Statistical Journal of the IAOS 36 (2020) 509–518
DOI 10.3233/SJI-190509
IOS Press

Representativity indicators for the survey on the use of information and communication technologies in Brazilian households

Mayra Pizzott Rodrigues dos Santos[a,*], Marcelo Trindade Pitta[a] and Denise Britz do Nascimento Silva[b]
[a]*Brazilian Network Information Center, São Paulo, Brazil*
[b]*National School of Statistical Sciences, Rio de Janeiro, Brazil*

Abstract. Most surveys are affected by nonresponse and, in the case of household surveys, this happens when individuals or households do not provide the requested information. This phenomenon needs to be considered when assessing the survey quality. R-indicators are a valuable tool to evaluate the impact of nonresponse on survey results. In this paper, the R-indicator is used to assess and analyze the representativeness of the sample for the ICT Households Survey, conducted by NIC.br. The ICT Households survey sample was originally designed to provide results for Brazil's regions, and it is not possible to ensure the quality of the sample for estimating statistics for smaller areas such as federative units (states). R-indicator methodology is usually employed as a measure of discrepancy between the selected/planned and realised samples. Here it is used to determine whether, in the planning stage, the sample would produce estimates with good precision for unplanned domains, such as federative units. The results of the estimated indicators revealed that the respondent and planned samples of ICT Households survey can be considered representative for Brazil and major region levels. At the federative unit level, however, there is evidence of a gap between the planned and respondent samples and the target population.

Keywords: Nonresponse, R-indicator, sample representativeness, ICT households survey

1. Introduction

The demand for statistics to plan and monitor public policies continues to grow and is challenging for the quality of the indicators. The present article introduces the R-indicator calculation for statistics of the Survey on the Use of Information and Communication Technologies in Brazilian Households – ICT Households, conducted annually by the Regional Center for Studies on the Development of the Information Society (Cetic.br), a department of the Brazilian Network Information Center (NIC.br), in order to assess the quality of estimates for territorial divisions considered in the survey design and other smaller unplanned domains not included within the planning scope.

Data produced by Cetic.br from surveys about the Internet, and information and communication technologies, for a variety of target populations (households, individuals, healthcare facilities, education, etc.) has been used since 2005 by government, the private sector and civil society to plan and monitor policies and programs aimed at improving the quality of the Internet and information and communication technologies in the country.[1] R-indicators, where R refers to "representativity", were developed as quality measurements for sample surveys that can provide information on the possible risk of producing biased estimates. Bethlehem et al. [1] present various meanings for representative sample and, when proposing an indicator

*Corresponding author: Mayra Pizzott Rodrigues dos Santos, Cetic.br, Avenida das Nações Unidas, 11541, 7th floor, 04578-000, São Paulo, Brazil. Tel.: +55 21 96623 1089; E-mail: mayrapizzottrs@gmail.com.

[1] More details can be found at: www.cetic.br.

for representativity, defines the concept as "the absence of selective forces".[2] These indicators measure how the composition of respondent samples differ from that of planned samples. They can be employed in various ways, such as the analysis of survey data after field work is completed. Another use is for monitoring during data collection, for example, in the field work, when collection efforts can be aimed at obtaining respondent samples whose composition is not so different from planned samples [1].

In sample surveys, the nonresponse phenomenon has two main consequences: a reduction in sample size; and changes in the composition of the originally planned and selected samples (thereby preventing effective implementation of sample designs). Whereas the sample size is directly related to the precision of estimates, modifications in sample composition can lead to bias. In this case, the probability of observing data from selected sample units is affected, and the adjustment for reducing nonresponse biases depends on the availability of auxiliary information about the missed units. If that information is not available, there is little that can be done [2].

According to Schouten et al. [3,4], R-indicators can be used to compare responses among different surveys that share the same target population; to compare responses in waves of repeated or longitudinal surveys; and to monitor responses to surveys during data collection and adapt data collection procedures based on historical data, registration data available and paradata.[3]

The present study proposes the use of R-indicators to assess the adequacy of a planned sample to produce estimates for areas or domains that were not considered in the ICT Households sample design.

This article aims at assessing the ICT Households sample as a source of information for obtaining accurate estimates for each Brazilian federative unit (state). The sample for this survey, whose design is stratified and clustered in various stages, was originally planned to provide estimates with controlled precision for Brazil's five major regions. Federative units were used in its sample plan as selection strata; however, the sample was allocated in each of these strata to minimize costs, not to control the precision of estimates for each federative unit. This imposed a reduction of the planned sample size in some federative units (mostly in the North and Northeast regions), and good precision is only guaranteed for regional and national estimates.

Since there is plan to produce estimates for each federative unit based on the respondent sample of the ICT Households , and the planned sample was designed to provide estimates for regions, the composition of the planned and respondent samples were compared with the target population from which the sample was taken, i.e., the population investigated in the 2010 Brazilian Census. This set of comparisons should provide evidence to detect if lack of representativeness, in relation to the production of federative unit estimates, was already present in the planned sample, even before the data collection, or if it was the result of differential nonresponse among the federative units.

