



High Frequency Mobile Phone Surveys of Households to Assess the Impacts of COVID-19

Guidelines on Sampling Design

The guidelines have been prepared by Kristen Himelein (Senior Economist / Statistician, Poverty and Equity Global Practice), Stephanie Eckman (Research Triangle Institute), Jonathan Kastelic (Survey Specialist, Development Data Group), Kevin McGee (Economist, Development Data Group), Michael Wild (Senior Statistician, Development Data Group) and Nobuo Yoshida (Lead Economist, Poverty and Equity Global Practice), with comments from Johannes G. Hoogeveen (Practice Manager, Poverty and Equity Global Practice).

Version: April 29, 2020.

1. Introduction

This note describes the sample considerations for using high frequency phones surveys to measure the economic implications of the 2020 COVID-19 pandemic. While face-to-face interviewing will not be possible given the risk of transmission to both interviewer and respondent, the implications of telephone-based data collection on the sample design must be carefully considered in planning response surveys. While the literature is limited for developing countries, there is evidence that mobile phone surveys can be effective (see the World Bank Listening to... projects, Ballivian et al 2015, Dabalén et al, 2016, Lau et al, 2019, and Leo et al, 2015 for examples) and telephone surveys were successfully used in Sierra Leone and Liberia in 2014 and 2015 to study the implications of the Ebola outbreaks there (Fu et al (2015) and Himelein and Kastelic (2015)).

1.1. Sampling Frame

The single most important methodological decision in designing a sample for a phone survey is the choice of the sampling frame. There are broadly three methods to create a sampling frame for telephone-based surveys:

1. A full or sub-sample from a representative survey (like HBS, LSMS-ISA etc.),
2. A sample from a list of phone numbers, and
3. A sample of numbers selected by Random Digit Dialing (RDD).

While the methods above are listed in their general order of preference, from most to least desirable, there are some advantage to number lists and RDD that may make them preferable in certain cases. The benefits and drawbacks are discussed in more detail below. Regardless of the sampling approach used, the high frequency telephone surveys may wish to sample only from mobile telephone numbers where possible, as land line numbers are frequently connected to businesses or fax machines.

1.1.1. Representative survey as sample frame

If there has been a recently conducted representative household survey that contains re-contact information for some or all household members, this approach is most likely optimal. The main benefit to this approach is that an abundance of household and person characteristics are available from the survey. These characteristics are important in reweighting the data (discussed below in section 6) and the survey may also provide re-contact information for multiple household members for individual-level surveys (discussed in section 3). The respondents have also (hopefully) participated recently in the main data collection exercise and therefore would be more likely to respond to the phone survey. Researchers need to be able to determine whether respondents have consented

(in the face-to-face survey) to be re-contacted, and if not, whether it is acceptable to recontact the respondent. Country regulations or human subjects concerns may prevent researchers from re-contacting a respondent who has not provided prior consent for recontact. Significant levels of non-consent to be re-contacted could contribute to bias which would need to be adjusted through weighting.

The main drawback to this approach is that the sample size for the phone survey is constrained by the sample size of the original, both on the aggregate level in terms of total respondents as well as for certain sub-groups of interest (female headed households, those working in the informal sector, certain geographies, etc.). In addition, using an existing survey in which the household were clustered effectively 'imports' those design effects into the phone survey data (discussed in section 3). The benefits of detailed information, however, likely outweigh the size and efficiency constraints and it would be preferable to supplement an existing survey with additional numbers from either a list or RDD rather than to abandon it entirely.

1.1.2. List-based sampling frames

List-based designs use numbers from government registries, telecommunications companies, marketing firms, or other sources. If possible, the lists should come from government registers or telephone companies, particularly if these sources can be expected to contain all working numbers assigned to individuals in the country. In cases where the country has several telephone companies, ideally all companies will provide complete lists to create a national sampling frame. Marketing companies often try to create representative lists for certain groups or areas but are limited by the source of their numbers (usually purchased from companies).

The main advantage of lists is that they can be expected to contain a high percentage of working numbers. Additionally, there is no upper limit on the sample size and no cluster effects to decrease precision in the data. The main drawbacks, however, are that it is difficult to tell how representative the list is of the full population and there is limited data for reweighting. With regard to reweighting, official lists may have only the name and phone number for respondents, perhaps including gender or location where the number was registered in a best-case scenario. In cases where a list from an alternate source is used, such as from a marketing firm, it is less likely than an official list to cover all possible respondents in the country, but more likely to contain information that can be used for weighting calculations. In addition to technical concerns, lists may also take time to obtain and therefore may not be possible in the context of a time-sensitive crisis response.

1.1.3. Random Digit Dialing

The third approach to sampling for telephone surveys is to randomly generate possible telephone numbers based on knowledge of how numbers are structured in a given country. This approach is common in the US (see, for example, the National Immunization Survey). The main benefits are that RDD is complete (all numbers have a chance of selection) and quick to implement, as no agreements on data access must be reached before beginning. The main drawback to this approach is the low efficiency: many of the generated numbers will be nonworking or unassigned. Calling and identifying these numbers can take up a high proportion of the data collection budget in RDD surveys, and efficiency will vary a great deal by country. For example, a recent RDD survey in Ghana conducted 1,076,258 initial outreaches with Interactive Voice Response (IVR), more than 85 percent of which went unanswered or were to invalid numbers, to reach a final sample size of 13,016 completed telephone interviews (L'Engle et al, 2018)¹. Often, data from the telephone companies can be used to increase efficiency, as with list-assisted RDD in the US (Tucker et al, 2002). One approach to reduce this impact would be to screen an RDD sample through commercial databases to remove non-working numbers. There are vendors who can provide those services in many countries.

¹ Also highlighting this point is the work done by the Pew Research Center, which has worked in more than 90 countries globally conducting public opinion research. In all but a handful of countries (Canada, the United Kingdom, France, Germany, Spain, the Czech Republic, Australia, Japan, and South Korea), surveys are conducted face-to-face despite the far higher costs because of the difficulties in collecting representative data by RDD.

1.2. Multi-frame designs

In some cases, the best approach may be to combine multiple sources into a dual-frame or multi-frame approach (Hartley, 1962). It is, for example, possible to combine a frame based on an existing survey or an RDD design with sector-specific surveys using an externally provided frame for groups, such as urban informal sector workers or tourism sector workers, that may be particularly affected by COVID-19 economic impacts. Another common approach is the combination of different lists of phone numbers, such as those provided by two different telecom companies. The main advantage to this approach is that boosts the efficiency of the design by, in principle, increasing the overall sample size or sample size for specific groups of interest. The main drawback is that it increases the complexity of the weight calculations and analysis. Because each sampling frame is used to select an independent sample, the questionnaire must contain questions to estimate the overlap (i.e. to identify if the individual is contained in the other sampling frames) of the two samples. With this information, the weights can be adjusted for multiple probabilities of selection and a final single weight can be calculated. For example, estimating the overlap between the frame of urban informal workers and an RDD frame would involve asking all cases contacted via RDD whether they live in an urban area and what sector they work in. In other multi-frame designs, estimating the overlap can be more challenging.

2. Sample Size Requirements

In general, the basic rule of sample design for telephone surveys is to take as many observations as financially possible, since a high non-contact/non-response rate can be expected and because it is more difficult to estimate the expected standard errors in a telephone survey compared to face-to-face. In the case of telephone surveys based on previous household surveys, attempting to contact all original respondents is the preferred method, combining multiple surveys if possible. In the case of list-based and RDD surveys, the sample size will be dictated by budget, and should incorporate stratification to the degree information is available. At the same time, it is important to understand the expected precision of a given sample size as it may be necessary to manage expectations of counterparts or may be an important factor in prioritizing questions during the design process.

2.1. Point Estimates vs. Difference/Change

Sample size requirement depend on analytic objectives. The number of observations for detecting statistically significant difference in an indicator between groups or over time may be substantially more than those required for reliable point estimates. A key element in rapid high frequency mobile data collection for situation monitoring is measuring changes in critical social and economic indicators. Therefore, sample sizes in these surveys should be large enough to meet the analytical and policy objectives.

2.1.1. Point Estimates

The sample size requirements for a point estimate follow the formulas below:

$$n = \frac{t_{\alpha}^2 \times \text{Var}(X)}{E^2} \times \text{design effect} \qquad n = \frac{t_{\alpha}^2 \times P(1-P)}{E^2} \times \text{design effect}$$

Continuous variable Proportion

Where t_{α} is a constant related to the confidence of the estimates (95% confidence intervals are common), $\text{Var}(X)$ is the variance of the continuous variable, P is the population prevalence, and E is the maximum acceptable margin of error. Point estimates take into account only the distribution of a single variable at a single point in time. Variables with higher variance, prevalence closer to parity, and small acceptable margins of error will all necessitate larger sample sizes.

The design effect is a measure of the statistical efficiency of the sample as compared to a simple random sample of the same size. The design effect is determined by the survey design along with any weight adjustments. Clustered designs tend to increase the design effect while stratification can reduce the design effect in some situations. When using a sample from a previous survey which used a clustered sample design, the design effects will be covered in the existing survey documentation for that survey. List-based samples and RDD samples usually have design effects

of 1 or slightly less than 1, if stratification variables correlated with the measures of interest are available (see Section 3).

