Statistical Journal of the IAOS 31 (2015) 215–225
DOI 10.3233/SJI-150897
IOS Press



# Assessing coverage of the 2010 Brazilian Census

Andréa Diniz da Silva*, Marcos Paulo Soares de Freitas and Djalma Galvão Carneiro Pessoa
*Brazilian Institute of Geography and Statistics, Avenida Republica do Chile 500, 10° andar Rio de Janeiro, Brazil*

**Abstract.** To assess census coverage the Brazilian Institute of Geography and Statistics – IBGE has been conducting a post enumeration survey – PES since 70's census. In 2010 the survey was conducted in a sample of enumeration areas in each of the 27 federation units, matching was performed for data from Census and PES and a reconciliation work was conducted on the unmatched housing units and persons. Finally, dual-system estimation was applied to estimate the 2010 Census net coverage, omission and erroneous inclusion. One of the biggest improvements of the 2010 Brazilian Census is the incorporation of new methodologies and technologies. Use of handheld devices in the 2010 Census and PES allowed improvement of quality and timeliness in the data collection process, and facilitated automatic matching of PES to the Census. An automatic matching step, based on the probabilistic linkage theory of Fellegi and Sunter, was added to the assisted matching and reconciliation already performed in the 2000 Census PES, making the three-step 2010 matching system. The paper gives an overview on the process and methods used to carry on the 2010 Brazilian Post Enumeration Survey, bringing key information on the preparation, data collection, sampling, matching and estimation. Results allow knowing the performance of the matching system in addition to 2010 Census coverage measuring as a hole.

Keywords: Post enumeration survey, PES, census evaluation, coverage rate, record linkage, dual system estimation

## 1. Introduction

Population and housing censuses are the main source of information about the population and living conditions of a country. At local levels, information that allows the knowledge of the reality of the population depends on the census to be updated. Census data are essential for the formulation and monitoring of policies and to guide decisions of investment in the public and private sectors. They also allow the identification of priorities in the areas of health, education, housing and public transportation, and are used to analyse the consumer market and labour force, among other uses. In some countries the census is used to define the distribution of government resources, delimit electoral zones and establish the number of people's representatives in the Parliament.

Statistical offices make great investments for their Census achieve high level of response. However, complete coverage of the population is hardly feasible and therefore part of the population is not counted by the census. Further, coverage errors do not occur the same way in all geographic areas (regions, municipalities or states). Besides, overcount and undercount are also affected by population characteristics such as age. Thus, the identification of the levels and patterns of undercount is essential in order that the information produced can be used appropriately.

There are several ways to measure the coverage of a census: comparison with other sources, such as sample surveys or administrative records data are some of the techniques used, however, demographic analysis and post enumeration surveys are the most used by statistical offices worldwide.

> "Two thirds of the countries or territories conduct a Post Enumeration Survey to evaluate the coverage of the census, among them 75% to evaluate also some content errors. In Africa, Asia and South America, almost 80% of the countries un-

*Corresponding author: Andréa Diniz da Silva, Brazilian Institute of Geography and Statistics, Avenida Republica do Chile 500, 10° andar Rio de Janeiro, Brazil, CEP 20031-170. Tel.: +55 21 2142 0314; Fax: +55 21 99212 0704; E-mail: adiniz@ibge.gov.br.

dertake a PES, versus only 40% of the Oceanian countries." [10]

Since the 1970 Census, IBGE has been evaluating census coverage. Considering the commitment to transparency regarding the quality of the statistics produced by the Institute s post enumeration survey was also a part of the 2010 Census.

The survey was conducted on a sample of enumeration areas of each of the 27 Federation units and the data were compared with those collected by the 2010 Census in the same areas. To ensure independence between the two collections and also the comparability of its data, special attention was paid to some issues, such as the independence between the teams of Census and PES, non-disclosure of the enumeration areas during the census data collection, use of same concepts and definitions, same operating conditions as well as same technological background.

## 2. Preparation

The 2010 PES planning began in December 2008 when the project was prepared and the schedule of the research was set up. During 2009, issues relating to the survey sample, questionnaires, manuals, training model, methodologies and tools to perform the matching of census data with the one collected by the PES and methods for estimating the coverage rates of the 2010 census have been studied and developed. In 2010, systems for implementing and monitoring the data collection and transmission were developed. In addition, a pilot PES was conducted after each of the two census rehearsals in order to test and improve all methods and processes planned.

The 2010 PES is part of the 2010 Census thus its core issues have been submitted to the Census Committee. In weekly meetings the group, comprising experts from the Directorates of Research, IT and Geosciences, discussed issues related to technical aspects of the 2010 Census. For technical matters relating specifically to the PES, a group was also formed with representatives from the front-office of the Researches Directorate, Coordination of Methods and Quality, Coordination of the Census in addition to the Coordinator of the 2010 PES. Besides, specific operational issues linked to the differentiated period of collection and the need for greater mobility of the supervisors, as the PES is a sample survey, were discussed by the 2010 Census Coordination of Planning and Budget.

During planning the following key aspects were considered:

- Concepts and definitions: question wording, concepts and definitions to be used in both census and PES, should be the same in order that the matching could be performed without facing problems of lack of comparability of the variables used.
- Operational conditions: availability of equipment, vehicles and additional financial resources under the same conditions as to collect the census is essential to avoid both undercount and overcount in the collection of PES.
- Technology: perform data collection in PES using handheld and the same systems used in the census for the transmission and monitoring the collection of PES is essential.
- Independence between census and PES: national coordination of PES directly linked to the front-office of the Directorate of Surveys and state coordinators directly linked to heads of State Offices was considered crucial to ensure the independence between the two surveys.

