Paradata on the 2010 Brazilian Census: Analysis of the fieldwork supervision process

Luciano Tavares Duarte[1]
IBGE, Rio de Janeiro, Brazil – luciano.duarte@ibge.gov.br
Denise Britz do Nascimento Silva[2]
ENCE, Rio de Janeiro, Brazil – denise.silva@ibge.gov.br
Jose Andre de Moura Brito[2]
ENCE, Rio de Janeiro, Brazil – jose.m.brito@ibge.gov.br

Abstract

The relevance of a Census is unquestionable for many National Statistical Systems due to its thematic range and territorial scope. Nonetheless, its complexity leads to challenges in ensuring timeliness and the quality of the results. The aim of this work is to identify potential causes of nonsampling errors associated to the data collection process of the 2010 Brazilian Census.

We used data obtained from the field work monitoring system that provided information about divergences observed between data collected by enumerators and supervisors. The latter carried out follow-up interviews in households selected by the supervision/monitoring plan. In addition, human resources databases containing socio-demographic information of enumerators and supervisors, such as gender, age and educational level, were also brought to bear to enhance the analysis. Moreover, in order to investigate associations between paradata and the socio-demographic profiles of the survey respondents, Census microdata was also used.

The statistical analysis employed generalized hierarchical models in which three nested hierarchical levels were taken into account, namely the supervisor, the enumerator and the respondent. The model response variable was defined as the occurrence of a discrepancy (or divergence) between the information collected by enumerators and their supervisors for at least one of the following respondent characteristics: sex, age and literacy.

The results indicate that the different hierarchical levels investigated are relevant to decompose the data variability and hence have to be considered in the analysis. We also found evidence of significant associations between the occurrence of divergences (on data collected by enumerators and supervisors) and socio-demographic characteristics of enumerators, supervisors and respondents. In addition, the results revealed notable regional differences regarding the divergences of each level and the associations with explanatory variables considered in the study.

Keywords: Hierarchical Models; Paradata; Census; Nonsampling Errors.

[1] Census Statistician - Brazilian Institute of Geography and Statistics - IBGE
[2] Researcher of the Brazilian Institute of Geography and Statistics - IBGE and Lecturer of the Graduate Program in Population, Territory and Public Statistics of the National School of Statistical Sciences – ENCE/IBGE.

Disclaimer: This paper is based on the MSc dissertation of the first author written at ENCE. The views expressed in this paper are those of the authors and do not necessarily represent the views of IBGE.

1. Introduction

Public statistics are extremely important for a country due to many aspects, play a key role in underpinning the development and the monitoring of public policies and are also indispensable for resource allocation in multiple areas. In addition, allow citizens and the private sector to take informed decisions and make a basic tool to assess government actions. They are also central for comparing the development of nations taking into account economic and social indicators (Holt, 2005). A Demographic Census is a key part of the public statistic framework, being a major source of socio-demographic information for a country due to its thematic range and national scope. Censuses are also essential for the production of relevant information on the many levels of territorial disaggregation in the country (IBGE, 2012).

In the last census operations (2010), the Brazilian Institute of Geography and Statistics (IBGE) adopted innovative methods aiming at improving its practices and keeping up to date with technological advances. The massive inclusion of new information and communication technologies (ICTs) in the data collection process led to a significant increase in options for operational control (IBGE, 2008). As a result, the production of management and operational information highly increased throughout the survey process.

The definition of paradata can then be examined in this context. Couper (1998) employed the term for the first time in quantitative surveys, and it refers to using information and measurements related to the process of collecting survey data, administrative and management data about the survey administration to assess and improve the quality of the survey process. This paper explores paradata for identifying issues and factors associated with the quality of the results and with the improvement of survey procedures and protocols. We analyze Brazilian Census paradata obtained from the fieldwork supervision process together with human resources information on enumeration staff and 2010 Census microdata related to the respondents (Duarte, 2014).

2. Possible error sources in the 2010 Demographic Census

Groves (1989) states that the total error associated with a survey process is a combination of of two components. The first, known as the systematic error or bias, may cause the true value of a parameter of interest to be overestimated or underestimated**.** In the case of sample surveys, the second component is associated with the variability, or precision, of the estimates (sampling error). The main hypothesis of this work is that there are many nonsampling error sources and factors associated with the Census collection operation that can cause flaws or uncontrolled variation, affecting the reliability of population and household census data. As with any other production process, human interference is considered an inherent source of variation for a survey, among other causes, accounting for a non-negligible part of the flaws or non-controlled variation. As a result, it is considered a potential source of nonsampling error (Biemer & Liyeberg, 2003).