In order to compare the planned and realised samples with the target population (data from 2010 Brazilian Census), the R-indicator was calculated at the level of the census enumeration areas (census tracts) for the pairs: population vs. planned sample; population vs. respondent sample; and planned sample vs. respondent sample.

Therefore, the indicator was obtained using information about the population (based on data from the 2010 Census) and about the planned and respondent samples for the ICT Households surveys in 2015 and 2016. In this case, the intent was to assess how representative the respondent and planned samples were of the population from which the samples were selected, to inform the decision about producing estimates at federative unit level.

The results of this study will be taken as input for a small area estimation feasibility study[4] in the ICT Households survey scope that, as already mentioned, only publishes results at regional and national levels.

2. Methodology

This section contains a summary of the data source, the R-indicator methodology, and the main procedures carried out for obtaining the results presented in Section 3.

[2] If there are no selective forces, every element in the population has the same probability of responding when selected in the sample (Bethlehem et al. [1, p. 2]). Therefore, when the probability of responding to a survey is $\rho_i = P(r_i = 1|s_i = 1)$, where r_i and s_i assume the value of 1 if element i was selected in the sample s and responded, and 0 otherwise. Then, there are no selective forces if ρ_i is a constant; i.e., if $\rho_i \equiv \rho$.

[3] Information regarding the survey's data collection process.

[4] An area is considered small if the size of the sample in the area is not large enough to produce estimates with the desired precision.

2.1. Data source

This study uses data from ICT Households 2015 and 2016 surveys, conducted under the auspices of CGI.br by NIC.br and Cetic.br. The primary objective of the survey is to measure ICT ownership and use among people aged 10 years or older in Brazil [5].

The sampling frame is composed of census enumeration areas defined for the 2010 Brazilian Census and the sample is selected according to a stratified multiple stage cluster design. The sample design was specified to produce estimates with controlled precision by geographical regions (North, Northeast, Southeast, South and Middle West) and the enumeration areas are stratified by federative units (26 states plus Brazil's capital city). For 9 states (Pará, Ceará, Pernambuco, Bahia, Minas Gerais, Rio de Janeiro, São Paulo, Paraná and Rio Grande do Sul), enumeration areas are also stratified according to administrative areas (metropolitan and non-metropolitan areas) totalizing 36 strata.

Strata are composed of municipalities. The largest municipalities[5] and state capitals are included in the sample with certainty. In this specific case, the enumeration areas are the primary sampling units (PSU) and are selected with probability proportional to the number of households. This is followed by other two sampling stages for selecting households and residents. All other municipalities are taken as primary sampling units (selected with probability proportional to number of residents aged 10 years or more) in which enumeration areas are selected as secondary sampling units (SSU), then households and residents are also selected to the sample. Therefore, the number of stages varies according to cities (municipalities) since some of them are always included in the sample whereas others are randomly selected.[6] Table 1 presents information about selection stages and corresponding sampling units for ICT Households.

2.2. R-indicators

R-indicators can be used to assess the degree to which respondent samples deviate from planned samples. If the response probabilities of all the sample units in the respondent sample are equal, given a set of auxiliary variables, one may assume that there would be no systematic differences between the composition of the respondent and planned samples. However, if the response probabilities vary, it is important to establish sampling units' profiles to inform data collection (enumeration intelligence) and to evaluate the extent to which the composition of the respondent sample is affected. This information can be obtained by defining a function of distance that measures how much the response probabilities of sampling units differ from the mean response probability [1]. These indicators were proposed by [2] and are the subject of a set of reports produced in the scope of the project "Representative Indicators for Survey Quality (RISQ)",[7] developed by European institutions.

Suppose that a probability sample s of size n is selected without replacement of a finite population. The sample can be represented by a vector of indicators $s = (s_1, s_2, \ldots, s_i, \ldots s_N)$, where s_i assumes a value of 1 when element i is selected in the sample and, 0 if not [6]. Survey responses can be represented by the vector of indicators $r = (r_1, r_2, \ldots, r_i, \ldots r_N)$, where r_i assumes a value of 1 if element i was selected in the sample and responded, and 0 otherwise. Each element i in the population has an unknown probability ρ_i of responding when selected for the sample, i.e., $\rho_i = P(r_i = 1|s_i = 1)$.

Since ρ_i is unknown, its value must be estimated. This is done using a vector of auxiliary variables that is available for all the elements in the sample. Various techniques for estimating response propensity can be used [1], such as logistic or probit models [7], or CHAID classification trees [8].

Therefore, let $\widehat{\rho}_i$ be an estimator of ρ_i based on a set of auxiliary information, and let $\widehat{\overline{\rho}}$ be the weighted average of the estimated response probabilities, such that:

$$\widehat{\overline{\rho}} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\rho}_i \frac{s_i}{\pi_i}, \tag{1}$$

where π_i is the first-order inclusion probability of sample unit i.