2.1.2. Difference / Change Measures

More commonly in crisis response surveys, however, analytical objectives are to measure the differences between groups or over time. These calculations rely on two (or more) distributions and therefore require additional information to determine the expected analytical power.

$$n = \left[\frac{4\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{D^2} \right] \times \text{design effect}$$

Factors influencing the target sample size for rapid high frequency mobile monitoring surveys include:

1. **Estimate of indicator prevalence or variance (σ^2):** As noted above higher variances or a prevalence closer to 50% require larger sample sizes to obtain a given level of precision as wider confidence intervals are more likely to overlap.
2. **Determination of the size of a plausible and policy relevant change between rounds (D):** The larger the difference or expected change between groups or rounds, the smaller the sample size required. However, those implementing telephone surveys often have less control over their final sample size than traditional face-to-face surveys, due to nonworking numbers and noncontacts. For that reason, this element is often calculated from a given sample size, which is to say for a given sample size, what is the minimum detectable effect that can be reliably picked up.
3. **Confidence (α) and Power ($1 - \beta$):** Power is the probability to find a significant effect if there truly is an effect. Confidence intervals are usually 95% (which means $\alpha = 0.05$) and common values for power are 90% or 80% (which means $\beta = 0.1$ or 0.2). If the expected direction of change is known, values for power calculations can be done as a one-tailed test (which would mean replacing $\alpha/2$ in the formula above with α), resulting in lower sample size requirements and small minimum detectable effects.
4. **Design effect:** When using previous survey with a clustered sample as a base for a phone survey, loss of efficiency due to clustering, and the effect on estimate precision from the sample survey design, must also be considered. In the case of samples selected from telephone lists or through random digit dialing, there is no clustering to consider. In both cases, stratification may increase precision.
5. **Estimate of the gain / loss of precision** resulting from the sample design for the phone survey: There are two elements common to telephone surveys that are important additional considerations. First, the high frequency surveys are being designed to be panel surveys in which the same numbers are called across many rounds. Using the same respondents across rounds increases the analytical power of the sample since unobservable but time invariant characteristics that end up in the error term are constant across rounds. Second, telephone surveys generally have higher levels of nonresponse than face-to-face surveys and collecting data over time inevitably leads to some sample attrition. Reweighting procedures, discussed below, can help reduce bias caused by nonresponse and attrition. However, the reweighting process usually increases the variance in the sample weights (with harder-to-interview respondents ending up with higher weights). The increased variance in the weights leads to increase the width of confidence intervals and decrease precision. Since it is very difficult to estimate the magnitude or impact of the increased sample weight variance during the design stage, analysts are advised to build in an additional design effect on top of their calculations.

Table 1 shows generalized simple random sample size requirement for 5, 10, and 15 percentage point minimum detectable effect at 80 and 90 percent power². A 50 percent initial prevalence and two-sided test are assumed, which produce the most conservative sample size. Approximately 519 observations are required per round or group to detect a change of 10 percentage points, assuming 90% power and simple random sampling. Under the same

² An online calculator for two independent proportion sample size calculations is located at <https://www.stat.ubc.ca/~rollin/stats/ssize/b2.html>

conditions, a 5-percentage point change between indicators increases the required number of observations to over 2000. This table highlights the need to manage expectations on what level of precision is possible from the survey. Moreover, the design effect must also be incorporated. In the first example above, if clustering leads to a design effect of 2.5, the number of observations required increases to approximately 1,300 in each comparison round or group. These calculations are for one analytical domain. If further disaggregation is desired, such as by levels of education, each analytic domain requires this number of observations.

Table 1. Sample size requirements to detect stated differences between two groups of equal size under a design effect of one.

Significance 95% (α)	Power (β)	Minimum Detectable Effect	Initial Prevalence	Base (SRS)	Comparison (SRS)	Total Sample Size
0.05	0.8	0.05	0.5	1565	1565	3130
0.05	0.8	0.10	0.5	388	388	776
0.05	0.8	0.15	0.5	170	170	340
0.05	0.9	0.05	0.5	2095	2095	4190
0.05	0.9	0.10	0.5	519	519	1038
0.05	0.9	0.15	0.5	227	227	454

Sample size calculations generally, and particularly in the case of high frequency phone surveys, rely on a large number of assumptions. Implementation constraints will also differ across populations and geography, and sample size and the sampling fraction, may be different from the start or may need to be modified during implementation to achieve a sufficient number of observations. Similarly, sample size targets per domain of inference may be larger in subsequent rounds than in the first round, if new indicators are added to the survey. This flexibility is one of the inherent benefits of having a high frequency approach, but it needs to be accounted for in the sample design as it is harder to add more observations later in the survey process.

3. Stratification

3.1. Analytical strata versus design strata

Stratification is common in survey designs because it allows analysts to guarantee a minimum number of observations for populations of interest and because it can provide improved precision over simple random sampling. The first objective, guaranteeing minimum sample sizes to do reliable analysis for certain populations, is achieved by creating analytical strata or domains of inference. These classifications can be thought of as the rows on a table in the final report, disaggregating results by geography, sex, age, sector of employment, etc. The second objective, improving the efficiency of the design, attempts to reduce the standard error at the global or domain of inference level, by creating homogenous groups based on auxiliary information, which are used as design strata. As design strata are more flexible than analytical strata, they can be constructed to meet the various survey objectives, with sampling fractions differing as needed.

The benefits of design stratification are related to the theory of optimal allocation, first introduced by Neyman. Under this approach, the total sample size is allocated to the design stratum based on the within-stratum variance, and, if the information is available, costs. However, whereas optimum allocation takes the number of strata as given, more recent approaches have extended this approach by applying allocation and stratification simultaneously, depending on the minimization of within domain variance (and costs). ‘Balanced’ sampling uses an algorithm to determine an optimal configuration of design strata for surveys with multiple objectives (Tillé, 2010). Further detail on the theory and implementation of balanced sampling is available in Annex 1.

3.2. Stratification in survey based on previous data collection

In the case of samples based on representative surveys, there is substantial information available to guide the stratification. In these cases, the analyst should carefully consider the survey objectives and prioritize groups for analysis. Of key interest should be variables likely to be relevant to COVID-19 surveys, such as age, health status and employment status. If these variables are available from the previous survey, they would be good choices for stratification. Given that extensive information is available, there may be a case to use an optimal allocation or balanced sampling, though the assistance of an experienced sampler would likely be required.

3.3. Stratification in list-based or RDD designs

For list-based sampling frames, samples should be selected with simple random sampling, explicitly or implicitly stratified where possible. Countries differ in the variables that will be available for stratification in these cases depending on the scope and completeness of information provided in the list. For RDD samples, stratification may not be possible, but should still be explored. In many countries, telephone numbers are clustered geographically (such as by area codes in the US). However, with the advent of cell phones, area codes refer only to where the number was initially registered and do not necessarily reflect the respondent's current residence. At minimum, if multiple mobile networks are being used, stratification should be done on network as there are usually differences in geography and potentially well-being.

4. Choice of respondent

Ultimately the choice of respondent will be dictated by whether the survey is seeking to capture individual level or household level information. In the case of list-based or RDD sampling frames, the list of numbers (hopefully) covers all members of the population with a phone number. In the case of surveys based on previous data collection, re-contact information is often captured only for the head of household or, if s/he does not have a phone, another member that does. There are, however, two reasons this individual may not be the most appropriate respondent for the survey. First, because household heads are predominantly male in most parts of the world and because phone ownership is skewed towards younger males, any sampling frame based solely on these groups would generate a biased estimate of the distribution of sex in the population, as well as any other variable that is correlated with sex. If the survey collects recontact information on multiple household members, then stratification on sex and perhaps age might be useful for individual level impact, assuming all members would be equally able to report on household level variables. If this information is not included, it might be possible to ask the respondent to speak to a female or younger member of the household, but the viability of this approach will depend on the cultural context.

Secondly, even in the case where the survey targets only household level information, the re-contact individual may not necessarily be the most knowledgeable about their households. Any inaccuracies in their reporting would be non-sampling error, as opposed to sampling error, which cannot be adjusted by reweighting. To address this respondent bias, interviewers may request the re-contact individual to pass the phone to one who is familiar with the household or make an appointment for a next call when the person is available. The added complexity, however, may increase non-response rates and should be carefully balanced with the likelihood of non-sampling error. Regardless of the approach chosen, the survey manual should clearly describe the respondent selection protocols and this process should be included in the interviewer training.

5. Threats to Representativeness

Both coverage and nonresponse can threaten the representativeness of a survey.

