 for 18–44 and 3.7 to 5.8 for 65+. Although the overall design effect is controlled to be 1.5, it may be higher for analytic subgroups within the 18–44 and 65+ domains. If this is a concern, the NLP problem can be changed to specify a larger minimum sample size per substratum. For example, if $n_{\min} = 800$ rather than 250, the solution for 65+ has total sample size of 5,320 with a cost of 31,014, which is 86% of the cost of the equal probability allocation 35,929. The ratio of largest to smallest weight then reduces from 5.8 to 4 and $deff_w(65+) = 1.28$.

3. An Example Using the Health and Retirement Study

The Health and Retirement Study (HRS, <http://hrsonline.isr.umich.edu>) is a longitudinal panel study that surveys a representative sample of more than 26,000 Americans over the age of 50 every two years. HRS periodically adds new age cohorts to its sample. The most recent were the Early Baby Boomers (EBBs) who were born 1948–1953 and the Middle Baby Boomers (MBBs) who were born 1954–1959 and who live in the 48 contiguous states

or the District of Columbia. Hispanics and non-Hispanics within those age groups were also domains of interest. This section reports analyses that were done using the EBBs and MBBs based on data from 2010 and 2011. The goal in this section is to determine a sample allocation to obtain specified numbers of HUs in several analytic domains.

The EBB and MBB cohorts were each about 8% of the total U.S. population in 2010. The EBB Hispanics and MBB Hispanics are each about 1.5% of the population. Thus, EBBs, MBBs, and especially Hispanics within those age cohorts are relatively rare groups. To attempt to obtain more age-eligible Hispanics in the sample, two steps were taken. First, segments (census blocks or combinations of blocks) were stratified by concentration of minority population. Second, age and race/ethnicity information on addresses in sample segments was purchased from MSG and used for sampling housing units.

Because HRS had an interest in oversampling both Black and Hispanic respondents in 2010–11, strata of segments were defined within geographic strata of PSUs. Four segment strata were created based on the proportion of Black and Hispanic persons within the block group that contained each segment, as shown in the 2000 decennial census. Only the segment strata whose population was more than 10% Black, more than 10% Hispanic, or both were sampled in 2010–11 and provide the data for the analysis in this section. We refer to these as *high minority* segments subsequently; they constitute about 35% of all segments in the sampling frame. The supplementary, online material provides a more complete description of the HRS 2010–11 design. In the discussion in the remainder of section 3, we use the following terminology:

- *Segment stratum*—strata into which area segments were classified based on decennial census data;
- *HU sampling substrata*—as described in section 3.1, these were used for sampling HUs based on the commercial list information; and
- *HRS analysis domains*—these are groups of persons to which target sample sizes are assigned.

3.1 Stratification Based on Commercial List Information

Listed HUs in the segments in 2010–11 were provided to MSG, which attached demographic information to listings for which it had data. For households that cooperated with the screener, we could determine how accurate the MSG information was by comparing to the data provided by the respondents. An important limitation of this analysis is that it covers only a subset of the segments in the entire frame. Whether the MSG information for cases that are in segments with less than 10% Black and less than 10% Hispanic populations have the same accuracy as those with higher concentrations of these groups could not be evaluated with the data we had. Nonetheless, the following is useful in the sense of illustrating how to formulate an allocation problem with multiple goals and constraints that is more complicated than the one in section 2.

We used MSG data on the two persons in the household as well as the Hispanic ethnicity status of the head of the household (assuming that the Hispanic ethnicity status of persons 1

and 2 were the same) to classify HUs in the 2010–11 sample segments into one of six HU substrata to be used for sampling:

1. 1 or more MBB Hispanic persons in the HU (MBB H)
2. 1 or more MBB non-Hispanic persons in the HU (MBB NH)
3. 1 or more EBB Hispanic persons in the HU (EBB H)
4. 1 or more EBB non-Hispanic persons in the HU (EBB NH)
5. no EBB or MBB persons in the HU (0 MBB EBB)
6. Unknown whether the HU contained an EBB or MBB person based on MSG data. Substrata 1–4 contained persons who were anticipated to be eligible for the EBB or MBB recruitment based on MSG data. As in the NSFG example, assume that a self-weighting sample of HUs will be selected within each substratum and that the sampling rates will be allowed to vary from one substratum to another.