Response propensity ρ_i is defined as the conditional probability of r_i, given the values of a vector X with m auxiliary variables:

$$\begin{aligned} \rho_i &= E(r_i = 1|\boldsymbol{X} = \boldsymbol{x}_i, s_i = 1) \\ &= P(r_i = 1|\boldsymbol{X} = \boldsymbol{x}_i, s_i = 1) \end{aligned}$$

According to [9], we model the response propensity using a logistic regression.

[5] In Brazil each city is a municipality.

[6] More information can be found at: https://www.cetic.br/media/docs/publicacoes/2/TIC_DOM_2016_LivroEletronico.pdf.

[7] https://www.cmist.manchester.ac.uk/research/projects/representative-indicators-for-survey-quality/.

Table 1
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 the number of households
Secondary sampling units (SSU)	Enumeration area	Probability proportional to the number of households	Households	Simple random sample of 15 households in each enumeration area
Tertiary sampling units (TSU)	Households	Simple random sample of 15 households in each enumeration area	Individuals	Simple random sample of one resident 10 years old or more
Final (or fourth) sampling units	Individuals	Simple random sample of one resident 10 years old or more	–	–

Schouten and Cobben [2] defined three R-indicators: one based on the standard deviation of the response probabilities (R_1); one taking into account the variance of the response probabilities (R_2); and another related to the proportional reduction of error (R_3).

Since, in the literature, its favoured version to evaluate the representativeness of a sample is the one obtained from the standard deviation of response probabilities (R_1), the estimator $\hat{R}_1(\rho)$ is employed throughout this paper. Estimates of the standard error (σ_R) of $\hat{R}_1(\rho)$ can be obtained as defined in [10,11]. R_1 has values in the interval [0, 1], where 1 indicates strong representativeness and lower values correspond to a less representative response.

$$\begin{aligned}\hat{R}_1(\rho) &= 1 - 2\hat{S}(\widehat{\rho}) \\ &= 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}\left(\widehat{\rho}_i - \widehat{\overline{\rho}}\right)^2}\end{aligned} \tag{2}$$

2.3. R-indicators in the case of ICT Households

ICT Households is a survey whose sample is designed to provide estimates with controlled precision for Brazil's five major regions. As the aim of the study is to assess whether ICT Households planned sample, and corresponding respondent sample, are representative at federative unit (state) level, this section presents the R-indicator methodology applied to the ICT Households survey framework.

In order to compare the planned and realised samples with the target population (data from 2010 Brazilian Census), the R-indicator was calculated at the level of the census enumeration areas for the pairs: population vs. planned sample; population vs. respondent sample; and planned vs. respondent sample. The results are presented for these comparisons and the estimated R-indicators are produced at national, regional and federative unit level.

Table 2
Categories of the auxiliary variables

Average number of residents in PPH	Average monthly income
1 to 3 residents	BRL 0.00 to BRL 650.00
4 or 5 residents	BRL 650.01 to BRL 950.00
6 or 7 residents	BRL 950.01 to BRL 1,500.00
More than 7 residents	More than BRL 1,500.00

It is important to highlight that it is not possible to link household or person level records from 2010 Census and ICT Households. The survey sampling frame is composed of census enumeration areas. Therefore, the work was carried out to evaluate the sample of enumeration areas, and the variable of interest (C_i) for the response propensity logistic model is defined as a response indicator such that:

$$C_i = \begin{cases} 1, \text{ if there is a response for the census enumeration area } i \\ 0, \text{ otherwise} \end{cases} \tag{3}$$

In the case of planned sample vs. respondent sample comparison, the response variable assumes the value of 1 if at least one household of the selected census enumeration area responded (so there is response for the census enumeration area), and 0 if none of the households in a enumeration area selected for the planned sample provided a response. For the population vs. respondent sample comparison, the response variable assumes the value of 1 if there is survey response for at least one household in a given census enumeration area and 0 otherwise.

Note that, in the case of population vs. planned sample comparison, the response variable assumes the value of 1 if the census enumeration area is present in the planned sample, and 0 if it is only found in the

Table 3
Allocation of the ICT Households survey sample

	ICT stratum	Sample		
		Census enumeration areas	Municipalities (cities)	Planned interviews
North	Rondônia	18	4	270
	Roraima	15	4	225
	Acre	15	4	225
	Amapá	15	6	225
	Tocantins	15	4	225
	Amazonas	38	8	570
	Pará – Belém MR	27	4	405
	Pará – Countryside	57	9	855
Northeast	Maranhão	71	12	1,065
	Piauí	36	7	540
	Ceará – Fortaleza MR	42	6	630
	Ceará – Countryside	55	8	825
	Pernambuco – Recife MR	41	6	615
	Pernambuco – Countryside	57	10	855
	Rio Grande do Norte	39	7	585
	Paraíba	45	11	675
	Alagoas	35	7	525
	Sergipe	28	6	420
	Bahia – Salvador MR	44	6	660
	Bahia – Countryside	122	19	1,830
Southeast	Minas Gerais – BH MR	63	8	945
	Minas Gerais – Countryside	146	27	2,190
	Espírito Santo	47	8	705
	Rio de Janeiro – RJ MR	136	13	2,040
	Rio de Janeiro – Countryside	53	7	795
	São Paulo – São Paulo MR	206	18	3,090
	São Paulo – Countryside	226	42	3,390
South	Paraná - Curitiba MR	42	6	630
	Paraná – Countryside	88	15	1,320
	Santa Catarina	82	13	1,230
	Rio Grande do Sul – Porto Alegre MR	50	7	750
	Rio Grande do Sul – Countryside	84	14	1,260
Center-west	Mato Grosso do Sul	32	5	480
	Mato Grosso	41	7	615
	Goiás	70	11	1,050
	Federal District	33	1	495

