2010 Brazilian Census Paradata: Analysis of the field work supervision process

Luciano Tavares Duarte
Denise Britz do Nascimento Silva
José André de Moura Brito

National School of Statistical Sciences
Brazilian Institute of Geography and Statistics

2010 Brazilian Population and Housing Census

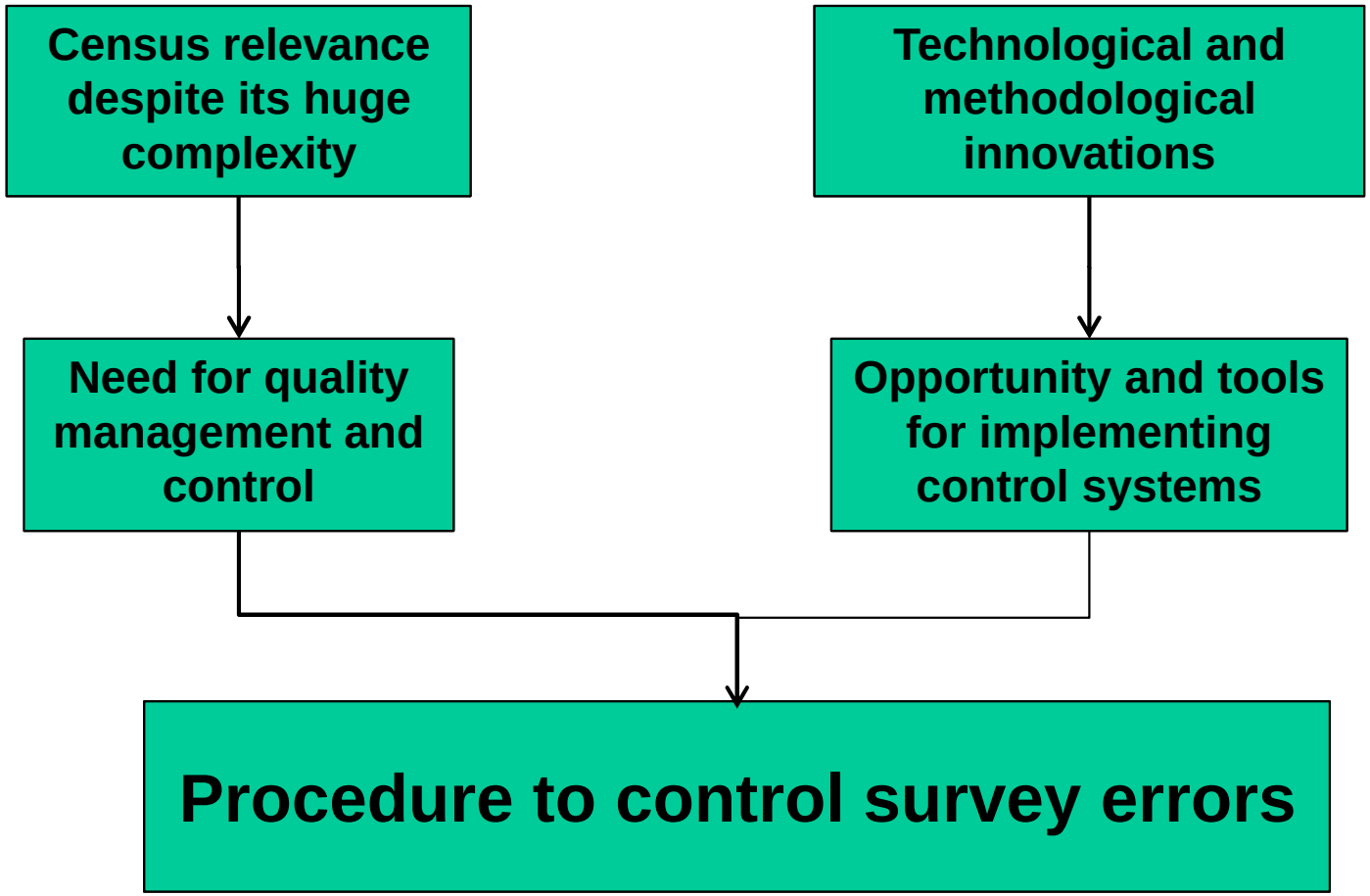
- The Population and Housing Census is certainly the most complex and massive operation conducted by a National Statistical Office in any country.

- The 2010 Brazilian Census collected basic population and housing characteristcs in the entire country for a single reference date: the night of July, 31st 2010.

- Census data allow analysis in terms of statistics on persons and households for a wide variety of geographical units ranging from the country as a whole to municipalities and neighbourhoods.

- The 2010 Brazilian Census incorporated a series of methodological and technological innovations, being the first fully digital national census of almost 200 million people.

Motivation


Census relevance despite its huge complexity
Technological and methodological innovations
Need for quality management and control
Opportunity and tools for implementing control systems
Procedure to control survey errors

MOTIVATION

There area several sources of non-sampling errors that can affect quality of census data.

OBJECTIVE

Analyse Census paradata to identify potential determinants of non-sampling errors associated to the data collection process of the 2010 Brazilian Census.

HOW?

Using data obtained from the field work monitoring system that provided information about divergences observed between data collected by enumerators and supervisors who carried out follow-up interviews in those households selected on the supervision/monitoring plan.

Databases

Paradata

- *Supervision*/monitoring system

  Divergences between data colected by enumerators and supervisors

- *Field Staff human resources data*

  Socio-demographic characteristics of enumerators and supervisors

- *Operational data*

  Time of interview, duration of field work, etc.