Table 6 provides an aggregated summary based on the classification using MSG data. The unweighted analyses show that of 28,164 addresses sent to MSG about 50% had matched data; 29% were correctly classified into one of HU substrata 1–5 while 21% had MSG data but were incorrectly classified. Fifty percent (unweighted) of the addresses had no MSG data. Table 6 also shows weighted estimates. The weights inflate the sample to the national level for the high minority segments only. An estimated 62% (=40% + 22%) had data on the MSG files. The weighted, correct classification rate was 40% compared to 29% unweighted. The fact that the unweighted rate is lower may be due to the fact that the HRS segment sample emphasizes areas with higher concentrations of minorities, which may be covered less well by the MSG proprietary address list than other types of areas.

The breakdown in the online supplement of the addresses for which MSG had no information (HU sampling substratum 6) is also informative. Twenty-five percent of those cases were unoccupied HUs; 49% had no EBBs or MBBs. Thus, almost three-quarters of the addresses in this substratum had no HUs that were of interest to HRS. Although MSG does code some HUs as unoccupied, we found that this code was seldom used and was not useful in identifying actual unoccupied units². All HUs coded as unoccupied by MSG were treated as Unknown. As shown later, an efficient sample design will sample substratum 6 at a low rate.

The following example illustrates how such an allocation might be determined. In this illustration, both EBB and MBB analytic domains of HUs are considered. In addition, we distinguish between Hispanic and non-Hispanic HUs. To set up the problem, we use the following notation:

D = analytic domain in HRS: 1=MBB/Hispanic; 2=MBB/non-Hispanic; 3=EBB/Hispanic; 4=EBB/non-Hispanic; in addition, define three other domains: 5=no MBB/EBB, 6=Unknown due to being a nonrespondent in HRS, 7=unoccupied HU;

²If addresses from the No-Stat file were included in the MSG file, the percentage of vacant addresses would rise substantially.

h = HU sampling substratum based on MSG information on individual addresses; $h = 1, \dots, 6$;

n_h = number of addresses (HUs) allocated to sampling substratum h ;

$p(h)D$ = proportion of HUs in sampling substratum h that are actually in HRS analysis domain D ;

$n(D) = \sum_h n_h p_h(D)$ = expected number of HUs eligible by being in HRS analysis domain D .

Table 7 shows estimates of $p_h(D)$ within each of the six sampling substrata based on how the HUs responded in the screener. The weights used for estimation inflated the sample to the universe of HUs in the high minority segment strata. The overall eligibility rate was 0.226, and the rates differed across MSG substrata. HU sampling substrata 1–4, that were anticipated to have at least 1 eligible person in some HRS domain, do have higher proportions of eligible HUs than HU sampling substrata 5 and 6 that were expected to have no eligibles or that had no data. The proportions of eligible HUs range from 0.521 to 0.696 in HU sampling substrata 1–4 and are 0.148 in sampling substratum 5 and 0.140 in sampling substratum 6.

Note that 0.230 of the HUs for which MSG had no data (sampling substratum 6) turned out to be unoccupied. The rates of unoccupied for sampling substrata 1–5 are much lower—ranging from 0.028 to 0.061. Inspecting these proportions, it is clear that the payoff from sampling substrata 5 and 6 will be very low. Another point to note is that the proportions of eligible households for the two rarest groups (MBB/Hispanic and EBB/Hispanic) are low. The proportions of households based on HRS screener interviews classified into these two domains were 0.021 and 0.038, whereas MBB NH and EBB NH domains were 0.078 and 0.090. Thus, considerably more households must be screened to locate eligible HUs in the MBB H and EBB H domains than for MBB NH and EBB NH domains.

3.2. Equal Probability Allocation

Using the proportions of eligible HUs in the total row of Table 7 (denoted as $p(D)$ in the following), we can estimate the number of HUs in an unstratified, equal probability sample that would have to be screened in order to obtain a specified number of sample eligibles in each domain. Suppose $n_{eq}(D)$ denotes the target number of sample HUs we want to obtain in domain D through equal probability sampling. The number of HUs that need to be screened is $n_{eq}(D)/p(D)$. Doing this calculation individually for each domain gives the approximate sample sizes in Table 8.

The targets in Table 8 are 1,500 for domains 1 (MBB H) and 3 (EBB H) and 3,000 for domains 2 (MBB NH) and 4 (EBB NH). For example, if an equal probability sample of HUs was selected, 38,352 ($\hat{=} 3000/0.078$) HUs would have to be screened in order to obtain 3,000 HUs in domain 2. The number to screen to hit the targets for all four domains would be determined by the domain with the smallest average proportion of eligible, $p(D)$. In our case, domain 1 (MBB H) has the smallest value at 0.021. To obtain 1,500 HUs in that domain, 72,951 ($\hat{=} 1500/0.021$) HUs would have to be screened to produce the desired samples for all domains. This would be wasteful because the desired numbers of eligibles

for substrata 2, 3, and 4 could be found by far fewer screening attempts because HUs that contain those age-ethnicity groups are more prevalent.