The questionnaires were designed to collect the necessary data for comparing the census and the PES and also to register the mobility of the population in between the Census and the PES data collection. The questions were selected among those asked in the Census short form, which was applied to all population. Once the PES sample was independent from the Census sample, to which the Census long form is applied, questions asked solely in the Census long form were not considered.

In total, 7 questions on the housing unit address (street type and name, housing unit number, complements and references information) and 1 to identify if people moved in between the census and the PES were asked. Questions on persons were 3 in the short form and 15 in the long one. The short form included complete name of the responsible of the household and its partner and the number of men and women living in. The long form included information on complete name, sex and age of the out movers; relation to the household responsible, name, sex, month and year of birth/age, month and year when the person moved in, race and literacy.

The preparatory activities included two pilot Post Enumeration Surveys that were undertaken joining the two 2010 Census dress rehearsals. The objective of the two operations was to refine the processes and procedures to be adopted in the 2010 data collection. The operation was used to evaluate and improve the material, techniques and time allocated for training, as well as the quality of the cartography and also the hand-

held device application. Besides, an assessment of the coverage levels and patterns in the enumeration areas could be done as the operation comprised all the PES steps: data collection, matching and reconciliation.

In order to ensure the independence between Census and PES the survey counted on staff working exclusively in the PES project, so that an independent team conducted the survey. In each State the PES team had a Coordinator who was supported by one or more assistants, according to the number of enumeration areas in the PES sample. Each enumerator was responsible for one enumeration area and one supervisor followed the work in 3 to 6 EA, according to the distance and accessibility of the EA.

The training started in September in order to allow enough time to the cascade strategy adopted and to be as close as possible to the data collection date. It was developed according to the 2010 Census methodology with emphasis on consolidating the PES concepts. The training was structured in two parts: self-study and classroom activities. The self-study comprised reading the manuals and completing exercises found on the Exercise Notebook. The classroom training was an important step in the training because it was the moment when the training activities could be developed and theoretical content could be deepened.

All training material was developed based on the material prepared for the 2010 Census in order to maintain not only the same concepts and definitions used, but also the same structure, language and layout. The video and manual were reproduced under the responsibility of the Centre for Documentation and Information Dissemination of IBGE and sent to State Units for distribution in local training.

Cascade training was implemented as follows:

- Regional Training: 4 poles (N, NE, SE, S), 27 Coordinator and 49 Assistant trained;
- State Training: 26 States and 1 Federal District, 1,179 Supervisors trained; and
- Local Training: around 1,700 localities, 3,877 enumerators trained.

## 3. Sampling

The sample was designed to allow measuring the coverage of occupied permanent private housing units and their residents in the 2010 Census. The design allows us to estimate the following rates:

- Net Coverage Error of occupied private housing units;
- Omission rate of occupied private housing units;
- Erroneous inclusion rate of occupied private housing units;
- Net Coverage Error of people living in occupied private housing units by age groups;
- Omission rate of people lining in occupied private housing units;
- Erroneous inclusion rate of people living in occupied private housing units.

In addition, the following levels of disaggregation were considered: Brazil, Regions, Federation Units, and each of these geographic levels, by housing unit status, urban and rural.

### *3.1. Target population*

The target population of the PES was composed of the permanent private occupied housing units and their residents. The survey was conducted in the 26 States and in the Federal District, in urban and rural areas. The PES did not target special enumeration areas such as campsites, military bases, ships, boats, indigenous areas; and institutions such as penitentiary institutions, asylums, orphanages, convents and hospitals. Coverage rates were estimated for occupied private housing units and people living there.

### *3.2. Sample design*

To estimate the coverage rates described above, two different sampling designs were considered, each of them being used for the estimation of specific rates.

The estimation of net coverage error and omission of occupied private housing units (OPHU) and the omission of residents in OPHU used data from a one-stage stratified sample of census enumeration areas, considered as primary sampling units (PSU). Firstly, the stratification of the PSU was performed considering the federation units (FU), for which it was desired to obtain the estimates. Then, within each federation unit, the PSU were stratified by census areas[1] and, finally, were stratified by the situation of the enumeration area: urban and rural. The enumeration areas were selected with probability proportional to the number of OPHU provided in the address file of the 2010 Census. In each enumeration area, all units were registered and for each OPHU the information was collected on the total number of residents per sex and date of occupation.

[1] Area that includes contiguous enumeration areas, within one or more municipalities.

To estimate net coverage error and the omission of residents in OPHU, a sample of OPHU was selected by simple random sampling within each enumeration area already selected at first stage. In each OPHU selected at second stage, the enumerator filled a questionnaire to collect information about all persons living in the housing unit at the time of the survey, along with information about people who was living in the housing unit at the census reference date but moved out or died after this date.

### 3.3. Sample size

The size of the sample of enumeration areas was determined considering the desired levels of precision for estimating the omission rate of OPHU. As the sampling design that was effectively adopted in the survey was not a Simple Random Sampling (SRS) of OPHU, it was necessary to adjust the size, considering the Design Effect (DEFF) that indicates how the sampling design is less efficient since it has higher variance than the SRS.