Census Data

- *Census Microdata*

  Socio-economic characteristics of respondents

Matching procedure of Census data and paradata

Scope of the Study

- Respondents reporting their own information
- Enumerators who had performed at least 5 completed interviews
- Supervisors who were responsible for managing 5 to 20 interviewers
- Data from 5 Brazilian States (one in each of the 5 country regions)
  - Amazonas (North)
  - Alagoas (Northeast)
  - Rio de Janeiro (Southeast)
  - Santa Catarina (South)
  - Mato Grosso (Central West)


Brazil
Roraima
Amapá
Amazonas
Pará
Maranhão
Ceará
Rio Grande do Norte
Paraíba
Pernambuco
Alagoas
Sergipe
Piauí
Acre
Rondônia
Mato Grosso
Tocantins
Bahia
Distrito Federal
Goiás
Regions
North
Northeast
Central-West
Southeast
South
Mato Grosso do Sul
Minas Gerais
Espírito Santo
São Paulo
Rio de Janeiro
Paraná
Santa Catarina
Rio Grande do Sul

Analysis of the divergence between data collected by Census enumerators and supervisors

Variable of Interest: Occurency of Divergence

$$Y = \begin{cases} 1 & if\ there\ is\ divergence \\ 0 & otherwise \end{cases}$$

$$Y \sim Bernoulli\ (p)$$

Y=1 **if there is divergence between information collected by enumerator and supervisor on at least one of the main socio-demographic questions:**

- ✓ **Age**
- ✓ **Sex**
- ✓ **Know how to read and write (literacy)**

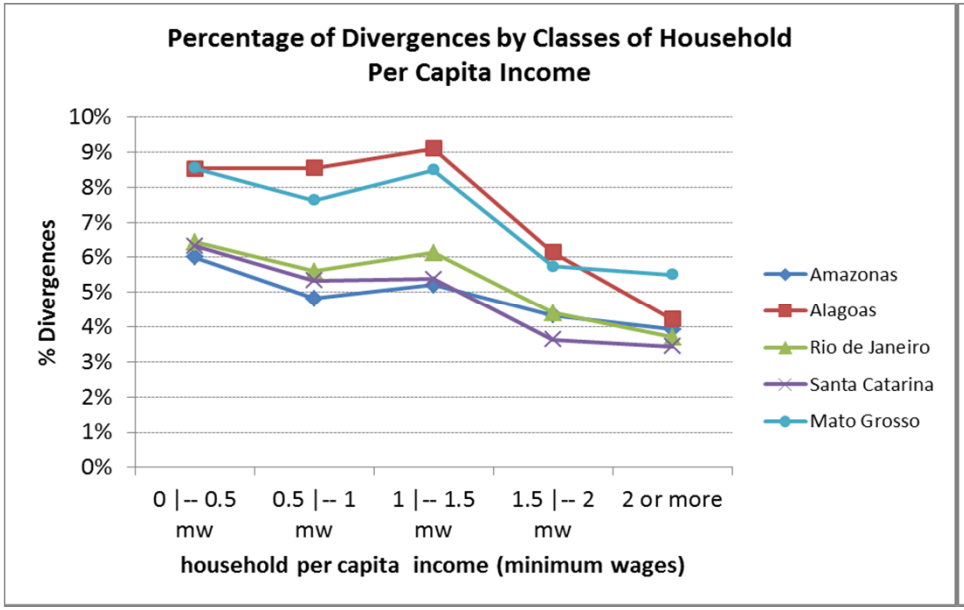
Data Collected by Supervisors in Follow-up Interviews for Households Selected by Census Supervision System

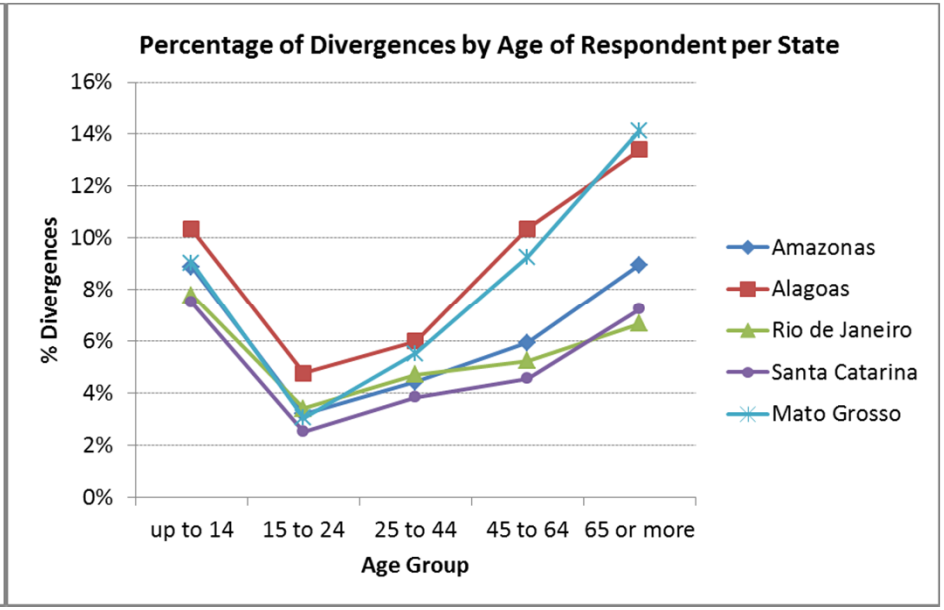
Percentage of Divergences on Main Questions per State

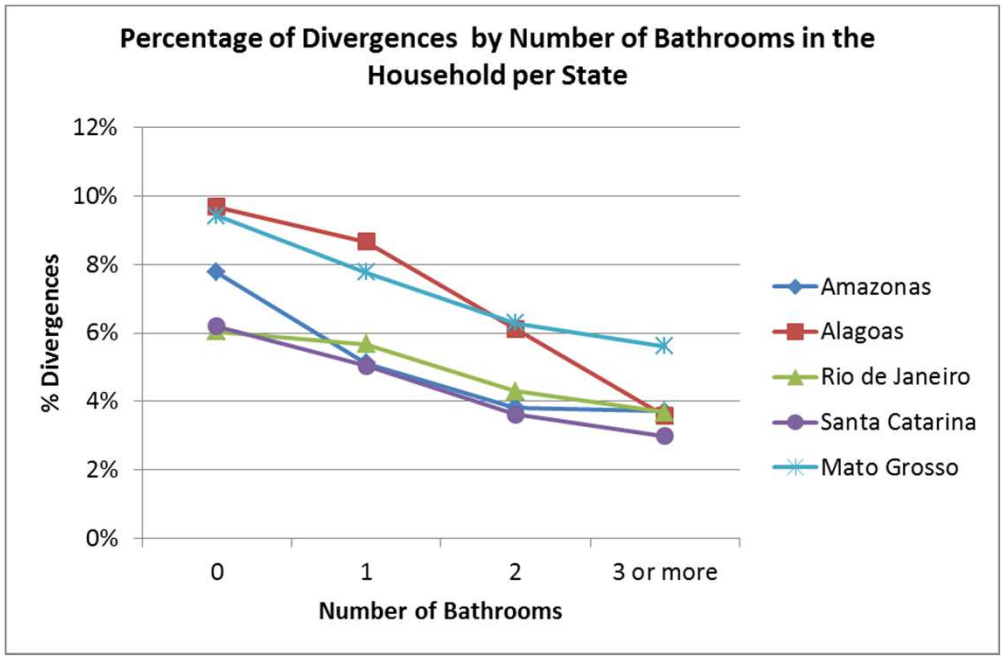
2010 Brazilian Census Supervision System