Source: CGI.br/NIC.br, Regional Center for Studies on the Development of the Information Society (Cetic.br), Survey on the Use of Information and Communication Technologies in Brazilian Households – ICT Households 2016.

census enumeration area frame (and has not been selected for the ICT Households sample). For evaluating the representativeness of the planned sample versus the survey frame (related to the target population), the weights assigned to enumeration areas not in the planned sample were set equal to 1 in order to implement the R-indicator methodology. The choice of auxiliary variables was made from those available in the census enumeration area frame. The variables should be the same for all models, despite of geographic areas or the year in the scope of analysis, assuring comparability of R-Indicators estimates across different years and domains. The final model (the propensity model[8]) is composed of two auxiliary variables related to the census enumeration area characteristics: the average number of residents in permanent private households (PPH[9]) in the enumeration area and the average monthly income of residents aged 10 years or older (with and without income) in the enumeration area. The quantitative variables were categorised as in Table 2. An Table 9 of model results (and estimated coefficients) are available in the appendix.

The response propensity model is defined as:

[8]Models with other auxiliary variables were tested but there was no evidence of statistical significance.

[9]PPH corresponds to permanent private households, i.e., households intended to serve exclusively as residences and that on 2010 Census reference date served as residences for one or more people [12].

Table 4
R-indicator estimates and corresponding $CI_{95\%}$, by survey year and type of comparison, Brazil, 2015–2016

Survey year	Comparison	R-indicator
2015	Population vs. planned sample	0.918 (0.880–0.955)
	Population vs. respondent sample	0.910 (0.872–0.947)
	Planned sample vs. respondent sample	0.909 (0.900–0.918)
2016	Population vs. planned sample	0.961 (0.912–1.009)
	Population vs. respondent sample	0.958 (0.914–1.003)
	Planned sample vs. respondent sample	0.913 (0.905–0.922)

CGI.br/NIC.br, Regional Center for Studies on the Development of the Information Society (Cetic.br), Survey on the Use of Information and Communication Technologies in Brazilian Households – Prepared by the authors.

Table 5
R-indicator estimates and corresponding $CI_{95\%}$, by type of comparison and region, Brazil, 2015

Region	Survey year: 2015 Comparison		
	Population vs. planned sample	Population vs. respondent sample	Planned sample vs. respondent sample
North	0.839 (0.694–0.983)	0.832 (0.693–0.972)	0.935 (0.915–0.954)
Northeast	0.895 (0.868–0.923)	0.893 (0.858–0.928)	0.933 (0.911–0.955)
Southeast	0.919 (0.865–0.972)	0.902 (0.852–0.953)	0.884 (0.869–0.898)
South	0.904 (0.824–0.984)	0.902 (0.821–0.983)	0.964 (0.956–0.972)
Center-West	0.863 (0.775–0.952)	0.861 (0.770–0.953)	0.947 (0.935–0.959)

CGI.br/NIC.br, Regional Center for Studies on the Development of the Information Society (Cetic.br), Survey on the Use of Information and Communication Technologies in Brazilian Households – Prepared by the authors.

Table 6
R-indicator estimates and corresponding $CI_{95\%}$, by type of comparison and region, Brazil, 2016

Region	Survey year: 2016 Comparison		
	Population vs. planned sample	Population vs. respondent sample	Planned sample vs. respondent sample
North	0.918 (0.762–1.073)	0.919 (0.765–1.073)	0.986 (0.972–1.000)
Northeast	0.911 (0.877–0.945)	0.908 (0.871–0.946)	0.919 (0.893–0.945)
Southeast	0.938 (0.887–0.988)	0.931 (0.881 - 0.981)	0.906 (0.895–0.916)
South	0.924 (0.824–1.024)	0.925 (0.824–1.026)	0.939 (0.919–0.959)
Center-West	0.901 (0.726–1.076)	0.847 (0.680–1.014)	0.755 (0.705–0.805)

Source: CGI.br/NIC.br, Regional Center for Studies on the Development of the Information Society (Cetic.br), Survey on the Use of Information and Communication Technologies in Brazilian Households – Prepared by the authors.

$$\begin{aligned} logit\left(\rho_{jki}\right) &= log\left(\frac{\rho_{jki}}{1-\rho_{jki}}\right) \\ &= \mu + \tau_j + \beta_k \begin{cases} j = 1, \ldots, 4. \\ k = 1, \ldots, 4. \\ i = 1, \ldots, n. \end{cases} \end{aligned} \tag{4}$$

where

μ represents the probability of success in the reference categories,

τ_j represents the effect of level j of the average number of PPH residents,

β_k represents the effect of level k of the average monthly income,

ρ_{jki} is the response probability of census enumeration area i at levels j and k.