The formulas used in each FU are the following:

$$n_{AAS} = \frac{N z_{\alpha 2}^2 CV_{00}^2}{Ner^2 + z_{\alpha 2}^2 CV_{00}^2}$$

and

$$n_{AC} = n_{AAS} DEFF$$

where

$n_{AAS}$ is the sample size of enumeration areas on simple random sampling at FU;

$N$ is the total number of enumeration areas at FU;

$CV_{00} = \frac{s}{\bar{y}}$ is the estimated coefficient of variation of the number of housing units omitted at FU, calculated based on data from the 2000 Census PES;

$$s = \frac{1}{n^{\cdot} - 1} \sum_{i=1}^{n^{\cdot}} (y_i - \bar{y})^2;$$

$$\bar{y} = \frac{\sum_{i=1}^{n^{\cdot}} y_i}{n^{\cdot}};$$

$n^{\cdot}$ is the number of enumeration areas selected for the 2000 Census PES sample at FU;

$y_i$ is the total number of housing units omitted in the enumeration areas $i$ selected for the 2000 Census PES sample;

Table 1
Enumeration areas in the sample of PES by Federation Units

| Federation Units (UF) | Number of enumeration areas in the sample | | |
|---|---|---|---|
| | Total | Urban | Rural |
| **Total** | **4,011** | **2,954** | **1,057** |
| Rondônia | 75 | 44 | 31 |
| Acre | 42 | 26 | 16 |
| Amazonas | 179 | 123 | 56 |
| Roraima | 76 | 59 | 17 |
| Pará | 62 | 37 | 25 |
| Amapá | 92 | 77 | 15 |
| Tocantins | 83 | 55 | 28 |
| Maranhão | 152 | 74 | 78 |
| Piauí | 58 | 30 | 28 |
| Ceará | 185 | 128 | 57 |
| Rio Grande do Norte | 116 | 82 | 34 |
| Paraíba | 113 | 73 | 40 |
| Pernambuco | 235 | 164 | 71 |
| Alagoas | 116 | 75 | 41 |
| Sergipe | 162 | 100 | 62 |
| Bahia | 222 | 129 | 93 |
| Minas Gerais | 345 | 264 | 81 |
| Espírito Santo | 131 | 105 | 26 |
| Rio de Janeiro | 173 | 161 | 12 |
| São Paulo | 393 | 351 | 42 |
| Paraná | 76 | 55 | 21 |
| Santa Catarina | 172 | 136 | 36 |
| Rio Grande do Sul | 186 | 138 | 48 |
| Mato grosso do Sul | 115 | 90 | 25 |
| Mato Grosso | 167 | 126 | 41 |
| Goiás | 88 | 68 | 20 |
| Distrito Federal | 197 | 184 | 13 |

$er$ is the maximum relative error desired when estimating the housing units omission rate;

$z_{\alpha 2}$ is the quantile in the standard normal distribution corresponding to the desired confidence level;

$(1-\alpha)$ is the level of confidence to estimate the omission rate of housing units with relative error $er$;

$n_{AC}$ is the sample size of enumeration areas under cluster sampling and

*DEFF* is the design effect, estimated based on the sample of 2000 Census PES.

The sample sizes were calculated to ensure the precision of the estimates in terms of FU: relative error of 0.20 with a confidence level of 95%. In total 4,011 enumeration areas were selected. The sample size for each FU is shown in Table 1.

Once settled the sample sizes of enumeration areas by FU, it was defined the sample size of housing units in each enumeration area, for the two stages sample. We decided to fix the sampling fraction and use the same selection procedure adopted to select the housing units in which the long form was applied in the 2010 Census: simple random sampling.

Studies conducted during the planning of the PES sample [24] indicated that the contribution of variabil-

ity within enumeration areas in the total variability is very small. At the occasion, fractions of 10%, 15% and 20% were evaluated, and the results in terms of coefficient of variation of the estimates were very close. Thus, there would be no need for a very large sample of housing units within each enumeration area, and the fraction of 10% was chosen. The same fraction was also adopted in 2010 PES.

## 4. Data collection

In most of the enumeration areas data collection was conducted from November 14 to December 23 2010, during around two weeks in each enumeration area. In order to keep proximity to the reference date of the 2010 Census and take profit of the recent mobilization of the population, the PES data collection started just after the census data collection was completed in the municipality. In the states capitals and a few other big towns, where the census data collection took longer due to a lower availability of the respondents, the PES data collection started after census data collection was completed in the enumeration area. This strategy allowed the necessary time to complete the operations of census data collection, including the phases of monitoring and revision, in order to ensure strict independence between the PES and the census.

The PES data collection team counted on 27 coordinators (one per state or federal district), 49 assistants, 1,179 supervisors and 3,877 enumerators. Permanent staff, appointed by the chief of the IBGE state office, occupied the positions of state coordinators and assistant. All of them attended the Census training and were previously in the Census project as training coordinator, cartography coordinator or other position related to preparatory activities, or had experience in household surveys. Supervisors and enumerators were hired on short-term contracts through the 2010 Census public contest. Supervisors were either hired for the PES or transferred from the census team to work in the PES once the Census work was completed. All the enumerators were transferred from the census team once the data collection was completed. For both, supervisors and enumerators, the sine qua non condition to work in the PES was to work in different enumeration area from the ones they worked in the Census.

The PES data collection was monitored through the Data Collection Management Indicators System (SIGC), which displayed the number of enumeration areas by status (not started, in progress and finished), the number of housing unit and people counted. A National Coordinator, a Survey Directorate staff, as well as the 27 state Coordinators, carried on the monitoring. The States Offices Chiefs also followed up all the data collection, through the system.