States	Follow-up Interviews	main questions checked on follow-up interviews - % of Divergence			
		Know how to read and write	Sex	Age	Divergence in at least one question
Alagoas	16,730	4.22	0.50	3.90	8.16
Mato Grosso	25,836	2.81	0.70	4.14	7.24
Rio de Janeiro	106,347	1.37	0.72	3.56	5.39
Amazonas	21,281	1.88	0.62	3.12	5.34
Santa Catarina	46,512	1.34	0.59	3.01	4.75
Brazil	1,237,827	2.44	0.61	3.46	6.20

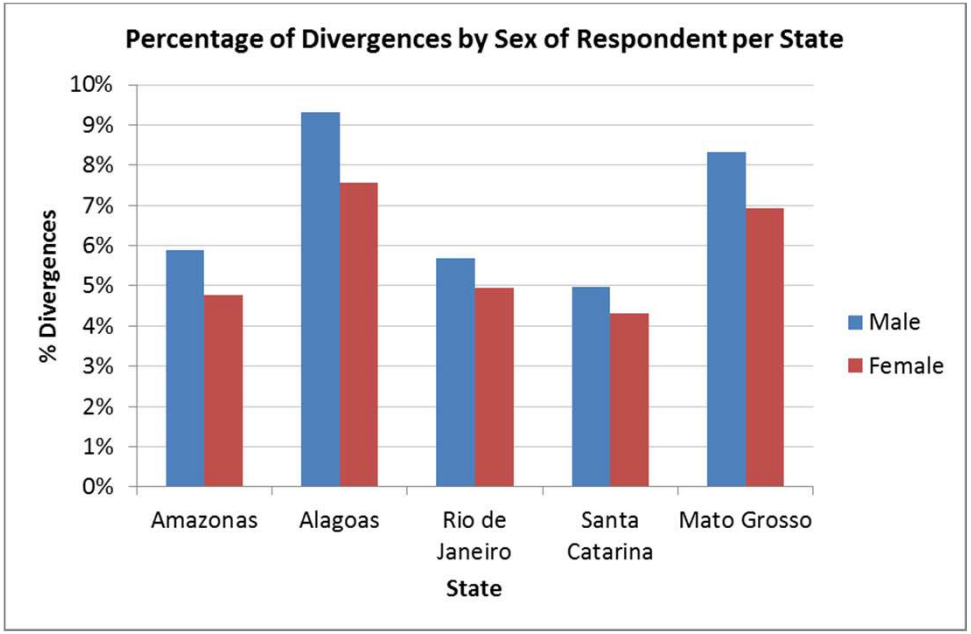
Empirical Evidence – Percentage of Divergence According to Respondent and Household Characteristics