3.3. Allocation based on Non-Linear Programming

A more cost-efficient allocation can be found by accounting for the fact that the proportion of eligible HUs differs among the sampling substrata. As in the NSFG example, we minimize costs subject to constraints that at least a minimum number of HUs be selected from each substratum and that a desired expected number of HUs is obtained in each domain.

Using the same notation as in Section 2, the expected cost of a sample HU selected at

random in sampling substratum h is $c_h = p_h c_{S+I} + (1-p_h) c_S$ where $p_h = \sum_{D=1}^4 p_h(D)$ is the proportion of HUs in sampling substratum h that are in the eligible domains 1–4. The expected cost of the sample is then the sum of the costs over the six sampling substrata:

$$C = \sum_{h=1}^6 n_h c_h. \quad (2)$$

To equalize the effective sample sizes in the NLP allocation with the equal probability targets, $n_{eq}(D)$ in Table 8, we constrain the effective size in domain D , $n_{eff}(D) = n(D)/deff_w(D)$, to be at least equal to $n_{eq}(D)$. Given this preparation, a formal statement of the optimization problem is as follows.

Objective: Find $\{n_h\}_{h=1}^H$ to minimize total cost in (2)

Subject to the constraints:

$n_h \geq n_{\min}$ for all sampling substrata, h ;

$n_{eff}(D) \geq n_{eq}(D)$

$n_h \leq N_h$ with N_h being the number of HUs in the population in sampling substratum h ; and $deff_w(D) \leq d_0(D)$, a fixed constant.

In this example, we used the following values. First, a minimum number ($n_{\min}$) of 1,000 HUs was to be selected from each of the six sampling substrata. Hours per case were used as a surrogate for costs— $c_S = 3$ and $c_{S+I} = 10$. The bound on the weighting design effect in all domains was $d_0(D) = 1.5$. The sample sizes of housing units based on the NLP solution that would produce these targets are shown in Table 9.

A total of 60,818 HUs would be sampled in the initial solution shown in Table 9. The sample size of eligible cases and the effective sample sizes in the initial solution are listed in the first two rows of Table 10. The NLP algorithm was unable to find a solution that achieved the target sizes in all domains while meeting all constraints. The initial solution hit the target for MBB H and EBB NH but produced larger effective sample sizes of eligibles in the MBB NH and EBB H domains than the targets of 3,000 and 1,500. To reduce the number interviewed to those targets, we can randomly deselect a proportion of the initial sample in those

domains. Defining $f(D) = n_{eq}^O(D) / n_{eff}(D)$, the number of sample cases retained after deselection is $n^*(D) = f(D)n(D)$; the number retained in stratum h is $n_h^*(D) = f(D)n_h(D)$. The effective sample size in substratum h after deselection is $n_{h,eff}^*(D) = f(D)n_{h,eff}(D)$ where $n_{h,eff}(D) = n_h(D) / def_f_w(D)$. The retention fractions, $f(D)$, and the numbers of eligible cases after deselection are shown in rows three and four of Table 10. The sample sizes in row 4 are those after deselection. These would be deflated by the design effects shown later in Table 12 to get effective sample sizes. For example, the effective sample size for MBB H after deselection is $3,282/1.09 = 3,011$. For comparison, the expected numbers of eligible cases in an equal probability sample are shown in the fifth row. In the initial NLP allocation, there are 13,131 eligible cases which are reduced to 10,874 through deselection. An equal probability allocation produces many more eligible cases than desired—16,483 vs. 9,000. For both the NLP and equal probability allocations deselection is done in a way to yield effective total sample size of 9,000 distributed among the four domains as shown in Table 8 and the last line of Table 10.

The overall percentage of eligible HUs in Table 9 in the sample after deselection is 19% ($=10,874/58,562$); in sampling substrata 1–4 the percentage is 50%. These values compare to the overall percentage of eligible HUs expected in an equal probability sample after deselection of 12% ($9,000/72,951$). The percentage eligible in substrata 1–4 for the equal probability allocation is 34%. That is, the NLP allocation increases the proportion of eligible HUs compared to equal probability sampling.