The 2010 Census control systems produced management information for evaluating the work of the entire field staff in an attempt of minimizing, still during data collection phase, the effects of possible mistakes that directly impact the quality of the data, especially those caused by human interference. The fieldwork supervision system stands out among the control systems, providing information on discrepancies between interviews conducted by Enumerators and follow-up interviews conducted by Supervisors. In addition to this management information from the control systems, human resources data on socio-demographic profiles of Enumerators and Supervisors involved in the fieldwork operation was linked to the occurrence (or not) of divergences. Moreover, another potential source of variation inherent to the enumeration process was identified: the respondents that effectively provided information during the interview (Weisberg, 2005). In the last Brazilian Census, a new feature of the questionnaire enabled the identification of who was the person that provided information about each resident, so it was possible to gather information on all the characteristics investigated in the census for every respondent.

Furthermore, the operational control information used in the collection process for safeguarding the integrity of the survey database could also constitute a valuable input to this study, such as the time and date in which each interview was conducted and the number of times the questionnaire was edited (Nicolaas, 2011). Due to the multiple potential factors which may cause nonsampling errors and the availability of information on the collection process, the analysis of the 2010 Brazilian Census paradata was welcome and contributes to the planning of future census operations and other sample surveys. This work studied the divergences between the information collected by Enumerators during data collection for the 2010 Brazilian Demographic Census and by Supervisors in follow-up interviews administered according to the fieldwork supervision system.

We considered there was a divergence when different answers were obtained by the Enumerator and the Supervisor for at least one of the following three main socio-demographic characteristics of the respondent in the interviewed household: sex, age and literacy. It is worth singling out that the divergence is used as a proxy of the occurrence of a defect in the collection process and, although it is believed that the information of the Supervisor is more reliable, divergences may certainly come from flaws of the Supervisor, not of the Enumerator. A limitation of the study is the fact that the divergence is considered a flaw without the possibility of detecting the real source of the error.

The main goal was to identify factors that are possibly related to divergences, i.e., that are assumed to be sources of variation. Three agents directly participated in filling out the information, being main suspects in the causes of divergences: the person who provided the information (Respondent), the Enumerator and the Supervisor. Based on this assumption, we searched for empirical evidences that the three agents influenced the answers of follow-up interviews and analyzed how much their characteristics may have an effect on the probability of a divergence, lowering or increasing its predicted value. In addition, we searched for evidences of possible variations associated with the socio-demographic characteristics of the agents, assumed to be associated with the event. The target population of this study [3] consisted of a subset of the follow-up interviews conducted in census enumeration sectors where only one Enumerator has done the job [4]. Data from the states of Alagoas (AL), Amazonas (AM), Santa Catarina (SC), Rio de Janeiro (RJ) and Mato Grosso (MT) were selected, each representing a geographic region of the country[5].

3. Application of generalized hierarchical models to the analysis of divergences

Regarding data structure, it is worth noting that each Enumerator was responsible for conducting one set of follow-up interviews [6], and each Supervisor was in charge of coordinating one group of Enumerators. Consequently, it is an aggregate, or hierarchical, data structure in which the first level corresponds to the Respondent, the second level to the Enumerator and the third level to the Supervisor. In order to obtain inferences that consider the hypothesis that there is a significant effect originated in the hierarchical levels of the data structure, we employed hierarchical logistic regression models (Hox, 2010; Raudenbush & Bryk, 2002). The conditional hierarchical logistic model adopted has the following expression:

[3] The criteria for choosing the states considered the percentage of divergences per state in each region and database sizes that enabled the application of the chosen statistical modeling procedure for data analysis.

[4] In enumeration sectors with more than one Enumerator, it was not possible to identify which of them conducted each interview, making it impossible to associate and interview with an Enumerator. These cases account for 11% of the interviews.

[5] Brazil is geopolitically divided into five regions: North, Northeast, West Central, Southeast and South.