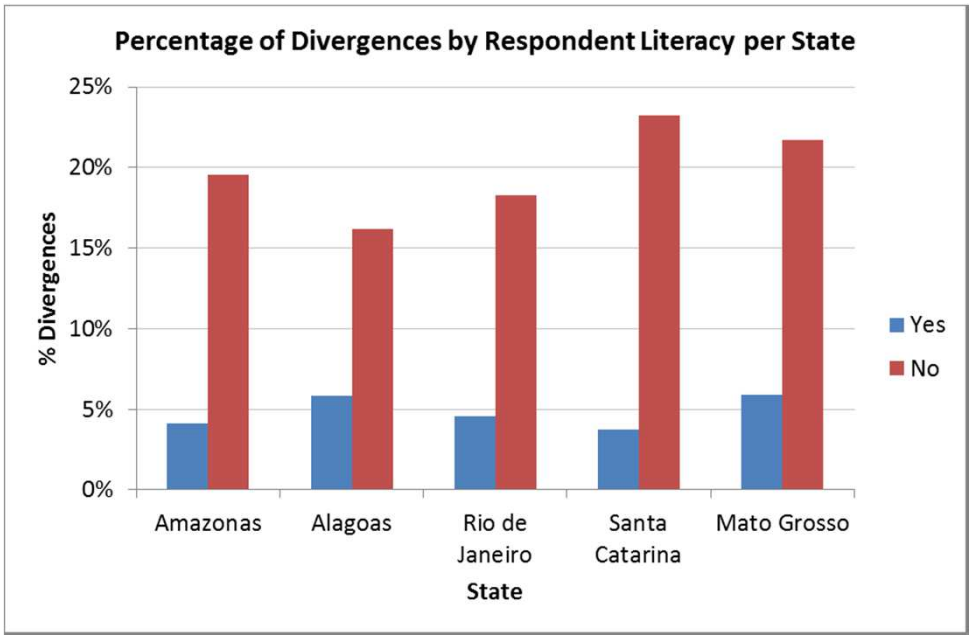




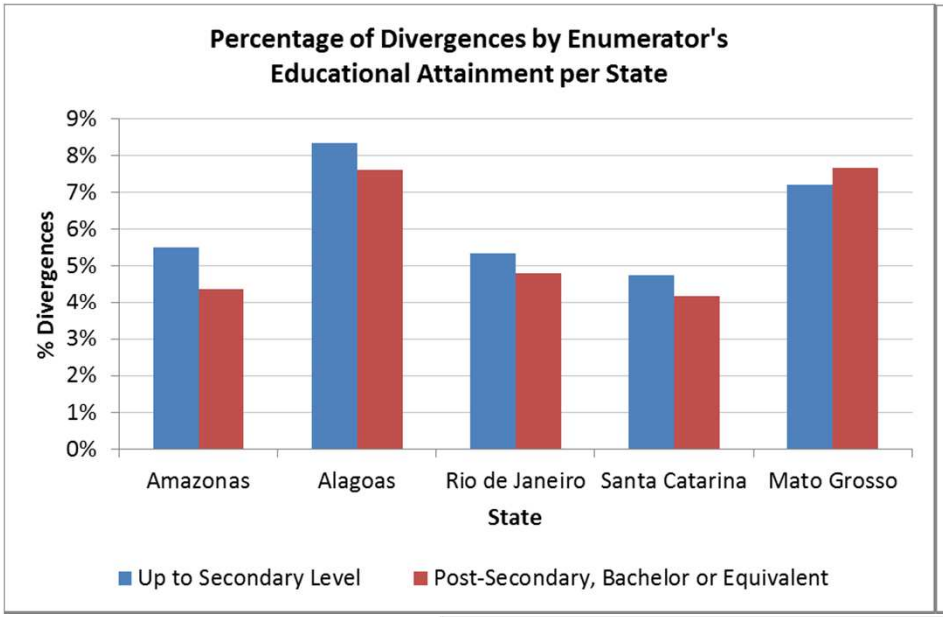


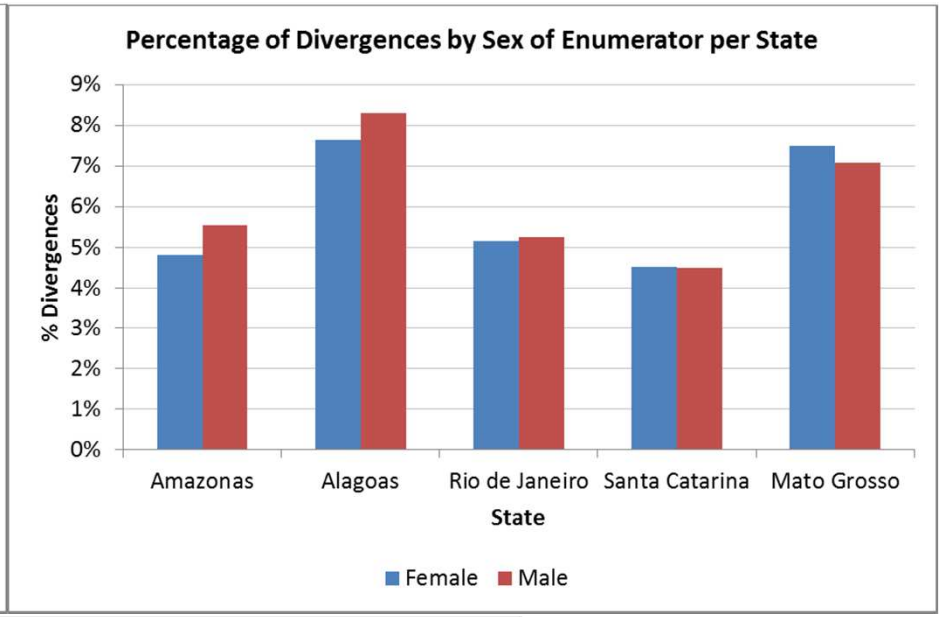
Empirical Evidence – Percentage of Divergence According to Respondent and Household Characteristics

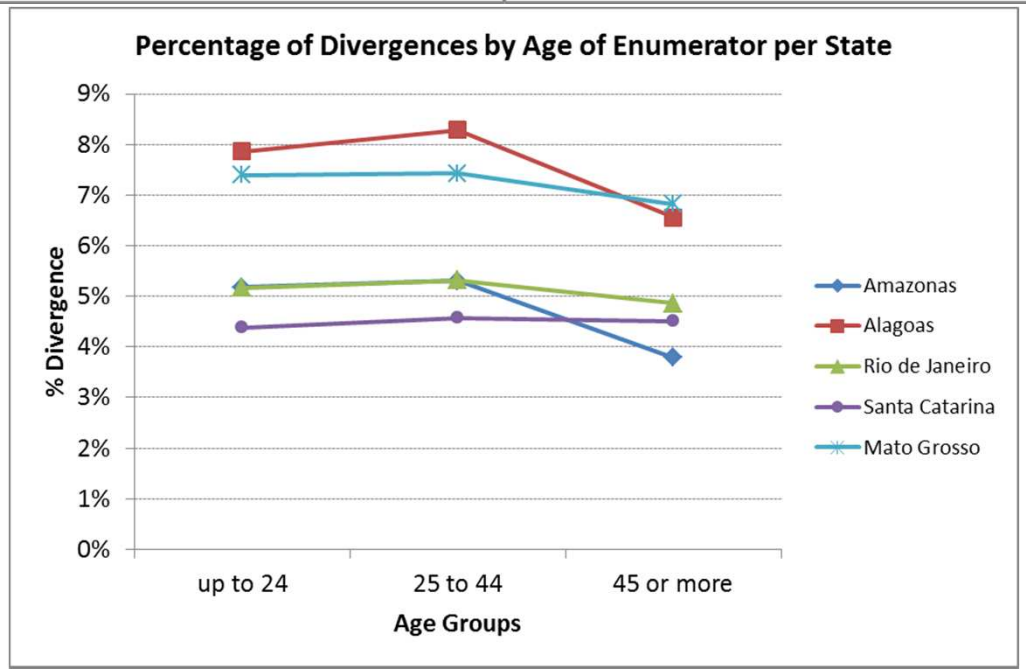




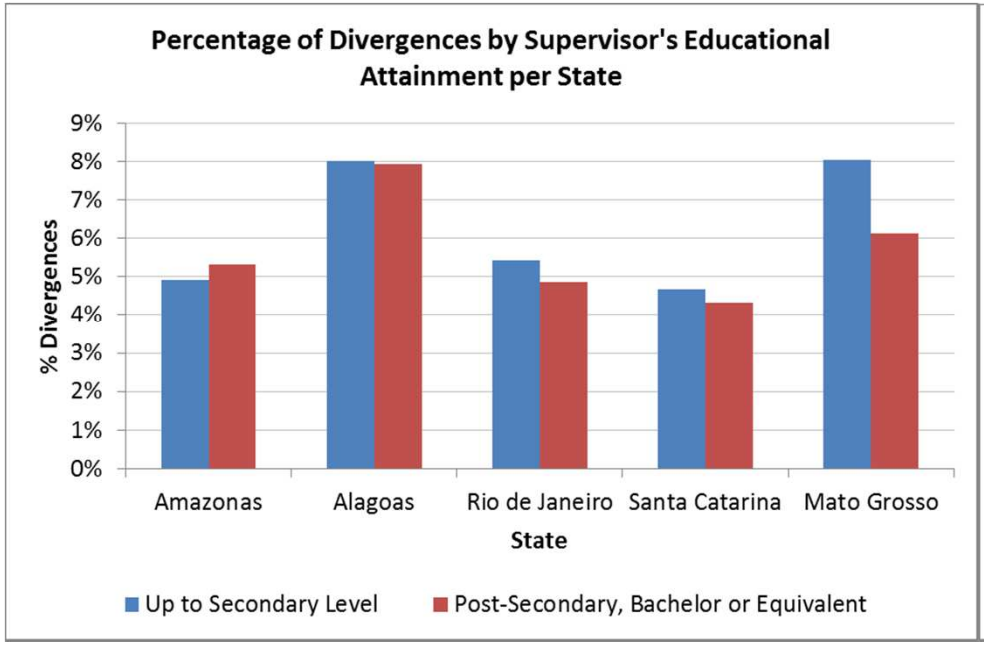
Empirical Evidence – Percentage of Divergence According to the Enumerator Socio-Demographic Characteristics

