DOI http://dx.doi.org/10.20947/S0102-30982016c0011

Análise de paradados do Censo Demográfico 2010: uma investigação de fatores associados a erros não amostrais do levantamento de dados

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

A relevância de um Censo Demográfico para o sistema de estatísticas públicas de uma nação é indiscutível do ponto de vista de sua abrangência temática e territorial. Em contrapartida, sua complexidade e dimensão levam a desafios na garantia da qualidade de seus resultados. O presente artigo tem por objetivo apresentar os possíveis fatores associados a erros não amostrais detectados na coleta das informações, mediante a análise de paradados e dos microdados do Censo Demográfico brasileiro de 2010. Os dados utilizados provêm das informações sobre a operação de coleta e administração da pesquisa oriundas, respectivamente, do sistema de gerenciamento de recursos humanos do pessoal de coleta e do sistema de supervisão da operação de coleta, ou seja, os paradados. Também se utilizam os microdados do universo do Censo Demográfico. Neste estudo foram analisadas as divergências observadas entre as informações coletadas pelos recenseadores e aquelas obtidas por supervisores nas reentrevistas realizadas em procedimentos de supervisão do trabalho de campo. Para análise de divergências detectadas entre os dados coletados por recenseadores e supervisores, foram empregados modelos hierárquicos generalizados. Os resultados mostram que existem diferenciais nas divergências associados à estrutura de coleta dos dados e às características dos recenseadores, supervisores e informantes, além de revelarem diferenças regionais. Fica evidente, sobretudo, uma forte influência das características do informante nas chances de ocorrência das divergências, em detrimento das características dos supervisores e recenseadores. Os resultados da modelagem estatística sugerem que as entrevistas realizadas com informantes do sexo masculino, analfabetos ou com baixa escolaridade, mais velhos e que vivem em domicílios com indicadores que refletem condições de vida menos satisfatórias apresentam aumento nas chances em favor da ocorrência de divergências entre respostas coletadas por recenseador e supervisor.

Palavras-chave: Censo Demográfico brasileiro. Paradados. Modelos hierárquicos. Erros não amostrais.

* Instituto Brasileiro de Geografia e Estatística (IBGE), Rio de Janeiro-RJ, Brasil (luciano.duarte@ibge.gov.br).
** Escola Nacional de Ciências Estatísticas (Ence), Instituto Brasileiro de Geografia e Estatística (IBGE), Rio de Janeiro-RJ, Brasil (denise.silva@ibge.gov.br).
*** Escola Nacional de Ciências Estatísticas (Ence), Instituto Brasileiro de Geografia e Estatística (IBGE), Rio de Janeiro-RJ, Brasil (jose.m.brito@ibge.gov.br).

Introdução

Para uma nação, é inquestionável a relevância das estatísticas públicas, sob vários aspectos. São a base para o desenvolvimento e monitoramento de políticas públicas, alocação de recursos nas mais diversas áreas, bem como subsidiam o setor privado na condução e tendências do mercado. Assim, a cada dia mais, as estatísticas públicas vêm se tornando uma ferramenta nas mãos dos cidadãos para o julgamento das ações governamentais, além de desempenharem papel fundamental na comparação do desenvolvimento dos países por meio de indicadores econômicos e sociais no cenário internacional (HOLT, 2005).

Nesse sentido, o Censo Demográfico no Brasil caracteriza-se como uma das mais importantes fontes de informação sociodemográfica do rol das estatísticas públicas produzidas no país, tanto por sua abrangência temática, quanto por sua cobertura em âmbito nacional. Além disso, tal instrumento tem papel fundamental no que concerne à produção de informações relevantes para os diversos níveis de desagregação territorial no país (IBGE, 2012).

Buscando sempre o aprimoramento de suas práticas, em consonância com os recentes avanços tecnológicos, o Instituto Brasileiro de Geografia e Estatística (IBGE) tem desenvolvido e adotado métodos inovadores nas últimas operações censitárias. Uma massiva inserção de novas tecnologias de informação e comunicação (TICs) no processo de coleta de dados gerou ganhos significativos no que diz respeito às possibilidades de controle da operação (IBGE, 2008). Como consequência, houve um substancial aumento na produção de informações gerenciais e operacionais do processo de pesquisa.

É nesse contexto que se insere a definição do termo paradados (*paradata*), utilizado pela primeira vez no âmbito de pesquisas quantitativas por Couper (1998). A expressão refere-se ao uso de dados relativos à operação de coleta e administração da pesquisa para avaliação e melhoria da qualidade do processo de pesquisa.

A importância do desenvolvimento do uso de paradados para os Institutos Nacionais de Estatística se reflete na vasta bibliografia já disponível sobre o tema e também no empenho de vários pesquisadores internacionais em focar sua pesquisa acadêmica nesta área de conhecimento. Em 2013, foi publicado um volume do *Journal of the Royal Statistical Society (Series A)* dedicado ao assunto.[1] A série intitulada *The use of paradata in social survey research* consiste, basicamente, em uma coletânea de artigos dedicados a recentes métodos aplicados à análise de paradados, à discussão sobre qualidade de dados e à detecção de erros de medição usando novas fontes de paradados. É importante ressaltar, entretanto, a falta de publicações cujo tema seja o uso de paradados na operação censitária, provavelmente devido às questões de sigilo das informações.[2]

[1] Disponível em: ‹http://onlinelibrary.wiley.com/doi/10.1111/rssa.2012.176.issue-1/issuetoc› Acesso em: jul. 2013.

[2] No caso do IBGE, o acesso a dados restritos pode ser solicitado por meio de contato inicial com o Centro de Documentação e Disseminação de Informações do IBGE (CDDI), pelo *e-mail* ibge@ibge.gov.br, que enviará as normas de acesso. A solicitação será avaliada pelo Comitê de Avaliação a Acesso a Dados não Desidentificados (CAD).

No presente artigo, explora-se parte deste conjunto de informações a fim de identificar questões e fatores associados à melhoria dos procedimentos e protocolos do levantamento e à qualidade dos resultados. São analisados os paradados do Censo Demográfico de 2010 referentes ao sistema de supervisão do processo de coleta de dados, combinando as informações de recursos humanos e microdados do universo referentes aos informantes.[3]

É importante destacar que três agentes tiveram participação direta no preenchimento das informações dos questionários do Censo Demográfico de 2010: a pessoa que prestou as informações – o informante –, o recenseador e o supervisor. O principal objetivo deste artigo é identificar possíveis fatores associados às divergências encontradas entre as informações registradas nos questionários pelo recenseador e aquelas obtidas pelo supervisor nos procedimentos de controle de qualidade da operação de coleta do Censo 2010. A hipótese é de que os três agentes constituíram-se como fonte de variabilidade não desprezível das divergências encontradas.

A seguir, discorre-se sobre as possíveis fontes de erro em levantamentos estatísticos e descreve-se o aporte metodológico aplicado ao estudo das divergências. Posteriormente, são apresentados os resultados da aplicação dos métodos estatísticos ao fenômeno das divergências. Por fim, são tecidas as principais considerações com base nos resultados alcançados por meio das análises estatísticas dos dados.

Possíveis fontes de erro em levantamentos estatísticos

Groves (1989) define que o erro total associado a um processo de pesquisa é derivado de duas componentes. Uma refere-se ao erro sistemático ou vício, que pode levar a superestimar ou subestimar o verdadeiro valor de um parâmetro populacional de interesse. A outra componente, em caso de pesquisas amostrais, está associada à variabilidade ou à precisão das estimativas.

Já a abordagem de Biemer e Lyeberg (2003) separa os erros associados à pesquisa em duas componentes: os erros amostrais e os não amostrais. Como mencionado, a principal hipótese deste trabalho é de que há várias fontes de erro não amostral e fatores relacionados ao processo de coleta que se apresentam como potenciais causas de falha ou variação não controlada, podendo afetar a confiabilidade dos dados populacionais e domiciliares de uma operação censitária.

Os autores afirmam que em uma pesquisa, como em qualquer outro processo de produção em que há interferência humana, tal fator é considerado fonte de variação intrínseca que contribui com uma parcela não desprezível dentre as diversas causas de falha ou variação não controlada, caracterizando-se assim como uma das possíveis fontes de erro não amostral (BIEMER; LYEBERG, 2003).