## 5. Matching system

One of the biggest improvements of the 2010 Brazilian Census is the incorporation of new methodologies and technologies developed or improved throughout the last decade. The use of handheld devices for the data collection, already experienced in the 2007 population count, was one of the successful innovations in the Census project, allowing improvement of quality and timeliness in the data collection process. The use of such technologies in the PES enabled immediate transfer of data collected and facilitated the implementation of automatic matching with the census data.

A matching system was designed aiming to find as much as possible the units that were enumerated by both Census and PES – the true matches. An accurate matching process was essential as the number of matches/unmatches had an effect on the coverage rates. The level of false matches was strongly controlled. The issue of undesirable false positive (false matches) was emphasized in the training, supervision and revision of the matching operation. The false negative (missed true matches) was minimised by successive steps in the matching system. Both false positives and false negatives were controlled and minimized in each step of the matching system.

The matching system comprised three stages: automatic linkage, assisted matching and reconciliation.

### *5.1. Automatic linkage*

The automatic linkage step was based on the probabilistic linkage theory [11], in which a probabilistic model is developed to identify the matched records, in our case, person or housing unit in the Census and Post Enumeration Survey data files.

In accordance with the basic ideas of the probabilistic linkage theory, scores were computed which depended on the agreement and the disagreement probabilities of selected variables in the pairs of records. For this, the probabilities of having agreement for matched and unmatched pairs were compared. It is expected that variables with greater discrimination power between matched and unmatched pairs will have higher frequency of agreement for the matched pairs.

*5.1.1. Notation*

Let us consider two sets $A$ and $B$ of entities, which in our case were housing units or persons, having non-empty intersection. We denote the generic elements in those sets by $a \in A, b \in B$. Associated with the sets $A$ and $B$, we observe a set of variables whose values are recorded in the files $\alpha(A)$ and $\alpha(B)$, respectively. The records corresponding to the entities $a$ and $b$ are denoted by $\alpha(a)$ and $\alpha(b)$, respectively. We define two subsets of the set $\alpha(A) \times \alpha(B)$ of pairs of records:

matched pairs:

$$M = \{(\alpha(a), \alpha(b)) \,|a = b\};$$

unmatched pairs:

$U = \{(\alpha(a), \alpha(b)) \,|a \neq b\}$. A comparison function $C$: $(\alpha(a), \alpha(b)) \rightarrow \gamma$, associates to each pair of records a comparison vector $\gamma$ having dimension equal to the number of variables used in the comparison, and components in the set $\{0, 1\}$.

For example, the following $\gamma$ vector could be obtained comparing housing units records with three variables:

$\gamma_1$ – the last name of the household head;

$\gamma_2$ – the first name of the household head;

$\gamma_3$ – the name of the street of the housing unit.

A simple agreement pattern is $\gamma = (1, 0, 1)$, where the value 1 of the vector component means an agreement and 0 a disagreement. There is also the flexibility of getting a more complex agreement pattern like $\gamma = (0.36, 0, 0.80)$ using a different comparison criterion.

*5.1.2. Likelihood ratio*

We consider the following conditional probabilities:

$P(\gamma|M)$ – the conditional probability of having an agreement pattern $\gamma$ for the two records, given that they are matched;

$P(\gamma|U)$ – the conditional probability of having an agreement pattern $\gamma$ for the two records, given that they are unmatched.

Let $P(M)$ be the probability of the two records $(\alpha(a), \alpha(b))$ be matched. Then, by the Bayes theorem we have:

$$P(M|\gamma) = \frac{P(\gamma|M)\,P(M)}{(P(\gamma))},$$

but

$$P(\gamma) = P(\gamma|M)\,P(M) + P(\gamma|U)\,(1 - P(M)),$$

therefore

$$P(M|\gamma) = \frac{1}{1 + \frac{P(\gamma|U)(1-P(M))}{P(\gamma|M)P(M)}}.$$

By this equation, we see that $P(M|\gamma)$ increases with the ratio

$$R(\gamma) = \frac{P(\gamma|M)}{P(\gamma|U)},$$

called likelihood ratio, and which will be used as a score for the record linkage.

*5.1.3. Linkage rule*

We will use the following linkage rule:

1. Order the comparison vectors of the records according to their likelihood ratio values $R(\gamma)$;
2. Choose an upper cut-point $W_1$ and a lower cut-point $W_2$ for $R(\gamma)$;
3. Declare matched pairs the elements of $\alpha(A) \times \alpha(B)$ having $R(\gamma)$ value greater than $W_1$, and unmatched pairs those having $R(\gamma)$ value smaller than $W_2$.

A linkage rule $F$:$\Gamma \rightarrow \{A_1, A_2, A_3\}$ associates to each comparison pattern $\gamma \in \Gamma$ one of the three actions: $A_1$-declare the pair to be matched; $A_2$-do not decide and $A_3$-declare the pair to be unmatched. The Fellegi-Sunter decision rule $F(\gamma)$ is determined by the cut-points $W_1$ and $W_2$:

$$F(\gamma) = \left\{ \begin{array}{ll} A_1 & \textit{if} \quad R(\gamma) \geqslant W_1 \\ A_3 & \textit{if} \quad R(\gamma) \leqslant W_2 \\ A_2 & \textit{otherwise} \end{array} \right\}$$

The values of $W_1$ e $W_2$ are determined from the fixed values for the error probabilities of linking an unmatched pair and of not linking a matched pair.