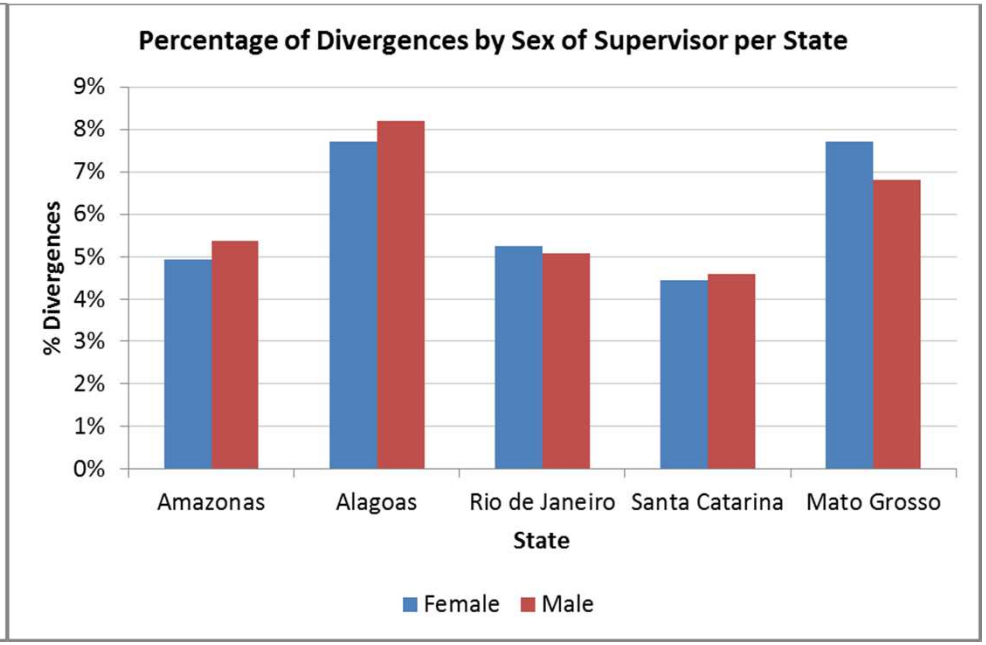


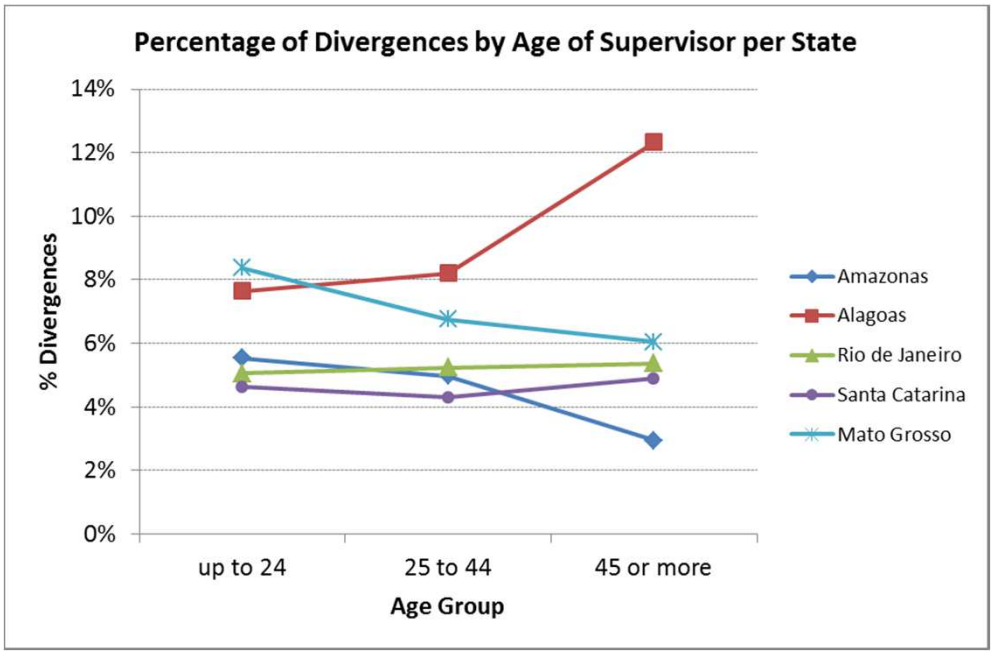




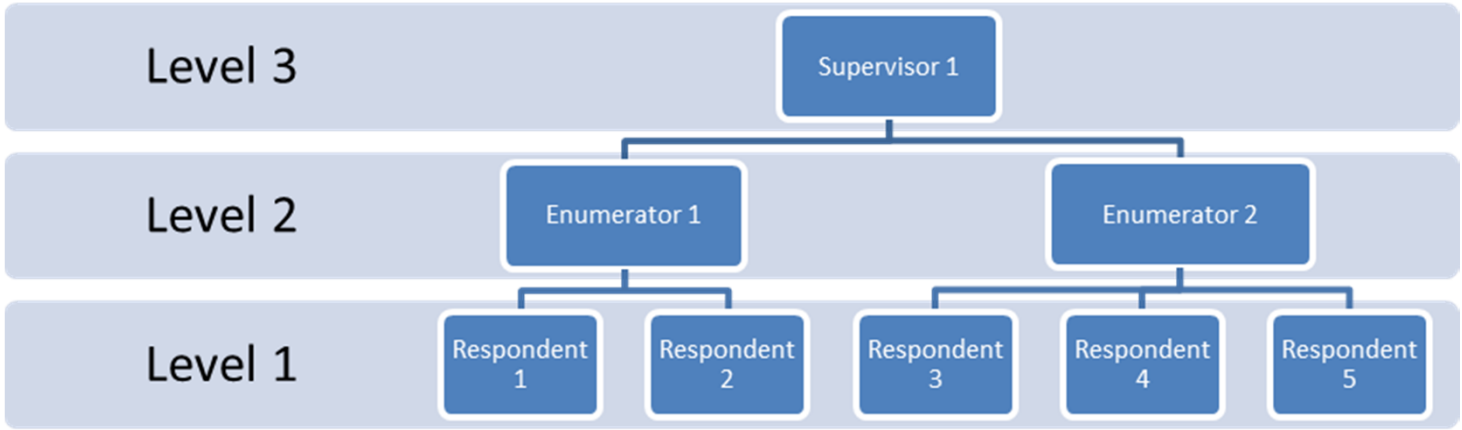
Empirical Evidence – Percentage of Divergence According to the Supervisor Socio-Demographic Characteristics







Hierarchical Data Structure



Hierarchical Models for Divergences

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

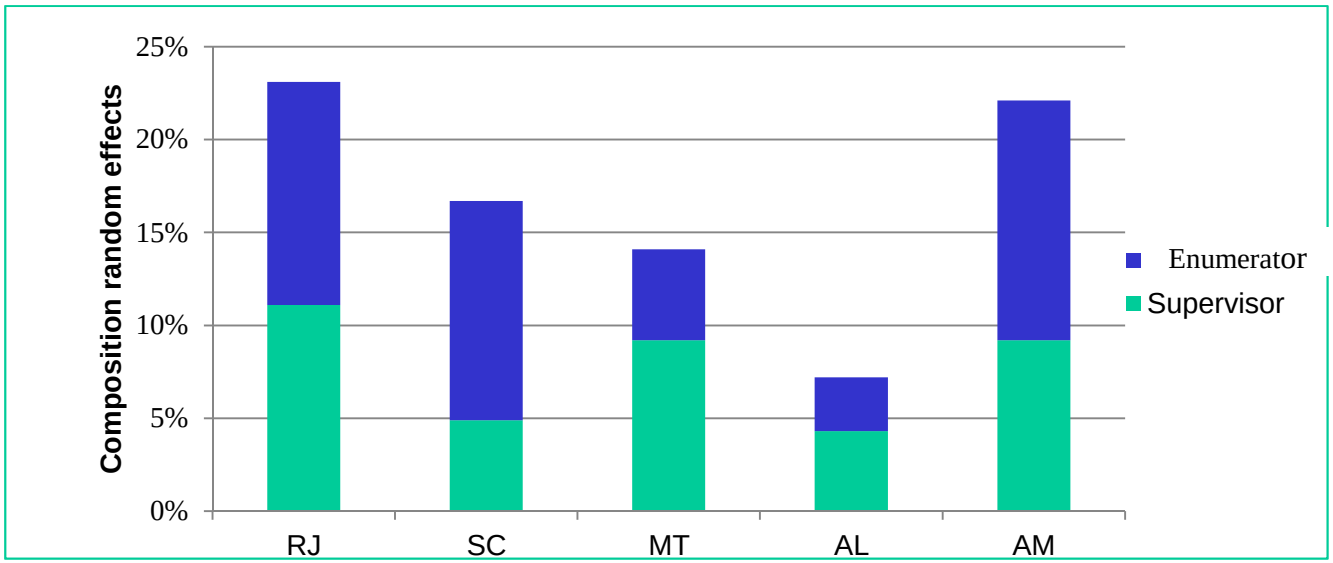
$$\beta_{0jk} = \beta_0 + u_{0k} + v_{0jk} \quad \begin{cases} \beta_0 - \text{ intercept} \\ u_{0k} \sim N\left(0, \sigma_{u_0}^2\right) - \text{Variance Component due to Supervisors} \\ v_{0jk} \sim N\left(0, \sigma_{v_0}^2\right) - \text{Variance Component due to Enumerators} \end{cases}$$

π_{ijk} is the probability of divergence between information collected by enumerator and supervisor on at least one of the main socio-demographic questions: *sex*, *age* and *literacy* for Respondent *i*, Enumerator *j* and Supervisor *k*.

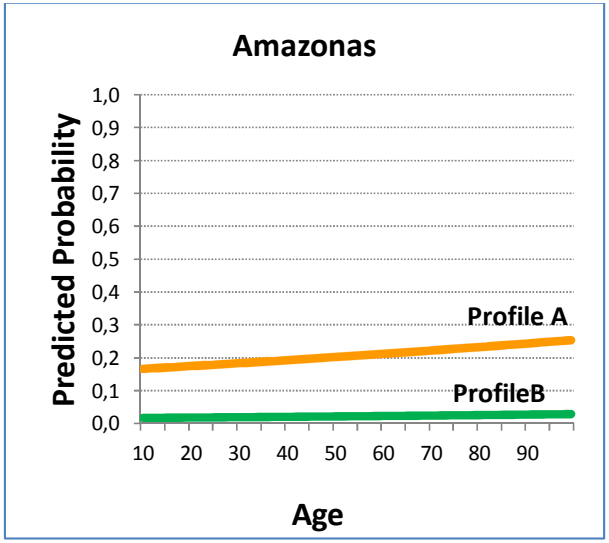
Effects	ODDS RATIOS				
	RJ	SC	MT	AL	AM
Level 1 – Respondent and Corresponding Household					
Age	1.007	1.011	1.018	1.017	1.006
Sex					
Male/ Female	1.258	1.292	1.188	1.258	1.265
Know to read and write					
Yes/ No	0.225	0.146	0.289	0.425	0.194
Race					
White / Non White	-	0,842	-	-	-
Form of reporting age					
Date of Birth / Declared age	0.616	0.505	0.303	-	-
Relation with household reference person					
Reference person or spouse /Other	0.887	0.737	-	-	-
$\log_{10}$(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					
Direct from provider/ Other form or do not have	-	-	-	-	0.633
Sewage Disposal					
Piped sewer system/ Other form	-	1.159	-	-	-
Type of family					
One person or nuclear family /Other type	0.839	0.757	0.796	-	-
Time of Interview					
6pm or before / After 6pm	0.896	-	-	-	-
Leve 2 – Enumerator					
Educational Attainment					
Up to Secondary / Bachelor	-	-	-	-	1.221
Level 3 – Supervisor					
Educational Attainment					
Up to Secondary /Bachelor	-	-	1.231	-	-
Age	-	-	-	1.014	-