Therefore, the response probability (ρ_{jki}) for census enumeration area i at levels j and k, is given by:

$$\rho_{jki} = \frac{exp\left(\mu + \tau_j + \beta_k\right)}{1 + exp\left(\mu + \tau_j + \beta_k\right)} \tag{5}$$

Since the R-indicator is being calculated for specific comparisons, the meaning of ρ_{jki} varies according to the corresponding comparison. In the case of the population vs. respondent sample comparison, ρ_{jki} is the response probability in census enumeration area i at levels j and k. When the R-indicator is used for comparing the population and planned sample, ρ_{jki} is the probability of census enumeration area i at levels j and k be selected for the planned sample.

Santos [13] incorporated the complex survey design in the estimation of propensity models and the same estimation approach is employed here. Calculations were carried out using the R survey package and *svyglm* function. To take into account the survey design and sample weights for the population vs. planned sample

Table 7
R-indicator estimates and corresponding $CI_{95\%}$, by type of comparison and federative unit, Brazil, 2015

Region	Federative Unit	Survey Year – 2015 Comparison		
		Population vs. planned sample	Population vs. respondent sample	Planned sample vs. respondent sample
North	Rondônia	0.745 (0.436–1.054)	0.745 (0.436–1.054)	1.000 (1.000–1.000)
	Acre	0.668 (0.517–0.819)	0.668 (0.517–0.819)	1.000 (1.000–1.000)
	Amazonas	0.758 (0.583–0.932)	0.742 (0.596–0.888)	0.662 (0.420–0.904)
	Roraima	0.549 (0.379–0.720)	0.549 (0.379–0.720)	–
	Pará	0.823 (0.616–1.031)	0.823 (0.616–1.031)	1.000 (1.000–1.000)
	Amapá	0.534 (0.405–0.662)	0.534 (0.405–0.662)	1.000 (1.000–1.000)
	Tocantins	0.731 (0.566–0.895)	0.731 (0.566–0.895)	1.000 (1.000–1.000)
Northeast	Maranhão	0.779 (0.668–0.890)	0.777 (0.660–0.894)	0.816 (0.723–0.909)
	Piauí	0.940 (0.743–1.138)	0.940 (0.743–1.138)	1.000 (1.000–1.000)
	Ceará	0.874 (0.787–0.962)	0.793 (0.676–0.909)	0.775 (0.622–0.928)
	Rio Grande do Norte	0.843 (0.656–1.030)	0.843 (0.656–1.030)	1.000 (1.000–1.000)
	Paraíba	0.799 (0.651–0.946)	0.799 (0.651–0.946)	1.000 (1.000–1.000)
	Pernambuco	0.846 (0.728–0.964)	0.852 (0.726–0.978)	0.976 (0.960–0.993)
	Alagoas	0.818 (0.721–0.915)	0.818 (0.721–0.915)	1.000 (1.000–1.000)
	Sergipe	0.880 (0.715–1.044)	0.880 (0.715–1.044)	1.000 (1.000–1.000)
	Bahia	0.922 (0.813–1.031)	0.922 (0.813–1.031)	1.000 (1.000–1.000)
Southeast	Minas Gerais	0.908 (0.800–1.017)	0.912 (0.800–1.024)	0.984 (0.971–0.997)
	Espírito Santo	0.599 (0.428–0.770)	0.599 (0.428–0.770)	1.000 (1.000–1.000)
	Rio de Janeiro	0.930 (0.810–1.051)	0.903 (0.777–1.029)	0.796 (0.749–0.843)
	São Paulo	0.873 (0.815–0.931)	0.801 (0.741–0.860)	0.754 (0.716–0.791)
South	Paraná	0.895 (0.764–1.027)	0.891 (0.761–1.020)	0.892 (0.862–0.922)
	Santa Catarina	0.821 (0.685–0.956)	0.816 (0.681–0.951)	0.956 (0.935–0.976)
	Rio Grande do Sul	0.895 (0.692–1.097)	0.895 (0.692–1.097)	1.000 (1.000–1.000)
Center-West	Mato Grosso do Sul	0.718 (0.588–0.848)	0.706 (0.569–0.842)	0.757 (0.626–0.888)
	Mato Grosso	0.758 (0.517–0.999)	0.758 (0.517–0.999)	1.000 (1.000–1.000)
	Goiás	0.801 (0.651–0.950)	0.804 (0.647–0.961)	0.849 (0.795–0.903)
	Distrito Federal	0.869 (0.614–1.123)	0.869 (0.614–1.123)	1.000 (1.000–1.000)

Source: CGI.br/NIC.br, Regional Center for Studies on the Development of the Information Society (Cetic.br), Survey on the Use of Information and Communication Technologies in Brazilian Households – Prepared by the authors.

and population vs. respondent sample comparisons, a weight equal 1 was assigned for those units (census enumeration areas) not selected for the sample and the corresponding sample weights for the others.