[3] Este artigo tem como base a dissertação de mestrado de Duarte (2014), na qual o uso dos dados foi viabilizado dentro das instalações do IBGE e por meio de autorização expressa para sua utilização para fins acadêmicos, com base em termo de compromisso firmado com a instituição, garantindo manutenção do sigilo das informações.

As principais fontes de erros não amostrais em pesquisas, segundo Biemer e Lyeberg (2003), são:

- erro de especificação – gerado, por exemplo, por problemas na definição e operacionalização de conceitos;
- erro de cadastro – que pode afetar a cobertura;
- erro de medida – oriundo do processo de trabalho do entrevistador, da resposta do entrevistado e do instrumento de medida (o questionário);
- erro de não resposta – parcial ou total;
- erro de processamento – que pode ocorrer nas etapas de captura dos dados, crítica, codificação, etc.

Os sistemas de controle do Censo 2010 produziram informações gerenciais com a finalidade de avaliar o trabalho de toda a equipe de campo, na tentativa de mitigar, ainda em tempo de coleta, efeitos de possíveis falhas que impactam diretamente na qualidade dos dados obtidos, sobretudo no que se refere ao efeito da interferência humana no processo. Dentre estes instrumentos, destaca-se o sistema de supervisão, que fornece informações sobre divergências entre entrevistas realizadas por recenseadores e reentrevistas feitas por supervisores (IBGE, 2013).

Além dos dados gerenciais dos sistemas de controle, há outra rica fonte de informações administrativas e de recursos humanos que permite conhecer e associar o perfil sociodemográfico dos recenseadores e supervisores que trabalharam na coleta dos dados à ocorrência de divergências no censo brasileiro. Adicionalmente, de acordo com Weisberg (2005), identifica-se outra potencial fonte de variação inerente ao processo de coleta: a pessoa (ou as pessoas) que efetivamente forneceu informações durante a realização da entrevista, que denominamos de informante. No último censo, uma das novidades no questionário foi a identificação de quem prestou as informações sobre cada um dos moradores, o que possibilitou conhecer todas as características investigadas nos questionários a respeito de cada informante.

Outras bases de dados também se caracterizam como importantes fontes de investigação. Trata-se de informações utilizadas para o controle operacional da coleta, com objetivo de dar segurança à integridade dos dados da pesquisa: dados de controle de data e hora em que cada entrevista foi realizada; número de vezes que o questionário foi editado; etc. (NICOLAAS, 2011).

Metodologia aplicada ao estudo das divergências

Tendo em vista a multiplicidade de potenciais fatores geradores de erros não amostrais e a disponibilidade de informações referentes ao processo de coleta, fez-se oportuna a análise dos paradados do Censo 2010, o que pode contribuir para o planejamento de futuras operações censitárias e pesquisas amostrais. Neste estudo foram analisadas as

divergências observadas entre as informações coletadas pelos recenseadores durante a realização da entrevista para o Censo 2010 e aquelas obtidas pelos supervisores por meio das reentrevistas geradas pelo sistema de supervisão.[4]

As reentrevistas foram feitas em um subconjunto de domicílios durante a operação de coleta do censo, com o objetivo de fornecer indícios de falhas de coleta cometidas pelos recenseadores. A seleção da amostra de unidades para verificação por reentrevista tinha como finalidade fornecer resultados indicativos de divergência para cada setor censitário. Por ser o setor censitário um recorte geográfico muito pequeno, a opção de fixar uma precisão mínima para realização de uma amostra probabilística de reentrevistas implicaria, em alguns casos, a seleção de todas as unidades do setor. Evidentemente, seria impraticável atribuir aos supervisores a tarefa de conferir todas as unidades domiciliares trabalhadas por seus recenseadores. Alguns procedimentos testados durante as operações pré-censitárias foram utilizados para a definição de critérios de seleção da amostra de unidades (questionários) a serem verificadas em cada setor censitário.

Visando tornar exequível o procedimento de campo de verificação pelos supervisores, foi elaborado um algoritmo para seleção de uma amostra não probabilística que atendesse, simultaneamente, à necessidade de avaliação de cobertura e à viabilidade de aplicação das reentrevistas. Uma das questões abordadas, por exemplo, foi a heterogeneidade dos tamanhos dos setores censitários (tanto em número de unidades quanto em área) e a necessidade de se controlar a distância entre as unidades a serem verificadas, limitando assim os deslocamentos realizados pelos supervisores.

Definição das fontes de dados e variáveis de estudo

Foi considerada divergência, no questionário, a ocorrência de diferença entre as respostas obtidas pelo recenseador e aquelas coletas pelo supervisor para, pelo menos, um dos três quesitos investigados sobre o informante no domicílio entrevistado: sexo, idade[5] e sabe ler e escrever. Tais quesitos foram escolhidos por fazerem parte de ambos os questionários utilizados no Censo 2010 (o questionário do universo e o da amostra). Além disso, tais questões são objetivas e, portanto, dificilmente gerariam dúvidas de entendimento por parte do entrevistado. Acredita-se ser pouco provável que o tipo de abordagem ou a maneira com que estas perguntas são realizadas pelo recenseador ou supervisor tenham forte influência nos resultados das respostas.

Vale destacar que a divergência é utilizada como *proxy* da ocorrência de falha no processo de coleta, pois, apesar de se acreditar que a informação do supervisor seja mais "confiável", muitas das divergências são oriundas de falhas do supervisor e não do recenseador. Uma limitação do estudo, portanto, é a questão de considerar a divergência

[4] Mais detalhes sobre o sistema de supervisão do Censo 2010, os critérios de seleção dos domicílios que foram visitados para as reentrevistas e sobre as bases de dados utilizadas neste estudo estão descritos em Duarte (2014).

[5] No Censo 2010 não se pergunta sobre idade diretamente ao entrevistado, mas esta é calculada a partir de perguntas sobre mês e ano de nascimento, ou idade completa ou presumida.

uma falha sem poder identificar a real fonte do erro. Essa limitação é inerente à fonte de informação, pois não é realizada uma reconciliação por um outro agente para verificar quem cometeu a falha de coleta, se o recenseador ou o supervisor.

Com base nesta suposição, decidiu-se buscar evidências empíricas sobre a influência que os três agentes (informante, recenseador e supervisor) exerceram nas respostas das entrevistas e reentrevistas, e com que grau de intensidade cada um deles pode ter atuado, de forma a gerar maior ou menor interferência nas probabilidades de divergência. Além disso, procuraram-se evidências de possíveis variações associadas às características sociodemográficas destes agentes como prováveis fatores explicativos para o fenômeno em estudo.

Tomou-se como população de estudo o subconjunto das reentrevistas aplicadas em setores censitários trabalhados por somente um recenseador.[6] Foi selecionado um estado por grande região geográfica: Alagoas, Amazonas, Santa Catarina, Rio de Janeiro e Mato Grosso. Esta escolha tomou por base os percentuais de divergências encontradas exclusivamente para os dados do informante e daquelas identificadas para os demais moradores (Gráfico 1). Observa-se que os percentuais de divergência nos estados variam entre 4% e 9%, sendo que os menos desenvolvidos apresentam os maiores valores de divergência. Para a análise selecionaram-se estados de diferentes portes populacionais, regiões e níveis de divergência. Foram escolhidos estados com baixos níveis de divergência, para as Regiões Sul e Norte, e com altos níveis de divergência, no Centro-Oeste e Nordeste. Não foram analisados os estados com os maiores percentuais de divergências, pois se considerou, também, o tamanho dos municípios estudados.

No que diz respeito à Região Sudeste, a escolha do Rio de Janeiro teve como principal motivação o papel estratégico da unidade estadual do IBGE deste estado no processo de treinamento e disseminação de instruções passadas ao pessoal de campo durante a fase de coleta de dados, além da proximidade com as equipes de acompanhamento centralizado na sede do IBGE localizada na cidade do Rio de Janeiro.

A base de dados final para análise das divergências foi constituída de 170.395 registros para os quais foi possível realizar a correspondência biunívoca entre as bases de dados da supervisão e dos microdados do universo do Censo 2010.[7] O total de registros – reentrevistas realizadas em setores coletados por somente um recenseador – processados para cada estado foi de: 15.836 no Amazonas (AM); 12.946 em Alagoas (AL); 82.741 no Rio de Janeiro (RJ); 37.856 em Santa Catarina (SC); e 21.016 em Mato Grosso (MT).[8]

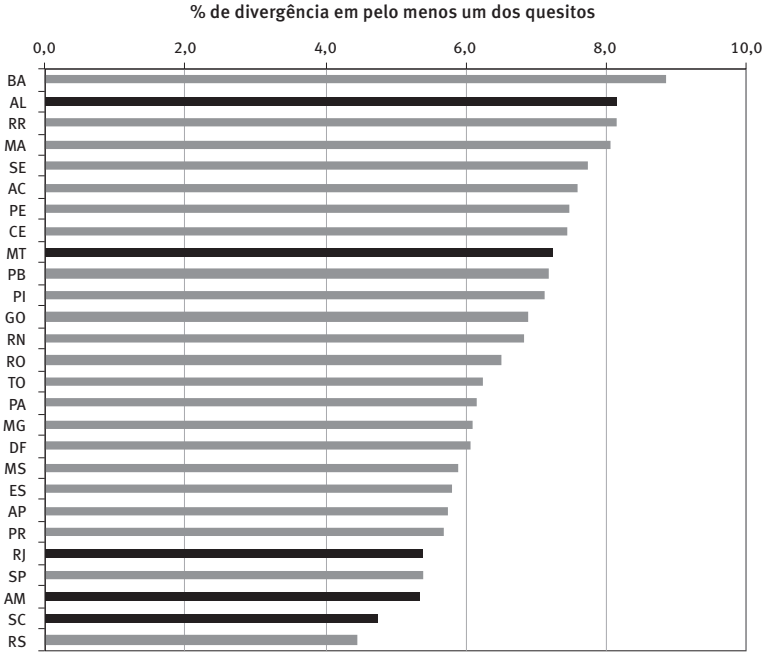
[6] Nos setores trabalhados por mais de um recenseador não foi possível identificar qual deles aplicou cada uma das reentrevistas, inviabilizando a associação reentrevista/recenseador (equivale a 11% do total de reentrevistas aplicadas).

[7] Um dos prováveis motivos para estes registros não terem sido pareados estava relacionado à possibilidade de que um recenseador, ao ser instruído a realizar correções em um questionário, erroneamente, excluía o questionário com erro e criava o registro de um novo questionário, o que ocorreu em cerca de 10% dos casos. Tal fato causava a exclusão da chave primária de identificação do questionário original, impossibilitando o pareamento com os dados da supervisão.

[8] Mais detalhes sobre a construção das bases estão disponíveis em Duarte (2014).

É importante ressaltar que a decisão de considerar apenas alguns estados no escopo do estudo tinha também como foco permitir a realização de uma investigação inicial na qual, caso necessário, as respectivas unidades estaduais do IBGE poderiam esclarecer dúvidas e colaborar na análise.

GRÁFICO 1
Porcentagens de divergências entre as informações coletas pelo recenseador e pelo supervisor na realização das reentrevistas, em pelo menos um dos quesitos do informante (sexo, idade e sabe ler e escrever), por unidade da federação
Brasil – 2010

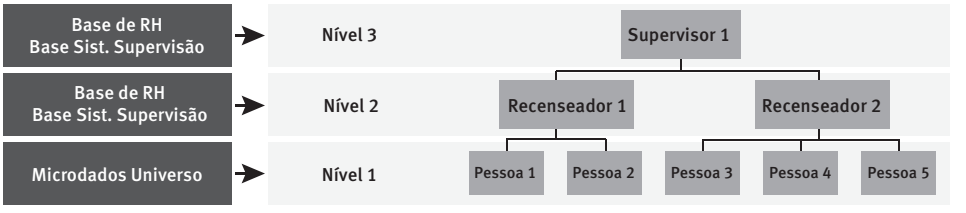


Fonte: IBGE. Censo Demográfico 2010 (microdados do universo, base de dados de recursos humanos do pessoal de coleta e base de dados do sistema de supervisão).

Modelagem estatística aplicada ao estudo das divergências

A estrutura de dados deste estudo é decorrente dos procedimentos adotados para reentrevista. No Censo 2010 foi definido que cada supervisor era responsável pela aplicação de um conjunto de reentrevistas em uma amostra dos questionários realizados pelos recenseadores sob sua coordenação, as quais foram aplicadas somente em domicílios com informante único; portanto, cada reentrevista corresponde a um mesmo e único informante. Trata-se, assim, de uma estrutura de dados agregados ou hierárquicos aninhados, em que o 1º nível corresponde ao informante, o segundo nível corresponde ao recenseador e o terceiro nível corresponde ao supervisor (Figura 1).

FIGURA 1
Estrutura hierárquica das bases de dados



Buscou-se, então, a abordagem metodológica adequada para obtenção de inferências que atendessem aos questionamentos hora levantados e que incorporassem a hipótese de que existe efeito significativo nos níveis de hierarquia na estrutura dos dados. Para isso, foram empregados modelos hierárquicos de regressão logística (HOX, 2010; RAUDENBUSH; BRYK, 2002).

A metodologia de análise de dados hierárquicos consiste na aplicação de métodos estatísticos para o tratamento de dados com padrões de variação complexos. Esta complexidade diz respeito, sobretudo, à formação de subgrupos dentro da população-alvo (SNIJDERS; BOSKER, 1999).

Em situações comuns de análise, uma das hipóteses a respeito dos dados é que todas as unidades são mutuamente independentes. Quando este pressuposto é violado, os resultados da aplicação de um modelo tradicional podem ser seriamente comprometidos (SNIJDERS; BOSKER, 1999). As estimativas de erro dos coeficientes de regressão do primeiro nível podem ser viesadas. Além disso, estimativas de erros dos coeficientes de covariáveis de níveis superiores tendem a ser subestimadas, trazendo prejuízos às inferências realizadas (BARTHOLOMEW et al., 2008).

Uma das justificativas para utilização de um modelo hierárquico consiste na necessidade de considerar a possibilidade de existência de correlação significativa entre as respostas de indivíduos pertencentes a um mesmo grupo em um mesmo nível de agregação. O interesse neste estudo foi de investigar fatores associados às divergências entre as informações coletadas pelo supervisor e pelo recenseador, levando à formulação de uma variável resposta dicotômica associada à ocorrência ou não de divergência. Com isso, assumiu-se que a ocorrência de divergência é uma variável aleatória Y com distribuição *Bernoulli*, para a qual se deseja estimar a probabilidade de ocorrência π, em que:

$$Y = \begin{cases} 1 & \text{ocorrência de divergência} \\ 0 & \text{não ocorrência de divergência} \end{cases}$$

Levando em conta as características da variável resposta e a estrutura de agregação dos dados (cada supervisor responsável por vários recenseadores e cada recenseador realizando entrevistas em vários domicílios), optou-se pela utilização de um modelo hierárquico logístico, que é dado pelas seguintes expressões:

$$Logit(\pi_{ijk}) = \underbrace{(\beta_{0jk})}_{\text{efeito aleatório}} + \underbrace{\overbrace{\sum_{q=1}^{Q}\beta_q X_{qjki}}^{\text{Informantes}} + \overbrace{\sum_{r=1}^{R}\gamma_r Z_{rjk}}^{\text{Recenseadores}} + \overbrace{\sum_{s=1}^{S}\delta_s W_{sk}}^{\text{Supervisores}}}_{\text{efeitos fixos/covariáveis}} \quad (1)$$

$$\beta_{0jk} = \beta_0 + u_{0k} + v_{0jk} \begin{cases} \beta_0 - \text{intercepto} \\ u_{0k} \sim N(0, \sigma^2_{u_0}) - \text{componente de variância atribuída aos supervisores} \\ v_{0jk} \sim N(0, \sigma^2_{v_0}) - \text{componente de variância atribuída aos recenseadores} \end{cases} \quad (2)$$

Em que:

π_{ijk} é a probabilidade de haver divergência em pelo menos um dos quesitos sexo, idade e sabe ler e escrever do informante i, associado ao recenseador j e ao supervisor k;

Q refere-se ao número de efeitos fixos e covariáveis associadas aos informantes;

R corresponde ao número de efeitos fixos e covariáveis associadas aos recenseadores;

S é o número de efeitos fixos e covariáveis associadas aos supervisores.