Those two errors correspond to the Type I and Type II errors of the Statistical Theory of Testing Hypothesis.

Consider $K$ comparison variables with a comparison pattern $\gamma = (\gamma_1, \ldots, \gamma_K)$. Usually, we assume the conditional independence:

$$P(\gamma|M) = P(\gamma_1|M) \ldots P(\gamma_K|M)$$

and

$$P(\gamma|U) = P(\gamma_1|U) \ldots P(\gamma_K|U).$$

We adopt the following notation: $m_k = P(\gamma_k = 1|M)$ e $u_k = P(\gamma_k = 1|U)$ for $k = 1, \ldots, K$.

Under the assumption of conditional independence, the likelihood ratio becomes:

$$R(\gamma) = \frac{P(\gamma|M)}{P(\gamma|U)} = \frac{P(\gamma_1|M)\dots P(\gamma_K|M)}{P(\gamma_1|U)\dots P(\gamma_K|U)}.$$

Instead, we usually take the base 2 log of $R(\gamma)$, getting the following score:

$$\begin{aligned}\log_2(R(\gamma)) &= \log_2\left(\frac{P(\gamma_1|M)}{P(\gamma_1|U)}\right) + \dots \\ &\quad + \log_2\left(\frac{P(\gamma_K|M)}{P(\gamma_K|U)}\right) \\ &= \log_2\left(\frac{m_1}{u_1}\right) + \dots + \log_2\left(\frac{m_K}{m_K}\right).\end{aligned}$$

In practice, the values of $m_k$ and $u_k$ are not known, therefore it is not possible to compute the values of the score R. However, these values could be estimated using the EM algorithm, which is a general method to derive Maximum Likelihood Estimators in the presence of missing values [9]. In this method, starting from a parametric model for the observed data, iterations are executed, each having two steps. In the E step the missing values are imputed and in the M step the Maximum Likelihood Estimator is computed for the completed data. The procedure stops when some fixed convergence criterion is reached.

To implement the described ideas we have used the library RecordLinkage [1] of R [27]. The first step in the implementation consisted in comparing all pairs of records in the data files from the Post Enumeration Survey and from the Census, using selected variables and a comparison criterion. The RecordLinkage library allows the choice of different comparison criteria for character string variables, using different similarity measures. We adopted the Jaro-Winkler criterion.

One of the functions of the RecordLinkage library implements the EM algorithm to estimate the conditional probabilities of the resulting comparison patterns. This function assumes that the values of the $\gamma$ vector components are either 0 or 1. In order to use the Jaro-Winkler criterion a threshold value was set for the similarity measure. Values of the similarity measure above this threshold were taken to be equal to 1, and 0 otherwise. This function returns the value of the conditional probability of having an observed comparison pattern given $M, U$ and the corresponding score. Using the fixed limits for type I and II errors, the pairs of records were classified in one of three classes: matched, unmatched and not classified.

The following variables were used in the housing unit record comparison: 1. type of street; 2. concatenated title and name of the street; 3. number value part of the address; 4. address complement; 5. first name of the household head; 6. last name of the household head; 7. total of men; 8. total of women.

The values of the variables 2, 5 and 6. were compared through the Jaro-Winkler similarity criterion adopting the thresholds 0.85, 0.90 and 0.90, respectively. For the remaining variables the exact match criterion was adopted.

The final linkage results obtained were of the $m{:}n$ association type, that is, to each record of the Census file more than one record in the Post Enumeration Survey file could be linked. An $1:1$ association was then obtained by solving the problem of maximizing the sum of scores of the pairs, subjected to the condition of an 1:1 association between records from the Census and Post Enumeration Survey files.

This optimization problem could be expressed in terms of a linear programing problem, and solved by means of the Simplex Algorithm using the library lpSolve [21] of R. For the usual file sizes to be linked, the method shown itself not applicable, due to RAM limitations of the R.

The final solution to this problem was obtained by means of the library clue [20] of R, that, instead, implements the Hungarian Algorithm, which was much more efficient than the use of the Simplex Algorithm.

For linking persons, the housing units in the Post Enumeration Survey file that were first matched were taken as blocks. The procedure used was similar to the one adopted for the linkage of housing units. Measures of similarity were computed using the following variables: first name of the person; last name of the person, and age of the person. Differently from the housing unit linkage, weights of agreement were fixed in advance, a weight being assigned to each agreement standard.

### 5.2. Assisted matching

The assisted step was held for all housing units and persons declared unmatched according to the rule described in 5.1.3 – Linkage rule.[2] The procedures

[2] "3.declare matched pairs the elements of $\alpha(A) \times \alpha(B)$ having $R(\gamma)$ value greater than $W_1$, and unmatched pairs those having $R(\gamma)$ value smaller than $W_2$."

Table 2
Distribution of Matched Private Occupied Housing Units by stage matching

| Region | Proportion of matched housing units in each step of matching system | | | | | |
|---|---|---|---|---|---|---|
| | Total | Automatic | | Assisted | Reconciliation | Unknown |
| | | Inside enumeration area | Neighbour enumeration area | | | |
| Brazil | 100 | 75.63 | 0.80 | 20.22 | 3.20 | 0.15 |
| North | 100 | 72.00 | 0.90 | 22.66 | 4.25 | 0.18 |
| Northeast | 100 | 72.46 | 0.87 | 22.95 | 3.68 | 0.04 |
| Southeast | 100 | 79.37 | 0.82 | 17.70 | 1.83 | 0.28 |
| South | 100 | 84.23 | 0.84 | 13.14 | 1.60 | 0.19 |
| Midwest | 100 | 73.41 | 0.43 | 21.28 | 4.78 | 0.10 |

Source: IBGE, Pesquisa de Avaliação da Cobertura da Coleta do Censo 2010.

included revision on the "unmatched" pairs through an application developed by the IBGE IT Directorate staff. The operators were trained to evaluate the pair and undo the match in case it was detected to be really false, as much as search for new true pairs. 15 operators and 5 supervisors were in charge of the job.