Confidence intervals for R-indicators were estimated considering the survey design and using R code available on RISQ website[10] that implements methodology developed by Shlomo et al. [11].[11]

3. Results

ICT Households is a national survey with a multistage stratified sample. The planned sample size is 33,210 households in 2,214 census enumeration areas and 350 cities. The geographic distribution of the sample in the country is subjected to cost restrictions and population representativeness. Table 3 presents the sample allocation of census tracts by geographic strata.

According to the sample plan, 15 households are selected per census enumeration area. It can be seen in Table 3 that the North and Center-West regions have less cities in the sample than other regions. In the North, this smaller sample is spread among more states than in the Center-West. In the planned sample, the average number of census enumeration areas by federative unit is 125 in the Southeast region whereas only 25 in the North. This uneven distribution is due to the pattern of population spread in the regions and the high cost of conducting face-to-face interviews in more remote areas of the country.

The R-indicators were calculated to compare three databases (population – sample frame, planned sample and respondent sample) for two survey occasions (2015 and 2016) and for three geographical levels (Brazil, regions and federative units). Table 4 presents the R-indicators and confidence intervals ($CI_{95\%}$) cal-

[10] https://www.cmi.manchester.ac.uk/research/projects/representative-indicators-for-survey-quality/tools/.

[11] The paper is available at http://hummedia.manchester.ac.uk/institutes/cmist/risq/shlomo-skinner-schouten-2011.pdf. Section 6 presents the development of confidence interval for R-indicators.

Table 8
R-indicator estimates and corresponding $CI_{95\%}$, by type of comparison and federative unit, Brazil, 2016

Region	Federative Unit	Survey Year – 2016 Comparison		
		Population vs. planned sample	Population vs. respondent sample	Planned sample vs. respondent sample
North	Rondônia	0.716 (0.432–0.999)	0.716 (0.432–0.999)	1.000 (1.000–1.000)
	Acre	0.705 (0.443–0.968)	0.705 (0.443–0.968)	1.000 (1.000–1.000)
	Amazonas	0.759 (0.561–0.956)	0.759 (0.561–0.956)	1.000 (1.000–1.000)
	Roraima	0.597 (0.427–0.767)	0.597 (0.427–0.767)	1.000 (1.000–1.000)
	Pará	0.825 (0.693–0.957)	0.831 (0.708–0.954)	0.977 (0.968–0.986)
	Amapá	0.535 (0.321–0.748)	0.535 (0.321–0.748)	1.000 (1.000–1.000)
	Tocantins	0.657 (0.432–0.883)	0.657 (0.432–0.883)	0.000 (0.000–0.000)
Northeast	Maranhão	0.807 (0.691–0.924)	0.799 (0.682–0.917)	0.704 (0.420–0.987)
	Piauí	0.728 (0.562–0.893)	0.730 (0.558–0.903)	–
	Ceará	0.917 (0.812–1.021)	0.839 (0.757–0.921)	0.768 (0.635–0.902)
	Rio Grande do Norte	0.878 (0.696–1.060)	0.878 (0.696–1.060)	1.000 (1.000–1.000)
	Paraíba	0.831 (0.684–0.978)	0.831 (0.684–0.978)	1.000 (1.000–1.000)
	Pernambuco	0.838 (0.752–0.924)	0.838 (0.752–0.924)	1.000 (1.000–1.000)
	Alagoas	0.726 (0.551–0.902)	0.726 (0.551–0.902)	1.000 (1.000–1.000)
	Sergipe	0.844 (0.675–1.012)	0.844 (0.675–1.012)	1.000 (1.000–1.000)
	Bahia	0.897 (0.792–1.002)	0.897 (0.792–1.002)	1.000 (1.000–1.000)
Southeast	Minas Gerais	0.911 (0.801–1.022)	0.911 (0.801–1.022)	1.000 (1.000–1.000)
	Espírito Santo	0.777 (0.430–1.123)	0.750 (0.403–1.096)	0.721 (0.551–0.891)
	Rio de Janeiro	0.938 (0.811–1.065)	0.947 (0.812–1.083)	0.958 (0.949–0.966)
	São Paulo	0.899 (0.827–0.971)	0.901 (0.834–0.968)	0.878 (0.862–0.893)
South	Paraná	0.846 (0.726–0.965)	0.843 (0.722–0.964)	0.950 (0.930–0.970)
	Santa Catarina	0.914 (0.714–1.115)	0.943 (0.690–1.196)	0.831 (0.749–0.913)
	Rio Grande do Sul	0.816 (0.696–0.936)	0.825 (0.699–0.952)	0.967 (0.961–0.973)
Center-West	Mato Grosso do Sul	0.703 (0.592–0.814)	0.632 (0.480–0.784)	0.527 (0.293–0.760)
	Mato Grosso	0.728 (0.529–0.927)	0.728 (0.529–0.927)	1.000 (1.000–1.000)
	Goiás	0.837 (0.567–1.108)	0.812 (0.558–1.065)	0.915 (0.893–0.937)
	Distrito Federal	0.861 (0.609–1.112)	0.867 (0.620–1.115)	0.806 (0.763–0.850)

Source: CGI.br/NIC.br, Regional Center for Studies on the Development of the Information Society (Cetic.br), Survey on the Use of Information and Communication Technologies in Brazilian Households – Prepared by the authors.

culated at the national level for each year of the ICT Households survey.