5.1. Coverage Issues

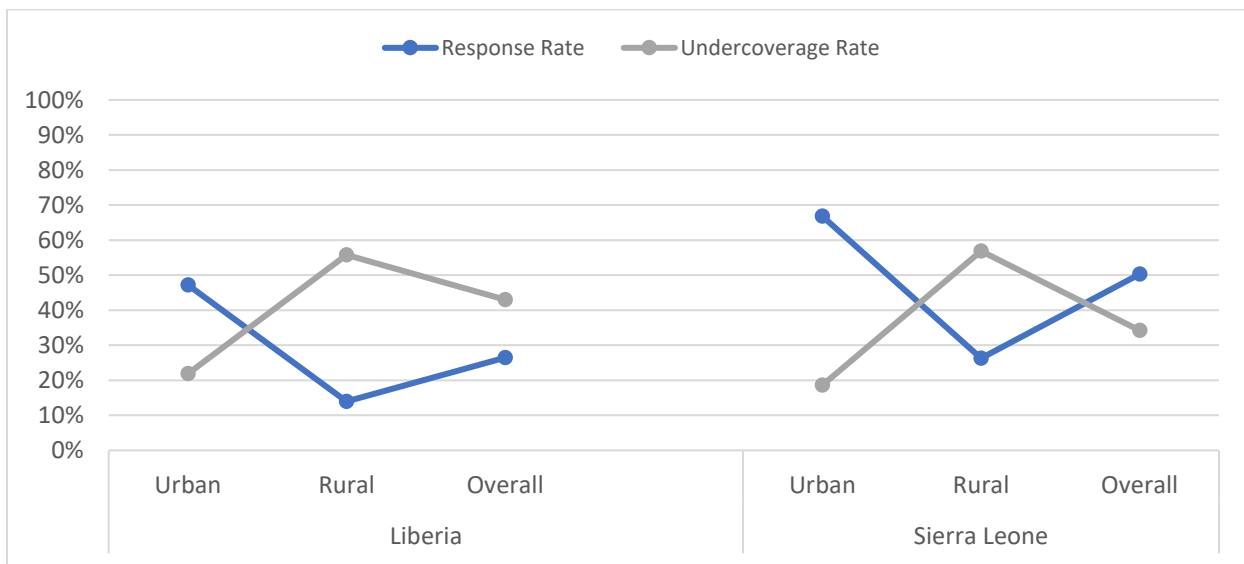
When the sampling frame does not match the population of interest, both undercoverage and overcoverage can occur. These error sources are more common in phone surveys than face-to-face surveys, so those carrying out a phone survey for the first time should think carefully about undercoverage and overcoverage.

5.1.1. Undercoverage

Undercoverage is a major concern for telephone surveys, particularly those implemented in the developing world. Individuals without cell phones and those with cell phones but living outside of areas with network coverage would be undercovered by a telephone survey. Inasmuch as these non-responding groups differ from the covered population, bias will be introduced in the resulting estimates. This bias can be mitigated somewhat with reweighting techniques based on observable traits (assuming a representative sample exists). However, unobservable characteristics likely also affect undercoverage and weights cannot compensate for bias due to these factors.

Households without telephones have no chance of selection. Figure 1 below shows the percent of households undercovered in the high frequency cell phone surveys conducted in Liberia and Sierra Leone during the Ebola crisis. In Liberia, more than 40 percent of households did not have a recontact phone number in the baseline Household Income and Expenditure Survey, including 55 percent in rural areas. Though the surveys never attempted to contact these households, they must be considered as non-response.

Figure 1. Undercoverage and Response Rates by Country and Urban/Rural for Ebola phone surveys



For list-based samples, those households not included in list-based sampling frames would also be undercovered. For example, if there are two large telecom companies in a country, but only one cooperated with a survey effort, everyone who has a telephone number only through the non-cooperating company will have no chance to participate in the survey. If a given telecom attracts customers from the capital city or with higher incomes, this undercoverage could introduce bias into the survey data.

Also related to undercoverage is the issue of phone sharing. In many contexts, households will have fewer cell phones than members and phones will be shared between respondents. In cross sectional surveys this issue manifests as the coverage errors described above but can introduce additional complication to panel analysis – such as the type undertaken with high frequency monitoring of crisis situations.

5.1.2. Overcoverage

Overcoverage exists when the sampling frame is more expansive than the target population. This is particularly important for **RDD surveys** since the random generation of phone numbers will often lead to numbers that are nonworking or assigned to ineligible respondents (e.g. businesses, government offices, people living outside the country etc.). A frame that consists of **telephone numbers provided by telecom companies** will less likely contain fewer nonworking numbers but will still face the challenge of numbers no longer in service as well as numbers assigned to an ineligible respondent. Significant overcoverage reduces the efficiency of the sample and may require significantly larger sample sizes due to high ineligibility rates. Overcoverage can be reduced if pre-screening services

are available. Such service can identify nonworking numbers and those assigned to businesses ahead of data collection.

In addition, overcoverage can come from multiplicity, in which some individuals have multiple numbers or SIM cards and would have a higher probability of being selected (assuming the sampling lists cover all carriers). Theoretically the sampling weights should take into account the number of eligible SIM cards used by the respondent. This information, however, would not be available in the sampling frame and would have to be included as a question in the survey itself, which would then introduce potentially troublesome measurement error into the weight calculations.

5.2. Non-Response

In addition to under- and over-coverage associated with telephone surveys, nonresponse is a critical issue. Due to the format of phone interviews, overall nonresponse rates are consistently and often substantially higher than for face-to-face surveys. While face-to-face surveys typically target a response rate of at least 85 percent, completion rates are much lower for telephone surveys. Response rates for telephone surveys using a sample from a face-to-face interview as a frame are generally higher than RDD, but still substantially lower than face-to-face surveys. Telephone surveys conducted during the Ebola epidemic in Liberia used previous face-to-face survey respondents as a frame and achieved response rates of 46 percent of households providing a phone number and 26 percent of the overall sample. In Sierra Leone these figures were 75 percent and 50 percent, respectively (see Figure 1). A higher nonresponse rate will result in a lower sample size (when given a fixed set of phone numbers) or require more call attempts to achieve a given sample size. Therefore, response rates can have a large impact on the required workload and/or output of a telephone survey.

In addition, bias can arise due to differential nonresponse among sub-groups of the sample. Groups that are more likely to have cell phones and more likely to respond to calls (such as younger, urban, better educated, and male respondents in a recent RDD survey conducted in Ghana (L'Engle et al, 2018)), will likely respond at a higher rate. Nonresponse bias occurs when the respondent distribution significantly differs from the overall sample for key characteristics. Like undercoverage, reweighting techniques can be used but will depend on observable characteristics. Nonresponse bias also decreases precision for a given sample size as there are diminishing returns to additional respondents with similar characteristics.

5.2.1. Sources of non-response

There are several different sources for nonresponse in telephone surveys:

1. **Invalid or disconnected telephone:** Depending on the quality of the frame used for the survey, there is often the potential for telephone numbers that were recorded wrongly, are no longer in service, or simply do not exist. Frames taken from a previous face-to-face survey that collected telephone numbers are less likely to encounter numbers assigned to ineligible respondents but will face invalid numbers (wrongly recorded in face-to-face survey) and numbers that have been disconnected since the face-to-face survey was conducted. Invalid or disconnected phone numbers are difficult to overcome. While interviewers can sometimes overcome incorrect addresses for face-to-face surveys (e.g. though asking others in the community), incorrect phone numbers cannot be easily corrected. The source and share of numbers that are invalid or disconnected will vary depending on the type of frame used for the survey. Invalid or disconnected numbers are sources of non-response when a previous survey is used as the frame, but not when the sample is selected from a telecom list or RDD. In those sample types, invalid or disconnected numbers are not eligible (i.e., they do not belong in the target population).

An important consideration is that in some countries, it can be challenging to distinguish a working from a non-working number. Countries differ with respect to their call outcome codes and their clarity. Sometimes there are clear error codes that are returned that identify non-working numbers, but this is not the case for all countries. As a result, many researchers include all numbers in the denominator, which depresses response rate calculations (Lau and di Tada 2018).

2. **Not able to contact respondent:** Even when telephone numbers are valid and active, an interviewer may still not be able to make contact with a respondent. There are various reasons why this might occur, such as the respondent is out of the network area, the phone is turned off or the battery is empty, the respondent is not carrying the phone, or the respondent is unable or unwilling to pick up the call (particularly if the number calling is unknown to them). These issues are likely to affect any telephone survey (regardless of the frame).
3. **Refusal:** Even if the interviewer makes successful contact with a respondent, the respondent may not agree to participate in the survey. This source of nonresponse is also present in face-to-face surveys and, depending on the context, could be higher or lower in telephone surveys.
4. **Survey break-off:** In some situations, a respondent agrees to participate, but does not complete the interview. Breakoff can be due to a refusal halfway through the interview, a dropped call, or another reason. Often these individuals are prioritized in the survey system for follow-up. Sometimes there can be a point in the interview where if an individual completes questions up to that point, the survey is flagged as partially complete and counts as a response.

5.2.2. *Strategies to Limit Nonresponse*

It is important to try and limit the rate of nonresponse to reduce the risk of nonresponse bias and reach sample size targets. There are some strategies that can be adopted to limit the various sources of nonresponse. The success of these strategies is highly dependent on the country context, and thus it is recommended to consult with national counterparts and to conduct a small pilot study prior to the start of fieldwork.