Medida para avaliação das componentes de variância

O coeficiente de correlação intraclasse (ρ) é a medida relativa de composição da variação não explicada de um fenômeno estocástico. No presente estudo, por meio do cálculo desse coeficiente, foi possível estimar o quanto da variação não explicada das divergências poderia ser atribuído aos recenseadores, o quanto poderia ser atribuído aos supervisores e uma outra parcela que diz respeito a outros fatores, entre eles o informante.

O cálculo dos coeficientes de correlação intraclasse toma por base as parcelas de componentes de variância de todos os níveis de hierarquia adotados no modelo, neste caso três níveis. No entanto, em sua concepção, o modelo logístico hierárquico não contém um parâmetro de escala que corresponda ao erro aleatório, que seria a componente de variância do primeiro nível (HOX, 2010). Isso se dá pelo fato de a função de ligação ser expressa pela esperança do quociente entre as probabilidades de ocorrência e não ocorrência do evento, o que impede a obtenção direta de uma componente de variância para o 1º nível de hierarquia.

Existem algumas formas para o cálculo do coeficiente de correlação intraclasse para modelos com variável resposta binária, por meio da padronização dos parâmetros de locação e escala, de acordo com a função de ligação utilizada (HOX, 2010, p. 133). Um dos métodos mais utilizados é o da variável latente, também conhecido como representação de modelo limítrofe (*threshold model*) (SNIJDERS; BOSKER, 1999, p. 223).

Suponha que a variável binária Y_{ijk} pode ser representada por uma variável latente contínua Y^*_{ijk}. Pode-se supor que Y^*_{ijk} é a propensão de $Y_{ijk} = 1$. Então, define-se a variável latente como uma variável contínua de forma que $Y^*_{ijk} > 0$ se a variável observada $Y_{ijk} = 1$ e $Y^*_{ijk} \leq 0$ se $Y_{ijk} = 0$.

$$Y_{ijk} = \begin{cases} 1 & Y^*_{ijk} > 0 \\ 0 & Y^*_{ijk} \leq 0 \end{cases} \quad (3)$$

Sob essas premissas, o modelo logístico hierárquico de três níveis pode ser representado de forma equivalente ao seguinte modelo linear hierárquico para a variável latente contínua Y^{*}_{ijk}:

$$Y^{*}_{ijk} = \underbrace{\beta_{0jk}}_{\text{efeito aleatório}} + \underbrace{\overbrace{\sum_{q=1}^{Q}\beta_q X_{qjki}}^{\text{Informantes}} + \overbrace{\sum_{r=1}^{R}\gamma_r Z_{rjk}}^{\text{Recenseadores}} + \overbrace{\sum_{s=1}^{S}\delta_s W_{sk}}^{\text{Supervisores}}}_{\text{efeitos fixos/covariáveis}} + u_{0jk} + v_{0jk} + \varepsilon^{*}_{ijk} \quad (4)$$

Em que:

ε^{*}_{ijk} é o termo de erro do 1º nível, o qual se supõe ter uma distribuição logística padrão, com média 0 e variância $\pi^2/_3 \cong 3{,}29$;

β_0 é o intercepto;

$u_{0k} \sim N(0,\sigma^2_{u_0})$ corresponde ao componente de variância atribuída aos supervisores;

$v_{0k} \sim N(0,\sigma^2_{v_0})$ é o componente de variância atribuída aos recenseadores.

Para o modelo de três níveis, a correlação entre as variáveis latentes para os indivíduos de um mesmo grupo nos níveis 2 (recenseador) ou 3 (supervisor) é dada, respectivamente, por (HOX, 2010, p. 133):

$$\rho_{v0} = \frac{\sigma^2_{v_0}}{\sigma^2_{u_0} + \sigma^2_{v_0} + \left(\frac{\pi^2}{3}\right)} = \frac{\sigma^2_{v_0}}{\sigma^2_{u_0} + \sigma^2_{v_0} + 3{,}29} \quad (5)$$

$$\rho_{u0} = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_{v_0} + \left(\frac{\pi^2}{3}\right)} = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_{v_0} + 3{,}29} \quad (6)$$

Analogamente, é possível obter a representação da variação conjunta dos níveis 2 e 3 (recenseador e supervisor) em uma mesma medida, por meio da soma dos coeficientes ρ_{v0} e ρ_{u0}:

$$\rho_{v0} + \rho_{u0} = \frac{\sigma^2_{v_0} + \sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_{v_0} + \left(\frac{\pi^2}{3}\right)} = \frac{\sigma^2_{v_0} + \sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_{v_0} + 3{,}29} \quad (7)$$

Para estimação das componentes de variância, foi escolhido o método de integração numérica de Laplace. Tal opção se deu, sobretudo, por conta da possibilidade de utilização das razões de verossimilhanças para o teste dos fatores aleatórios nos modelos (SNIJDERS; BOSKER, 1999). O método de Laplace por integração numérica se baseia na aproximação da função de verossimilhança real. Apesar desses métodos apresentarem boas aproximações para as estimativas, há que se levar em conta a alta carga de processamento computacional demandada para execução de seus algoritmos. Neste artigo, a estimação dos modelos foi implementada no pacote estatístico SAS.[9]

[9] SAS Enterprise Guide, versão 4.1 – PROC GLIMMIX, SAS Institute Inc., Cary, NC, USA.

Avaliação de adequabilidade do ajuste dos modelos

A comparação entre dois modelos M_0 e M_1 com números de parâmetros q e p, respectivamente, pode ser feita com base na estatística da razão de verossimilhanças de dois modelos, dada por:

$$\Delta D = 2\,[l(b_1;y) - l(b_0;y)] \quad (8)$$

Em que:

$l(b_0;y)$ é o valor do logaritmo da verossimilhança para o modelo M_0;

$l(b_1;y)$ é o valor do logaritmo da verossimilhança para o modelo M_1;

ΔD segue uma distribuição qui-quadrado com p-q graus de liberdade (DOBSON; BARNETT, 2008).

$$\Delta D \sim X^2_{p-q} \text{ onde } q < p < N \quad (9)$$

Snijders e Bosker (1999) propuseram uma medida de adequabilidade do ajuste para modelos hierárquicos logísticos, a qual consiste em uma extensão do método de McKelvey e Zavoina (1975, apud SNIJDERS; BOSKER, 1999) baseado na variação explicada da variável latente Y^*_{ijk}.

No caso do modelo logístico hierárquico com três níveis, considerado neste estudo, a variância de Y^*_{ijk} pode ser decomposta nas seguintes parcelas:

σ^2_ε – variância do primeiro nível hierárquico fixada em $\pi^2/_3 \cong 3{,}29$;

$\sigma^2_{u_0}$ – componente de variância atribuída ao segundo nível hierárquico (supervisores);

$\sigma^2_{v_0}$ – componente de variância atribuída ao terceiro nível hierárquico (recenseadores);

σ^2_F – variância atribuída aos efeitos fixos do modelo, que corresponde à variância dos valores preditos da variável resposta binária Y.

Dessa forma, para o modelo hierárquico de três níveis, a proporção da variância explicada pelos efeitos fixos é determinada pela seguinte expressão:

$$R^2_{MZ} = \frac{\sigma^2_F}{\sigma^2_\varepsilon + \sigma^2_{u_0} + \sigma^2_{v_0} + \sigma^2_F} \quad (10)$$

Resultados

Primeiramente, foram realizados os testes para os efeitos aleatórios a fim de avaliar estatisticamente a hipótese de que havia diferença na variação aleatória nas divergências entre grupos de recenseadores e entre grupos de supervisores.

Vale lembrar que, além dos níveis hierárquicos, foram obtidas informações provenientes das diferentes bases de dados referentes aos informantes, variáveis individuais e outras sobre as características dos domicílios, bem como algumas características dos recenseadores e dos supervisores. Os dados dos informantes e de seus domicílios foram obtidos nos microdados do universo do Censo Demográfico 2010 e as informações dos recenseadores e supervisores são provenientes da base de dados de recursos humanos da equipe de coleta.

À medida que os modelos eram avaliados, algumas variáveis apresentaram categorias estatisticamente significativas de acordo com os testes aplicados, enquanto outras não. Nesses casos, foram consideradas algumas combinações de agrupamento em busca de melhores estimativas para os modelos. O Apêndice contém uma lista de todas as variáveis incluídas inicialmente na modelagem, sendo que a descrição detalhada de suas categorias e as formas de agrupamento estão disponíveis em Duarte (2014).