We can also compare the weights and design effects due to weighting before and after deselection. Table 11 shows the weights for the NLP allocation. Because the retention rates, $f(D)$, vary by domain, the weights after deselection vary among HUs in a given HU sampling substratum as shown in the last four columns of Table 11. On the other hand, the values of $def_f_w(D)$ change very little after deselection. In Table 12 the range of $def_f_w(D)$ is 1.09 to 1.40 before and after deselection. The def_f_w for the full sample is 1.22 for the NLP allocation before deselection and after 1.29 after deselection. The equal probability allocation, of course, has $def_f_w(D) = 1$ in each of the four domains, and weighting def_f 's of 1.07 for MBB, 1.01 for EBB, and 1.05 for all domains combined.

The costs of the NLP and equal probability allocations can also be compared. In the NLP allocation all n_h initial sample cases in substratum h will be screened. After deselection $n_h^* f(D) p_h(D)$ cases will be interviewed in substratum h and domain D . The total cost of the NLP allocation is then

$$C_{NLP} = \sum_h n_h \left[c_S + c_I \sum_D f(D) p_h(D) \right] \quad (3)$$

For the equal probability allocation, cases can be randomly retained within each domain to obtain the targets in Table 8. The total cost of screening in the equal probability sample will be $c_S n_{eq}$ while the cost of interviewing cases in domain D is $c_I n_{eq}(D) f_{eq}(D)$ leading to a total cost for the equal probability allocation with deletion of

$$C_{eq} = n_{eq}c_S + c_I \sum_D n_{eq}(D)f_{eq}(D).$$

The cost of the NLP allocation after deselection in Table 13 is 92% of the cost of the equal probability allocation. Screening cost in the NLP allocation is 83% of that of the equal probability allocation. On the other hand, interviewing cost is 21% larger in the NLP allocation because more cases must be interviewed to obtain the desired effective sample sizes.

A final point relevant to this example and the one in the previous section is that the overall sampling rates must be feasible. This is a concern in a multistage sample where the rates at the last stage of sampling times the product of the rates at previous stages must equal the desired overall rates. This may not be possible if the products of the selection probabilities at steps prior to the last are smaller than the optimal rates in some PSU-segments. This will typically be an issue only if an existing sample of PSUs and other units prior to the last is used. If all stages of sampling are designed at once, the rates can be controlled.

4. Conclusion

Interest in the use of commercial address lists for household sampling in the U.S. has been growing over the last decade. The lists have the potential to reduce costs by eliminating or reducing the need to have field personnel compile lists of housing units. Vendors of the lists also can provide limited demographic information about the persons who live at some addresses. Although this information is incomplete and inaccurate in many cases, the demographic data can still be useful for stratifying addresses for targeted sampling. In this article, we specifically addressed the use of the lists for sampling certain age-ethnicity groups. Stratifying using somewhat error-prone data can still improve the efficiency of locating members of some domains as we have demonstrated. As noted in Brock-Roth, Montaquila, & Han (2012), the demographic information on the lists is not all equally good; the age and Hispanicity designations we used may be some of the more accurate variables. Utility of the lists also varies depending on the ages of persons in a household. In particular, older, more financially stable persons are better covered by the lists. We also found that use of a list was not effective for all domains studied. In particular, the list was less efficient than equal probability sampling for targeting a common domain, like 18–44 year olds, than a less prevalent but better covered domain, like persons 65 or older. In a more complicated allocation with sample size goals for several domains, as in the example in section 3, gains from using the list may be marginal.

Implementing the procedures illustrated here requires that relevant data be available. Estimates of some key parameters are required to use our approach. In our first example, estimates of the proportion of housing units that have at least one person in a target domain, of the average number of persons per housing unit that are in the domain, and of the population numbers of HUs are needed. These must be available for each substratum formed based on the commercial list information. In the second example, estimates of the proportion of HUs in a substratum that are in each domain are required. When designing a sample from

scratch, these may be difficult to obtain. This problem is akin to the issue of needing variance components for a sample allocation when a sample that collected the variables desired from a new survey has not been done before.

If a survey organization has done a previous household survey in which the demographics that define domains of interest have been collected, the sample can be provided to a vendor for matching as was done in Shook-Sa, Currivan, McMichael, & Iannacchione (2013). Commercial list vendors generally maintain historical files so that a sample can be matched retrospectively against a file constructed near the date of the survey. Estimates from the matched sample can then be used to design a new survey. The matching is likely to be far from perfect, as illustrated in Amaya, Skalland, & Wooten (2010), and any survey should include a stratum for address listings that are unmatched or have no demographics.