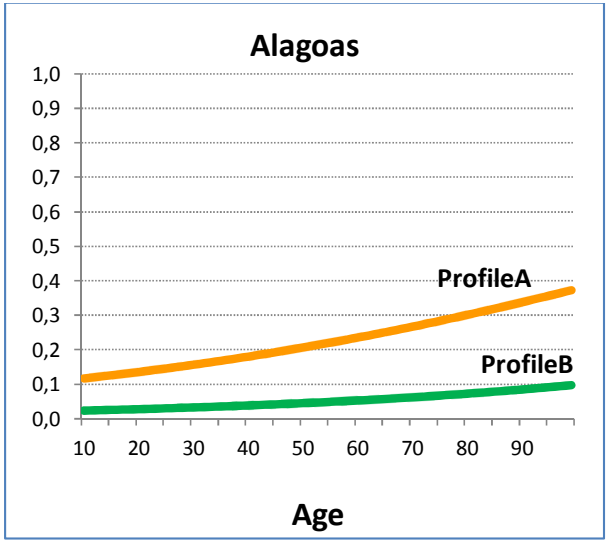
Intraclass Correlation Coefficient

State	Random Effect		
	Supervisor (ρ_{u_0})	Enumerator (ρ_{v_0})	Total(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

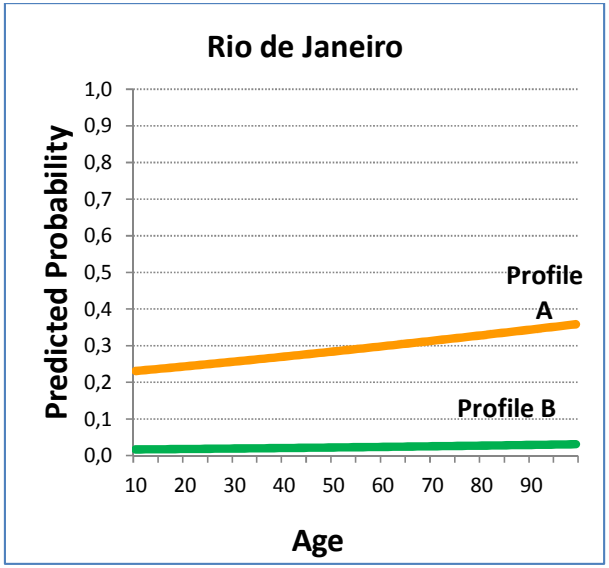


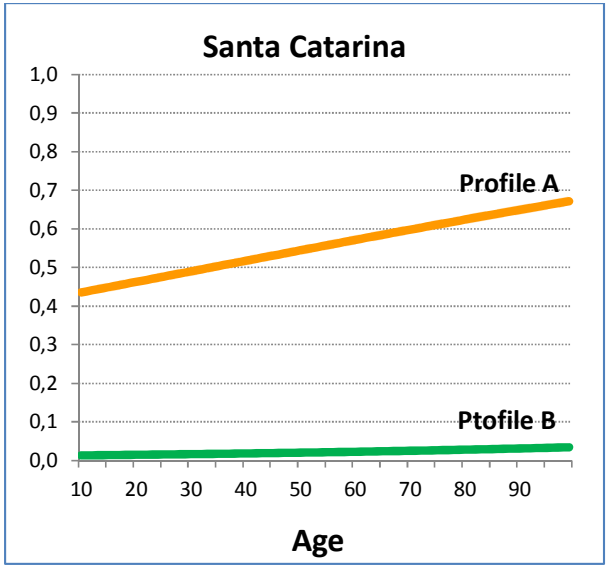
Predicted Probabilities According to Respondent Age