A Tabela 1 apresenta as razões de verossimilhanças entre os modelos não condicionais em relação ao modelo nulo, sem considerar os efeitos aleatórios com seus respectivos níveis de significância. O modelo M1 corresponde ao modelo logístico nulo, contendo somente o parâmetro β_0 (intercepto). O modelo hierárquico não condicional M2 acrescenta, a M1, o efeito aleatório de recenseador v_{0jk}. No modelo hierárquico condicional M3 adiciona-se o efeito aleatório de supervisor u_{0k}, tal que:

M1 – $Logit\ (\pi_{ijk}) = \beta_0$

M2 – $Logit\ (\pi_{ijk}) = \beta_0 + v_{0jk}$

M3 – $Logit\ (\pi_{ijk}) = \beta_0 + u_{0k} + v_{0jk}$

TABELA 1
Deviances e razões de verossimilhanças dos modelos hierárquicos da divergência em pelo menos um dos quesitos do informante (sexo, idade e sabe ler e escrever)
Estados selecionados – 2010

Modelos	Amazonas	Alagoas	Rio de Janeiro	Santa Catarina	Mato Grosso
Razões de verossimilhanças (ΔD)					
M1 – Sem efeito aleatório (nulo)	6399,95	7199,91	33722,27	13916,01	11029,09
M2 – Com efeito de recenseador	6211,56	7178,76	32499,04	13745,41	10779,58
M3 – Com efeitos de recenseador e supervisor	6159,72	7154,47	32116,58	13711,87	10578,19
Decomposição das razões de verossimilhanças e níveis de significância					
$\Delta D_{(M2-M1)}$	188,39 (***)	21,15(***)	1223,23(***)	170,6(***)	249,51(***)
$\Delta D_{(M3-M2)}$	51,84(***)	24,29(***)	382,46(***)	33,54(***)	201,39(***)

Fonte: IBGE. Censo Demográfico 2010 (microdados do universo, base de dados de recursos humanos do pessoal de coleta e base de dados do sistema de supervisão).
(***) Níveis de significância >0,001.

Para todos os estados, há evidências de que ambos os efeitos aleatórios de recenseador e supervisor sejam estatisticamente significativos segundo o teste da razão de verossimilhanças com nível de significância menor que 0,001. Isso evidencia a hipótese de que uma parcela da variação das divergências nas respostas das reentrevistas pode ser atribuída ao efeito dos recenseadores, assim como há uma parte significativa que pode ser imputada aos supervisores.

Avaliação da composição da variação não explicada das divergências

Uma vez constatada a significância estatística dos fatores aleatórios referentes aos efeitos de recenseadores e supervisores, obtiveram-se as medidas relativas da composição

da variância que permitem avaliar a participação destes dois fatores na variação não explicada das divergências.

Foram então calculadas as estimativas das componentes de variância para os efeitos aleatórios de recenseador e supervisor, desconsiderando os efeitos fixos dos três níveis hierárquicos. Com base nestas componentes, foram obtidos os coeficientes de correlação intraclasse (ρ).

A Tabela 2 traz os coeficientes para os efeitos aleatórios de supervisor e recenseador e a soma correspondente aos coeficientes dos dois agentes para o modelo não condicional $Logit\ (\pi_{ijk}) = \beta_{0jk}$. Observa-se que os valores dos coeficientes não apresentam um padrão entre os estados no que se refere a uma maior parcela de variação das divergências atribuídas ao recenseador ou ao supervisor, o que evidencia relevante diferencial entre os estados selecionados quanto à composição da variabilidade das divergências. As somas das parcelas de recenseador e supervisor também são muito distintas, variando de 7% em Alagoas até 23% no Rio de Janeiro. Além de a soma das duas componentes ser heterogênea entre os estados, as parcelas da variação explicada são diferenciadas entre os agentes para os estados, também ratificando a validade de estudar o fenômeno regionalmente.

TABELA 2
Coeficientes de correlação intraclasse (ρ) de recenseador e supervisor dos modelos hierárquicos da divergência em pelo menos um dos quesitos do informante (sexo, idade e sabe ler e escrever) Estados selecionados – 2010

Estados	Efeito aleatório		
	Supervisor (ρ_{u_0})	Recenseador (ρ_{v_0})	Soma ($\rho_{u_0} + \rho_{v_0}$)
Rio de Janeiro	0,111	0,120	0,231
Santa Catarina	0,049	0,118	0,168
Mato Grosso	0,092	0,049	0,142
Alagoas	0,043	0,029	0,071
Amazonas	0,092	0,129	0,221

Fonte: IBGE. Censo Demográfico 2010 (microdados do universo, base de dados de recursos humanos do pessoal de coleta e base de dados do sistema de supervisão).

Mesmo constatadas estas diferenças, é importante observar que, em nenhum dos estados, a soma dos coeficientes ρ explica a maior parte do total da variação. O Rio de Janeiro apresentou o maior percentual de variação associada à estrutura de coleta (23%). Ou seja, isso significa que 77% da variação não explicada entre as divergências pode estar relacionada a outros fatores, sobretudo ao informante, que se supõe ser uma das principais fontes de variação não controlada.

Avaliação dos efeitos fixos como fontes de variação não explicada das divergências

Além de fontes de variação aleatória, foram obtidas estimativas de possíveis efeitos fixos na explicação das divergências. Para isso foram testadas, na condição de variáveis explicativas, diversas características associadas aos três níveis hierárquicos considerados neste estudo. Foram mantidas no modelo as variáveis explicativas com

níveis de significância inferiores a 5%, segundo o teste de Wald. No que se refere aos informantes, foram incluídas inicialmente todas as variáveis disponíveis no questionário do universo do Censo e, para os recenseadores e supervisores, avaliaram-se todas as variáveis disponíveis nos bancos de dados de recursos humanos do IBGE (a lista de variáveis se encontra no Apêndice). Os resultados apresentados só incluem as variáveis que mostraram efeitos estatisticamente significativos. Com base no valor das estimativas dos parâmetros dos efeitos fixos, obtiveram-se as estimativas das razões de chances da ocorrência de divergências.

A Tabela 3 apresenta estas razões de chances em favor da ocorrência de divergências relativas aos efeitos fixos considerados significativos (α=5%) no modelo condicional, após a inclusão das variáveis (listadas na Tabela) associadas aos três níveis hierárquicos. Observa-se que os resultados obtidos mostram tendências parecidas quanto às razões de chances estimadas para todos os estados. Isso indica que há evidências de que os efeitos significativos comuns encontrados para essas unidades da federação apontam para uma mesma direção em relação ao aumento ou à redução nas razões de chances. Esta é uma constatação relevante na aceitação da coerência dos resultados entre os estados.

As três variáveis do informante (idade, sexo e sabe ler e escrever) utilizadas para avaliação das divergências entre recenseador e supervisor se destacam como efeitos significativos em todos os modelos ajustados. Nenhum outro atributo exclusivamente do informante foi estatisticamente significativo para todos os estados. Observa-se que as razões de chances em favor das divergências para o efeito de idade do informante variam de 1,006 (AM) a 1,018 (MT), o que significa um aumento esperado de 1% a cada ano de acréscimo na idade do informante, no Amazonas, e de quase 2%, em Mato Grosso. Os demais estados têm resultados também dentro desta faixa.

Cabe relembrar que a análise refere-se às divergências ocorridas nas informações prestadas pelo próprio informante. Portanto, parece coerente assumir que existe uma tendência no aumento da vantagem de divergência de informantes com idade mais avançada. Ou seja, os resultados sugerem que, com a elevação da idade, também cresça a probabilidade de prestar informações diferentes.

Em relação à variável sexo do informante, observa-se que o sexo masculino apresenta razões de chances entre 1,19 (MT) e 1,29 (SC). Ao avaliar o conjunto dos cinco estados, estima-se que a vantagem em favor da ocorrência de divergência para informantes do sexo masculino seja de 20% a 30% maior do que para informantes do sexo feminino (pontos centrais dos intervalos de confiança).

Embora esta tendência nos efeitos significativos comuns se mostre aparentemente coerente, é notória a quantidade de variáveis significativas não comuns entre os estados. Por exemplo, o efeito de residência em domicílio com energia elétrica apresenta-se como significativo no modelo somente para o Amazonas. Em contrapartida, o logaritmo da renda domiciliar *per capita* é determinante para os demais estados. Nota-se, entretanto, que tanto a renda como a disponibilidade de energia elétrica representam a mesma dimensão

socioeconômica dos indivíduos nos diferentes estados e permitem mensurar o efeito associado a esta dimensão na ocorrência de divergências.