1. **Removing invalid numbers from the frame:** One strategy is to make sure to identify and remove any phone numbers that can already be identified as invalid from the frame. The method to identify invalid numbers will vary by country, but typically what can be reviewed is (i) the number of digits in the phone number and (ii) whether the phone number fits a fixed pattern for phone numbers within the country. For RDD surveys, these validity characteristics should be automatically incorporated into the generation algorithm. For other frames, reviewing the list of numbers for these hallmarks is a trivial procedure that can save time and effort trying to call an obviously invalid number. As discussed above, in many countries there are firms that will remove nonworking and business numbers from a sample. It should be noted, however, that removing invalid numbers from a survey-based frame does not reduce non-response, though it can save time and resources, potentially allowing for higher interview completions generally.
2. **Pre-contact attempt through SMS:** Respondents can be sent messages through SMS prior to an interviewer attempting to make a call. This pre-contact can combat nonresponse in multiple ways. At a minimum, it informs the respondent of the proposed timing and the purpose of the call. With this warning, respondents may monitor their phone more closely or be more likely to answer. Informing the respondent in the SMS what number they will be contacted from will also improve the likelihood that they will answer the call. The respondent should also be informed when they can expect the phone call if a reasonable estimate can be provided. That way respondents who are willing to participate may be more careful about keeping their phone on, keeping it charged, and carrying their phone during the time they can expect a call from the interviewer. Providing the respondent with this information can (i) increase the chance that they are successfully contacted and (ii) reduce the number of contact attempts required to reach them. One methodological study from Australia found that sending an SMS prior to calling substantially increased response and cooperation rates of respondents (Dal Grande et al, 2016). The SMS messages can be sent through an automated system or manually. In the case of countries with multiple languages, it may be necessary to have multiple versions of the SMS.
3. **Respondent incentives:** One additional way to increase response rates is to provide the respondent with a small reward for their participation in the survey. The type and level of incentive varies by country, but in many developing country contexts an easy incentive to provide is mobile credit or airtime which can be transferred directly to the number the respondent was called on. However, increased response rates from the incentive must be weighed against the potential for response bias from the respondent as a result of the incentive.

(Stecklov et al, 2018). While offering an incentive has been shown to reduce survey nonresponse, studies have found that increasing the incentive amount does not also increase response rates (Ballivian et al. 2015, Gibson et al. 2019, Lau et al. 2019).

6. Weights

As in all surveys, weights are required to correct estimates for different probabilities of selection. In the case of samples based on representative surveys, the original sampling weights form the base of the weight calculations. They should be adjusted for any subsampling done for the phone survey. In list-based and RDD samples, weights are required if the sampling fraction varies between strata.

The most important function of weights in phone surveys, however, is to reweight the phone survey data to be closer to a representative population. **Phone surveys generally do not yield representative data.** At best, there is undercoverage for all households or individuals that do not have access to a mobile phone, currently estimated to be about one-third of the world's population.³ High frequency telephone surveys also generally have substantial issues with non-response and attrition, which further threaten the representativeness of the results. The reweighting strategies described below are the main methods available to an analyst to adjust the results to match observable characteristics. This process, however, does not guarantee unbiased results, and therefore should be thought of as 'adjustments' rather than 'corrections' for bias.

6.1. Weights as a compensation for non-response

Weights can be adjusted either by using information from the sample or from auxiliary data on the target population. It is generally recommended that analysts use a sample-based technique first followed by population-based techniques if and only if there is high quality auxiliary information available.

6.1.1. Sample-based reweighting techniques

Two major techniques are commonly used to minimize the impacts of non-response by reweighting the observations to match the original sample on a known set of observable characteristics:

1. **Weighting class adjustments:** Divide the sample (both respondents and nonrespondents) into cells, for example age group x gender x urban/rural. Increase the weight of the respondents in each cell by the inverse of the response rate for that cell. Weighting class adjustments rely on having exact totals for each cell. Cells that exist in the sample but contain no respondents will need to be collapsed with other neighboring cells.
2. **Propensity score adjustments:** This method is more common when there are too many variables to use a simple weighting class adjustment. A modeling approach such as logistic or probit regression can be used to predict the probability that each case responded, given the observable characteristics. The predictive variables in the model could be age, gender, urban/rural, but also para-data about the survey, such as the number of calls made to each case (see Section 7.3 for more on para-data). The inverse of the predicted probability is then used as a weight adjustment. The scores can also be grouped into classes to avoid extreme weight adjustments, at the cost of reduced bias adjustment. Further detail on propensity score adjustments is provided in Annex 2 (see Sections A2.3 and A2.4).

6.1.2. Population-based reweighting techniques

Post-stratification adjusts weights to known population totals generated by auxiliary data. Post-stratification can reduce variances, but the primary goal is to reduce coverage errors using high quality auxiliary data. These adjustments require high quality auxiliary data on the characteristics of a population and then aligns the weights to those estimates (Little, 1993). Further detail on post stratification and the related technique of raking is provided in Annex 2 (see Sections A2.1 and A2.2, respectively).

³ The Mobile Economy 2020 : https://www.gsma.com/mobileeconomy/wp-content/uploads/2020/03/GSMA_MobileEconomy2020_Global.pdf

6.2. Other considerations

In addition, the weights for a telephone survey should include an adjustment for the number of eligible phone numbers held by the respondent. Each phone number increases that respondent's probability of inclusion. In surveys based on lists from telecom companies, this calculation is more complicated if not all telecom companies participate as only additional phone numbers from participating telecoms would impact the weights.

One consequence of using these types of adjustments described above is that the variance of the weights will likely increase, decreasing the precision of estimates. Winsorizing trims outlier weights by replacing them with the highest non-outlier weight – for example replacing the weights in the 99th and 100th percentiles with the highest value in the 98th percentile. Variance is reduced, but at the cost of introducing a small amount of bias.

Note that adjustments for the number of phones held by the respondent and trimming are done before any population-based adjustments.

6.3. Number of weight variables

An important consideration for the 'high frequency' aspect of the telephone surveys is that each combination of data will have its own set of weights. For example, after the first round of a survey conducted using an existing survey as the baseline, there will be one set of weights in addition to those associated with the baseline survey. After the second round, there will be three sets: cross sectional weights for round 1, cross sectional weights for round 2, and panel weights for rounds 1 and 2. Any questions asked only in round 1 would use the cross sectional weights for round 1, and similarly any questions asked only in round 2 will use the appropriate cross sectional weights for round 2. Any questions in both rounds will use panel weights for those two rounds. As additional rounds are added, the number of associated sets of weights increases substantially (Himelein, 2014). Not all combinations of the weights will be used, but in cases where questions are rotated in and out (appearing for example in rounds 1, 3, 5, and 7) it becomes challenging. As an example, more than 20 sets of weights were used in the analysis of the 5 rounds of Liberia data from the Ebola surveys. While it is not hard to calculate the weights for the various combinations, analysts should take care to ensure the correct set is applied.

7. Overlap with Implementation Issues

There are several important issues around survey design and implementation that are interrelated to the sampling strategy.

7.1. Questionnaire design

Effective design of a questionnaire is a critical element to the success of any survey, but there are some aspects that are especially relevant for sampling. At a minimum, if the sample is list-based or RDD, the questionnaire must capture the number of active telephone numbers the respondent has, so that this information can be incorporated into the weight calculations. Additionally, if multiple frames are used, the questionnaire must include questions to estimate the overlap for the weights to be accurately calculated. For RDD surveys and surveys using lists of telephone numbers from telecom companies, it is also important that the questionnaire captures basic demographic information to assess the coverage of the sample. The profile of the successfully interviewed sample can be compared with other data sources with representative coverage of the general population (e.g. censuses, face-to-face surveys, etc.) to identify under- and over-coverage in the final sample and attempt to adjust for it in the weights.

In addition, the analytic objective and sample size requirements have implications for questionnaire design. Since comparisons require larger numbers of observations than point estimates, it may be possible to randomize point estimate questions across the surveys to have more indicators for the same length questionnaire.

7.2. Informed Implementation and Responsive and Adaptive Design

Design strata can also be used to assign higher values to collecting observations in hard-to-observe strata, allowing for an adaptive approach to minimizing the impacts of non-response. For example, in the Liberia high frequency

surveys, a female household head working in agriculture in rural Nimba county has a base weight one hundred times larger than the weight of a male wage employee in urban Grand Gedeh. If the CATI software allows for the calculation of weights and non-response corrections in real time, this information can be fed into the implementation procedures to identify ‘high-value targets,’ which can then be prioritized for repeated attempts to interview. Even if the weights cannot be calculated in real time, for studies based on previous surveys there are benefits to creating a targeting system that prioritize cases with higher starting base weights.

7.3. Collection of Meta-Data and Para-Data

In order to effectively manage and monitor data collection via a telephone survey, it is critically important to capture detailed meta- and para-data on the survey process (contact attempts, interview result, etc.). These data serve as an input into the weight calculation and allow for a better understanding of coverage. It is important that any meta- or para-data included in weight adjustments should also be correlated to outcomes of interest. If the meta- or para-data is only correlated with response, then this information would be good to use in an adaptive design but using it for weighting could introduce additional bias for outcomes.

The meta-data that should be collected include a complete log of contact attempts including date/time and result of each attempt. The result of an attempt should be recorded with some granularity (e.g. fully completed interview, partial interview, refusal, rescheduled, no answer, wrong number, etc.). If multiple numbers are available for the same respondent/household, then which number was contacted in each attempt should also be logged. The number of telephone numbers available for a respondent or household is also an important piece of information to capture, because the probability that a household will be successfully contacted is higher if there are more numbers available.

7.4. Survey Management System

A strong survey management system is essential to the success of telephone surveys, including a carefully documented flow of assignments to interviews. The system should ensure a smooth flow of assignments to interviewers and sorting of assignments following contact attempts. For example, completed interviews should be sent for data quality review, ultimately unsuccessful contact attempts are logged and removed from an interviewer’s workload, reschedules logged and returned to interviewer’s workload close to the rescheduled date, etc. A clear rule for the required number of contact attempts to make with an assignment before classifying as unsuccessful should be established and integrated into the management system. The complexity of the system implemented will depend on the software used and the capacity of the implementation agency/firm.

The survey management system should also include a robust monitoring system to limit non-sampling error as much as possible. In the current context where telephone interviews cannot be conducted at a central location (i.e. at a call center) effective monitoring of interviewer effort and performance is especially critical. From a sampling perspective, it is important to ensure that interviewers are making the necessary attempts to contact all assigned cases and adhering to the established rules for number of attempts to make. The monitoring system should at a minimum include callbacks to a subsample of *successful and unsuccessful* respondents/households by an independent monitoring team. Systems, however, should be intuitive and not overly complicated as the options for piloting and troubleshooting will be limited by the accelerated timeline to get to the field in crisis situations. Reliance on complex automation, in the absence of the required comprehensive testing or time to sufficiently train the implementing partner, may undermine survey integrity.

Cited References

- Ballivian, A., Azevedo, J., Durbin, W., Rios, J., Godoy, J., & Borisova, C. (2015). Using mobile phones for high-frequency data collection. *Mobile Research Methods*, 21.
- Capacci, S., M. Mazzocchi, and S. Brasini. (2018). "Estimation of unobservable selection effects in on-line surveys through propensity score matching: An application to public acceptance of healthy eating policies." *PLoS ONE* 13(4): e0196020. <https://doi.org/10.1371/journal.pone.0196020>
- Dabalén, A, Etang, A, Hoogeveen, J, Mushi, E, Schipper, Y, and von Engelhardt, J. (2016). *Mobile Phone Panel Surveys in Developing Countries: A Practical Guide for Microdata Collection*. World Bank Publications.
- Dal Grande, E., Chittleborough, C. R., Campostrini, S., Dollard, M., & Taylor, A. W. (2016). Pre-Survey Text Messages (SMS) Improve Participation Rate in an Australian Mobile Telephone Survey: An Experimental Study. *PLoS one*, 11(2), <https://doi.org/10.1371/journal.pone.0150231>
- Deville, J. C., Särndal, C. E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, 88(423), 1013-1020.
- Fu, N., Glennerster, R., Himelein, K., Rosas Raffo, N., Suri, TK. (2015). *The socio-economic impacts of Ebola in Sierra Leone: results from a high frequency cell phone survey- round one*. Washington, D.C.: World Bank Group. <https://hubs.worldbank.org/docs/ImageBank/Pages/DocProfile.aspx?nodeid=27412469>.
- Gibson DG, Wosu AC, Pariyo GW, et al. Effect of airtime incentives on response and cooperation rates in non-communicable disease interactive voice response surveys: randomised controlled trials in Bangladesh and Uganda *BMJ Global Health* 2019;4:e001604. <https://gh.bmj.com/content/4/5/e001604>.
- Hartley, H. O. (1962, September). Multiple frame surveys. In *Proceedings of the social statistics section, American Statistical Association* (Vol. 19, No. 6, pp. 203-206).
- Himelein, K. (2014). Weight Calculations for Panel Surveys with Subsampling and Split-off Tracking. *Statistics and Public Policy*, 1(1), 40-45.
- Himelein, K. and Kastelic, JG. (2015). *The socio-economic impacts of Ebola in Liberia: results from a high frequency cell phone survey*. Washington, DC ; World Bank Group. <https://hubs.worldbank.org/docs/ImageBank/Pages/DocProfile.aspx?nodeid=24050332>
- Kelly L'Engle, E. S., Adimazoya, E. A., Yartey, E., Lenzi, R., Tarpo, C., Heward-Mills, N. L., Lew, K. & Ampeh, Y. (2018). Survey research with a random digit dial national mobile phone sample in Ghana: methods and sample quality. *PLoS one*, 13(1).
- Lau, C. Q., Cronberg, A., Marks, L., & Amaya, A. (2019, December). In Search of the Optimal Mode for Mobile Phone Surveys in Developing Countries. A Comparison of IVR, SMS, and CATI in Nigeria. In *Survey Research Methods* (Vol. 13, No. 3, pp. 305-318).
- Lau, Charles, and Nicolas di Tada. 2018. "Identifying Non-Working Phone Numbers for Response Rate Calculations in Africa." *Survey Practice* 11 (2). <https://doi.org/10.29115/SP-2018-0020>.
- Charles Q Lau, Ansie Lombaard, Melissa Baker, Joe Eyerman, Lisa Thalji, How Representative Are SMS Surveys in Africa? Experimental Evidence From Four Countries, *International Journal of Public Opinion Research*, Volume 31, Issue 2, Summer 2019, Pages 309–330, <https://doi.org/10.1093/ijpor/edy008>.
- Lee, S. (2006). "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics*. 22 (2): 329–349.
- Leo, B., Morello, R., Mellon, J., Peixoto, T., & Davenport, S. T. (2015). Do mobile phone surveys work in poor countries?. *Center for Global Development Working Paper*, (398).
- Little, R. J. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, 88(423), 1001-1012.

- Rosenbaum, P. R., and D. B. Rubin. (1983). "The Central Role of the Propensity Score in Observational Studies for Casual Effects." *Biometrika* 70 (1): 41-55.
- Rosenbaum, P.R., and D.B. Rubin. (1984). "Reducing Bias in Observational Studies using Subclassification on the Propensity Score." *Journal of the American Statistical Association*. 79: 516-524.
- Schonlau M., A. van Soest, A. Kapteyn, and M. Couper (2006). "Selection Bias in Web Surveys and the Use of Propensity Scores." RAND Labor and Population Working Paper series 229. RAND Pittsburgh, PA.
- Stecklov, G., Weinreb, A. and Carletto, C. (2018), Can incentives improve survey data quality in developing countries?: results from a field experiment in India. *J. R. Stat. Soc. A*, 181: 1033-1056. doi:10.1111/rssa.12333
- Terhanian, G., J. Bremer, R. Smith, and R. Thomas. (2000). Correcting Data from Online Survey for the Effects of Nonrandom Selection and Nonrandom Assignment. Research paper: Harris Interactive.
- Tillé, Y. (2010), Balanced sampling by means of the cube method, Presentation at the Euskal Estatistika Erakundea, XXIII Seminario Internacional de Estadística, viewed 4 April, 2020, https://www.eustat.eus/productosServicios/52.2_balanced_sampling.pdf
- Tucker, C., Lepkowski, J. M., & Piekarski, L. (2002). The current efficiency of list-assisted telephone sampling designs. *Public Opinion Quarterly*, 66(3), 321-338.
- Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73(361), 40-46.

Other Relevant References

- Bülow, E. (2009) "Use and Theory of Random Digit Dialing in Sweden" Available at <http://probability.univ.kiev.ua/school09/papers/bulow.pdf>.
- Heckel, C. and Wiese, K. (2012) "Sampling Frames for Telephone Surveys in Europe" in Häder, S., Häder, M. and Kühne, M. (eds). *Telephone Surveys in Europe: Research and Practice*. Berlin: Springer.
- Kuusela, V. and Simpanen, M. (2012) "Finland" in Häder, S., Häder, M. and Kühne, M. (eds). *Telephone Surveys in Europe: Research and Practice*. Berlin: Springer.
- McBride, O., Morgan, K. and McGee, H. (2012) "Recruitment using mobile telephones in an Irish general population sexual health survey: challenges and practical solutions" *BMC Medical Research Methodology* 12, 45 DOI: 10.1186/1471-2288-12-45.
- Mohorko, A., de Leeuw, E., and Hox, J. (2013). "Coverage Bias in European Telephone Surveys: Developments of Landline and Mobile Phone Coverage across Countries and over Time" *Survey Methods: Insights from the Field*. Retrieved from <http://surveyinsights.org/?p=828>.
- Vicente, P. and Reis, E. (2009) "The Mobile-only Population in Portugal and Its Impact in a Dual Frame Telephone Survey" *Survey Research Methods* Vol.3, No.2, pp. 105-111.

Annex 1. Code reference for sample implementation

An implementation of the stratification approach described in section 3 is available in the statistical (open-source) software R through the [SamplingStrata](#) package. One of the interesting features of this approach, is that it can be applied for multiple domains as well as for multiple target variables simultaneously. However, it requires the variables used for stratification to be categorical. In its most recent version, it also allows for the creation of spatial stratification. The outcome of this approach is the required sample size by domain as well as the within stratification.

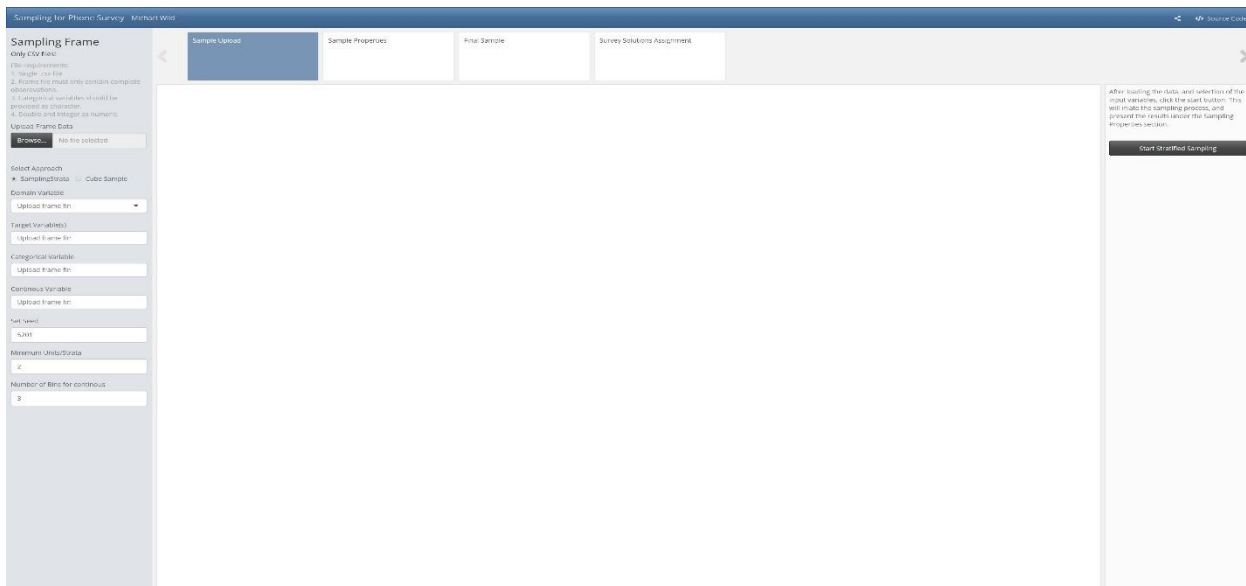
Also this approach is available in R, through the [sampling](#) packages and its *samplecube* function. A more recent implementation can be found in the [BalancedSampling](#) package, which also allows for the creation of spatially balanced samples. The theoretical motivation is the same as for the first approach, however it also allows for continuous variables to be used. Its aim is to receive the “same means in the population and the sample for all the auxiliary variables” (Tillé, 2010). In cases where certain (geographic) domains (i.e. provinces) are required for the estimation, a separate balanced sample needs to be created for each domain. One important difference of this approach is that the sample size is required to be calculated beforehand (i.e. through the formula described under Section 0) to provide individual inclusion probabilities and is not a result of the design process. Nevertheless, the efficiency gains may allow for a reduction in overall sample size.

To support both approaches for the current COVID19 initiative, Michael Wild has generated an R package containing a graphical user interface (GUI), available for local installation. The package is can be installed by executing `devtools::install_github("michael-cw/SurveySolutionsCOVID19tools", build_vignettes = T, force = T)`, and allows (after installation) to launch a GUI, which is used to collect the inputs for the above mentioned packages. This allows also R users with only basic knowledge of R (and the required packages) to apply this approach.

STEP 1: START THE GUI

After installation of the package, run the following commands in your Rstudio GUI⁴

```
library(SurveySolutionsCOVID19tools)
suso_covid19_samplingApp()
```



⁴ You may also use the native R interface, or any other GUI, however the underlying guide refers only to Rstudio.

This will open the application in your default browser (recommended browser: MS Edge or Google Chrome) with the start screen as shown in Figure 2.

STEP 2: UPLOAD SAMPLING FRAME

Uploading the frame is done through clicking on Browse... The frame file has to be in .csv format, and should ideally contain only the variables used for sampling, which are:

1. The target variable(s), i.e. employment status. In the case of the stratification module, multiple target variables are allowed, in the case of balanced sampling only a single variable can be chosen.
2. The domain variables (only for stratification, in case of balanced sampling you have to upload the frame and sample for each domain separately)
3. The variables used for stratification/balancing (see methods about requirements)

Another important requirement is that none of the variables used in any of the two approaches contains any missing values.

After uploading the available variables can be selected from the corresponding input fields to the left. The full data set can be inspected in the middle part of the application (**Error! Reference source not found.**).

The screenshot shows the 'Sampling for Phone Survey' application interface. The main window is titled 'Sampling Frame' and contains a table of data entries. The table has columns for various variables: CT, REG, COM, Nom, HApoly, PO0BMTOT, PO0BWTOT, Pop020, Pop2040, Pop4065, Pop65P, H00PTOT, H00P01, H00P02, H00P03, and H00P04. The table displays 10 entries, with the first entry being: 1, 1, 4, 261, Zurich, 8781, 175836, 187437, 57324, 131422, 108178, 66349, 186880, 94787, 55019, 17596, 1944.

On the left side, there is a sidebar with the following sections:

- File requirements:**
 1. Single .csv file
 2. Frame file must only contain complete observations.
 3. Categorical variables should be provided as character.
 4. Double and Integer as numeric.
- Upload Frame Data:**
 - Browse... sample_frame.csv
 - Upload Complete
- Select Approach:**
 - SamplingStrata (selected)
 - Cube Sample
- Domain Variable:** REG
- Target Variable(s):** Pop020, Pop2040, Pop4065, Pop65P
- Categorical Variable(s):** Select variable(s) below
- Continuous Variable(s):** Select variable(s) below
- Set Seed:** 6363
- Minimum Units/Strata:** 2
- Number of Bins for continuous:** 3

On the right side, there is a panel titled 'Start Stratified Sampling' with a table of CV1, CV2, CV3, and CV4 values for each entry. The values are: 1: 0.05, 0.05, 0.05, 0.05; 2: 0.05, 0.05, 0.05, 0.05; 3: 0.05, 0.05, 0.05, 0.05; 4: 0.05, 0.05, 0.05, 0.05; 5: 0.05, 0.05, 0.05, 0.05; 6: 0.05, 0.05, 0.05, 0.05; 7: 0.05, 0.05, 0.05, 0.05.

Figure 3: Data Upload

STEP 3A: SAMPLING STRATA

SamplingStrata requires the specification of several input parameters:

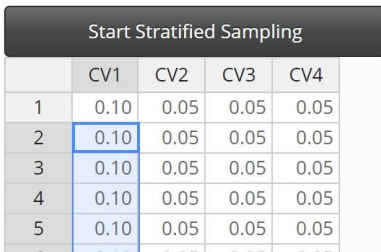
- i. Domain Variable
- ii. Target Variable(s)
- iii. Categorical variables used in the stratification
- iv. Continuous variables used for the stratification

Domain Variable

Select the (single) variable specifying the desired domain for the estimation. These can be geographic domains (i.e. provinces) or socio-economic domains (i.e. gender). The more domains you provide, the larger the sample size will be. If you only require the desired precision at the national level, your domain variable should include only a single value (i.e. 1).

Target Variable

After selection of the domain variable, you need to specify the variable of interest, which is: The variable for which you require estimates at the desired level of precision for each of the provided domains. After having done that, you will see the CV table to the right as shown in figure 4.



	CV1	CV2	CV3	CV4
1	0.10	0.05	0.05	0.05
2	0.10	0.05	0.05	0.05
3	0.10	0.05	0.05	0.05
4	0.10	0.05	0.05	0.05
5	0.10	0.05	0.05	0.05

Figure 4: CV table sampling strata with 10 % CV for target variable 1.

For each target variable, the table will contain a separate column, in the same order as the specified variables. The number of rows is determined by the number of desired domains. Each value in this table can be modified. This means, you can specify a separate CV for each domain and variable. In the following we change the desired CV from 5% to 1% for the first target variable only.

Stratification Variables

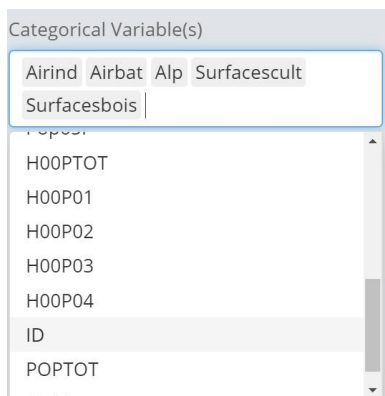


Figure 5: Selection of stratification variables

After setting the target variables, it is now time to select the stratification variables. Currently the stratification only works for categorical variables, which need to be provided as numeric inputs. However, you may also provide continuous variables, which are transformed to categorical. The transformation is described further down below.

Let's start with a set of categorical stratification variables for now, which can be selected as shown in Figure 5

That's it. You can now start the stratification by clicking on the *Start Stratified Sampling* button. A progress bar in the lower right corner will inform you, when the optimization is finished.