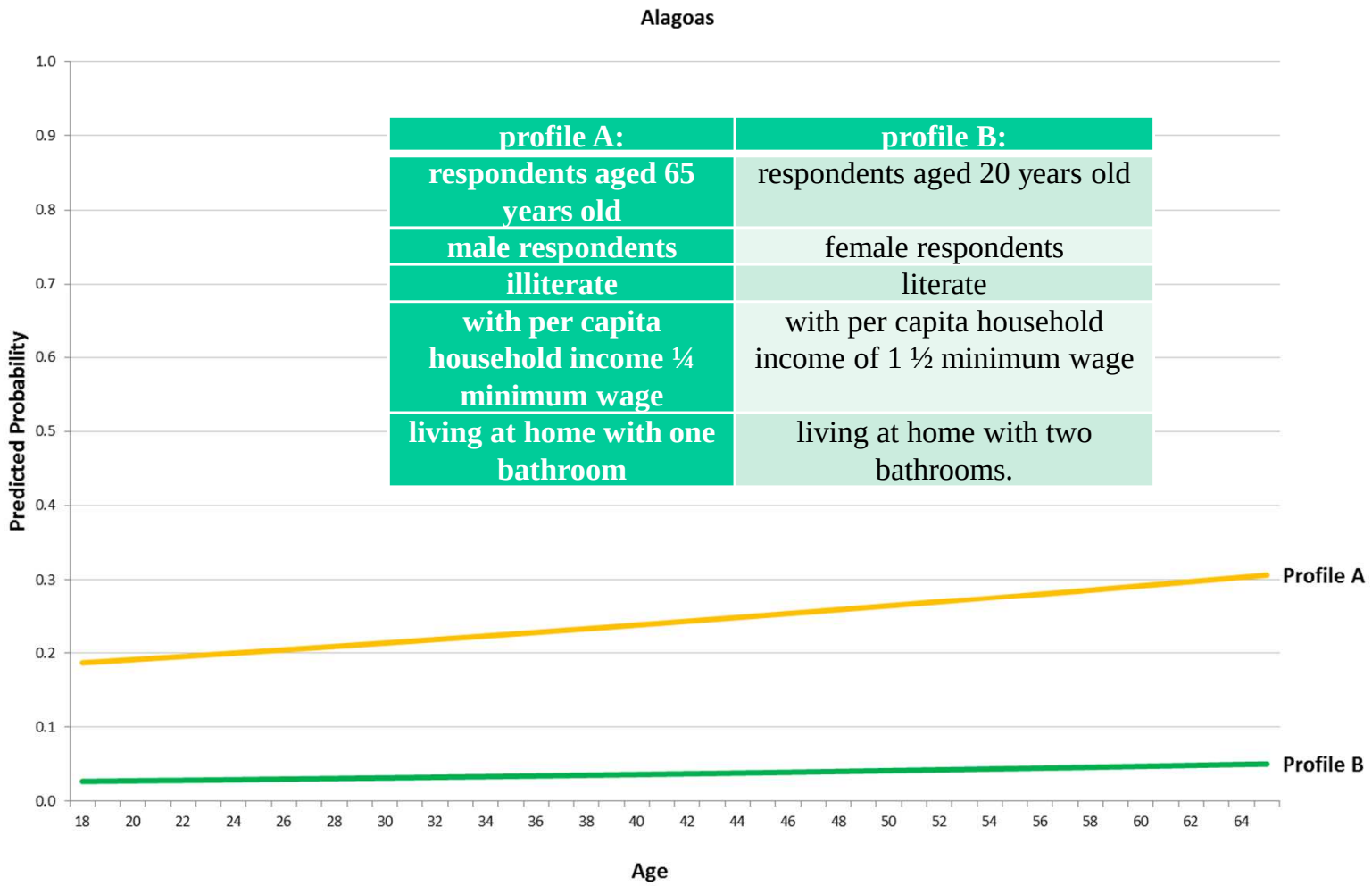


profile A:	profile B:
male respondents	female respondents
illiterate	literate
with per capita household income ¼ minimum wage	with per capita household income of 1 ½ minimum wage
living at home with one bathroom	living at home with two bathrooms.





Predicted Probabilities According to Supervisor Age



profile A:	profile B:
respondents aged 65 years old	respondents aged 20 years old
male respondents	female respondents
illiterate	literate
with per capita household income ¼ minimum wage	with per capita household income of 1 ½ minimum wage
living at home with one bathroom	living at home with two bathrooms.

Conclusions

- It is noticeable that model incorporates many more explanatory variables (fixed effects) associated with respondent characteristics than those related to the enumerators or supervisors.
- Modelling results indicate that odds in favour of divergence increase when respondents are older men living in poorer households.
- Socio-Demographic characteristics of enumerators and supervisors did not show a consistent effect on the probability of divergence for all states.
- This works provides new evidence on how the hierarchical management of the field work process is associated with the probability of divergence in different country regions.
- It constitutes the first initiative of combined use of paradata and Census data to contribute to improving future Census and surveys in Brazil.

Main References

AGRESTI, A. **An Introduction to Categorical Data Analysis**. NY, John Wiley & Sons: 1996.

BARTHOLOMEW, D. J. STEELE, F. GALBRAITH, J. MOUSTAKI, I. **Analysis of Multivariate Social Science Data**. 2ª Edição. ed. Boca Raton, FL, Chapman & HAll/CRC: 2008.

BIANCHINI, Z. M. **A Qualidade na Produção de Estatísticas no IBGE.** Textos para discussão - Diretoria de Pesquisas - número 14 – IBGE: 2004.

BIANCHINI, Z. M. & ALBIERI, S. – **Qualidade na produção de informações: Desenvolvimento e revisão de metodologias no IBGE**. VIII Reunión sobre Estadística Pública - Modelos para ***el*** Desarrollo de los Sistemas **Nacionales *de*** Estadística en Latinoiamerica y el Caribe - IASI-INEGI -Aguascaleientes – **México: 21-22** de maio de 2008.

BIEMER, P. P. e LIYEBERG, L. E. Quality assurance and quality control in Surveys. In: _____ **International Handbook of Survey Methodology**. New York, Psycology Press/EAM: 2009.

BIEMER, P.P. & LIYEBERG, L.E. **Introduction to Survey Quality**. New York, John Wiley & Sons: 2003.

COUPER, M. P. **Measuring survey quality in a CASIC environment**. In: Proceedings of the Section on Survey Research Methods of the American Statistical Association: 1998. Disponível em: <http://www.amstat.org/sections/srms/proceedings/papers/1998_006.pdf> Acesso em: 07/2013.

COUPER, M. P. KREUTER, F. **Using paradata to explore item level response times in surveys.** J. R. Statist. Soc. A 176, Part 1, pp. 271–286: 2013.

GOLDSTEIN, H. **Multilevel Statistical Models**. 4ª. ed. [S.l.]: John Wiley & Sons, 2011.

GROVES, R. M. **Survey errors and survey costs.** NY, John Wiley & Sons: 1989 .

HOX, J. J. **Multilevel Analysis: Techiniques and Aplications**. NY, 2ª Edição. Routledge: 2010.

NICOLAAS, G. **Survey Paradata: A review.** National Centre for Social Research (NatCen/ESRC), January 2011. Disponível em <http://eprints.ncrm.ac.uk/1719/1/Nicolaas_review_paper_jan11.pdf> Acesso em: 02/2013.

RAUNDENBUSH, S. W. BRYCK, A. S. **Hierarchical Linear Models**. 2ª Edição. ed. [S.l.]: SAGE Publications, 2002.

SNIJDERS. T, BOSKER. R, **Multilevel Analysis: An introduction to basic and advanced multilevel modeling**, SAGE, Londres: 1999.

STEELE, F. e DURRANT, G.B. **Alternative Approaches to Multilevel Modelling of Survey Non-Contact and Refusal**. International Statistical Review: 2011 , 79, 1, pg. 70–91.

STERN, M.J. et al. **Toward Understanding Response Sequence in Check-All-That-Apply Web Survey Questions: A Research Note with Results from Client-Side Paradata and Implications for Smartphone**. Survey Practice: 2012. vol. 5 n.4, ISSN: 2168-0094.

WEISBERG, H. F. **The total survey error aproach**. Chicago: Chicago Press: 2005.

WEST, B.T. **An examination of the quality and utility of interviewer observations in the National Survey of Family Growth**. J. R. Statist. Soc.A (2013) 176, Part1, pp. 211–225.

Thank you!!!

Luciano Tavares Duarte
luciano.duarte@ibge.gov.br

Denise Britz do Nascimento Silva
denise.silva@ibge.gov.br
José André de Moura Brito
jose.m.brito@ibge.gov.br