TABELA 3
Estimativas dos efeitos fixos nos modelos hierárquicos da divergência em pelo menos um dos quesitos do informante (sexo, idade e sabe ler e escrever)
Estados selecionados – 2010

Variáveis significativas em pelo menos um estado	Razões de chances				
	RJ	SC	MT	AL	AM
1º nível (informante e domicílio)					
Idade	1,007	1,011	1,018	1,017	1,006
Sexo					
Masculino / feminino	1,258	1,292	1,188	1,258	1,265
Sabe ler e escrever					
Sim/ não	0,225	0,146	0,289	0,425	0,194
Cor ou raça					
Branca / não branca	-	0,842	-	-	-
Forma de declaração da idade					
Data de nascimento / idade declarada	0,616	0,505	0,303	-	-
Relação com o responsável pelo domicílio					
Responsável ou cônjuge / outra condição	0,887	0,737	-	-	-
$\log_{10}$(renda domiciliar *per capita*)	0,874	0,841	0,898	0,898	-
Número de banheiros	0,872	0,857	0,895	0,827	0,888
Origem dos dados					
Básico / amostra	-	-	1,272	-	-
Responsabilidade pelo domicílio é de:					
Apenas um morador / mais de um morador	-	-	1,175	-	-
Ignorado (1)/ mais de um morador	-	-	1,094	-	-
Tem energia elétrica no domicílio					
De companhia distribuidora/ outra forma ou não tem	-	-	-	-	0,633
O esgoto do banheiro ou sanitário é lançado (jogado) em:					
Rede geral de esgoto ou pluvial / outra forma	-	1,159	-	-	-
Espécie da unidade doméstica					
Unipessoal ou nuclear / outra forma	0,839	0,757	0,796	-	-
Horário de coleta do questionário					
Até 18 h / após 18h	0,896	-	-	-	-
2º nível (recenseador)					
Escolaridade do recenseador					
Fundamental ou médio / superior (comp. ou incomp.)	-	-	-	-	1,221
3º nível (supervisor)					
Escolaridade do supervisor					
Médio / superior (comp. ou incomp.)	-	-	1,231	-	-
Idade do supervisor	-	-	-	1,014	-

Fonte: IBGE. Censo Demográfico 2010 (microdados do universo, base de dados de recursos humanos do pessoal de coleta e base de dados do sistema de supervisão).
(1) A categoria ignorado é não significativa, no entanto, foi mantida por não se adequar à agregação de nenhuma das outras duas categorias.
(-) Não estatisticamente significativo para o Estado ao nível de 5%.

Os resultados sugerem que as entrevistas realizadas com informantes do sexo masculino, analfabetos, mais velhos e que vivem em domicílios com indicadores que refletem condições de vida menos satisfatórias mostram aumento na chance da ocorrência de divergências entre recenseador e supervisor. Por exemplo, no Mato Grosso, o incremento de um ano na idade do informante eleva a chance da ocorrência de divergência em quase 2%. Em Santa Catarina, o fato de o informante saber ler e escrever reduz a vantagem em favor da ocorrência de divergência em 84% e, em Alagoas, o acréscimo de um banheiro no domicílio é indicativo de redução na chance da ocorrência de divergência da ordem de 17%.

Percebe-se, também, que diferenciais de perfil dos recursos humanos disponíveis nos diferentes estados analisados apresentam associação distinta em relação à incidência de divergências. No Amazonas, estima-se um crescimento de 22% na vantagem em favor das divergências para recenseadores com menor nível de instrução, o que é coerente com o esperado. Por outro lado, para os supervisores, não havia expectativa de que o nível de escolaridade fosse significativo no modelo. No entanto, para o estado do Mato Grosso, os resultados do modelo indicam um aumento de 23% na chance de divergências quando os supervisores possuem menor nível de instrução. O estado de Alagoas é o único que registra ampliação na chance de divergências para supervisores mais velhos. Os resultados referentes ao 3º nível apontam para a necessidade de maior aprofundamento metodológico no sentido de investigar possíveis interações entre as características de recenseador e supervisor.

Qualidade do ajuste dos modelos

Para avaliação da qualidade do ajuste no que se refere aos efeitos fixos dos modelos, foram calculados os valores dos coeficientes *pseudo* R^2 pelo método de McKelvey e Zavoina (SNIJDERS; BOSKER, 1999) (Tabela 4). Todos os modelos estimados apresentaram baixa proporção de explicação atribuída aos efeitos fixos em relação à variação total estimada para a variável latente Y^*, variando entre 10% e 15%.

TABELA 4
Coeficientes de determinação *pseudo* R^2 de McKelvey e Zavoina (R^2_{MZ}) para os modelos hierárquicos da divergência em pelo menos um dos quesitos do informante (sexo, idade e sabe ler e escrever) Estados selecionados – 2010

Estados	$\sigma^2_{u_0}$	$\sigma^2_{v_0}$	σ^2_f	σ^2_R	R^2_{MZ}
Amazonas	0,3867	0,5450	0,5962	3,29	0,12
Alagoas	0,1507	0,1019	0,3876		0,10
Rio de Janeiro	0,4759	0,5149	0,7631		0,15
Santa Catarina	0,1956	0,4668	0,4463		0,10
Mato Grosso	0,3544	0,1886	0,6671		0,15

Fonte: IBGE. Censo Demográfico 2010 (microdados do universo, base de dados de recursos humanos do pessoal de coleta e base de dados do sistema de supervisão).

Cabe esclarecer que não havia a expectativa de alcançar altos percentuais de explicação pelas componentes fixas dos modelos, uma vez que já foi observada e evidenciada a significância do efeito da estrutura hierárquica sobre a variação total estimada do fenômeno. Além disso, foram identificados poucos fatores determinantes da variação nos níveis associados ao recenseador e supervisor.

Análise das probabilidades estimadas para os modelos

Com base nos modelos ajustados para os estados analisados, obtiveram-se as probabilidades estimadas para as divergências segundo as variáveis quantitativas estatisticamente significativas. Para isso, foram arbitrados dois cenários com distintas situações, buscando contrastar combinações de categorias e valores dos efeitos dos modelos.

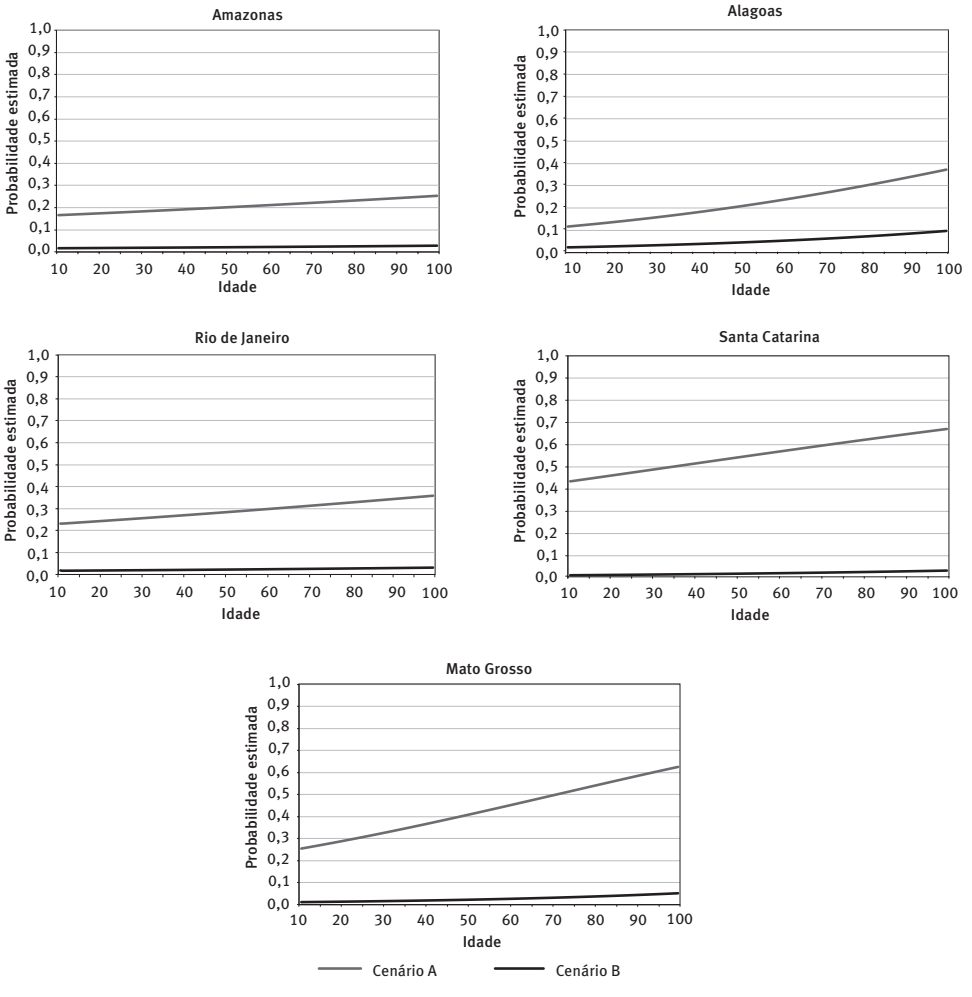
No cenário A, são privilegiados valores e categorias que apresentaram maiores proporções de divergências observadas e, no cenário B, aqueles com menores proporções de divergências. Por exemplo, no estado de Alagoas, para as estimativas de probabilidade segundo idade, o cenário A é constituído por informantes hipotéticos cujas idades variam entre 10 e 100 anos, do sexo masculino, analfabetos, residentes em domicílio com apenas um banheiro, com renda domiciliar *per capita* de ¼ de salário mínimo e cujos dados da reentrevista foram verificados por supervisor com 45 anos de idade. Já o cenário B corresponde a informantes hipotéticos cujas idades variam entre 10 e 100 anos de idade, do sexo feminino, alfabetizadas, residentes em domicílio com dois banheiros, com renda domiciliar *per capita* de 1 ½ salário mínimo e cujos dados da reentrevista foram verificados por supervisor com 20 anos de idade.

Ao analisar o Gráfico 2, um aspecto comum para todos os estados é a pequena variação nas probabilidades estimadas para os perfis definidos pelo cenário B, o qual apresentou alguma variação, visualmente perceptível por meio da análise gráfica, apenas para Alagoas.

É notória a ocorrência de comportamento similar das tendências (decréscimo ou aumento) das probabilidades de divergência, para ambos os cenários A e B nos estados. Outro aspecto comum, como era esperado, é que os patamares das probabilidades estimadas no cenário A são sempre bem maiores do que o nível mantido pelo cenário B.

Os resultados dos modelos estimados evidenciaram que parte da variação não explicada na ocorrência das divergências pode ser atribuída aos diferenciais entre supervisores e entre recenseadores. A participação desta parcela da variação não é homogênea entre os estados estudados e poucas das características destes agentes (recenseadores e supervisores) são estatisticamente significativas na explicação do fenômeno. Ou seja, apesar de se verificar a existência de significativas diferenças nas probabilidades de divergências entre recenseadores e entre supervisores, para maioria dos estados, não é possível apontar quais características dos agentes estão relacionadas a estas diferenças. A parcela de variação atribuída a outros fatores é predominantemente maior para todos os estados. Esta parcela corresponde à variabilidade entre informantes, além de outras fontes de variação não identificáveis por meio dos dados disponíveis.

GRÁFICO 2
Probabilidades estimadas a partir dos modelos hierárquicos da divergência em pelo menos um dos quesitos do informante (sexo, idade e sabe ler e escrever), segundo idade do informante e cenários (A e B)
Estados selecionados – 2010



Fonte: IBGE. Censo Demográfico 2010 (microdados do universo, base de dados de recursos humanos do pessoal de coleta e base de dados do sistema de supervisão).

Ao contrário do que se observou com recenseadores e supervisores, dentre as características dos informantes elencadas no estudo, obteve-se um número expressivo de variáveis estatisticamente significativas. Embora não se possa afirmar que o informante é responsável por maior parte da variação não explicada, o número de variáveis associadas ao mesmo, que se mostraram significativas nos modelos estimados, indica que as características dos informantes e de seus domicílios parecem ter maior influência na probabilidade de divergências do que as características dos agentes de coleta.

Vale aqui lembrar que as divergências não permitem apontar onde existem falhas nos dados finais do censo, mas identificam exatamente onde ocorreram falhas de processo,

indicando as unidades para as quais, em diferentes momentos de abordagem, as informações coletadas diferiram entre si por algum motivo: seja por parte do recenseador, supervisor, informante ou qualquer outra fonte de interferência no processo de coleta. É importante ressaltar que os casos de divergência foram revistos durante o processo de coleta do Censo, pois a identificação de divergências era parte do processo de controle de qualidade da pesquisa.

Considerações finais

O artigo procura identificar fatores associados a fontes de erro não amostral que pudessem ter impactado no processo de coleta e consequentemente na qualidade dos resultados do Censo Demográfico de 2010. Nesse sentido, buscaram-se elementos que trouxessem evidências de possíveis causas de falhas na coleta por meio da análise dos paradados da pesquisa. A abordagem das divergências entre informações coletadas por recenseadores e supervisores foi escolhida como foco de análise na investigação de fatores associados a erros não amostrais.

A trajetória de análise que culminou na identificação de possíveis fatores associados a erros não amostrais apontou tendências, semelhanças e diferenças no que se refere à explicação do fenômeno das divergências. Os resultados indicam não haver expressiva relação entre o perfil dos recenseadores e supervisores com o fenômeno das divergências, bem como que os efeitos das características dos agentes para alguns dos estados podem ser atribuídos às diferenças regionais. Essas diferenças podem estar associadas a vários fatores, tais como aspectos particulares na gestão da coleta centralizada nas unidades Estaduais, treinamento, disponibilidade de recursos humanos, etc. Ou seja, existem indicativos de que haja fontes de variação associadas a níveis hierárquicos superiores ao nível de supervisor.

É pequena a quantidade de informações disponíveis sobre os agentes de coleta, se comparada ao elenco de variáveis sobre os informantes. A investigação evidencia o poder de associação exercido sobre as divergências de forma diferente pelos três níveis hierárquicos estudados em diferentes regiões do país. Por outro lado, são observadas semelhanças e tendências que apontam em direção a uma melhor compreensão do fenômeno, até então nunca investigado sob a ótica da influência da estrutura hierárquica na coleta de dados do Censo brasileiro.

Os resultados do artigo podem ser utilizados para o planejamento das etapas de treinamento da equipe de campo e supervisão de coleta. Adicionalmente, dadas as restrições encontradas para análise de divergências, espera-se que o estudo estimule iniciativas, em futuras operações censitárias e outras pesquisas do IBGE, para que a apropriação de paradados ocorra não só como recurso para controle técnico e operacional da coleta, mas também como forte aliada para análise e interpretação *post hoc* das informações produzidas.

Além disso, outras fontes de paradados podem aprimorar a qualidade da coleta. Uma delas é a base de dados do entorno dos domicílios do Censo 2010. Esta base consiste de

dados obtidos por meio da observação de recenseadores e supervisores sobre características gerais de segmentos de logradouros dos setores urbanos. Tais informações são relevantes para o estudo de falhas de cobertura.

Cabe ressaltar, também, o ineditismo deste estudo no qual, pela primeira vez no Brasil, informações referentes ao processo de coleta de uma pesquisa deste porte foram analisadas com base em informações sobre os agentes de coleta de dados, os informantes e o processo de trabalho, utilizando concomitantemente bases de dados de recursos humanos, do processo de coleta e dos microdados da pesquisa.

Referências

BIEMER, P. P.; LIYEBERG, L. E. **Introduction to survey quality**. New York: John Wiley & Sons, 2003.

BARTHOLOMEW, D. J.; STEELE, F.; GALBRAITH, J. I. **Analysis of multivariate social science data**. 2. ed.Boca Raton, FL: Chapman & HAll/CRC, 2008.

COUPER, M. Measuring survey quality in a CASIC environment. In: SECTION ON SURVEY RESEARCH METHODS OF THE AMERICAN STATISTICAL ASSOCIATION. **Proceedings...** Alexandria, VA: ASA, 1998. Disponível em: ‹http://www.amstat.org/sections/srms/proceedings/papers/1998_006.pdf›.

DOBSON, A. J.; BARNET, A. G. **An introduction to eeneralized linear models**. 3. ed. [S.l.]: Chapman & Hall/CRC, 2008.