This study used a particular commercial vendor, MSG, which aggregated commercial data information from four other vendors. There are other list vendors that may be worth considering when doing household sampling. For example, Aristotle (<http://aristotle.com/>) offers lists based on voter registrations from all 50 states; age and gender data for persons on these lists may be quite reliable. Valassis Communications Inc. (<http://www.valassislists.com>) specializes in direct mail advertising and claims to have virtually complete coverage of all U.S. mailing addresses. For complete coverage of all residential addresses, using a commercial list(s) in an area sample with some checks for completeness is still prudent. However, as noted by Eckman & O'Muircheartaigh (2011), devising effective field checks is difficult.

Our results for sample allocations are specific to a case where the goal is to achieve target sample sizes for domains for which a commercial list has information. We set up allocation algorithms to obtain those sample sizes (in expectation) while limiting the increase in variances of overall estimates. The goal of obtaining sample sizes large enough to permit separate domain analyses is not unusual in household surveys. For example, the 1999–2002 and 2003–2006 National Health and Nutrition Examination Surveys (National Center for Health Statistics, 2013) set target expected sample sizes for low income persons, adolescents, the elderly, Blacks, and Mexican Americans. The National Household Survey on Drug Abuse (Gfroerer, Eyerman, & Chromy, 2002) in the past has had equal target sample sizes in the age groups, 12–17, 18–25, and 26 and older.

A point worth noting is that it makes a difference whether only screening costs are considered as opposed to screening plus subsequent interviewing cost. If only screening cost is the concern, using a commercial list is clearly efficient. If both screening and interviewing costs are accounted for, then differential sampling based on the list information is less attractive. The cost analyses presented here may also be more pertinent to a cross-sectional survey than to a longitudinal survey where persons are retained for a number of interviews. If the cost of screening in a longitudinal survey can be amortized over all future cycles, saving money on initial screening may be less of a concern. However, given how budgets are typically set, the need for near-term savings may over-ride long-term thinking.

Sample size solutions depend on how the allocation problem is formulated and what the goals of a survey are. Our approach applies where sample sizes are to be controlled. Another way of setting up a problem would be to minimize the variance of an estimator (or of some weighted combination of important estimators) subject to a fixed budget constraint as in Bethel (1989), Leaver, et al. (1996), and Valliant & Gentle (1997). The varying goals and constraints of different surveys can usually be accommodated through careful formulation of the allocation problem. These problems generally do not have closed form solutions, as is the case for simple textbook problems like Neyman or cost-constrained allocations (e.g., see Cochran 1977). However, software tools are now readily available for solving the mathematical programming problems that are used for the type of constrained optimization problems that we have examined.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially funded by the National Institute on Aging and the Social Security Administration [grant no. 3-U01-AG-009740-23-S1], and the National Center for Health Statistics [contract no. 200-2010-33976]. We are grateful to the referees for their well-informed comments that improved the presentation.

References

- Amaya A, LeClere F, Fiorio L, English N. Improving the Utility of the DSF Address-based Frame through Ancillary Information. *Field Methods*. 2014; 26(1):70–86.
- Amaya, A., Skalland, B., Wooten, K. What's in a Match?. *Survey Practice*. 2010. Retrieved from <http://surveypractice.wordpress.com/2010/12/08/what%E2%80%99s-in-a-match/>
- Bethel J. Sample Allocation in Multivariate Surveys. *Survey Methodology*. 1989; 15:47–57.
- Brock-Roth, S., Montaquila, J., Han, D. The ABS frame: Quality and Considerations. *Survey Practice*. 2013. Retrieved from <http://surveypractice.org/index.php/SurveyPractice/article/view/73/pdf>
- Cochran, W. *Sampling Techniques*. New York: John Wiley & Sons; 1977.
- Deming, W. *Sample Design in Business Research*. New York: John Wiley & Sons; 1960.
- Eckman, S. Errors in Housing Unit Listing and their Effects on Survey Estimation. College Park, MD, USA: University of Maryland; 2010. Retrieved from <http://drum.lib.umd.edu/handle/1903/10302>
- Eckman S, O'Muircheartaigh C. Performance of the Half-Open Interval Missed Housing Unit Procedure. *Survey Research Methods*. 2011; 5(3):125–131.
- Gabler S, Haeder S, Lahiri P. A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering. *Survey Methodology*. 1999; 25(1):105–106.
- Gfroerer, J., Eyerman, J., Chromy, J. *Redesigning an Ongoing National Household Survey: Methodological Issues*. Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies; 2002.
- Groves R, Mosher W, Lepkowski J, Kirgis N. Planning and Development of the Continuous National Survey of Family Growth. *Vital Health Statistics*. 2009; 1(48)
- Howden, L., Meyer, J. Age and Sex Composition: 2010. U.S. Department of Commerce; 2011 May. Retrieved from <http://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf>
- Iannacchione V. The Changing Role of Address-Based Sampling in Survey Research. *Public Opinion Quarterly*. 2011; 75(3):556–575.
- Iannacchione V, Staab J, Redden D. Evaluating the Use of Residential Mailing Lists in a Metropolitan Household Survey. *Public Opinion Quarterly*. 2003; 67(2):202–210.
- Kish, L. *Survey Sampling*. New York: John Wiley & Sons; 1965.

- Kish L. Weighting in Deft2. *The Survey Statistician*. 1987 Jun.
- Leaver, S., Johnson, W., Baskin, R., Scarlett, S., Morse, R. Proceedings of the Section on Survey Research Methods. American Statistical Association; 1996. Commodities and Services Sample Redesign for the 1998 Consumer Price Index Revision; p. 239-244.
- National Center for Health Statistics. The National Health and Nutrition Examination Survey: Sample Design 1999–2006. *Vital and Health Statistics*; Washington: 2012. Retrieved from www.cdc.gov/nchs/data/series/sr_02/sr02_155.pdf
- National Center for Health Statistics. Analytic Note Regarding 2007–2010 Survey Design Changes and Combining Data Across Other Survey Cycles. Hyattsville, Maryland: 2013. Retrieved from http://www.cdc.gov/nchs/nhanes/search/nhanes13_14.aspx
- RTI International. National Survey on Drug Use And Health: Sample Redesign Issues and Methodological Studies. Research Triangle Park North Carolina; USA: 2012 Apr. Retrieved from www.samhsa.gov/data/NSDUH/NSDUHMethodsSIMS2012.pdf
- Shook-Sa B, Currivan D, McMichael J, Iannachione V. Extending the Coverage of Address- based Sampling Frames: Beyond the USPS Computerized Delivery Sequence File. *Public Opinion Quarterly*. 2013; doi: 10.1093/poq/nft041
- U.S. Census Bureau. 2010 Census Redistricting Data (Public Law 94-171) Summary File. Washington, DC, USA: Department of Commerce; 2011. Retrieved from <http://www.census.gov/prod/cen2010/doc/pl94-171.pdf>
- Valliant R, Gentle J. An Application Of Mathematical Programming to Sample Allocation. *Computational Statistics and Data Analysis*. 1997; 25:337–360.
- Valliant, R., Dever, JA., Kreuter, F. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer; 2013.
- Varadhan, R. *alabama: Constrained Nonlinear Optimization*. 2010. R package version 2010.10-1 Retrieved from <http://CRAN.R-project.org/package=alabama>
- Victor R, Haley R, Willett D, Peshock R, Vaeth P, Basit M, Hobbs H. The Dallas Heart Study: A Population-based Probability Sample for the Multidisciplinary Study of Ethnic Differences in Cardiovascular Health. *Journal of Cardiology*. 2004; 93(12):1473–1480. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15194016>.
- Waksberg J, Judkins D, Massey J. Geographic Oversampling in Demographic Surveys of the United States. *Survey Methodology*. 1997; 23:61–71.
- Winston, W., Venkataramanan, M. *Introduction to Mathematical Programming*. Pacific Gove CA: Duxbury Press; 2003.

Estimated total numbers of persons from NSFG 2011–2012 screener compared to numbers identified by MSG

Table 1

Age group	Est. Total Persons in Age Range Identified by NSFG (1000s)	% of NSFG persons	Est. Total Persons in Age Range Identified by MSG (1000s)	% of MSG persons	Ratio of Total Persons (MSG/NSFG)%	% of persons in 2010 census
MSG has HU record						
but no ages on persons			53,966	36.8%		
Under 18	58,072	23.2%				24.0%
18–24	22,286	8.9%	2,021	1.4%	9.1%	9.9%
25–29	17,527	7.0%	3,725	2.5%	21.3%	6.8%
30–34	17,348	6.9%	7,227	4.9%	41.7%	6.5%
35–39	14,987	6.0%	6,306	4.3%	42.1%	6.5%
40–44	15,222	6.1%	9,515	6.5%	62.5%	6.8%
45+	104,383	41.8%	63,952	43.6%	61.3%	39.5%
Total	249,825	100.0%	146,711	100.0%	58.7%	100.0%
Subgroups						
18–44	87,369	35.0%	28,794	19.6%	33.0%	36.5%
65+	30,674	12.3%	22,891	15.6%	74.6%	13.1%

