ANÁLISE DO USO DO INFORMANTE *PROXY* NA PNAD CONTÍNUA[1]

Elizabeth Belo Hypólito[2]
Denise Britz do Nascimento Silva[3]

1 INTRODUÇÃO

Embora pouco frequentes no Brasil, estudos sobre erros não amostrais têm grande valor, pois, como destacado por Biemer (2010), são importantes para a compreensão da incerteza nas estimativas, interpretação dos resultados e construção da confiança e da credibilidade da pesquisa.

Uma das possíveis fontes de erro não amostral em pesquisas domiciliares é a realização de entrevistas de forma indireta, ou seja, por meio de um informante *proxy*, o qual responde às perguntas do questionário em nome da pessoa selecionada. O senso comum sugere que não há ninguém melhor do que a própria pessoa para reportar seus dados e suas opiniões, exceto nos casos de crianças e pessoas com doenças físicas ou mentais que as impeçam de responder (Moore, 1988). Assim, a entrevista direta é sempre preferível. No entanto, o uso de informante *proxy* é utilizado em muitas pesquisas para aumentar a taxa de resposta, uma vez que nem todas as pessoas selecionadas podem ser facilmente contatadas ou estão disponíveis para participar (Biemer e Lyberg, 2003).

Pesquisas de força de trabalho, por disporem de prazos relativamente curtos para a coleta e divulgação das informações, além de orçamentos limitados, costumam utilizar amplamente esse recurso. No Brasil, mais de 50% das entrevistas da Pesquisa Nacional por Amostra de Domicílios (PNAD) Contínua, principal fonte de informações conjunturais sobre a inserção da população no mercado de trabalho, são realizadas por *proxy* (Hypólito, 2020). Internacionalmente, cerca de 65,0% das informações da pesquisa canadense sobre força de trabalho são coletadas por *proxy* (Statistics Canada, 2018). Em 2018, esse percentual foi superior a 50,0% em seis dos 35 países reportados no relatório de qualidade das pesquisas a respeito da força de trabalho da União Europeia, Croácia (51,5%), Espanha (51,0%), Eslováquia (51,1%), Eslovênia (54,2%), Macedônia do Norte (53,8%) e Sérvia (53,1%) (Eurostat, 2020).

Embora seja um importante recurso para a redução da não resposta, o uso do informante *proxy* levanta alguns questionamentos sobre o grau de confiabilidade dos dados coletados. É razoável que, mesmo o *proxy* com amplo conhecimento sobre a pessoa selecionada, não saiba responder todas as informações requisitadas pela pesquisa e, consequentemente, contribua para

1. DOI: http://dx.doi.org/10.38116/bmt72/nt4

2. Professora e pesquisadora na Escola Nacional de Ciências Estatísticas do Instituto Brasileiro de Geografia e Estatística (Ence/IBGE). *E-mail*: <bhypolito@gmail.com>.

3. Professora e pesquisadora da Ence/IBGE. *E-mail*: <denisebritz@gmail.com>.

o aumento da não resposta de item e do erro de medida. Define-se que há não resposta de item quando a informação referente a uma pergunta específica do questionário está ausente por recusa ou desconhecimento do informante, porque foi perdida durante a captura dos dados ou até mesmo apagada durante a etapa de crítica por conter valores inconsistentes. O erro de medida ocorre quando há um desvio entre a informação verdadeira e a que é registrada pela pesquisa.

Ao longo das últimas décadas, diversos estudos foram realizados sobre os possíveis efeitos do uso de entrevistas indiretas na qualidade dos dados coletados, especialmente no que se refere a variáveis numéricas e a perguntas que envolvem percepção, interpretação ou respostas abertas que necessitam de codificação *a posteriori*. Wolfgang, Byrne e Spratt (2003), Dawe e Knight (1997), Coder e Feldman (1984), Coder (1980) sugerem que o uso de *proxy* contribui para o aumento da não resposta de item. Thomsen e Villund (2011), Wolfgang, Byrne e Spratt (2003), Dawe e Knight (1997) e Boehm (1989) indicam que o *proxy* pode aumentar o erro de medida. Contudo, Biggs (1992) e Moore (1988), em duas importantes revisões de literatura, apontam que a maioria dos estudos realizados sobre o tema apresentam limitações e, portanto, não devem ser considerados conclusivos em relação à diferença de qualidade entre respostas obtidas de forma direta e indireta.

Considerando a relevância do tema para a produção de estatísticas oficiais, assim como a escassez de estudos realizados no Brasil, este artigo tem como objetivo central ampliar o conhecimento sobre o uso entrevistas indiretas na PNAD Contínua, analisando a taxa de *proxy* da pesquisa e buscando compreender seus impactos sobre a qualidade dos dados coletados.

2 FONTE DE DADOS

A PNAD Contínua, implantada pelo IBGE em 2012, produz indicadores trimestrais sobre a inserção da população na força de trabalho no Brasil e em suas principais regiões administrativas. Por meio de trimestres móveis, o instituto produz um conjunto restrito de indicadores mensais para o Brasil. Além disso, fornece informações anuais para outros temas, como educação, características gerais dos moradores, acesso à internet, rendimentos de todas as fontes e outras formas de trabalho – como trabalho infantil, trabalho para o próprio consumo –, entre outros temas permanentes.

O plano amostral da pesquisa é conglomerado em dois estágios, com estratificação das unidades primárias de amostragem (UPA).[4] Além disso, possui um esquema de rotação de domicílios, no qual cada domicílio selecionado permanece na amostra por cinco trimestres, sendo os seus moradores entrevistados uma vez a cada trimestre. Em todos os trimestres, há domicílios na primeira, segunda, terceira, quarta e quinta entrevista e, de um trimestre para

4. A UPA é definida como um setor censitário ou um conjunto de setores censitários contendo no mínimo sessenta domicílios particulares permanentes (DPP). O IBGE (2010) define o setor censitário como "uma área contínua, contida em área urbana ou rural, cuja dimensão, número de domicílios e de estabelecimentos permitem ao recenseador cumprir suas atividades em um prazo determinado, respeitando o cronograma de atividades".

o seguinte, há sobreposição de 80% dos domicílios. Mais informações podem ser obtidas em IBGE (2014; 2020a; 2020b).

Para lançar luz às questões de interesse deste estudo, foram utilizados microdados da PNAD Contínua, os quais podem ser obtidos na página do IBGE, bem como algumas variáveis do processo de pesquisa. Essas últimas não são divulgadas pelo instituto e foram solicitadas por meio de processo específico para acesso a dados individualizados não desidentificados para uso em programas de pós-graduação por funcionários do IBGE.[5] As variáveis foram as seguintes: o informante de cada parte da pesquisa; no caso de informante *proxy* que era morador do domicílio, o seu número de ordem; a marca de imputação das variáveis de trabalho e de rendimentos de trabalho; a chave longitudinal para pareamento de pessoas; e a identificação do entrevistador.

É importante enfatizar que a PNAD Contínua é uma pesquisa por amostra de domicílios e não de pessoas. Assim, o domicílio é entrevistado cinco vezes, mesmo quando há alteração parcial ou total do conjunto de pessoas nele residente. No entanto, pessoas que permanecem no mesmo endereço durante o período de realização da pesquisa podem ser acompanhadas por meio de uma chave construída com base no nome da pessoa selecionada, o número de ordem na lista de residentes do domicílio, o sexo e a data de nascimento.

Todas as análises foram realizadas para o conjunto das pessoas de 14 anos ou mais de idade, as quais constituem a população em idade de trabalhar definida pela pesquisa. Além disso, elas são, a princípio, consideradas capazes de responder suas próprias informações sem a necessidade imediata de um informante *proxy*. Em relação ao período, foram escolhidos o segundo e o terceiro trimestres de 2017.

Cabe destacar que este estudo tem como foco a qualidade dos dados coletados pela PNAD Contínua, considerando o escopo das unidades amostrais. Nenhuma inferência foi realizada para a população de pesquisa. Portanto, o desenho amostral não foi levado em conta nas análises.

3 TAXA DE *PROXY*

No manual de entrevista da PNAD Contínua, destaca-se que, para os casos em que não é possível realizar a entrevista diretamente com a pessoa selecionada, "o entrevistador deve, criteriosamente, verificar se, dentre os moradores presentes, existem pessoas capacitadas a prestar, com segurança, as informações referentes aos ausentes ou que possam contatá-los para obtê-las" (IBGE, 2016, p. 22).

No fim do questionário, há um quesito destinado ao entrevistador, no qual deve ser registrado quem foi o informante: própria pessoa, *proxy* morador do domicílio e *proxy* não morador. Quando o informante é um *proxy* morador, o entrevistador registra, também, o número de ordem dessa pessoa na lista de moradores do domicílio, permitindo identificá-la e caracterizá-la. Para o *proxy* não morador, nenhuma informação adicional é coletada.

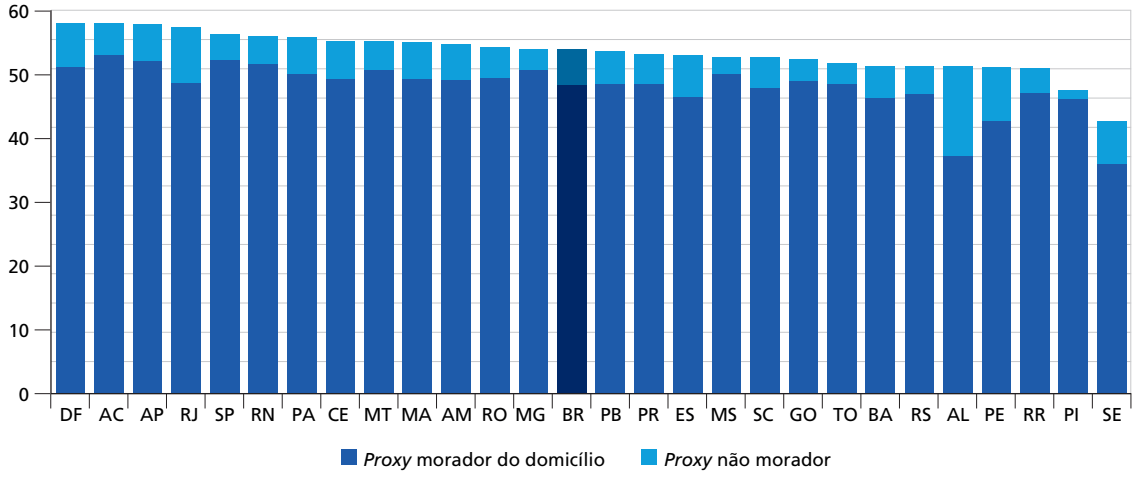
5. Os dados foram usados na tese de doutorado de Hypólito (2020), a qual deu origem a este artigo.

O primeiro passo para a compreensão da dinâmica do uso de informante *proxy* foi calcular a taxa de ocorrência desse fenômeno. Ela foi definida como o percentual de entrevistas realizadas indiretamente, ou seja, que foram respondidas por *proxy*, sobre o total de entrevistas com identificação de informante.

No terceiro trimestre de 2017, a PNAD Contínua coletou informações sobre 456.138 pessoas com 14 anos ou mais de idade. Dessas, 48,3% tiveram suas entrevistas respondidas por *proxy* morador e 5,5% por *proxy* não morador do domicílio, levando a uma taxa de *proxy* total de 53,8%. Em relação às Unidades da Federação (UFs), apenas duas delas apresentaram taxas totais inferiores a 50%, Piauí, 47,5%, e Sergipe, 42,6% (gráfico 1).

Chama atenção o fato de 14,1% das entrevistas de Alagoas terem sido realizadas com informante *proxy* não morador do domicílio selecionado (gráfico 1). Infelizmente, a pesquisa não capta nenhuma informação sobre esse informante. Identificá-lo em categorias como parente da pessoa de referência, empregado doméstico, vizinho etc., poderia trazer ganhos aos estudos de qualidade e permitir ações de melhoria da coleta. Um exemplo desse tipo de categorização pode ser encontrado na pesquisa de avaliação do Censo Americano de 2000 (Wolfgang, Byrne e Spratt, 2003).

GRÁFICO 1
Taxa de *proxy* das entrevistas de trabalho e rendimento por UF, segundo o tipo de informante *proxy* – Brasil (2017)[1]
(Em %)



Fonte: PNAD Contínua/IBGE.
Elaboração das autoras.
Nota: [1] Terceiro trimestre de 2017.

A taxa variou consideravelmente em relação às características das pessoas selecionadas, sendo mais elevada para os homens e as pessoas mais jovens, em especial os filhos das pessoas de referência. Por exemplo, para os homens que eram cônjuges, a taxa foi igual a 75,1%; e para as mulheres nessa condição, 44,3% (tabela 1).

TABELA 1
Taxa de *proxy* das entrevistas de trabalho e rendimento por sexo, segundo a condição no domicílio – Brasil (2017)[1]
(Em %)

Condição do informante *proxy* no domicílio	Taxa de *proxy*		
	Total	Homens	Mulheres
Pessoa de referência	**35,6**	43,9	24,0
Cônjuge	**52,6**	75,1	44,3
Filho de 14 a 17 anos	**82,2**	84,1	80,1
Filho de 18 a 23 anos	**77,5**	80,5	73,7
Filho maior de 23 anos	**74,8**	78,2	70,2
Outro parente	**76,9**	80,5	73,8
Outra condição	**73,0**	74,2	71,8

Fonte: PNAD Contínua/IBGE.
Elaboração das autoras.
Nota: [1] Terceiro trimestre de 2017.

No que diz respeito à relação entre a pessoa selecionada que teve a entrevista respondida de forma indireta e o correspondente informante *proxy*, morador do domicílio, observou-se que 44,0% dessas entrevistas correspondiam à pessoa de referência respondendo por seu cônjuge e seus filhos. Outros 26,9% das entrevistas eram referentes a cônjuges respondendo pela pessoa de referência e pelos filhos desta (tabela 2). Esses resultados sugerem que a escolha do informante, na maioria das vezes, atendeu aos requisitos do manual, ou seja, o *proxy* foi uma pessoa que, aparentemente, tinha capacidade para responder pelos demais moradores.

TABELA 2
Taxa de *proxy* das entrevistas de trabalho e rendimento, segundo a condição no domicílio – Brasil (2017)[1]
(Em %)

Condição do informante *proxy* no domicílio	Condição da pessoa que teve suas informações fornecidas por *proxy* morador do domicílio				
	Pessoa de referência	Cônjuge	Filho	Outro parente	Outra condição
Pessoa de referência	-	22,0	22,0	7,5	0,5
Cônjuge	16,6	-	10,3	1,7	0,1
Filho	5,9	3,1	4,2	1,3	0,1
Outro parente	1,7	0,4	0,7	1,4	0,0
Outra condição	0,1	0,0	-	0,0	0,1

Fonte: PNAD Contínua/IBGE.
Elaboração das autoras.
Nota: [1] Terceiro trimestre de 2017.

4 IMPACTOS DO USO DO INFORMANTE *PROXY* NA QUALIDADE DAS INFORMAÇÕES PESQUISADAS

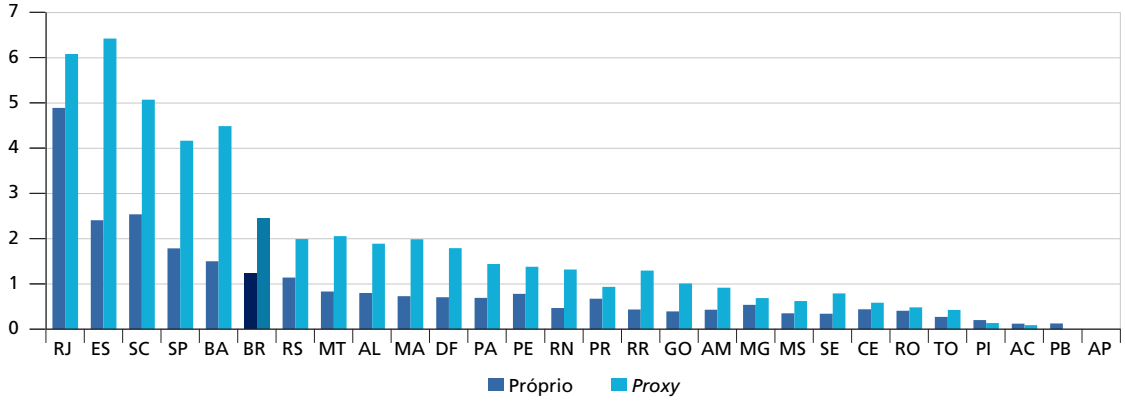
O passo seguinte foi avaliar os impactos do uso do informante *proxy* na qualidade das informações coletadas, observando se a não resposta de item e o erro de medida foram mais elevados para os registros obtidos por entrevista indireta do que para os registros obtidos com a própria pessoa.

Na PNAD Contínua, a não resposta de item é tratada por meio de imputação. Em geral, por abordarem informações sensíveis, os quesitos de rendimentos costumam apresentar taxas de imputação superiores as de outros itens da pesquisa. Além disso, por serem dados quantitativos, podem apresentar uma dificuldade adicional para o informante *proxy*. Portanto, espera-se que a não resposta de item e, consequentemente, a taxa de imputação de rendimentos seja mais elevada quando a resposta é obtida de forma indireta.

Para verificar essa hipótese, primeiramente, calculou-se a taxa de imputação dos quesitos de rendimento habitual do trabalho principal. A taxa foi definida como o número de pessoas ocupadas recebendo pagamento em dinheiro pelo trabalho principal e que tiveram seus registros de rendimento com marca de imputação sobre o número de pessoas ocupadas recebendo pagamento em dinheiro pelo trabalho principal.

A taxa de imputação da variável de rendimento habitual em dinheiro do trabalho principal das pessoas ocupadas tem sido consideravelmente pequena desde o início da pesquisa. No terceiro trimestre de 2017, foram imputados 4.344 registros de um total de 227.592, levando a uma taxa de 1,9%. Embora pequena, a taxa variou bastante quando comparados os valores calculados para as pessoas que responderam diretamente à pesquisa e para aquelas que tiveram suas informações obtidas indiretamente. Enquanto para o primeiro grupo ela foi de 1,2%; para o grupo com respostas fornecidas por *proxy* morador, ela foi igual a 2,3%; e para o grupo com respostas de *proxy* não morador, 3,5%. Para o total de entrevistas indiretas, a taxa de imputação foi igual a 2,4% (gráfico 2).

GRÁFICO 2
Taxa de imputação do rendimento habitual, em dinheiro, do trabalho principal, por UF, segundo o tipo de informante – Brasil (2017)[1]
(Em %)



Fonte: PNAD Contínua/IBGE.
Elaboração das autoras.
Nota: [1] Terceiro trimestre de 2017.

Outro possível efeito do uso do *proxy* na qualidade dos dados coletados seria o erro de medida, gerando divergências entre respostas fornecidas para duas edições consecutivas da pesquisa. Por exemplo, suponha que, por desconhecimento da informação, um *proxy* emite uma resposta incorreta para um determinado item e, na edição seguinte, a pesquisa é respondida pela própria pessoa selecionada, que fornece a informação verdadeira. Logo, haveria uma inconsistência entre as duas respostas, gerada por um erro de medida na primeira ocasião.

Espera-se que a divergência seja menor nos casos em que a própria pessoa responde as informações nas duas edições da pesquisa. Ademais, acredita-se que, para os casos em que há alteração de informantes entre os trimestres, ou seja, para *proxies* diferentes ou próprio em uma entrevista e *proxy* na outra, a taxa de divergência seja mais elevada do que nos casos em que o mesmo *proxy* ou a própria pessoa selecionada responde as duas entrevistas.

Considerou-se como divergentes os casos em que, para uma mesma pessoa selecionada, a resposta para um determinado quesito no segundo trimestre foi diferente da resposta fornecida para esse quesito no terceiro trimestre. A taxa de divergência foi definida como o total de registros divergentes sobre o total de registros com respostas nos dois trimestres. Essa foi calculada para as seguintes variáveis: cor ou raça, nível de instrução, posição na ocupação, código de ocupação e código de atividade.

Como o propósito é identificar a possível ocorrência de erro de medida em ao menos uma das duas entrevistas, alguns filtros foram aplicados a essas variáveis, evitando assim que alterações reais nos dados fossem classificadas como divergências. Para o nível de instrução, foram considerados apenas os casos em que a pessoa não frequentava escola em nenhuma das duas ocasiões, mas havia frequentado anteriormente, visto que, para esse grupo, uma mudança de curso no período de três meses entre as entrevistas seria pouco provável. Para as variáveis de trabalho, foram analisadas apenas as divergências que ocorreram nas informações de pessoas ocupadas em um único trabalho nos dois trimestres e que, no terceiro trimestre, tinham declarado estar naquele trabalho há mais de três meses, o que indica que se trata do mesmo trabalho.

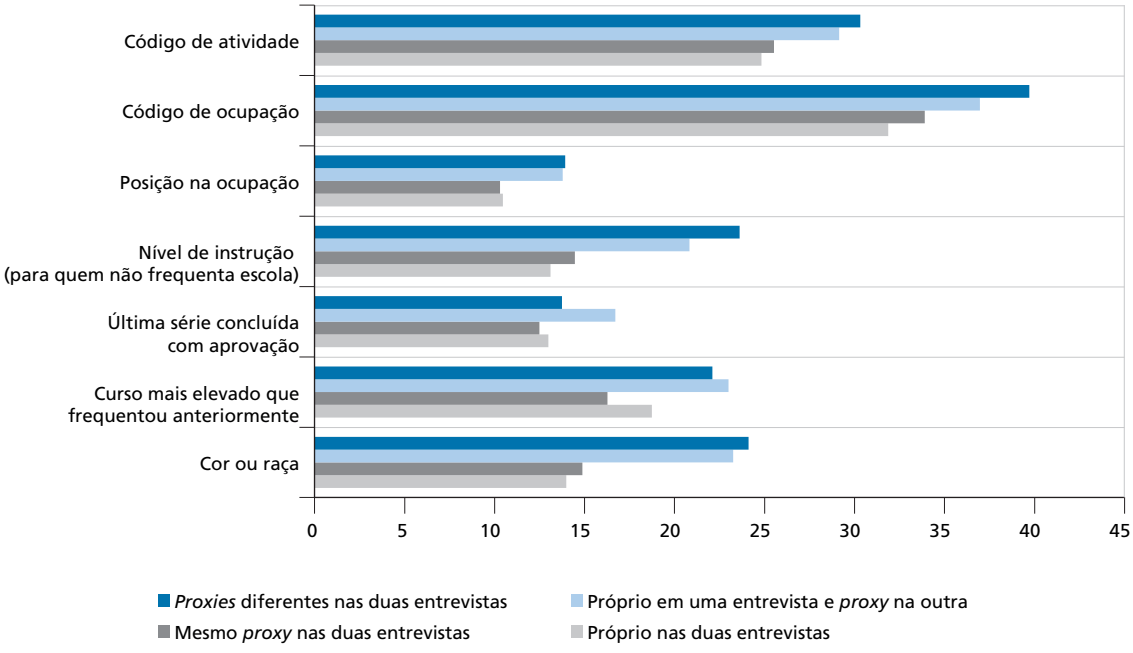
A taxa de divergência foi calculada para o número total de entrevistas e por tipo de informante: própria pessoa nos dois trimestres; mesmo informante *proxy* morador do domicílio; diferentes *proxies* moradores do domicílio; e própria pessoa em um trimestre e *proxy* no outro ou vice-versa. Os valores foram mais elevados quando informantes diferentes (*proxies* diferentes ou *proxy* em uma entrevista e próprio na outra) forneceram as respostas do que quando a mesma pessoa respondeu as duas entrevistas (a própria pessoa ou o mesmo *proxy*). De forma geral, as maiores taxas ocorreram quando *proxies* diferentes responderam as duas entrevistas (gráfico 3).

As variáveis de ocupação e atividade apresentaram as taxas mais elevadas entre todas as analisadas. Esse resultado é compreensível visto que o processo de captação desses dados pode envolver diversos atores. Primeiramente, o informante descreve com suas palavras a

informação que lhe é solicitada. O entrevistador escuta, interpreta e tenta realizar a codificação ainda durante o momento da coleta. Caso não encontre o código adequado ou tenha dúvidas, escreve o que ouviu. Nessa situação, a codificação é feita no escritório, podendo ser realizada por outro profissional.

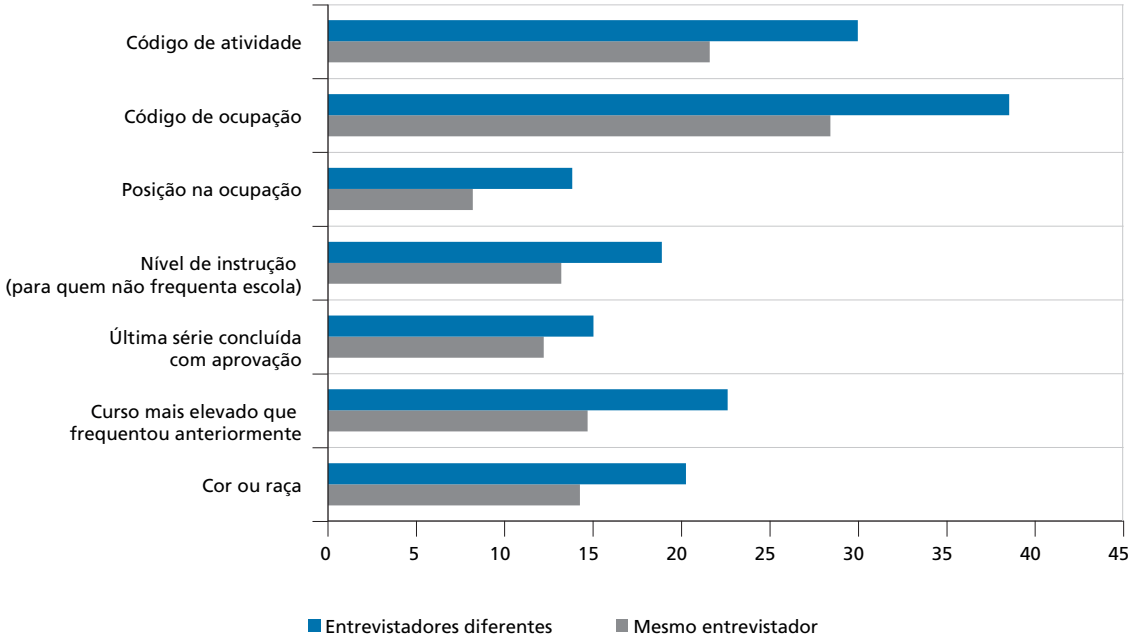
Chama atenção a divergência ocorrida quando a própria pessoa respondeu nos dois trimestres, a qual foi superior a 11,0% para todas as variáveis. Esses resultados sugerem que o tipo de informante não é a única variável que influencia na divergência dos dados. Dessa forma, outra possível fonte de variação associada à coleta, quer seja, a mudança de entrevistadores entre os dois trimestres, também foi avaliada. Para todas as variáveis, os valores obtidos para os casos em que houve alteração de entrevistadores foram superiores aos obtidos quando o entrevistador foi o mesmo (gráfico 4).

GRÁFICO 3
Taxa de divergência entre entrevistas pareadas, por variáveis da pesquisa, segundo o tipo de informante – Brasil (2017)[1]
(Em %)



Fonte: PNAD Contínua/IBGE.
Elaboração das autoras.
Nota: [1] Segundo e terceiro trimestres de 2017.

GRÁFICO 4
Taxa de divergência entre entrevistas pareadas, por variáveis da pesquisa, segundo o tipo de entrevistador – Brasil (2017)[1]
(Em %)



Fonte: PNAD Contínua/IBGE.
Elaboração das autoras.
Nota: [1] Segundo e terceiro trimestres de 2017.

Para compreender os efeitos conjuntos de respondentes e entrevistadores, foram ajustados modelos logísticos (Dobson e Barnett, 2008) para estimar a probabilidade de ocorrência de divergência em três variáveis, cor ou raça, nível de instrução e código de ocupação. Aqui serão apresentados apenas os resultados para o modelo referente ao código na ocupação. Os demais resultados podem ser obtidos em Hypólito (2020).

A variável resposta do modelo foi definida como 1, se houve divergência; e 0, caso contrário. Como variáveis explicativas para o fenômeno, consideraram-se: o tipo de informante (própria pessoa nos dois trimestres; mesmo informante *proxy* morador do domicílio; diferentes *proxies* moradores do domicílio; e própria pessoa em um trimestre e *proxy* no outro ou vice-versa); o indicador de troca de entrevistadores entre as duas ocasiões da pesquisa (1 se houve troca e 0, caso contrário); os grupamentos de UF;[6] o nível de instrução; e a posição na ocupação.

A estratégia usada para a modelagem foi a inserção das variáveis explicativas uma a uma. O modelo com uma variável a mais foi comparado ao modelo imediatamente anterior pelo teste da diferença das *deviances* e, também, pelo AIC. Todos os testes realizados durante o processo de modelagem usaram nível de significância igual a 5%. Os modelos foram ajustados no *software* R.

6. Como os coeficientes estimados para as UFs resultaram em valores muito próximos entre si, foi necessário agrupá-las de acordo com a semelhança em relação à taxa de divergência. Para mais detalhes, ver Hypólito (2020).

A categoria de referência foi definida como pessoas moradoras nas UFs do grupo 1, ou seja, Rio de Janeiro, Piauí, São Paulo, Minas Gerais, Rio Grande do Sul, Amazonas, Paraíba, Amapá, Mato Grosso do Sul, Sergipe, Pará e Pernambuco, com escolaridade até nível médio incompleto, empregadas ou conta própria, que responderam as próprias informações nas duas entrevistas, as quais foram conduzidas pelo mesmo entrevistador. Com exceção do intercepto, os coeficientes estimados foram todos positivos, indicando que as chances de divergência aumentam para a alteração de qualquer uma das características do grupo de referência, desde que fixadas as demais covariáveis (tabela 3).

Por exemplo, residir no Mato Grosso, UF que compõe sozinha o grupo 5, gera um aumento médio de 84,8% nas chances de divergência em comparação a residir no grupo 1; ter as entrevistas respondidas por *proxies* diferentes resulta em aumento médio de 35,7% em comparação a ter as entrevistas respondidas pela própria pessoa; ter as entrevistas realizadas por entrevistadores diferentes aumenta as chances em 43,4% em comparação a tê-las realizadas pelo mesmo entrevistador; ser um empregador amplia as chances de divergência em 77,8% em relação a ser empregado ou conta própria; e ter nível de instrução superior completo aumenta as chances em 53,0% em relação a não ter instrução ou ter até o nível médio incompleto (tabela 3).

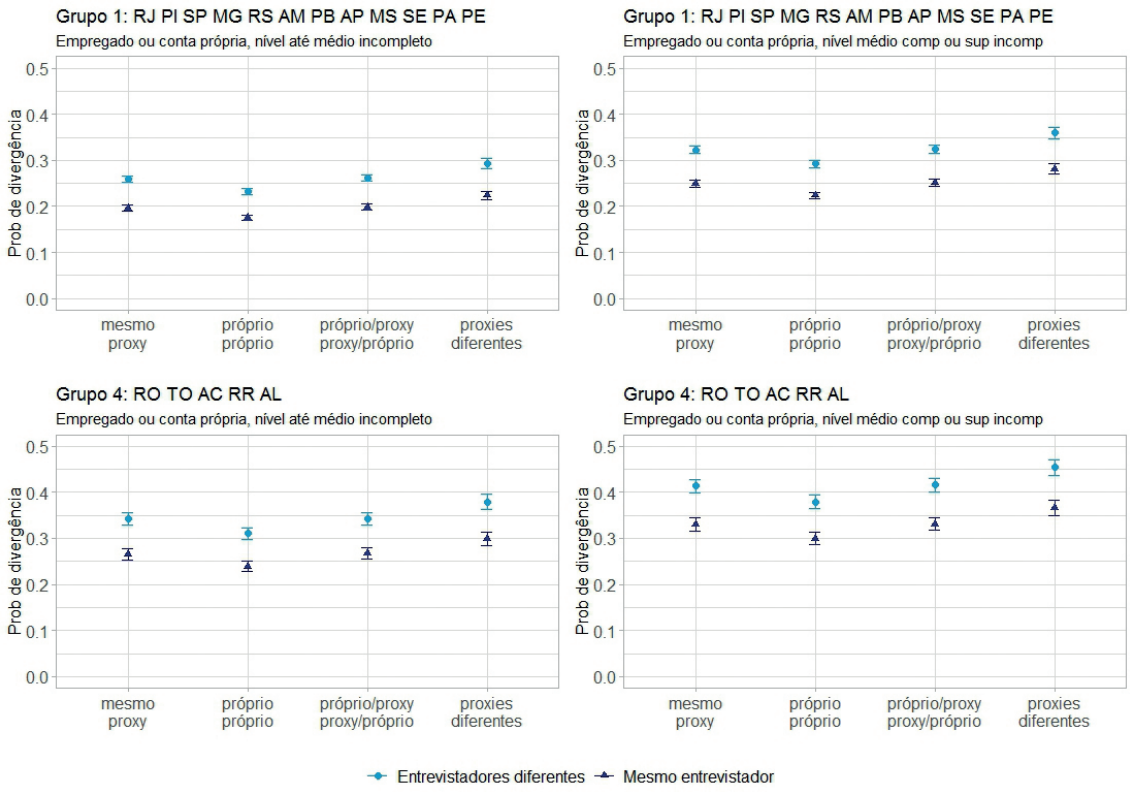
TABELA 3
Coeficientes estimados e razões de chances do modelo ajustado para a ocorrência de divergência no código de ocupação (2017)[1]

		Coeficiente (*c*)	Desvio- padrão	Teste de Wald		100* {exp(c)-1}
				z	*p*-valor	
	Intercepto	-1,55	0,02	-77,68	0,00	-78,80
Grupo de UFs	1: RJ PI SP MG RS AM PB AP MS SE PA PE (referência)	–	–	–	–	–
	2: ES DF SC BA PR RN	0,24	0,02	13,50	0,00	26,60
	3: GO MA CE	0,30	0,02	12,92	0,00	35,60
	4: RO TO AC RR AL	0,39	0,03	13,67	0,00	48,00
	5: MT	0,61	0,05	13,19	0,00	84,80
Informante	Próprio (referência)	–	–	–	–	–
	Mesmo *proxy* morador	0,14	0,02	7,60	0,00	15,30
	Proxies diferentes	0,31	0,03	11,59	0,00	35,70
	Próprio e *proxy*	0,15	0,02	7,74	0,00	16,20
	Proxy não morador	0,15	0,06	2,72	0,01	16,20
Entrevistador	Mesmo entrevistador (referência)	–	–	–	–	–
	Entrevistadores diferentes	0,36	0,02	23,35	0,00	43,40
Posição na ocupação	Empregado ou conta própria (referência)	–	–	–	–	–
	Empregador	0,58	0,03	16,81	0,00	77,80
	Trabalhador familiar auxiliar	0,10	0,05	2,16	0,03	10,80
Nível de instrução	Sem instrução até médio incompleto (referência)	–	–	–	–	–
	Médio completo ou superior incompleto	0,31	0,02	18,50	0,00	35,80
	Superior completo	0,43	0,02	21,18	0,00	53,00

Fonte: PNAD Contínua/IBGE.
Elaboração das autoras.
Nota: [1] Segundo e terceiro trimestres de 2017.

O gráfico 5 compara as probabilidades preditas de divergência para empregados ou conta própria que tinham nível de instrução até ensino médio incompleto com os que tinham nível médio completo ou superior incompleto, em dois grupos de UFs, o de referência e o grupo 4, composto por Rondônia, Tocantins, Acre, Roraima e Alagoas. Considerando o mesmo grupo e o mesmo par de informantes, as probabilidades estimadas foram mais elevadas para aqueles com nível médio completo ou superior incompleto, cujas entrevistas foram conduzidas por entrevistadores diferentes, enquanto as mais baixas foram estimadas para aqueles sem instrução ou com nível até ensino médio incompleto, que responderam as próprias informações para o mesmo entrevistador nas duas ocasiões. Na comparação entre os grupos de UFs, podemos observar que, para a mesma posição na ocupação, nível de escolaridade, par de informante e de entrevistadores, as probabilidades foram sempre mais elevadas para as UFs do grupo 4.

GRÁFICO 5
Probabilidades estimadas para a divergência na variável código de ocupação (2017)[1]



Fonte: PNAD Contínua/IBGE.
Elaboração das autoras.
Nota: [1] Segundo e terceiro trimestres de 2017.

5 CONSIDERAÇÕES FINAIS

Os resultados mostraram que a maior parte das entrevistas das pessoas de 14 anos ou mais de idade da PNAD Contínua foi realizada indiretamente. Embora a taxa de *proxy* seja bastante elevada, ela é consistente com as taxas observadas em algumas pesquisas internacionais de força de trabalho.

No que diz respeito às possíveis consequências do uso do informante *proxy*, a taxa de imputação foi consideravelmente baixa para a variável de rendimento habitual do trabalho, uma das mais sensíveis da pesquisa. Ainda assim, para informações obtidas com *proxy* moradores, a taxa foi o dobro da calculada para respostas fornecidas pela própria pessoa. Para o caso de *proxy* não moradores, ela foi o triplo desse valor. É importante destacar que o fato da imputação ser baixa, isoladamente, não garante a qualidade dos dados coletados. O informante pode não saber a informação solicitada e ainda assim fornecer uma resposta nos valores válidos para a pesquisa, porém equivocada.

Desse modo, o trabalho também abordou a divergência entre informações obtidas para o mesmo quesito em dois trimestres consecutivos. As divergências foram mais elevadas quando houve alternância de informantes entre os trimestres do que quando a própria pessoa ou o mesmo *proxy* respondeu nas duas ocasiões. Esse resultado é bastante coerente uma vez que, no caso de próprio e *proxy*, é esperado que o primeiro saiba mais sobre si mesmo do que qualquer outro informante e que, portanto, haja um percentual de divergências razoável quando cada um responde em um trimestre. No caso de *proxys* diferentes, eles podem ter graus de conhecimento distintos sobre a informação solicitada. Por exemplo, entre um cônjuge e um filho de 17 a 23 anos, é possível que o primeiro tenha mais conhecimento sobre as características de trabalho da pessoa de referência.

O tipo de informante não foi a única variável de coleta que influenciou na divergência. Para todas as variáveis analisadas, a taxa foi mais elevada quando houve troca de entrevistadores. Para avaliar o efeito conjunto de informantes e entrevistadores, foram utilizados modelos logísticos. Para a variável código da ocupação, as probabilidades mais elevadas foram estimadas para os casos em que houve alternância de informantes e de entrevistadores. As mais baixas foram para os casos em que a própria pessoa respondeu nos dois trimestres, diante do mesmo entrevistador.

Embora os resultados apresentados neste estudo indiquem que o uso de informante *proxy* aumente a não resposta de item e o erro de medida, a coibição de entrevistas indiretas não seria uma medida cabível neste momento. Isso porque, tendo em vista o curto prazo de coleta e divulgação da PNAD Contínua, além das restrições orçamentárias comuns à realização de pesquisas domiciliares, a exigência de que a entrevista seja realizada com a própria pessoa poderia aumentar consideravelmente a não resposta de unidades. Além disso, considerando que algumas pessoas são mais difíceis de serem localizadas do que outras, essa medida poderia introduzir vieses na pesquisa. Dessa forma, outras alternativas devem ser consideradas para a redução dos erros não amostrais.

Uma alternativa para mitigar os efeitos do uso do *proxy* nos códigos de ocupação e de atividade seria a investigação dependente, a qual pode ser proativa ou reativa. Na primeira, as

respostas obtidas no passado são previamente fornecidas ao informante durante a entrevista corrente. Na segunda, a pergunta é feita e, caso a resposta atual difira consideravelmente da passada, a informação anterior é fornecida para verificação. Cabe destacar que, a despeito do receio de muitos pesquisadores de que a investigação dependente, especialmente a proativa, possa gerar aquiescência por parte do informante ou fraude de dados por parte do entrevistador, estudos como os de Jäckle (2008), Lynn *et al*. (2006) e Sala e Lynn (2004) têm concluído que seu uso traz ganhos na qualidade de dados longitudinais.

A análise de experiências de institutos internacionais de estatística com a investigação dependente, assim como a realização de testes em campo, pode enriquecer a discussão sobre uma possível aplicação da técnica na PNAD Contínua. Por exemplo, a pesquisa americana sobre força de trabalho, *Current population survey*, utiliza a investigação dependente para informações referentes às características gerais dos moradores e do domicílio, além de variáveis de trabalho, como ocupação, atividade e duração da desocupação (US Census Bureau, 2006). Além dessa pesquisa, a do Reino Unido, *UK labour force survey*, também utiliza a técnica para uma gama de variáveis (Office for National Statistics, 2007).

Por fim, é importante destacar que os resultados obtidos, além de permitirem melhorias nas próprias pesquisas, podem ser de grande utilidade para a transparência e o aumento da confiança dos usuários nos dados produzidos. Para tal, podem ser disponibilizados em relatórios de qualidade com periodicidade a ser definida de acordo com a necessidade.

REFERÊNCIAS

BIEMER, P. P. Total survey error: design, implementation and evaluation. **Public Opinion Quarterly**, Oxford, 2010.

BIEMER, P. P.; LYBERG, L. E. **Introduction to survey quality**. New Jersey: John Wiley Sons, 2003.

BIGGS, B. **Self/proxy respondent rules and data quality**. Ottawa, 1992.

BOEHM, L. **Reliability of proxy response in the current population survey**. [S.l.], 1989.

CODER, J. **Some results from the 1979 income survey development program research panel**. American Statistical Association, Proceedings of the Survey Methods Research Section, p. 540-545, 1980.

CODER, J.; FELDMAN, A. **Early indications of item nonresponse on the survey of income and program participation**: proceedings of the survey methods research section. United States: American Statistical Association, 1984, p. 693-697.

DAWE, F.; KNIGHT, I. A study of proxy response in the labour force survey. **Survey Methodology Bulletin**, 1997.

DOBSON, A. J.; BARNETT, A. G. **An introduction to generalized linear models** (Third Edition). United States: CRCPress, 2008.

EUROSTAT. **Quality report of the European Union Labour Force Survey 2018-2020 edition**. [S.l.], 2020.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. **Journal of the Royal Statistical Society**, v. 28, n. 1, p. 100-108, 1979. Disponível em: <https://bit.ly/3CuFwWR>. Acesso em: 9 ago. 2018.

HYPÓLITO, E. B. **Erros não amostrais em pesquisas domiciliares**: impactos na qualidade. 2020. Tese (Doutorado) – Escola Nacional de Ciências Estatísticas (Ence), Rio de Janeiro, RJ, 2020.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo