Combining Quota and Probability Sub-Sampling Within Enumeration Areas to Produce More Reliable Estimates

Isabela Bertolini Coelho[1], Marcelo Trindade Pitta[1], Pedro Luis do Nascimento Silva[2]
[1]Brazilian Network Information Center (NIC.br), Avenida das Nações Unidas, 11541 - 7th floor, SP 04578-000, Brazil
[2]National School of Statistical Sciences, Rua André Cavalcanti, 106, RJ 20231-050, Brazil

Abstract

Traditional surveys face increasing challenges due to rising non-response rates and the diminishing resources available to survey organizations. A recently proposed solution involves the combination of non-probability sample surveys (often cheaper) with probability sample surveys (more expensive), using the latter as a reference to weight the former. Considering a special case in which a single survey was designed and carried out by simultaneously using the two sampling approaches in a single field operation, this paper compared the use of quasi-randomization and sample matching methods to assign weights to the non-probability sample. The quasi-randomization method provided the closest point estimates and smaller standard errors (on average) when compared to the benchmark estimates.

Key Words: non-probability sampling; quota sampling; probability sampling; combining samples; survey weighting

1. Introduction

Traditional surveys for official and public statistics often use probability sampling designs. These face increasing challenges due to rising non-response rates and the diminishing resources available to survey organizations. At the same time, they face increasing demand for more timely and disaggregated data. This challenging scenario implies that surveys need to produce more with less.

This happens at the same time when new data sources emerge as potential alternatives: mobile phones, satellite images and other remote sensors, social networks, business transactions, etc. all generate data which were not designed to yield estimates or make inference such as those obtained from traditional surveys or censuses. Even though some of these data do not conform to production standards adopted for most official statistics, one cannot ignore their existence.

Consequently, use of alternative data sources or combinations of data sources, and data collection using mixed modes have been considered aiming to satisfy the increasing demands without increasing the costs, while trying to maintain the quality standards.

A recently proposed solution involves the combination of non-probability sample surveys with probability sample surveys to obtain the required estimates, by using the latter as a reference to weight the former. This is an interesting option: non-probability sample surveys are often cheaper to conduct than strict probability sample surveys, implying that sometimes larger samples can be collected, though often at the cost of some bias. The main idea is to combine the (cheaper) non-probability sample survey with a traditional (more expensive) probability sample to mitigate this bias of the non-probability sample by assigning pseudo-probability weights to it - (Elliott and Valliant 2017), (Valliant 2019), (Elliott 2009), (Dever 2018).

For reference, considering the experience of the Brazilian Network Information Center (NIC.br) with some data collection service providers, an interview obtained through standard probabilistic area household sampling costs about three times as much as an interview obtained through quota sampling. This type of cost ratio may vary from country to country and from survey to survey, but this serves to illustrate the reasons why some surveys still use quota or other cheaper non-probability sampling designs. Collection service providers also argue that quota samples enable faster turnaround time, since returns are not needed to find the selected residents if they are not at home when the household is recruited into the survey and a resident is sampled for interview.

Most of the available references consider the case where neither the probability nor the non-probability surveys were planned for this kind of survey combination exercise. In the scenarios discussed in the literature, both samples are independent and carried out by different institutes or for different purposes, which means that most of the time the target surveys use different questionnaires, reference periods, sampling frames, etc.

In this article we consider the special case in which a single survey was designed and carried out using the two approaches simultaneously in a single field operation. In 2011, the design adopted for the Survey on the Use of ICT in Brazilian Households - hereafter called the ICT Households Survey (ICTHS for short) - used strict probability sampling up to the selection of enumeration areas (EAs). Then within each sampling stratum or municipality, half of the sample EAs had sub-sampling of households and individual respondents carried out by traditional probability sampling methods, and the other half had sub-sampling of households and individual respondents carried out using quota sampling.

This design was adopted as a transition strategy between a period (2005-2010) when the survey used only quota sampling for households and individual respondents in the last sampling stage, to a new period when the survey started using probability sampling for households and individual respondents in the last sampling stage – 2012 and forward. This enabled this single year (2011) survey to be a bridge between the past and future series after a major survey redesign took place. One of the reasons for moving the survey away from using quota sampling for households and individual respondents in the last sampling stage was the evidence of bias in the estimates of some key parameters of interest observed in the 2010 edition of the survey.

Our aim here is to combine the data from the two half-samples collected in 2011 to increase the precision of the survey estimates, and hence possibly enable production of estimates for more detailed domains than those that were used to publish estimates obtained solely from the ‘full probability’ half-sample. Even though the data is now out of date, this is a case where both samples were collected with the same questionnaire and for the same reference period, using the same fully standardized methods, except for the approach used to sub-sample households within enumeration areas. Therefore, in this instance the common support assumption required for the survey combination approach is easier to justify.