Source for Census 2010 percentages: Howden & Meyer (2011)

Table 2

Proportion of HUs by commercial list substratum with 1 or more persons and expected number of persons in the age ranges 18–44 and 65+ based on NSFG screener responses

HU sampling substratum	Proportion of HUs with 1+ persons in the age range $P_h(D)$		Expected cost (hours) per HU $C_h(D)$		Expected number of persons per HU in age range $a_h(D)$		Estimated national counts of HUs (1000s)	
	18–44	65+	18–44	65+	18–44	65+	18–44	65+
1 List has record; 0 persons in age group	0.207	0.062	4.45	3.43	0.292	0.071	45,070	52,800
2 List has record; 1+ persons in age group	0.720	0.671	8.04	7.70	1.249	0.947	27,359	19,629
3 List has record; no age info	0.465	0.122	6.25	3.85	0.785	0.159	29,780	29,780
4 No list record	0.372	0.102	5.60	3.71	0.596	0.128	27,990	27,990
Total	0.409	0.176			0.671	0.236	130,199	130,199

Table 3

Non-Linear Programming Allocation solution for ages 18–44 in the NSFG example

Substratum	n_h (HUs)	Proportion of sample HUs	Expected cost	Exp. no. of eligible sample persons	Optimal sampling rate	Weight for each HH in substratum
1	466	0.14	2,072	136	0.0000103	96,724
2	1,628	0.50	13,088	2,033	0.0000595	16,808
3	691	0.21	4,321	543	0.0000232	43,100
4	485	0.15	2,715	289	0.0000173	57,765
Total	3,269	1.00	22,196	3,000		

Table 4

Non-Linear Programming Allocation solution for ages 65+ in the NSFG example

Substratum	n_h (HUs)	Proportion of sample HUs	Expected cost	Exp. no. of eligible sample persons	Optimal sampling rate	Weight for each HH in substratum
1	564	0.12	1,936	40	0.0000148	67,687
2	2,933	0.62	22,574	2,779	0.0000856	11,688
3	682	0.14	2,628	109	0.0000229	43,636
4	567	0.12	2,106	73	0.0000203	49,338
Total	4,746	1.00	29,244	3,000		

Table 5

Comparisons of NLP and equal probability sample sizes and costs in the NSFG example.

Age subgroup	HU sample sizes			Expected costs			Ratio of NLP to equal probability
	NLP	Equal probability	Ratio of NLP to equal probability	Type of cost	NLP	Equal probability	
18-44	3,269	2,980	1.10	Screening	9,808	8,941	1.10
				Interviewing	12,388	8,535	1.45
				Total cost	22,196	17,477	1.27
65+	4,746	8,489	0.56	Screening	14,239	25,468	0.56
				Interviewing	15,006	10,461	1.43
				Total cost	29,244	35,929	0.81

Table 6

Summary of classification results using MSG data.

MSG classification	Unweighted		Weighted	
	No. of HUs	Percentage	No. of HUs	Percentage
Correctly classified	8,109	29%	15,504,526	40%
Incorrectly classified	5,979	21%	8,509,220	22%
No available data	14,076	50%	14,930,076	38%
Total	28,164	100%	38,943,822	100%

Table 7

Estimated proportion of HUs,

		HRS analysis domain based on responses to screener (D)							All eligible (HRS Domains 1- 4)
HU sampling substratum (h)		1 MBB/Hispanic	2 MBB/Non-Hispanic	3 EBB/Hispanic	4 EBB/Non-Hispanic	5 No MBB/EBB	6 Unknown due to NR	7 Unoccupied HU	
1	MBB H	0.325	0.090	0.095	0.010	0.323	0.122	0.034	0.521
2	MBB NH	0.017	0.525	0.003	0.063	0.231	0.117	0.045	0.607
3	EBB H	0.038	0.033	0.536	0.070	0.167	0.129	0.028	0.677
4	EBB NH	0.001	0.054	0.014	0.627	0.158	0.112	0.034	0.696
5	No MBB/EBB	0.017	0.044	0.036	0.052	0.679	0.111	0.061	0.148
6	Unknown 0.016		0.051	0.027	0.046	0.554	0.077	0.230	0.140
Total		0.021	0.078	0.038	0.090	0.553	0.099	0.122	0.226