### *5.3. Reconciliation*

The last step of the matching system was made in the reconciliation stage. State coordinators were in charge of this task. They were advised to double check the data collected on the unmatched housing units and people, by both Census and PES, and search for new matches, as they knew better the field and could detect some systematic errors in the enumeration that could not be done by the clerical staff. Besides, new true matches could also be found while carrying out field checks, especially in rural areas where the addressing system is not standardised.

The reconciliation phase was an opportunity to complete the matching process, however its main objective was to double-check the information on housing units and people when any discrepancy was found in the data registered by Census or PES. Such discrepancies could be a different number of residents of a housing unit in Census and PES data collection or a complete housing unit not found in either the Census or the PES. The reconciliation was performed by the supervisors who worked in the data collection and staff of other household surveys. Supervision was implemented in two stages, the first one in charge by the local branch chief and the second one by the state coordinator.

The matching system has been fully implemented, with its three stages, immediately after the completion of data collection in each enumeration area. The performance of each matching step to the total number of matched housing units can be seen in Table 2.

Distribution of Matched Private Occupied Housing Units by stage matching Brazil and Regions

## 6. Estimation

For each federation unit, omission rates of housing units and persons were estimated. In the following sections, we present the sample weights, the estimation method employed, each rate estimator and their respective variance.

### *6.1. Sample weights*

Sample weights were calculated taking into account the probabilities of selection and adjustments for nonresponse.

In the estimation of rates that consider all housing units listed in the enumeration area, in other words, making use of sample data at one stage, the sample weight used was calculated for each enumeration area. The formulas used in each *FU* are the following:

$$w_{hi} = \frac{1}{m_h}\frac{N_h}{N_{hi}}$$

and

$$w^*_{hi} = \frac{1}{m_h}\frac{N_h}{N_{hi}}\frac{\sum\limits_{i\in s_{ht}} w_{hi}}{\sum\limits_{i\in s_{hr}} w_{hi}}$$

where

$w_{hi}$ is the basic weight sample of enumeration area $i$ of the stratum $h$;

$w^*_{hi}$ is the final weight sample of enumeration area $i$ of the stratum $h$;

$m_h$ is the number of census tracts selected for sample in the stratum $h$;

$N_h$ is the number of OPHU provided in geographical operating base of the 2010 Census at the stratum $h$;

$N_{hi}$ is the number of OPHU provided in geographical operating base of the 2010 Census at enumeration area $i$ of the stratum $h$;

$s_{ht}$ is the total set of selected enumeration areas for sample in the stratum $h$ and

$s_{hr}$ is the total set of selected enumeration areas for sample in the stratum $h$ with held collect;

To estimate the rates that consider only the housing units in the two stages sample in each enumeration area, the sample weight used was calculated taking into account the probabilities of selection of two stages in each *FU*, by the following formula:

$$w_{hij} = \frac{1}{m_h}\frac{N_h}{N_{hi}}\frac{\sum\limits_{i\in s_{ht}} w_{hi}}{\sum\limits_{i\in s_{hr}} w_{hi}}\frac{N_{hi}}{n_{hi}} = \frac{1}{m_h}\frac{N_h}{N_{hi}} \frac{\sum\limits_{i\in s_{ht}} w_{hi}}{\sum\limits_{i\in s_{hr}} w_{hi}} f_{hi}$$

where

$N_{hi}$ is the number of OPHU listed in the first stage of the 2010 PES at enumeration area $i$ of the stratum $h$;

$n_{hi}$ is the number of OPHU at the second stage's sample of the 2010 PES at enumeration area $i$ of the stratum $h$ and

$f_{hi}$ is the effective sampling fraction of the second stage of the 2010 PES at enumeration area $i$ of the stratum $h$.

### 6.2. Estimation method

The method used to estimate the omission rate was the Dual System Estimation [19], considered the most effective among a few different types of methodologies. Originally developed for estimating the size of closed populations (population unchanged throughout the study time) in biometric studies, this statistical model, which aims to estimate the level of coverage of the Census, is based on the technique of capture-recapture estimation. This technique was developed to estimate the size of animal populations and requires the use of independent sources of population to confront the information obtained by the sources.

The method makes the following assumptions:

- Closed population: The population remains unchanged. This fact does not occur in reality, since during the period between the Census and the PES, occur births, deaths and migrations. However, it is assumed that these changes are minimal, because there are questions in the PES questionnaire that allow treating appropriately such changes, so that it can be considered that the population is closed. The PES early collection occurred immediately after ending the collection of Census, reducing the elapsed time between the collection of the two surveys.
- Independence between the two surveys: It is of fundamental importance to the success of the survey and use of the proposed estimators in the method adopted. The 2010 Census, respected this requirement, at least with regard to the use of exclusive teams in planning and fieldwork of the PES. It is important to say that the PES collection at the selected enumeration area only began after the completion of the Census collects.
- No erroneous inclusions in Census: In practice, there are some types of undue inclusions, which must be removed from the total population. Examples: creation of housing units and people, duplication, registration of people who moved after the reference date of the Census and etc.
- Exact matching between Census and PES: In the comparison of information can occur failures or lack of sufficient information to enable the units to be coded rightly. To circumvent this problem, exists the reconciliation phase, where the divergent information are clarified by returning to the surveyed housing unit. Therefore, it is said that reconciliation allows a is a exact matching between Census and PES.