The screenshot shows the 'Sampling Frame' interface. The 'Sample Upload' tab is active, displaying a table with 10 entries. The table columns include CT, REG, COM, Nom, HApoly, P008MTOT, P008WTOT, Pop020, Pop2040, Pop4065, Pop55P, H00PTOT, H00P01, H00P02, H00P03, and H00P04. The right sidebar contains a 'Start Stratified Sampling' button and a table with columns CV1, CV2, CV3, and CV4. A progress bar and a search bar are also present. A notification box at the bottom right indicates 'Searching for Optimal Number of Strata ...'.

Figure 6: Stratification in progress

Attention: The application uses a genetic algorithm for the optimization and depending on the number of domains/target variables, this may require substantial computational resources. The function supports parallel execution; however, the availability depends on the number of (logical) CPU cores. If your system has 4 or less cores, the optimization will be carried out sequentially, and may take significantly longer to complete.

After the optimization is finished, it will run the evaluation of the results through simulation and present CV and bias, plus the sample sizes for each of the target variables as well as across the domains.

The screenshot shows the 'Sampling Frame' interface with the 'Sample Properties' tab active. The main area displays two box plots: 'Overall CV' and 'Overall Bias'. The 'Overall CV' plot shows the distribution of CV values for 6 domains. The 'Overall Bias' plot shows the distribution of bias values for 6 domains. On the right, a 'Simulation Results' table is visible, showing Mean, SE, CV, and bias for each domain. A 'Download Sampling Report' button is also present.

Domain	Mean	SE	CV	bias
1.00	212,064.73	14,590.02	0.07	0.03
2.00	141,129.18	8,327.97	0.06	0.05
3.00	166,366.87	10,540.14	0.06	0.07
4.00	278,988.84	12,217.26	0.04	0.01
5.00	283,787.96	14,536.91	0.05	0.02
6.00	302,365.05	16,711.98	0.06	0.06
1.00	206,884.24	12,648.63	0.06	0.01
2.00	140,473.14	8,535.12	0.06	0.04
3.00	165,560.16	10,819.61	0.07	0.06
4.00	278,862.01	12,604.57	0.05	0.01

Figure 7: Stratification results

Continuous Variables

In case were continuous variables are provided, a transformation to a categorical format is achieved by using the function:

```
SamplingStrata: var.bin()
```

which requires the specification of the number of desired categories to apply a K-means clustering. The default for this is 3. Changing this parameter to an unreasonable number of categories may result in nonconvergence of the optimization. In case you require a more granular categorization it is recommended to do this before uploading the data, with the software package of your choice.

Seed

For the purpose of creating reproduceable samples in both, the final sample as well as the random seed for the optimization, it is recommended to provide a seed value. Using this seed with the application, will allow you to always get exactly the same sample every time you run the stratification (assuming all inputs are the same). Therefore, it is recommended, to write down the seed together with the sample after creation of the final sample.

Minimum Number of units per stratum

For the final estimation it is helpful if you have at least 2 units in each stratum, however an increase of this parameter is recommended. Nevertheless, be careful, since increasing it too much may result in non-convergence of the optimization.

Evaluation of results

Selecting the Sample Properties section allows you to view the quality of the specified design, and if all restrictions on your CV are met. Currently the screen shows the CV for each variable, it's bias for the variables, and across domains, as well as total and domain sample sizes and number of strata. If you require this for a report, you may very well take a screenshot now.

Download

The download file is **.zip** compressed, and contains three files:

1. The original frame file, updated with the stratification IDs
2. The design file, containing all information about the design (i.e. domain, stratum id, sample size etc.)
3. The final sample including the weights.

The file name itself contains time and data, as well as the used seed. By using the seed in the application, the optimization and simulation will be reproduceable

STEP 3B: BALANCED SAMPLING

Switching to Balanced Sampling requires selection of Cube Sample after uploading the file. Which will also result in a slightly different set of inputs.

The screenshot shows a form titled "Balanced Sampling Interface" with the following sections and inputs:

- Select Approach:** Radio buttons for "SamplingStrata" (unselected) and "Cube Sample" (selected).
- Target Variable Type:** Radio buttons for "Categorical" (selected) and "Continuous" (unselected).
- Target Variable:** A dropdown menu with the text "Select variable bellow".
- Required CV:** A text input field containing "0.05".
- Sample Size:** A text input field containing "0".
- Balancing Variables:** A text input field with the text "Select variable bellow".
- Set Seed:** A text input field containing "1092".

Figure 8: Balanced Sampling Interface

Target Variable

The first required step is the selection of a *single* target variable, either continuous or categorical. If the latter, the categorical variable requires to be numeric, and coded with 0 and 1. After selection, and specification of the desired CV, the sample size window will contain the required sample size. This is only the theoretical one, in case you require more (i.e. to compensate for nonresponse), you may increase the value.

After specifying Target Variable and CV, the application automatically calculates the simple random sample size as a recommendation. Upwards and downwards adjustments are possible.

Balancing Variables

In the final step you need to specify the balancing variables (i.e. the variables for which you require the means to be equal to the means of your frame population). Having done so, allows you to start the cube sampling algorithm.

Seed & Download

Same as above

The screenshot shows a section of the interface with the following inputs:

- Target Variable Type:** Radio buttons for "Categorical" (unselected) and "Continuous" (selected).
- Target Variable:** A dropdown menu with the value "P00BMTOT".
- Required CV:** A text input field containing "0.05".
- Sample Size:** A text input field containing "1927".

Figure 9: Target variable, CV and Sample Size

Annex 2. Implementation of Weighting Procedures

A2.1. Weighting class or cell weighting adjustments and post-stratification

Due to the high rates of non-responses and attrition in phone surveys, the distribution of responses is often quite different from that of the initial sample. Table A1 illustrates such an example. The right panel shows how the target population is distributed in terms of two characteristics A and B while the left panel shows how the actual responses to a phone survey are distributed. Since they are so different, without compensating weights, summary statistics on A and B from the phone survey are very different from those of the target population. Furthermore, other statistics from the phone survey are also likely to be different.

Table A1. Respondent vs. Sample distributions

Respondent counts				
	B1	B2	B3	Total
A1	20	40	40	100
A2	50	140	310	500
A3	100	50	50	200
A4	30	100	70	200
Total	200	330	470	1000

Sample counts				
	B1	B2	B3	Total
A1	80	40	55	175
A2	60	150	340	550
A3	170	60	200	430
A4	55	165	125	345
Total	365	415	720	1500

A cell weighting or weighting class approach assigns a weight to each cell in the sample of the phone survey so that the weighted total of each cell becomes identical to the initial sample. In the case of list-based or RDD designs where respondents are selected with simple random sampling, the weighted total is the simple count of observations in each cell. In the case of a selection from an existing survey, the original weights carry forward into the new survey, as well as weights to compensate for any subsampling from the original survey. In these cases, the value of the sample cell is the sum of the weights instead of the count.

In example in table A1 and assuming a simple random sample, the combination of A1 and B1 has 20 respondents. But, the same cell in the initial sample represents 80 observations of the target population. To make them consistent, all observations in this cell in the sample distribution will have a weight of 4 ($80/20$). The cell weighting repeats this for all possible combinations of A and B. After the weighting exercise, the weighted distribution of the respondents will become identical to that of the initial sample. If instead of simple random sampling, this example was a subsample from an existing survey and the values represented the sum of the weights, instead of assigning each observation a weight of 4 in the A1/B1 cell, the weight of each observation would be multiplied by 4. This approach preserves the relative relationship between the respondents within the cell but adjusts the total cell sum to total target population.

A challenge of the cell weighting is to calculate weights for all possible combinations of target features for sampling and if the number of categories in each feature increases, the number of combinations can increase dramatically, along with the associated computational burden – though there are software packages to assist in the implementation.

A2.2. Calibration Rake/RIM weighting

In some cases, rake or rim weighting is used to reduce the computational burden of the cell weighting, though a more common application is to reweight samples when the cell-level values are unavailable, but row and column totals are provided. Raking is a commonly used approach to calibrate weights to population totals after nonresponse adjustments have been performed. This is an iterative procedure that focuses on one feature at a time to make the marginal distribution of the sample in terms of that feature identical to that of the target population, then proceeding to the next features, and repeating the process until convergence is achieved. The process is illustrated below.

Step 1. Calculate the weighted totals of the cells from the survey.

	B1	B2	B3	B4	Total
A1	79,586	125,489	22,566	4,581	232,222
A2	97,089	185,057	22,689	5,422	310,257
Total	176,675	310,546	45,255	10,003	542,479

Step 2. Compare those totals against the total from the auxiliary data

	B1	B2	B3	B4	Total
A1					281,839
A2					317,818
Total	179,897	359,794	47,973	11,993	599,657

Step 3. Rake across. Divide the total for row A1 in the auxiliary data by the total from the survey data ($281,839 / 232,222 = 1.214$) and replace for row A2 ($317,818 / 310,257 = 1.319$). Multiple the values in each cell of the respective rows. The total in the rows now match those in the auxiliary data.

	B1	B2	B3	B4	Total	
A1	96,590	152,301	27,387	5,560	281,839	1.214
A2	99,455	189,567	23,242	5,554	317,818	1.024
Total	196,046	341,868	50,629	11,114	599,657	

Step 4. Rake down. Divide the total in the auxiliary data by the new total after the first rake for column B1 ($179,897 / 196,046 = 0.784$). Repeat the procedure for columns B2, B3, and B4.

	B1	B2	B3	B4	Total
A1	88,634	160,287	25,951	6,000	280,871
A2	91,263	199,507	22,022	5,993	318,786
Total	179,897	359,794	47,973	11,993	599,657
	0.918	1.052	0.948	1.079	

Step 5. Repeat process until convergence is reached.

	B1	B2	B3	B4	Total
A1	88,927	160,865	26,028	6,019	281,839
A2	90,970	198,929	21,945	5,974	317,818
Total	179,897	359,794	47,973	11,993	599,657

Step 6. Divide the raked totals by the weighted totals from the survey data. These adjustments should then be applied to the weight of each observation in the cell.

	B1	B2	B3	B4
A1	1.117	1.282	1.153	1.314
A2	0.937	1.075	0.967	1.102

Weighting class and rake weighting are useful when there is a limited amount of information available for reweighting. With surveys using subsamples from existing datasets, more information is available that could feasibly

be implemented using one of these approaches. A more common approach for those situations is to use a propensity weight.

A2.3. Propensity score weighting (PSW)

The propensity score weighting assumes that whether to participate in a phone or web survey depends on some observable features of a household or individual. This approach was originally developed to make a control group comparable with a treatment group (Rosebaum and Rubin 1983 and 1984) but has been recently applied to make statistics from a phone or web survey comparable to those of a nationally representative survey (e.g., Terhanian, et al. , 2000, Schonlau, et al., 2006, Lee, 2006, and Cappaci et al., 2018). Profiles of voluntary participants for a web survey are often concentrated to specific groups and very different from a nationally representative one. As a result, summary statistics can be very different from those of a nationally representative survey. But, reweighting by the propensity score can make the statistics from the web survey comparable to the nationally representative ones.

To carry out PSW, we need to have a household survey that is representative for a target population and a phone or web survey that include voluntary participants. The goal of the PSW is to estimate a set of new weights so that weighted average of summary statistics in a phone or web survey are very similar to summary statistics of the household survey. The procedures below describe the PSW approach in cases where the sample is selected from an existing baseline and where the propensity score model is being developed using data collected in survey itself, as would be the case in a list-based or RDD approach.

A2.3.1. Survey as baseline (based on Himelein, 2014).

1. Construct a variable in the baseline dataset which denotes if the household responded in telephone survey. This variable is the dependent variable in the regression.
2. Identify covariates in the baseline dataset that may explain the likelihood of a respondent participating in the phone survey. Note that characteristics like ownership of a mobile phone will perfectly predict failure, but other asset and dwelling characteristic variables are often strong predictors. Sector of employment, education, remoteness (as proxied by distance to major infrastructure) also tend to be useful variables. In contrast to standard analysis, the model does not have to be a structural model. The coefficients themselves have no interpretation and it is a rare case when omitted variable bias is actually an asset to the analyst.
3. Perform a logistic regression model to determine the likelihood of non-response or attrition based on the household and/or individual characteristics.
4. Divide the continuous measure of the likelihood of attrition into deciles and collapse to the mean. This value should then be applied to the weights as the non-response adjustment.
5. Perform a simple check. Depending on how the logistic regression is specified (i.e. whether 1 = non-response or 1 = remain), it may be necessary to take the reciprocal of the non-response adjustment. A simple check is to make sure that the non-response adjustment is greater than 1 for those respondents that had higher levels of non-response (i.e. the value of the weights is increasing to compensate for those with similar characteristics but did not respond).

A2.3.2. List-based or RDD designs

1. Appropriate covariates are identified with the assumption that the condition of strongly ignorable treatment assignment is met either exactly or approximately (i.e. non-response is more or less orthogonal to the impact being measured). These covariates, which include demographic, behavioral, attitudinal, and topic-specific variables, are included in both the telephone survey and the auxiliary data (hopefully a recently collected representative dataset).
2. Data from the household survey and a phone or web survey are merged.
3. The appropriate propensity score model is estimated using logistic regression, and respondents from the household survey and phone/web surveys are sub-classified based on their propensity scores. It must be noted at this point, that there must be sufficient overlap in each subgroup between the household and phone/web

surveys. If there is not, then the condition of strong ignorability is not met, and the weighting procedure will fail. This is a consideration when constructing a propensity model. As a result, each subgroup has sample households from both the household survey and the phone/web survey.

4. Then for each subgroup, calculate what is the share of households in the household survey belonging to this subgroup, what is the share of households in the phone/web survey belonging to this subgroup, and the ratio of the former to the latter. A weight for the phone/web survey is adjusted by multiplying this ratio with the original weight.
5. Some prefers to do further adjustment using the rake weighting with the propensity stratification as one feature and traditional demographic variables as others (e.g., Terhenian, et al., 2000).

A more precise mathematical description of this process is available in Lee (2006).

A2.4. Detailed description of Propensity Score Weighting (PSW) in Lee (2006)

Suppose that there are two samples: (a) a phone survey sample (s^p) with n^p units each with a base weight of d_j^p where $j = 1, \dots, n^p$, and (b) a reference survey (usually a nationally representative household survey) sample (s^R) with n^R units each with a base weight of d_j^R where $j = 1, \dots, n^R$. First, the two samples are combined into one, $s = (s^p \cup s^R)$ with $n = n^p + n^R$ units. We calculate propensity scores from the combined data, s . The propensity score of the i th unit is the likelihood of the unit participating in the phone survey ($g=1$) rather than the reference survey ($g=0$), where $i = 1, \dots, n$, given auxiliary variables. Propensity scores are defined as $e(x_i) = \Pr(i \in s^p | x_i, i = 1, \dots, n)$ and estimated in a logistic regression as in the following equation using covariates observed commonly in the phone and the reference survey, x_{obs} .

$$\ln \left[\frac{e(x)}{1 - e(x)} \right] = \alpha + \beta^T f(x)$$

where $f(x)$ is some function of the covariates. Critical assumptions in doing this are (a) that given a set of covariate values, a person must have some nonzero probability of being in the phone survey and (b) that probability must be estimable from the combined sample, s

Based on the predicted propensity score, $\hat{e}(x_{obs})$, the distribution of the phone sample units is rearranged so that s^p resembles s^R . Mechanically, this is first done by sorting s by $\hat{e}(x_{obs})$ and portioning s into C subclasses, where each subclass has about the same number of units. Alternatively, one might use only s^R in this subclassification. However, the objective of this study is to evaluate the current practice, which uses s . Based on Cochran (1968), the conventional choice is to use five subclasses based on quintile points. Ideally, all units in a given subclass will have about the same propensity score or, at least, the range of scores in each class is fairly narrow. In the c th subclass, denoted as s_c , there are $n_c = n_c^p + n_c^R$ units, where n_c^p is the number of units from s^p , and n_c^R from s^R . The total number of units s remains the same because

$$\sum_{c=1}^C (n_c^p + n_c^R) = \sum_{c=1}^C n_c = n$$

Second, we compute the following adjustment factor:

$$f_c = \frac{\sum_{k \in (s_c^R)} d_k^R / \sum_{k \in (s^R)} d_k^R}{\sum_{j \in (s_c^p)} d_j^p / \sum_{j \in (s^p)} d_j^p}$$

Where s_c^R and s_c^p are the sets of units in the reference sample and phone survey sample of the c th subclass.

The adjusted weight for unit j in class c of the phone survey sample becomes

$$d_j^{p.PSA} = f_c d_j^p$$

The estimator for the mean of a study variable, y , from s^p becomes

$$\bar{y}^{p.PSA} = \frac{\sum_c \sum_{j \in (s_c^p)} d_j^{p.PSA} y_j}{\sum_c \sum_{j \in (s_c^p)} d_j^{p.PSA}}$$

Note that the reference sample is not used for estimating the mean. It is needed only in the adjustment process and, thus, is required to have only the covariate data, not necessarily the variables of interest. The same reference sample can be used for the adjustment of more than one web survey as long as its target population coincides with that of the phone survey samples and the temporal circumstances are equivalent.

Additional references for Annex 2

Capacci, S., M. Mazzocchi, and S. Brasini. (2018). "Estimation of unobservable selection effects in on-line surveys through propensity score matching: An application to public acceptance of healthy eating policies." PLoS ONE 13(4): e0196020. <https://doi.org/10.1371/journal.pone.0196020>

Lee, S. (2006). "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics*. 22 (2): 329–349.

Rosenbaum, P. R., and D. B. Rubin. (1983). "The Central Role of the Propensity Score in Observational Studies for Casual Effects." *Biometrika* 70 (1): 41-55.

Rosenbaum, P.R., and D.B. Rubin. (1984). "Reducing Bias in Observational Studies using Subclassification on the Propensity Score." *Journal of the American Statistical Association*. 79: 516-524.

Schonlau M., A. van Soest, A. Kapteyn, and M. Couper (2006). "Selection Bias in Web Surveys and the Use of Propensity Scores." RAND Labor and Population Working Paper series 229. RAND Pittsburgh, PA.

Terhanian, G., J. Bremer, R. Smith, and R. Thomas. (2000). Correcting Data from Online Survey for the Effects of Nonrandom Selection and Nonrandom Assignment. Research paper: Harris Interactive.