[6] Follow-up interviews were conducted only for households with a single respondent therefore each follow-up interview corresponds to one respondent.

$$Logit\left(\pi_{ijk}\right) = \underbrace{\beta_{0jk}}_{\text{random effect}} + \underbrace{\overbrace{\sum_{q=1}^{Q}\beta_q X_{qjki}}^{\text{Respondents}} + \overbrace{\sum_{r=1}^{R}\gamma_r Z_{rjk}}^{\text{Enumerators}} + \overbrace{\sum_{s=1}^{S}\delta_s W_{sk}}^{\text{Supervisors}}}_{\text{fixed effects/covariates}} \quad (1)$$

$$\beta_{0jk} = \beta_0 + u_{0k} + v_{0jk} \quad \begin{cases} \beta_0 - \text{ intercept} \\ u_{0k} \sim N\left(0, \sigma_{u_0}^2\right) - \text{Variance component attributed to Supervisors} \\ v_{0jk} \sim N\left(0, \sigma_{v_0}^2\right) - \text{Variance component attributed to Enumerators} \end{cases}$$

where π_{ijk} is the probability of a divergence occurring in at least one of the socio-demographic characteristics for Respondent *i*, associated with Enumerator *j* and Supervisor *k*.

4. Results

We first obtained estimates for the variance components of the Enumerator and Supervisor random effects, without taking into account the fixed effects of the three hierarchical levels. Table 1 presents the Intraclass Correlation Coefficients (ρ) for the random effects, as well as the sum corresponding to the coefficients of both agents for the unconditional model: $Logit\left(\pi_{ijk}\right) = \beta_{0jk}$.

Table 1 – Estimated Intraclass Correlation Coefficients (ρ)

States	Random Effects		
	Supervisor (ρ_{u_0})	Enumerator (ρ_{v_0})	Sum(ρ_{u_0}+ρ_{v_0})
RJ	0,111	0,120	0,231
SC	0,049	0,118	0,168
MT	0,092	0,049	0,142
AL	0,043	0,029	0,071
AM	0,092	0,129	0,221

Sources: 2010 Census - Microdata, Fieldwork Supervision Database and Human Resources Database of the Collection Staff

The values of ρ coefficients of the five states show no regular pattern regarding a larger part of the variation of the divergences being accredited to the Enumerator or the Supervisor, indicating a notable difference in the variability composition of divergences for the selected states. The sum of the parts of the Enumerator and the Supervisor are very different as well, varying between 7% in Alagoas and 23% in Rio de Janeiro. In addition to the sum of the two components varying among the states, the parts of explained variation differ among the agents for each state, proving evidence that was important to take a regional approach for the analysis. Despite these differences, it is noteworthy to observe that the sums of ρ coefficients do not account for the greater part of the total variation for any of the states. Rio de Janeiro presented the highest percentage of variation associated with the data collection hierarchical structure (23%), i.e., 77% of the unexplained variation among the divergences may be associated with other factors, especially the Respondent, which is assumed to be one of the main sources of uncontrolled variation.

Table 2 displays the odds ratios in favor of the occurrence of divergences related to the fixed effects considered significant (α=5%) in the conditional model (Equation 1) after the inclusion of variables (characteristics) associated to the three hierarchical levels. The results show similar trends for the estimated odds ratios for all states, providing evidence that the significant effects lead to common movements regarding the increase or decrease of the odds ratios, an important account to confirm the coherence of the results when comparing different states. Although the trend in the significant common effects seems to be coherent, the number of significant non-common variables among the states is evident.

Table 2 – Odds Ratios for fixed effects by States

Effects	Odds Ratios				
	RJ	**SC**	**MT**	**AL**	**AM**
1st Level (Respondent and Household)					
Age	1.007	1.011	1.018	1.017	1.006
Sex					
Male / Female	1.258	1.292	1.188	1.258	1.265
Know to read and write					
Yes/no	0.225	0.146	0.289	0.425	0.194
Color or race					
White / Non white	-	0.842	-	-	-
Reported age					
Date of birth / Stated age	0.616	0.505	0.303	-	-
Relation with household reference person					
Reference person or spouse/other	0.887	0.737	-	-	-
log (per capita household income)	0.874	0.841	0.898	0.898	-
Number of bathrooms	0.872	0.857	0.895	0.827	0.888
Type of questionnaire					
Short / Long form	-	-	1.272	-	-
Reference person in household					
Only one / More than one	-	-	1.175	-	-
Not reported/ More than one*	-	-	1.094	-	-
Electricity					
From provider/Other form or non-existent	-	-	-	-	0.633
Sewage disposal					
Piped sewage system / Other form	-	1.159	-	-	-
Type of housing unit					
One-person or nuclear family / Other form	0.839	0.757	0.796	-	-
Time of interview					
Before 6 pm /After 6 pm	0.896	-	-	-	-
2nd Level (Enumerator)					
Educational Attainment of Enumerator					
Elementary or high school/ Some college or college degree	-	-	-	-	1.221
3rd Level (Supervisor)					
Educational Attainment of Supervisor					
High school / Some college or college degree	-	-	1.231	-	-
Age of Supervisor	-	-	-	1.014	-

Sources: 2010 Census - Microdata, Fieldwork Supervision Database and Human Resources Database of the Collection Staff
(*) the *not reported* category, although not significant in the model; was keep as factor level since it could not be aggregated with any other of the two categories.