After PES fulfilment, the data from the two surveys were compared and the units classified into four[5] categories, as described in Fig. 1.

| PES | Census | | |
|---|---|---|---|
| | Units listed | Units non-listed | Total |
| Units listed | $a$ | $b$ | $a+b$ |
| Units non-listed | $c$ | $d$ | $c+d$ |
| Total | $a+c$ | $b+d$ | $t$ |

Fig. 1. Results of the comparison of the data from two surveys.

where

$a$ is the number of units listed in both surveys;

$b$ is the number of units listed only in the PES;

$c$ is the number of units listed only in the Census;

$d$ is the number of units non-listed in both surveys and

$t$ is the total units of the population.

From Fig. 1, it is observed that the proportion of omitted units or not listed in the Census is given by $R = \frac{b+d}{t}$.

Since no one knows the number of units not listed in both surveys ($d$), this ratio should be calculated by the expression $R = \frac{b}{a+b}$.

The justification for this, is that you can use another expression, whose elements are known, is given below:

In a binomial distribution, $S$ is the number of successes and $N$ is the population size. So, $p = \frac{S}{N}$ is the probability of success in the population.

Whereas the amount of units listed in each study is a random variable with a binomial distribution, and that the two are independent variables, we have:

$p_1 = \frac{(a+c)}{t}$ is the probability of a unit being listed in the Census;

$p_2 = \frac{(a+b)}{t}$ is the probability of a unit being listed in the PES;

$p_{1'} = \frac{(b+d)}{t}$ is the probability of a unit not being listed in the Census;

$P_{12} = P_1 P_2 = \frac{a}{t}$ is the probability of a unit being listed in both surveys and

$P_{1'2} = P_{1'} P_2 = \frac{b}{t}$ is the probability of a unit being listed in the PES and not being listed in the Census.

Whence:

$$a = \frac{(a+c)\,(a+b)}{t};$$

$$b = \frac{(b+d)\,(a+b)}{t}$$

and

$$t = \frac{(a+c)\,(a+b)}{a} = \frac{(b+d)\,(a+b)}{b}.$$

Thus, one can derive an expression of the ratio of units not listed in the Census, as shown above:

$$R = \frac{b+d}{t} = p_{1'} = \frac{b+d}{\frac{(b+d)(a+b)}{b}} = \frac{b}{(a+b)}$$

The estimator for this ratio, in other words, to the omission rate of the Census, from the capture-recapture method is given by $\hat{R} = \frac{\hat{b}}{(\hat{a}+\hat{b})}$.

Although it is assumed that there are no units erroneously listed by the Census, as stated earlier, this failure occurs. Thus, was also estimated the rate of erroneous inclusion of units in the Census, for a correct estimating of the total population.

The erroneous inclusion rate is given by $IR = \frac{\hat{I}}{CENSUS}$ and net coverage rate is given by $TL = \frac{\hat{R}-TI}{1-TI}$.

The expression of the adjusted population is given by

$$\textit{Adjusted Population} = \frac{CENSUS}{1-TL},$$

where

$\hat{b}$ is the estimate of the total units non-listed in the Census;

$\left(\hat{a}+\hat{b}\right)$ is the estimate of the total units in the population from the PES;

$\hat{I}$ is the estimate of the total units erroneously included in the Census and

*CENSUS* is the total units in the population listed in the Census.

## Final remarks

The 2010 Census Post Enumeration Survey represent a great advance towards the previous ones not only due to complete independence from census operation, once its coordination was allocated outside census area, but also due to incorporation of new methodologies and technologies. Use of handheld devices and automatic matching allowed improvement of quality as a hole, bringing both more accuracy and timeliness for estimates.

## References


[1] Andreas Borg and Murat Sariyar (2012). RecordLinkage: Record Linkage in R. R package version 0.4-1. http://CRAN.R-project.org/package=RecordLinkage

[2] Australian Bureau of Statistics. (2007). Census of population and housing: details of undercount. Australia: Australian Bureau of Statistics.

[3] M.G. Borges and A.D. Silva, As Pesquisas de Avaliação e sua utilização no ajuste dos resultados de um Censo e em estimativas e projeções demográficas. Instituto Brasileiro de Geografia e Estatística, Diretoria de Pesquisas. Rio de Janeiro, 2010.

[4] L.N. Costa, Estudos e pesquisas de avaliação dos Censo Demográficos de 1970 a 1990. Textos para discussão Número 34. Instituto Brasileiro de Geografia e Estatística, Diretoria de Pesquisas. Rio de Janeiro, 1990.

[5] P. Christen, Febrl: An Open Source Data Cleaning, Deduplication and Record Linkage System with a Graphical User Interface. In: 14th ACM SIGKDD international conference on knowledge discovery and data mining, 2008.

[6] W.G. Cochran, Sampling techniques, 3rd Edition, John Wiley & Sons, New York, 1977.