We compare some methods presented in the literature for weighting the quota half-sample. This enabled assessing the feasibility and relative merits of some of the approaches recently proposed for weighting a survey that uses quota sampling methods for household / respondent sub-sampling. It also provided an opportunity to assess the strengths and weaknesses of a mixed design approach to conducting the future ICT Households Surveys in Brazil, possibly with larger sample sizes, made possible by the lower per-unit collection costs associated with quota sampling for household sub-sampling within EAs.

The article is divided in four sections, besides this Introduction. Section 2 describes the data and sample design for the 2011 ICTHS. Section 3 describes the methods for weighting the quota half-sample. Section 4 presents the numerical results and Section 5 contains our conclusions and indication of future work.

2. Data

Since 2005, the Regional Center for Studies on the Development of the Information Society (Cetic.br), a department of the NIC.br, has been collecting data about access, use and appropriation of Information and Communication Technologies (ICT) for several segments of the Brazilian society. The ICTHS measures the access and use of ICT by private permanent household residents in Brazil (Brazilian Internet Steering Committee 2012).

The sample design adopted since 2005 is a stratified multistage sample of households, using selection with probabilities proportional to size in some sampling stages. From 2005 to 2010 the last stage of the sample selection used quota sampling to select households within sampled EAs, considering some characteristics of individual respondents to fulfill known quotas (provided by official statistics estimates).

In 2011, Cetic.br decided to change the sample design to apply strict probability sampling in all sampling stages, including the sub-sampling of households and individual respondents within EAs. In order to make a smooth transition to the new approach, and in order to build a bridge in the time series of indicators available for public analysis after the change, Cetic.br opted to split the 2011 sample in two half-samples:

- half of the sample was carried out with the sample design as used before 2011, this is to say, using quota sampling for selecting households and individual respondents within EAs;

- half of the sample would use the strict probability sampling in all stages, including for sub-sampling households and individual respondents within EAs.

2.1 Sample design description

The target population for the ICTHS comprises private permanent households in Brazil and their residents 10 years or older. To cover the population, given the lack of up to date frames of households, the sample selection used a frame of EAs as defined for the Population Census of 2010, available from the Brazilian Institute for Geography and Statistics (IBGE). This is an area frame, with data from the 2010 Population Census associated with each of the EAs, besides the geographic information of location and boundaries, as well as description of the routes taken by the census interviewers to canvass each EA during the Census.

Considering the resources available at the time, the total sample size of ICT Households survey was set at 25,000 households / individuals (one eligible individual is selected per household). This total size was divided equally between the two half-samples – i.e. 12,500 households for the quota half-sample and 12,500 for the full probability half-sample.

Brazil's territory is divided into 27 federative units – 26 states and a Federal District where the capital city of Brasilia is located. The survey stratification was defined solely based on geography. First, all 27 state capital cities were included with certainty in the sample, thus turning each of these capital cities into strata (called "capitals"). Next, in each of nine states (Pará, Ceará, Pernambuco, Bahia, Minas Gerais, Rio de Janeiro, São Paulo, Paraná, and Rio Grande do Sul) a second stratum was created within the state (besides their capital cities), by grouping the municipalities that comprise each of the metropolitan areas centered around the state capitals (called "metropolitan"). In each of these nine states, the remaining non-metropolitan municipalities were included in a third stratum referred to as "non-metropolitan". Next, in each of the 13 remaining states, and in a group formed by the states of Acre, Amapá, Rondônia, Roraima, and Tocantins, a stratum was defined by grouping the remaining non-capital municipalities (called "interior").

Hence, the geographic stratification produced a total of 27 (capitals) + 9 (metropolitan) + 9 (non-metropolitan) + 14 (interior) = 59 strata for sampling of primary sampling units (PSUs). Within the capitals and metropolitan strata, EAs were the PSUs. Within the non-metropolitan and interior strata, municipalities were the PSUs, with EAs as the secondary sampling units (SSUs). These 59 strata are grouped, for analysis purposes, into 32 ICTHS Strata, as described in Table 1.

Overall, the sample included 317 municipalities across the country. To determine the number of EAs to select in each municipality, the number of households to sample in each EA was set to 10. EAs were distributed in municipalities proportionally to the total population aged 10 years old or older. Finally, 1,250 EAs were selected in these 317 municipalities for each of the two half-samples (Table 1).

Table 1 – Stratification and sample size allocation used in ICTHS 2011

Regions	ICT Strata	Number of units		
		Municipalities	EAs	Interviews
Southeast	Espírito Santo	8	56	560
	Minas Gerais – Capital + Metropolitan	10	76	760
	Minas Gerais – Interior	15	156	1,560
	São Paulo – Capital + Metropolitan	20	180	1,800
	São Paulo – Interior	27	200	2,000
	Rio de Janeiro – Capital + Metropolitan	17	136	1,360
	Rio de Janeiro – Interior	10	64	640
Northeast	Alagoas	7	52	520
	Salvador – Capital + Metropolitan	7	56	560
	Bahia – Interior	15	120	1,200
	Ceará – Capital + Metropolitan	7	56	560
	Ceará – Interior	10	72	720
	Maranhão	11	88	880
	Paraíba	8	60	600

	Pernambuco – Capital + Metropolitan	9	60	600
	Pernambuco – Interior	9	72	720
	Piauí	7	52	520
	Rio Grande do Norte	7	52	520
	Sergipe	5	40	400
South	Paraná – Capital + Metropolitan	8	52	520
	Paraná – Interior	12	96	960
	Santa Catarina	12	84	840
	Rio Grande do Sul – Capital + Metropolitan	10	60	600
	Rio Grande do Sul – Interior	12	92	920
North	Amazonas	7	56	560
	Pará – Capital + Metropolitan	5	40	400
	Pará – Interior	6	76	760
	Rondônia/Roraima/Acre/Amapá/Tocantins	10	72	720
Center-West	Distrito Federal	1	44	440
	Goiás	9	84	840
	Mato Grosso	9	52	520
	Mato Grosso do Sul	7	44	440
Total		317	2,500	25,000

Source: Adapted by the authors from Brazilian Internet Steering Committee, 2012.

As mentioned above, the two half-samples only differed in the selection of households and residents to be interviewed within the sampled EAs. All things considered, the design used to obtain the samples of households and residents can be described as a stratified multistage sample in three or four stages, depending on the stratum.

The number of stages in the sampling design depends primarily on the role given to the selection of municipalities. Several municipalities were sampled with probability equal to one (certainty municipalities). In this case, these municipalities functioned as strata for sampling EAs and the design had three sampling stages: EAs, households and residents (for households having more than one eligible resident). The three-stage probability sampling design can be described as:

- first stage, selection of census EAs with probability proportional to the population 10 years or older in 2010;

- second stage, selection of households through inverse simple random sampling (Vasconcellos, Silva, and Szwarcwald 2005);

- third stage, simple random sampling of eligible resident to answer the survey questionnaire, after compiling a list of all household residents, also called as Kish Grid (Kish 1949).

In the 23 strata defined above where the municipalities are not included with certainty in the sample, they are the PSUs, and the design had four stages: municipalities, EAs, households and residents. The four-stage probability sampling design can be described as:

- first stage, selection of municipalities with probability proportional to the population 10 years or older in 2010;

- second stage, selection of census EAs with probability proportional to the population 10 years or older in 2010;

- third stage, selection of households through inverse simple random sampling (Vasconcellos, Silva, and Szwarcwald 2005);

- fourth stage, simple random sampling of eligible resident to answer the survey questionnaire, after compiling a list of all household residents, also called as Kish Grid (Kish 1949).

For the quota half-sample, the initial sampling stages were the same as for the probability sample. The only difference is in the household / respondent selection stage. The selection of municipalities was done just once and the municipalities in both samples are the same. The selection of census EAs was carried out to obtain 2,500 EAs. This sample was then split randomly, half for the probability sample design (described above) and half for the quota design, always within each municipality. To ensure samples were balanced across the country, even numbers of EAs were allocated to each sampled municipality, so that the random split would generate samples of exactly the same size in terms of municipalities and EAs for each of the half-samples.

For the quota sample, the selection of households and respondents within EAs was based on population profile quotas by sex, age group, level of education, and economic activity status, according to official figures from the 2000 Population Census and the 2009 National Household Sample Survey (PNAD) also conducted annually by IBGE.

In some aspects, the procedure used to select households for the quota half-sample was based on a procedure similar to systematic sampling once it started by selecting a household at random within a census EA and in this household, the interviewer would try to carry out the survey. If the selected household did not meet an established quota, was ineligible or empty, or refused to take part in the survey, the interviewer would go to the neighboring address (physically located next door) and attempt to carry out the survey. If an interview was obtained in the first household, the interviewer would skip the following three addresses in the address list available and visit the fourth address to establish contact and attempt an interview. This procedure continued until all pre-established quotas were fulfilled.

As a very well-known disadvantage of the non-probability samples, this procedure makes it impossible to calculate the exact inclusion probabilities for both households and residents selected via in the quota half-sample. This happens because the inclusion of each household in the sample depends on characteristics of its residents, availability and agreement for answering the survey, as well as on the results of previous attempts and choice of respondent at interviews (already fulfilled quotas). Another difficulty arises from the fact that interviewers operating the quota subsampling method typically do not need to document details of their efforts to obtain the required number of interviews, which makes it hard even to compare the relative costs of this approach with those applying the strict probability subsampling protocols within sampled EAs.

In summary, the sampling design can be described as shown in Table 2.