The results show that R-indicators for the population vs. planned sample comparison are higher than those for the population vs. respondent sample. The same occurs in general for the case of planned sample vs. respondent sample, but differences among the indicators when taking into account the corresponding confidence intervals are not statistically significant. The estimated R-indicators are higher than 0.90, therefore the three comparisons indicate that samples can be considered as representative at national level.

Tables 5 and 6 present estimated R-indicators to evaluate sample representativity for Brazilian regions. It can be noted that all the regions have R-indicators higher than 0.75 for 2015 and 2016 surveys, indicating that compositions of the planned and respondent samples did not differ much from the population and the composition of the respondent sample did not differ much from the planned sample. The specific R-indicator estimates for the planned sample vs. respondent sample comparison are higher than those obtained for the population vs. planned sample and population vs. respondent sample, except for the Southeast region in 2015 and Center-West region in 2016, but the differences are not significant.

The values of estimated R-indicators for the 27 Brazilian federative units are displayed by regions (Tables 7 and 8). The R-indicators obtained from the comparison planned sample vs. respondent sample are generally higher than 0.8, pointing for a successful data collection process and correct implementation of the planned sample design. In the case of comparisons population vs planned sample and population vs respondent sample, for both years, R-indicators obtained for various states in the North, Center-West and Northeast regions also show some states in intermediate representativity level (from 0.60 to 0.85). The states of Roraima (the northernmost and least populated state of Brazil, located in the Amazon region) and Amapá (located in the North region and bordered by French Guiana) have the lower representativeness in relation to the other Brazilian federative units. The federative units with R-indicators greater than or equal to 0.90 for

Table 9
Estimated coefficients and corresponding statistics of the propensity model for planned versus respondent sample comparison, Brazil, 2015

		Estimate	Std. Error	t value	Pr ($>\|t\|$)	
Brasil	Intercept	3.722	0.645	5.772	0.000	***
	BRL 650.01 to BRL 950.00	1.030	0.619	1.665	0.096	.
	BRL 950.01 to BRL 1,500.00	−0.145	0.652	−0.222	0.824	
	More than BRL 1,500.00	−2.290	0.611	−3.747	0.000	***
	4 or 5 residents	0.298	0.596	0.500	0.617	
	6 or 7 residents	11.956	0.832	14.377	< 2e-16	***
	More than 7 residents	11.518	1.190	9.679	< 2e-16	***
North	(Intercept)	41.338	1.111	37.218	< 2e-16	***
	BRL 650.01 to BRL 950.00	−0.112	0.255	−0.437	0.663	
	BRL 950.01 to BRL 1,500.00	−20.800	1.082	−19.231	< 2e-16	***
	More than BRL 1.500.00	−20.859	1.079	−19.327	< 2e-16	***
	4 or 5 residents	−18.701	1.109	−16.857	< 2e-16	***
	6 or 7 residents	−18.476	1.268	−14.576	< 2e-16	***
	More than 7 residents	−19.060	1.496	−12.743	< 2e-16	***
Northeast	(Intercept)	21.760	1.013	21.484	< 2e-16	***
	BRL 650.01 to BRL 950.00	0.248	0.501	0.495	0.621	
	BRL 950.01 to BRL 1,500.00	0.960	0.920	1.043	0.298	
	More than BRL 1,500.00	−20.622	0.828	−24.921	< 2e-16	***
	4 or 5 residents	2.142	1.295	1.654	0.099	
Southeast	(Intercept)	2.937	0.706	4.160	0.000	***
	BRL 650.01 to BRL 950.00	1.361	0.620	2.195	0.029	*
	BRL 950.01 to BRL 1,500.00	0.656	0.846	0.776	0.438	
	More than BRL 1,500.00	−1.776	0.674	−2.634	0.009	**
	4 or 5 residents	0.128	0.685	0.187	0.851	
South	(Intercept)	22.516	0.675	33.356	< 2e-16	***
	BRL 650.01 to BRL 950.00	−0.023	0.218	−0.107	0.915	
	BRL 950.01 to BRL 1,500.00	−18.192	1.295	−14.053	< 2e-16	***
	More than BRL 1,500.00	−19.715	0.596	−33.092	< 2e-16	***
	4 or 5 residents	0.279	1.123	0.248	0.804	
Center-West	(Intercept)	21.782	0.782	27.840	< 2e-16	***
	BRL 650.01 to BRL 950.00	0.258	0.461	0.560	0.577	
	BRL 950.01 to BRL 1,500.00	−19.246	1.181	−16.298	< 2e-16	***
	More than BRL 1,500.00	−19.319	0.668	−28.921	< 2e-16	***
	4 or 5 residents	0.944	0.872	1.082	0.282	
	6 or 7 residents	0.907	1.250	0.726	0.470	

Source: CGI.br/NIC.br, Regional Center for Studies on the Development of the Information Society (Cetic.br), Survey on the Use of Information and Communication Technologies in Brazilian Households – Prepared by the authors.

the two years and in the two comparisons are Minas Gerais and Rio de Janeiro (both located in the Southeast and with greater sample sizes).

4. Final remarks

ICT Households is an annual survey, created in 2005, that plays an important role in the Brazilian public statistics system. It is the primary source of public data on information and communication technologies in Brazilian households, apart being national in scope. The definition and implementation of quality measurements, such as the R-indicator, make it possible to enhance its survey process and the information value.

The present study sought to assess, based on R-indicators, the planned and respondent samples of the ICT Households survey for the years 2015 and 2016. The results provide evidence that the respondent and planned samples of the ICT Households survey can be considered representative at the National and regional levels, due to R-indicator values greater than 0.80, as expected according to the survey sample design. When assessing the feasibility of producing federative unit estimates based on the planned and respondent ICT Households sample, it was found a lack of representativeness for some states.

The outcomes of the present study (R-indicator estimates) can be considered to formulate small area estimation methods for the production of survey estimates with controlled precision for all federative units. R-indicator values can be used to determine weights in composite estimators which, in turn, are obtained

from a linear combination of direct estimates and those based on synthetic estimators [14]. The lower the representativeness of the sample by federative unit (smaller values for the R-indicator), the lower can be the weight of the direct estimator of the federative unit in the linear combination. Consequently, greater weight should be associated to a synthetic estimator, that is usually obtained for more aggregated geographical levels for which the sample was actually designed to produce estimates with good precision (regional level, for example). Therefore, the evidence acquired in the present study can constitute the basis for the development of estimation methods for broadening the use and relevance of ICT Households survey data.

References

[1] Bethlehem J, Cobben F and Schouten B. Indicators for the representativeness of survey response. Statistics Canada Symposium, Gatineau, Canada, 2008.

[2] Schouten B and Cobben F. R-indexes for the comparison of different fieldwork strategies and data collection modes. Discussion Paper 07002, Statistics Netherlands, Voorburg, The Netherlands, 2007.

[3] Schouten B, Shlomo N and Skinner C. Indicators for representative response. Paper presented at Q2010, Helinski, Finland, 2010.

[4] Schouten B, Bethlehem JG, Beullens K, Kleven O, Loosveldt G, Luiten A, Rutar K, Shlomo N, Skinner C. Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. International Statistical Review 2012; 80: 382-399.

[5] Brazilian Network Information Center – NIC.br. Survey on the use of information and communication technologies in Brazilian households – ICT households 2016. Brazilian Internet Steering Committee (CGI.br); 2017 [updated 2017 Nov 23; cited 2019 Feb 27]. Available from https://www.cetic.br/publicacao/pesquisa-sobre-o-uso-das-tecnologias-de-informacao-e-comunicacao-nos-domicilios-brasileiros-tic-domicilios-2016.

[6] Cobben F and Schouten B. An empirical validation of R-indicators. Discussion Paper 08006, Statistics Netherlands, Voorburg, The Netherlands; 2008.

[7] Agresti A. Categorical data analysis (2nd ed.). New Jersey: John Wiley & Sons, 2002.

[8] Kass GV. An exploratory technique for investigating large quantities of categorical data. Applied Statistics. 1980; 29(2)): 119-127.

[9] Skinner C, Shlomo N, Schouten B, Zhang L and Bethlehem J. Measuring survey quality through representativeness indicators using sample and population-based information. NTTS Conference, Brussels, Belgium, 2009.

[10] de Heij V, Schouten B and Shlomo N. RISQ Manual 2.1. Tools in SAS and R for the computation of R-indicators, partial R-indicators and partial coefficients of variation. Representativity Indicators for Survey Quality; 2015 [updated 2017 Dec 7; cited 2019 Mar 5]. Available from http://hummedia.manchester.ac.uk/institutes/cmist/risq/RISQ-manual-v21.pdf.

[11] Shlomo N, Skinner C and Schouten B. Estimation of an indicator of the representativeness of survey response. Journal of Statistical Planning and Inference. 2011; 142(1): 201-211.

[12] Brazilian Institute of Geography and Statistics – IBGE. Metodologia do Censo Demográfico 2010 (2nd ed.). Série Relatórios Metodológicos, v.41, 2016.

[13] dos Santos MPR. Indicadores para monitoramento de representatividade no caso da Pesquisa Nacional por Amostra de Domicílios Contínua. Dissertation (Master's degree in population, territory and public statistics) – National School of Statistical Sciences (ENCE), Rio de Janeiro; 2017 [updated 2018 Apr 5; cited 2019 Feb 15]. Available from http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2017/Mayra_Dissertacao.pdf.

[14] Rao JNK. Small area estimation. Wiley Series in Survey Methodology. John New Jersey: John Wiley & Sons, Inc, 2003.