The variable indicating if the respondent lives in a household with electricity, for instance, is significant only in the state of Amazonas. On the other hand, the logarithm of per capita household income is a key factor for the other states. However, both income and availability of electricity represent a similar socio-economic feature of the individuals in different states and allow the model capture the effect of this dimension on the occurrence of divergences.

5. Conclusions

The results suggest that respondents who are male, illiterate, older and living in households with indicators reflecting poor life conditions present higher odds in favor of the occurrence of divergences on data collected by Enumerator and Supervisor. In the state of Mato Grosso, for instance, a one-year increase in the age of the respondent is associated with an estimated rise of almost 2% in the odds of a divergence. In Santa Catarina, if the respondent is literate, there is an 84% decrease in the estimated odds of a divergence and in Alagoas, an extra bathroom in the household leads, on average, to a reduction of roughly 17% in the estimated odds. There is also evidence that socio-demographic profiles of Enumerators and Supervisors (the field staff) present a distinct association with the occurrence of divergences in different states. In the state of Amazonas, Enumerators with a lower educational level are associated with a 22% increase in the odds in favor of divergences, an effect that was already anticipated. However, only in the State of Mato Grosso, the model results indicated a 23% increase in the odds of divergence when Supervisors had a lower education level. Older Supervisors are associated with an increase in the odds of divergences for the state of Alagoas. The results for the 3rd level indicate the need to enhance the methodology in order to investigate possible interactions among the characteristics of Enumerators and Supervisors.

Regarding the fixed effects of the three levels, the variables of the respondent prevail notably as predictors in comparison with the small number of significant variables related to the characteristics of Enumerators and Supervisors. Despite the differences between the states, we consider having achieved the proposed goal in this study. In fact, the purpose of this initial investigation was to identify possible predictors for information divergence, without an expectation of creating a single model for all the states or for the country as a whole. This work reveals the association between each hierarchical level and the divergences in different regions of the country. On the other hand, it discovered similarities and trends that lead to a better understanding of the phenomenon, which had never been investigated taking into account the hierarchical structure of the Brazilian Census paradata.

6. References

Biemer, P. P., & Liyeberg, L. E. (2003). Introduction to Survey Quality. New York: John Wiley & Sons.

Couper, M. (1998). Measuring survey quality in a CASIC environment. In: Proceedings of the Section on Survey Research Methods of the American Statistical Association. Available at: <http://www.amstat.org/sections/srms/proceedings/papers/1998_006.pdf>

Duarte, L. T. (2014). Análise dos Paradados do Censo Demográfico 2010 : investigações de fatores associados a erros não amostrais detectados na coleta das informações. 247 p. Master's thesis, Escola Nacional de Ciências Estatísticas, Rio de Janeiro.

Groves, R. M. (1989). Survey errors and survey costs. New York: John Wiley & Sons.

Holt, D. (2005). Methodological Issues in the Development and Use of Statistical Indicators for International Comparisons**.** Business Survey Methods Division, Statistics Canada, Survey Methodology.

Hox, J. (2010). Multilevel Analysis: Techniques and Applications, Second Edition (Quantitative Methodology Series). New York: Routledge.

IBGE (2008). Censos 2007 - Inovações e impactos nos sistemas de informações estatísticas e geográficas do Brasil. Rio de Janeiro.

_____ (2012). Censo Demográfico 2010. Resultados gerais da amostra. Rio de Janeiro.

Nicolaas, G. (2011). Survey Paradata: A Review. National Centre for Social Research (NatCen), January 2011. Accessed 7 February 2013. Available at: <http://eprints.ncrm.ac.uk/1719/1/Nicolaas_review_paper_jan11.pdf>

Raudenbush, S. W., & Bryck, A. S. (2002). Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd ed. Thousand Oaks: Sage Publications.

Weisberg, H. F. (2005). The total survey error aproach. Chicago: Chicago Press.