DUARTE, L. T. **Análise dos paradados do Censo Demográfico 2010:** investigações de fatores associados a erros não amostrais detectados na coleta das informações. 247 f. Dissertação (Mestrado em Estudos Populacionais e Pesquisas Sociais) – Escola Nacional de Ciências Estatísticas, Rio de Janeiro, 2014. Disponível em: ‹http://www.ence.ibge.gov.br/index.php/pos-grad-mest-dissert/mest-dissertacoes2014›. Acesso em: 29 nov. 2016.

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

HOLT, D. **Methodological issues in the development and use of statistical indicators for international comparisons**. Business Survey Methods Division, Statistics Canada, Survey Methodology, 2005.

HOX, J. J. **Multilevel analysis**: techiniques and aplications. 2. ed. New York: Routledge, 2010.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Censos 2007**. Inovações e impactos nos sistemas de informações estatísticas e geográficas do Brasil. Rio de Janeiro: IBGE, 2008.

_____. **Censo Demográfico 2010**. Metodologia do Censo Demográfico 2010. Rio de Janeiro: IBGE, 2013 (Série Relatórios Metodológicos, v. 41).

_____. **Censo Demográfico 2010**. Resultados gerais da amostra. Rio de Janeiro: IBGE, 2012.

NICOLAAS, G. **Survey paradata**: a review. National Centre for Social Research (NatCen), January 2011. Disponível em: ‹http://eprints.ncrm.ac.uk/1719/1/Nicolaas_review_paper_jan11.pdf›. Acesso em: 07 fev. 2013.

RAUDENBUSH, S. W.; BRYK, A. S. **Hierarchical linear models**: applications and data analysis methods. 2. ed. Thousand Oaks, CA: Sage Publications, 2002.

SNIJDERS. T.; BOSKER, R. **Multilevel analysis**: an introduction to basic and advanced multilevel modeling. Londres: SAGE, 1999.

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

Apêndice
Relação das variáveis consideradas no desenvolvimento dos modelos estatísticos

Variáveis associadas à pessoa que prestou as informações

Relação de parentesco ou de convivência com a pessoa responsável pelo domicílio, sexo, forma de declaração da idade (data de nascimento ou idade declarada), idade calculada em anos (0 a 140), cor ou raça, sabe ler e escrever e valor do rendimento mensal total em julho de 2010.

Variáveis associadas ao domicílio da pessoa que prestou as informações

Situação do domicílio (urbano e rural), alteração de espécies (entre ocupado, vago e fechado), data da entrevista, hora da entrevista, tempo total de entrevista em minutos, se domicílio próprio, alugado ou cedido, número de banheiros de uso exclusivo dos moradores no domicílio, tipo de esgotamento sanitário, forma de abastecimento de água, destino do lixo, existência de energia elétrica, número de pessoas moradoras em 31 de julho de 2010, responsabilidade pelo domicílio (apenas um ou mais de um morador), rendimento domiciliar *per capita* em julho de 2010 e espécie da unidade doméstica (arranjo domiciliar).

Variáveis associadas aos recenseadores e supervisores

Sexo, idade em anos completos, estado civil, número de dependentes no salário família, número de dependentes na Secretaria de Fazenda Federal e nível de escolaridade (o nível mínimo de escolaridade exigido para os supervisores era ensino médio completo).

Sobre os autores

Luciano Tavares Duarte é estatístico e mestre em Estudos Populacionais e Pesquisas Sociais pela Escola Nacional de Ciências Estatísticas (Ence). Tecnologista de informação geográfica e estatísticas do Instituto Brasileiro de Geografia e Estatística (IBGE).

Denise Britz do Nascimento Silva é doutora em Estatística pela University of Southampton e mestre em Estatística pela Universidade Federal do Rio de Janeiro. Professora da Pós-graduação e da Graduação da Escola Nacional de Ciências Estatísticas (Ence), do IBGE e coordenadora de Graduação em Estatística da mesma.

José André de Moura Brito é pós-doutor em Otimização pela Universidade Federal Fluminense, doutor e mestre em Engenharia de Sistemas e Computação (Otimização) pela Universidade Federal do Rio de Janeiro (UFRJ). Professor da Pós-graduação e da Graduação da Escola Nacional de Ciências Estatísticas (Ence), do IBGE.

Endereço para correspondência

Luciano Tavares Duarte
Av. Dom Helder Câmara, 6001, Bloco 6, Apto. 1402, Pilares
20771-002 – Rio de Janeiro-RJ, Brasil

Denise Britz do Nascimento Silva
Escola Nacional de Ciências Estatísticas
Rua André Cavalcanti, 106, sala 401, Bairro de Fátima
20231-050 – Rio de Janeiro-RJ, Brasil
José André de Moura Brito
Escola Nacional de Ciências Estatísticas
Rua André Cavalcanti, 106, sala 503-C, Bairro de Fátima
20231-050 – Rio de Janeiro-RJ, Brasil

Abstract

Paradata analysis of the 2010 Population Census: investigation of factors associated with nonsampling errors in the data collection stage

The relevance of a population census for a national statistical system is undeniable for its thematic and territorial coverage. Nonetheless, the complexity and size of a census operation lead to challenges for ensuring timeliness and quality of the results. This paper presents potential factors associated with non sampling errors detected in the data collection stage based on the analysis of Brazilian 2010 Population Census microdata and paradata. Data obtained from the field work monitoring system, called paradata, is used to provide information about divergences observed between data collected by enumerators and supervisors, also it is used the census microdata. The latter carried out follow-up interviews in households selected by the supervision/monitoring plan. Human resources databases containing socio-demographic information of enumerators and supervisors is also brought to enhance the analysis. The statistical modeling utilized is generalized hierarchical models, in which the response variable is defined as the occurrence of a discrepancy (or divergence) between the information collected by enumerators and their supervisors. The results indicate that the different hierarchical levels investigated are relevant to decompose data variability and hence have to be considered in the analysis. However, respondents' characteristics have markedly more influence on the chances of a divergence than those of enumerators' and supervisors'. In addition, there is evidence that respondents who are male, illiterate (or with low educational level), 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.

Keywords: Brazilian Population Census. Paradata. Hierarchical models. Nonsampling errors.

Resumen

Análisis de los paradatos del Censo Demográfico 2010: una investigación de los factores asociados a errores no muestrales en la etapa de relevamiento de los datos

La relevancia de un censo para el sistema de estadísticas públicas de una nación es indiscutible desde el punto de vista de su cobertura temática y territorial. Por otra parte, su complejidad y dimensión conducen a desafíos para garantizar la calidad de sus resultados. Este artículo tiene como objetivo presentar los posibles factores asociados a errores no muestrales detectados durante el relevamiento de los datos, mediante el análisis de los paradatos y microdatos del Censo Demográfico brasileño de 2010. Los paradatos se refieren a informaciones sobre la

operación de relevamiento y la administración de la investigación originarias, respectivamente, del sistema de gestión de recursos humanos del personal dedicado a la recogida y del sistema de supervisión de la operación de recolección. Este estudio analizó las divergencias observadas entre las informaciones recogidas por los encuestadores y las informaciones obtenidas por los supervisores en las reentrevistas de los procedimientos de supervisión del trabajo de campo. Para el análisis de las divergencias entre informaciones recogidas por los encuestadores y supervisores se utilizaron modelos jerárquicos generalizados. El estudio muestra que hay diferencias en las discordancias asociadas con la estructura de relevamiento de los datos, con las características de los encuestadores, supervisores e informantes, y revelan diferencias regionales. Queda evidente, sobretodo, una fuerte influencia de las características del informante en las posibilidades de ocurrencia de divergencias, en detrimento a las características de los supervisores y encuestadores. Los resultados del modelo estadístico sugieren que las entrevistas realizadas con informantes del sexo masculino, analfabetos o con bajo nivel educativo, mayores y que viven en hogares con indicadores que reflejan condiciones de vida menos satisfactorias, presentan chances adicionales en favor de la ocurrencia de divergencia entre las respuestas recogidas por el encuestador y el supervisor.

Palabras-clave: Censo Demográfico brasileiro. Paradatos. Modelos Jerárquicos. Errores no muestrales.

Recebido para publicação em 02/02/2016
Recomendado para publicação em 22/08/2016
Aceito para publicação em 01/12/2016