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 8

Target number of housing units in the four HRS analysis domains and number required to be screened in an equal probability sample

Number of HUs	HRS Analysis Domain				Total
	1	2	3	4	
Target number of sample HUs, $n_{eq}(D)$	1,500	3,000	1,500	3,000	9,000
Total number of HUs to screen in an equal probability sample	72,951	38,352	39,960	33,467	

Nonlinear programming solution for number of households to be screened to obtain target sample of eligible HUs in the HRS example

Table 9

		NLP Sample allocation before deselection			NLP solution after deselection							Equal probability sample				
	HU sampling substratum	Substratum sample sizes		Proportion of sample HUs	Expected no. of sample eligible HUs after deselection					Indigible + NR +Unocc	Total Eligible after deselection + inelig + NR + unocc	% of initial sample HUs that are eligible and retained	Expected allocation before deselection	Proportion of sample HUs	Expected no. of eligibles after deselection	% of initial sample HUs that are eligible and retained
		nh (HU(s))			MBB H	MBB NH	EBB H	EBB NH	All eligible							
1	MBB H		3,461	6%	1,125	225	208	33	1,592	1,659	3,251	46%	1,055	1%	453	43%
2	MBB NH		3,279	5%	56	1,239	5	206	1,507	1,288	2,795	46%	4,653	6%	1,504	32%
3	EBB H		1,241	2%	47	30	419	87	583	400	984	47%	1,149	2%	438	38%
4	EBB NH		1,566	3%	2	61	14	981	1,058	476	1,534	68%	5,049	7%	1,640	32%
5	No MBB/EBB		28,134	46%	488	882	630	1,460	3,460	23,962	27,422	12%	33,080	45%	2,762	8%
6	Unknown		23,138	38%	375	845	398	1,056	2,674	19,902	22,577	12%	27,966	38%	2,203	8%
Total			60,818	100%	2,094	3,282	1,675	3,823	10,874	47,688	58,562	19%	72,951	100%	9,000	12%

Notes: The expected number of eligible HUs after deselection in each (sampling substratum)/(HRS analysis domain) is computed as $\eta_{ll}p_h(D) f(D)$.

Estimated national counts of HUs, N/h , are based on the 2010–11 HRS samples.

Expected numbers of eligible cases and effective eligible cases before and after deselection to reduce the sample to the desired effective sample sizes.

Table 10

Domain	MBB H	MBB NH	EBB H	EBB NH	Total
Eligibles in NLP solution					
Expected no. in initial allocation $n(D)$	2,094	4,557	2,657	3,823	13,131
Expected effective no. in initial allocation $n\,eff(D)$	1,500	4,165	2,379	3,000	11,044
Retention fractions $f(D)$	1.00	0.72	0.63	1.00	
Expected no. after deselection $n^*(D)$	2,094	3,282	1,675	3,823	10,874
Equal probability sample					
Expected no. in initial allocation	1,500	5,706	2,738	6,539	16,483
Retention fractions $f_{eq}(D)$	1	0.53	0.55	0.46	
Effective no. after deselection in both allocations	1,500	3,000	1,500	3,000	9,000

Table 11

Weights in the NLP allocation before and after deselection

HU sampling substratum	Estimated national counts of HUs	Weights before deselection	Weights after deselection			
			MBB H	MBB NH	EBB H	EBB NH
1 MBB H	563,204	163	163	226	258	163
2 MBB NH	2,484,317	758	758	1,052	1,202	758
3 EBB H	613,278	494	494	686	784	494
4 EBB NH	2,695,279	1,721	1,721	2,389	2,730	1,721
5 No MBB/EBB	17,660,498	628	628	871	996	628
6 Unknown	14,930,076	645	645	896	1,023	645
Total	38,946,652					

Table 12

Weighting design effects for eligible units before and after deselection

		NLP allocation		Equal probability allocation after deselection
Domain		Before deselection	After deselection	
1	MBB H	1.40	1.40	1.00
2	MBB NH	1.09	1.09	1.00
3	EBB H	1.12	1.12	1.00
4	EBB NH	1.27	1.27	1.00
1 & 2	MBB	1.19	1.28	1.07
3 & 4	EBB	1.31	1.23	1.01
All combined		1.29	1.26	1.05

Table 13

Comparison of costs for NLP solution with deselection to equal probability allocation with deselection

Type of cost	NLP	Equal probability	Ratio of NLP to equal probability cost
Screening	182,455	218,853	0.83
Interviewing	76,121	63,000	1.21
Total cost	258,576	281,853	0.92