[7] P.S. Coelho and F. Casimiro, Post enumeration survey of the 2001 Portuguese population and housing censuses, *REVSTAT – Statistical Journal* **6**(3) (2008), 31–252.

[8] M. Dauphin and A. Canamucio, Design and implementation of a post-enumeration survey: developing country example. Washington, D.C.: International Statistical Programs Center, Bureau of the Census, 1993.

[9] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximun Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society. Series B* **39** (1977), 1–38.

[10] J.M. Durr, The 2010 round of population and housing censuses in the world. In: 19° SINAPE, São Pedro, SP, 2010.
[11] I.P. Fellegi and A.B. Sunter, A theory for record linkage, *Journal of the American Statistical Association* **64**(328) (1969), 1183–1210.
[12] M.P.S. Freitas, B.F. Cortez and T.M. Dantas, Pesquisa de avaliação da cobertura da coleta do censo 2010 – plano amostral. Instituto Brasileiro de Geografia e Estatística, Diretoria de Pesquisas. Rio de Janeiro, 2011.
[13] Instituto Brasileiro de Geografia e Estatística – IBGE. (2009). Censo demográfico 2010: manual do recenseador do censo 2010. Rio de Janeiro.
[14] Instituto Brasileiro de Geografia e Estatística – IBGE. (2010). Pesquisa de avaliação da cobertura da coleta do Censo 2010: censo experimental de rio claro. Rio de Janeiro.
[15] Instituto Brasileiro de Geografia e Estatística – IBGE. (2011). Pesquisa de avaliação da cobertura da coleta do censo 2010: coleta dos dados e reconciliação. Rio de Janeiro.
[16] Instituto Nacional de Estatística. (2003). Inquérito de qualidade: Censos 2001: XIV recenseamento geral da população: IV recenseamento geral da habitação Instituto Nacional de Estatística. Lisboa.
[17] M.A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association* (1989).
[18] D. Kerr, A review of procedures for estimating the net undercount of censuses in Canada, the United States, Britain and Australia. Canada: Statistics Canada, 1998.
[19] K.J. Krotki, Developments in dual system estimation of population size and growth. 1st Edition, The University of Alberta Press Edmonton, Alberta, 1978.
[20] Kurt Hornik (2014). clue: Cluster ensembles. R package version 0.3-48. URL http://CRAN.R-project.org/package=clue .
[21] Michel Berkelaar and others (2014). lpSolve: Interface to Lp_solve v. 5.5 to solve linear/integer programs. R package version 5.6.10. http://CRAN.R-project.org/package=lpSolve
[22] Office for National Statistics United Kingdom, 2011 United Kingdom census coverage assessment and adjustment methodology. In: UNECE/Eurostat Meeting on Population and Housing Censuses, Geneva, 2008.
[23] L.C.S. Oliveira et al., Censo demográfico 2000: resultados da pesquisa de avaliação da cobertura da coleta. Textos para discussão Número 9. Instituto Brasileiro de Geografia e Estatística, Diretoria de Pesquisas. Rio de Janeiro, 2003.
[24] L.C.S. Oliveira, M.P.S. Freitas and Z. Bianchini, Pesquisa de avaliação da cobertura da coleta do censo demográfico do ano 2000 – definição do desenho amostral. Instituto Brasileiro de Geografia e Estatística, Diretoria de Pesquisas. Rio de Janeiro, 1999.
[25] D.G.C. Pessoa, F.F. Farias and V.L. Xavier, Pareamento Automático na Pesquisa de Avaliação da Cobertura da Coleta do Censo Demográfico de 2010. Textos para discussão Número 41. Instituto Brasileiro de Geografia e Estatística, Diretoria de Pesquisas. Rio de Janeiro, 2012.
[26] B. Pink, Measuring net undercount in the 2006 population census. Australia: Australian Bureau of Statistics, 2007.
[27] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
[28] A.D. Silva and M.G. Borges, Avaliação da cobertura e ajuste dos dados do Censo: estudo da experiência do Reino Unido como subsídio para a discussão de um projeto brasileiro. Instituto Brasileiro de Geografia e Estatística, Diretoria de Pesquisas. Rio de Janeiro, 2010.
[29] A.D. Silva et al., Inovações no sistema de pareamento de domicílios e pessoas para a Pesquisa de Avaliação da Cobertura da Coleta do Censo 2010. In: 19° SINAPE, 2010, São Pedro, SP, 2010.
[30] A.D. Silva, 2010 Brazilian census post enumeration survey. In: 58th World Statistics Congress of the International Statistical Institute, Dublin, 2011.
[31] A.D. Silva et al., Study of record linkage software for the 2010 Brazilian census post enumeration survey. The Survey Statistician: the newsletter of the International Association of Survey Statisticians, No 65, 2012.
[32] United Nations, Principles and recommendations for population and housing censuses, revision 2. Statistical Papers, Series M n° 67/Rev.2, 2008.
[33] United Nations, Manual on census evaluation: post enumeration surveys. Department of Economic and Social Affairs. Statistics Division. New York, 2009.
[34] United Nations, Post enumeration surveys: operational guidelines. Technical Report. Department of Economic and Social Affairs. Statistics Division. New York, 2010.
[35] P.J. Waite, Evaluation of census quality and coverage. UNECE-Eurostat Work Session on Population Censuses. Geneva, 2004.
[36] V.L. Xavier, A system developed for solving the matching problem in the Brazilian census post enumeration survey. 58th World Statistics Congress of the International Statistical Institute. Dublin, 2011.