Table 2: 2011 ICT Household Survey Design

Selection Stages	Strata composition			
	Federative units or administrative areas (metropolitan and non-metropolitan)		Municipalities (state capital and largest municipalities)	
	Sampling Unit	Selection Procedure	Sampling Unit	Selection Procedure
Primary sampling units (PSU)	Municipality	Probability proportional to population size	Enumeration area	Probability proportional to population size
Secondary sampling units (SSU)	Enumeration area	Probability proportional to population size	Households	*Probability sampling* – inverse sampling of 10 households in each enumeration area *Quota sampling* - based on persons' profile
Tertiary sampling units (TSU)	Households	*Probability sampling* – inverse sampling of 10 households in each enumeration area *Quota sampling* - based on persons' profile	Individuals	*Probability sampling* - Simple random sample of one resident 10 years old or more *Quota sampling* - based on persons' profile

Final (or Fourth) sampling units	Individuals	*Probability sampling* - Simple random sample of one resident 10 years old or more *Quota sampling* - based on persons' profile		

2.2 Study variables

The ICTHS collects more than 50 indicators for households and residents in each edition of the survey. Most of them follow the international recommendation (ITU 2014) and others are adapted to Brazilian context. Among these indicators, four household level indicators were selected for analysis:

- Proportion of households with computer (which includes desktop computer, notebook, or tablet) – denoted by y_1;
- Proportion of households with computer, by type of computer – denoted by y_2;
- Proportion of households with Internet access – denoted by y_3; and
- Proportion of households with Internet access, by connection speed – denoted by y_4.

The types of computer considered for y_2 are: Desktop (y_{21}); Notebook (y_{22}) and Tablet (y_{23}). The ranges of connection speeds considered for y_4 are: up to 256 Kbps (y_{41}); above 256 Kbps up to 1 Mega (y_{42}); above 1 Mega up to 2 Mega (y_{43}); above 2 Mega up to 4 Mega (y_{44}); above 4 Mega up to 8 Mega (y_{45}); above 8 Mega (y_{46}); and does not know/ did not answer (y_{47}).

The results and evaluations of methods obtained here were made with reference to estimates and corresponding standard errors for the indicators defined by the above variables.

3. Methodology

Combining non-probability samples and probability samples has been the subject of several recent publications (Elliott 2009; Elliott and Valliant 2017; Rafei, Flannagan, and Elliott 2020; Valliant 2019; Valliant and Dever 2011; Dever 2018; Buelens, Burguer, and Van Den Brakel 2015). As discussed in these articles, combining samples to obtain improved results requires a set of characteristics are present in both samples. Those characteristics are described by Valliant (2019) as the 'common support', and include the following:

- Both samples should address the same target population;
- Both samples should collect a range of variables in the same way (same questions);

- Both surveys should be carried out for the same reference period; and
- Samples to be combined should have no intersection of respondents.

The survey considered in this article is a special case in which all these requirements were fulfilled by design. It is perhaps a rare case which makes this study on the use of methods for combining the samples attractive.

We selected two approaches to apply for combining the probability and quota half-samples: Quasi-Randomization (QR) and Sample Matching (SM). These methods were chosen because their implementation is relatively simple, and both enable obtaining a single set of weights for all the units in the non-probability half-sample. Superpopulation and other model-based approaches lead to increased complexity when modeling is required for many target survey variables and were therefore not considered here.

3.1 Quasi-Randomization (QR)

This method estimates pseudo-inclusion probabilities (PIPs) for the respondents of the non-probability sample, to use their reciprocals as weights for point and variance estimation. The basic idea of the method is to estimate the probability of being selected and answer the survey based on auxiliary variables $(x's)$ related to the profile of the respondents. This method considers the hypothesis that, given the auxiliary variables $(x's)$, the inclusion probabilities are independent of the target survey variables $(y's)$, that means the inclusion probabilities are Missing at Random (MAR).

To model and estimate those PIPs, data from both samples, the probability and non-probability sample, are bound together and the inclusion probabilities are estimated through a binary regression model considering the survey design. The probability sample, used as a reference sample, must represent the whole target population, being adjusted for non-response and expanding to known population totals (Valliant 2019). The process used to estimate the PIPs can be described in three steps:

1. Combine the samples in one file (top-down) and create an indicator variable z taking value 0 for the probability sample and 1 for the non-probability sample;
2. A column with weights used in this file will have the weights from the probability sample (for its cases) and a weight of one for all cases in the non-probability sample;
3. Use the combined data set to fit a logistic regression model, considering the complex sample survey design, to estimate the inclusion probabilities of being in the non-probability sample.

As result, the model estimates a pseudo-inclusion probability for each case in the file. In the sequence, the reciprocals of these estimated PIPs are used as weights for weighting the non-probability sample cases. These estimated weights are then used in all complex sample data analysis carried out using the non-probability sample. In this paper, estimates produced by this method were compared with estimates obtained from the reference probability sample, which are considered as the benchmark.

The variables from ICTHS 2011 used to fit the logistic regression model included the following.

- **Type of family** - classification based on number of residents within the household and sex of the respondent into 4 categories: one household member and male; one household member and female; 2 household members (spouse or not); and 3 or more residents.

- **Region** - division of Brazil into five macro-regions, according to the IBGE, namely Center-West, Northeast, North, Southeast and South.
- **Social class** - economic classification based on a scoring system that divides households into four 'socioeconomic' classes: A, B, C and D/E.
- **Economic activity status** - refers to work status with seven alternative answers, as shown in Table 3, which can be grouped into two categories: Economically active population and Economically inactive population.
- **Level of education of the head of household** - refers to completion of specific stages of formal education, divided into eleven subcategories, ranging from Illiterate or Pre-school to Tertiary Education or above.
- **Sex of the head of the household** – with codes for female and male according to respondent's declaration. There were some cases in which the respondent does not answer this question (coded as 98).
- **Number of residents** with 10 years old or more within the household.
- **Household area status** – with codes for urban or rural, depending on where it is located and as defined for the 2010 Population Census.
- **Age of head of the household** - expressed in years, as calculated on the day of the interview.

Table 3. Classification of economic activity status for the ICTHS respondents.

Code	Occupational status	Economic activity status
1	Working, even if with no formal registration	Active
2	Working as an apprentice, assistant etc.	
3	Worked or attempted to work in the previous week	
4	Unemployed	
5	Housewife not working	Inactive
6	Retired, pensioner	
7	Student not working	

Source: Brazilian Internet Steering Committee 2012.

3.2 Sample Matching (SM)

The sample matching approach aims to reduce selection bias by matching the non-probability sample to a control group using one or more characteristics of auxiliary data (Baker et al. 2013). This method can be applied in two different ways:

1. Treating the values of response variables *y* in the reference sample as missing values, and imputing these values based in the non-probability sample; or
2. Treating the weights in the non-probability sample as missing values and imputing them based in the weights present in the reference sample.

The first option is not the case in our study, since both samples have the full set of target survey variables available. Hence, we adopted the second approach, by

borrowing weights from the reference sample to the non-probability sample, using two distinct imputation methods: K-Nearest-Neighbor (KNN) and Hot Deck (HD).

As described in Elliott and Valliant (2017), a variation of matching is to match units in the non-probability sample with those in the probability sample. Each unit in the non-probability sample is then assigned the weight of its matching record in the probability sample. The matching is done using auxiliary variables ($x's$) related to the profile of the respondents. Here we used the same variables considered when modeling for estimating the PIPs in the QR approach.

KNN algorithms have been used as a machine learning tool for different purposes and can be used for classification or regression. In general, the method predicts a missing response variable by aggregating the observed values from K nearest neighbors (Buelens, Burguer, and Van Den Brakel 2015) defined by a distance function. In our study, for each case in the non-probability sample, the KNN method assigned the median of the weights of the K adjacent cases in the reference sample.

The HD method is common in survey practice for imputation because it selects, for each case requiring imputation, one donor at random from a pre-defined number of cases considered similar in the reference sample (Little and Rubin 2002). For our application, for each case in the non-probability sample, the HD method was used to randomly select one donor from a number of similar cases in the reference probability sample and assign the donor record's weight to the receiving case in the non-probability sample.

3.3 Variance estimation

In order to estimate the variance of each estimate for each of the study's target indicators, the *jackknife* approach, as proposed in Valliant (2019, pages 9 and 10), was used in order to include the variation due to estimated weights or pseudo-weights. As our case of study has a PSU even for the quota sample, which is not common, we use the PSUs to form deletion groups. The algorithm comprises the steps described below:

Step 0 - Delete one PSU from the reference sample and one PSU from the non-probability sample;
Step 1 – Re-calibrate the reference probability sample weights to known population totals adjusting for the deleted PSUs;
Step 2 - Apply the three methods described above (QR, SM-KNN, SM-HD) for weighting the non-probability sample;
Step 3 - Calibrate the resulting non-probability sample weights to known population totals;
Step 4 - Calibrate the combined sample weights to known population totals;
Step 5 - Estimate the target indicators from the reference, non-probability and combined samples and store them;
Step 6 - Repeat steps 1 to 5 until all the PSUs have been deleted once from each sample.

The result of this procedure is a database with as many replicate estimates as there are PSUs in both samples. Based on these replicates we estimate the variance through the Equation 3 – page 10 – in Valliant (2019).

$$v_j(\hat{\theta}) = \frac{G-1}{G}\sum_{g=1}^{G}\left(\hat{\theta}_g - \hat{\theta}\right)^2,$$

where G is the number of PSUs, $\hat{\theta}_g$ is the estimate based in the sample without PSU g, and $\hat{\theta}$ is the estimate based on the whole sample. We present the results for the selected indicators using the estimated standard errors (SE) expressed in percentage, since all our target indicators are percentages:

$$SE_j(\hat{\theta}) = 100 \times \sqrt{v_j(\hat{\theta})}.$$

4. Results

This section presents the results obtained through the application of the methods described above. It is expected that non-probability samples would produce biased estimates, even when they are planned as the one described: probability sampling in all stages, except for the sub-sampling of households and respondents within the EAs. To illustrate this potential issue, Table 4 shows the unweighted frequency profile of units observed in the two samples.

Table 4: Sample frequency profiles of respondents in the two half-samples

Profile characteristics	Sample design	
	Quota	Probability
Household Region		
Southeast	35%	35%
Northeast	31%	31%
South	15%	15%
North	10%	10%
Center-West	9%	9%
Household area status		
Urban	86%	85%
Rural	14%	15%
Household size		
One person	7%	12%
Two people	18%	24%
Three people	25%	26%
Four people	24%	21%
Five people	14%	10%
Six people and more	12%	8%
Respondent - Sex		
Male	48%	42%
Female	52%	58%
Respondent - Age		
10 to 15 years old	13%	8%
16 to 24 years old	19%	16%
25 to 34 years old	20%	20%
35 to 44 years old	17%	16%
45 to 59 years old	19%	22%
60 years old or older	13%	19%
Respondent - Level of education		
Illiterate/Pre-school	8%	9%

Elementary	51%	46%
Secondary	29%	32%
Tertiary	11%	13%
Respondent - Economic activity status		
Economically active	55%	53%
Economically inactive	45%	47%

For stratification characteristics such as region and household area status, both half-samples present the same percentage of respondents in each possible category. When other characteristics are considered, the frequency distributions exhibit non-negligible differences between the half-samples.

Given the choice of marginal distributions used in establishing the quotas, the quota half-sample frequency profiles are closer to known population profiles concerning sex, age groups, level of education and economic activity status. An important difference is observed between the frequency profile of respondents by household size. This is likely due to random variation and the collection protocol employed in the probability sample. Comparing these frequency profiles between the two half-samples can indicate the direction of the bias that the estimates from the quota half-sample may have.

In the sequence, we assigned two different sets of initial weights to the quota half-sample without using the probability half-sample: the first set considers the quota sample as a simple random sample (SRS) (Quota_w1) whereas the second uses the EA weights (which followed the probability sampling protocol used to select EAs), and treat only the sub-sampling of households and respondents within EAs as if it had been done by SRS (Quota_w2). The first set of weights corresponds to the approach of ignoring the design actually used and treating the resulting sample as if obtained from a stratified sampling procedure with proportional allocation along each of the margins of the variables used to define quotas. This was in fact the approach used by the ICTHS to weight the survey before the 2011 edition. The second set of weights was used to obtain estimates from the quota half-sample in the 2011 edition of ICTHS.

Table 5 shows these estimates for the target study indicators (Section 2.2) obtained using these two sets of weights. We also computed pooled-sample estimates by combining the half-samples, dividing their corresponding weights by two and then calibrating the polled-sample weights using raking to the known population totals. We used the function *svyby* from the *survey* package available in R to obtain all these weighted point estimates.

Table 5. Point estimates for selected indicators (%)

Method →	**Probability**	**Quota_w1**		**Quota_w2**	
Indicator ↓	**half-sample**	**Quota_w1**	**Pooled**	**Quota_w2**	**Pooled**
y_1	42.6	43.7	43.1	41.7	42.1
y_{21}	76.6	78.0	77.3	78.5	77.6
y_{22}	41.4	39.5	40.4	38.9	40.2
y_{23}	1.5	1.3	1.4	1.2	1.4
y_3	35.8	36.4	36.1	34.5	35.2
y_{41}	5.4	4.0	4.7	4.6	5.0
y_{42}	28.3	29.7	29.0	29.5	28.9
y_{43}	16.2	15.9	16.0	16.2	16.2

y_{44}	6.1	6.9	6.5	7.0	6.5
y_{45}	5.5	4.9	5.2	4.4	5.0
y_{46}	8.5	10.3	9.4	9.9	9.2
y_{47}	19.9	20.1	20.0	20.7	20.3

As explained in Section 3, we used the QR, SM-HD and SM-KNN approaches to estimate weights for the non-probability half-sample. The variables used to fit the model for the QR method, shown in Section 3.1, were selected through *stepwise* regression and the resulting model includes the type of family, social class, economic activity status, number of residents with 10 years old or more within the household, household area, sex and age of the head of the household. The coefficient estimates, their standard errors and corresponding p-values are shown in Table 6. The same variables were used for locating nearest neighbors in the SM-HD and SM-KNN approaches.

Table 6. Coefficient estimates, standard errors (in %) and p-values.

Coefficient		Estimate	Std. Error (%)	p-value
Intercept		-9.39	22.42	<2e-16
Type of family	2	-0.46	9.89	0.000
	3	-0.04	7.74	0.580
	4	0.19	8.60	0.027
Region	Northeast	-0.04	18.25	0.810
	Southeast	-0.47	17.96	0.009
	South	-0.18	19.08	0.339
	Center-West	-0.05	19.05	0.801
Social Class	B	0.32	14.13	0.024
	C	0.40	15.11	0.008
	D/E	0.24	16.02	0.127
Economic activity status	2	0.10	16.25	0.527
	3	0.73	17.71	0.000
	4	0.17	7.88	0.033
	5	-0.31	4.90	0.000
	6	-0.36	5.46	0.000
	7	0.14	4.71	0.003
Sex of head of household	Female	0.08	3.79	0.046
	Did not answer	0.27	29.24	0.353
Number of residents		0.15	1.70	<2e-16
Area		-0.11	11.17	0.339
Age of head of household		0.01	0.13	0.339

In order to obtain estimates for the study target indicators (Section 2.2), we used the original calibrated weights for households from the probability sample, and for those households from the non-probability sample, we used the reciprocals of the estimated PIPs provided by the fitted model. The resulting weights were used to estimate the target indicators, and these estimates were compared with the estimates from the probability sample, the benchmark for the results. We also computed the pooled half-samples combined estimates as described above.

For the SM-KNN, we used the *kNN* function from the *VIM* package in R for numerical variables based on a variation of the Gower Distance (Templ et al. 2020). In our case, the method assigned for a case in the non-probability sample, the median of the weights of the $K = 5$ nearest-neighbor cases in the reference sample. For the SM-HD imputation, we used the *hotdeck* function also available in the *VIM* package.

The variables area, number of residents, type of family, region, social class, economic activity status, level of education of head of household, sex of the respondent and age of the respondent were used for sorting the data before imputation. The stratum variable was used to define groups for the imputation, so that recipients should always be imputed using donors belonging to the same stratum. The target point and standard error indicator estimates obtained by each method are shown in Tables 7 and 8.

Table 7. Point estimates for selected indicators (%)

Method →	**QR**		**HD**		**KNN**	
Indicator ↓	**QR**	**Pooled**	**HD**	**Pooled**	**KNN**	**Pooled**
y_1	42.8	42.7	42.5	42.5	43.1	42.8
y_{21}	77.8	77.2	78.4	77.5	78.5	77.6
y_{22}	40.4	40.9	39.7	40.6	39.1	40.2
y_{23}	1.5	1.5	1.5	1.5	1.3	1.4
y_3	35.8	35.8	35.6	35.7	35.7	35.8
y_{41}	3.9	4.7	4.0	4.7	3.9	4.6
y_{42}	28.7	28.5	30.0	29.1	30.1	29.2
y_{43}	15.9	16.1	15.4	15.8	15.9	16.0
y_{44}	6.9	6.5	6.7	6.4	7.0	6.6
y_{45}	5.0	5.3	5.0	5.2	4.7	5.1
y_{46}	10.9	9.7	10.5	9.5	10.1	9.3
y_{47}	20.6	20.2	19.8	19.8	20.3	20.1

Table 8. Standard errors of estimates for selected indicators (%)

Method →	**Probability**	**QR**		**HD**		**KNN**	
Indicator ↓		**QR**	**Pooled**	**HD**	**Pooled**	**KNN**	**Pooled**
y_1	10.1	5.4	5.4	11.1	5.9	6.7	5.7
y_{21}	9.6	6.6	7.2	10.9	9.0	6.8	7.0
y_{22}	12.7	8.6	10.0	15.3	14.0	9.6	9.5
y_{23}	3.3	1.7	2.3	2.6	2.2	2.8	2.4
y_3	10.1	5.3	5.9	11.5	8.0	6.6	6.1
y_{41}	5.3	4.3	2.8	9.3	4.6	4.2	3.3
y_{42}	13.9	7.8	9.5	13.3	10.7	9.1	8.6
y_{43}	10.1	5.2	5.6	11.1	8.6	7.3	7.8
y_{44}	6.4	5.1	4.6	8.8	5.7	6.1	5.0
y_{45}	6.3	3.6	3.7	7.6	4.9	5.6	7.5
y_{46}	11.8	4.5	7.1	10.2	8.5	4.1	3.4
y_{47}	11.1	8.6	8.4	13.1	8.8	8.5	8.0

To compare the results of the point estimates for the selected indicators we calculated the mean of the square of the differences (MSD) between the probability and non-probability half-sample point estimates for each method. We also calculated the average standard error (ASD) for each method. The results are shown in Table 9.

Table 9. Summary statistics for point and standard error estimates for alternative estimation methods

Method	MSD	ASD
Quasi-randomization	1.00	5.57
Hot Deck imputation	1.31	10.41
K Nearest Neighbor imputation	1.61	6.45

These summaries reveal that the QR performed best for this exercise, providing the closest point estimates and smaller standard errors (on average) when compared to the benchmark estimates coming from the probability half-sample.

5. Conclusion

As already mentioned, one of the challenges of using non-probability samples is the fact that weights cannot be easily computed. The methods considered here seek to assign weights for the non-probability sample using the probability sample as a reference. Comparing with the benchmark, the QR approach yielded the best performance in our application. Another advantage of QR is that weights are not outcome dependent and can be applied to estimate any population parameter in the same way as is done with traditional probability sampling weights. This characteristic is useful for the ICTHS since it produces estimates for over 20 household ICT indicators and numerous domains.

In order to increase the level of precision of the estimates or even to produce them for more disaggregated levels, the NIC.br team is studying if there is a more efficient way to allocate the sample of enumeration areas between quota and strict probability sub-sampling methods. The goal is to increase the overall sample size, while keeping the quality of the indicators produced and the overall cost approximately the same. Currently, the idea is to increase the part of the sample using quota sub-sampling in harder-to-reach EAs or in those EAs with lower response rates, such as those living in apartment buildings or in higher income areas. This should enable increasing the sample size in some regions where the R-indicator (Coelho, Pitta, and Silva 2020; Dos Santos, Pitta, and Silva 2020) is small, such as the states in the North region of Brazil, where data collection is more expensive.

References

Baker, Reg, J. Michael Brick, Nancy A. Bates, Mike Battaglia, Mick P. Couper, Jill A. Dever, Krista J. Gile, and Roger Tourangeau. 2013. "Summary Report of the AAPOR Task Force on Non-Probability Sampling." *Journal of Survey Statistics and Methodology* 1 (2): 90–105. https://doi.org/10.1093/jssam/smt008.

Brazilian Internet Steering Committee. 2012. *Survey on the Use of Information and Communication Technologies in Brazil: ICT Households and Enterprises 2011*.

Buelens, Bart, Joep Burguer, and Jan Van Den Brakel. 2015. *Predictive Inference for Non-Probability Samples: A Simulation Study*.

Coelho, Isabela Bertolini, Marcelo Trindade Pitta, and Pedro Luís do Nascimento Silva. 2020. "Estimating State Level Indicators from ICT Household Surveys in Brazil." *Statistical Journal of the IAOS* Preprint: 1–14. https://doi.org/10.3233/SJI-190511.

Dever, Jill A. 2018. “Combining Probability and Nonprobability Samples to Form Efficient Hybrid Estimates: An Evaluation of the Commom Support Assumption.” In *2018 Federal Committee on Statistical Methodology (FCSM) Research Conference*, 15.

Elliott, Michael R. 2009. “Combining Data from Probability and Non- Probability Samples Using Pseudo-Weights.” *Survey Practice* 2 (6): 1–7. https://doi.org/10.29115/sp-2009-0025.

Elliott, Michael R., and Richard Valliant. 2017. “Inference for Nonprobability Samples.” *Statistical Science* 32 (2): 249–64. https://doi.org/10.1214/16-STS598.

ITU, International Telecommunication Union. 2014. “Manual for Measuring ICT Access and Use by Households and Individuals.”

Kish, Leslie. 1949. “A Procedure for Objective Respondent Selection within the Household.” *Journal of the American Statistical Association* 44 (247): 380–87. https://doi.org/10.1080/01621459.1949.10483314.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd editio. Vol. 151. New York: John Wiley & Sons. https://doi.org/10.2307/2982783.

Rafei, Ali, Carol A.C. Flannagan, and Michael R. Elliott. 2020. “Big Data for Finite Population Inference: Applying Quasi-Random Approaches to Naturalistic Driving Data Using Bayesian Additive Regression Trees.” *Journal of Survey Statistics and Methodology* 8 (1): 148–80. https://doi.org/10.1093/jssam/smz060.

Santos, Mayra Pizzott Rodrigues Dos, Marcelo Trindade Pitta, and Denise Britz do Nascimento Silva. 2020. “Representativity Indicators for the Survey on the Use of Information and Communication Technologies in Brazilian Households.” *Statistical Journal of the IAOS* 36 (2): 509–18. https://doi.org/10.3233/SJI-190509.

Templ, Matthias, Alexander Kowarik, Andreas Alfons, Gregor Cillia, and Bernd Prantner. 2020. “Package VIM.” CRAN R.

Valliant, Richard. 2019. “Comparing Alternatives for Estimation from Nonprobability Samples.” *Journal of Survey Statistics and Methodology*, no. March. https://doi.org/10.1093/jssam/smz003.

Valliant, Richard, and Jill A. Dever. 2011. *Estimating Propensity Adjustments for Volunteer Web Surveys*. *Sociological Methods and Research*. Vol. 40. https://doi.org/10.1177/0049124110392533.

Vasconcellos, Mauricio Teixeira Leite de, Pedro Luis do Nascimento Silva, and Célia Landmann Szwarcwald. 2005. “Sampling Design for the World Health Survey in Brazil.” *Cadernos de Saúde Pública* 21 (S): S89–99.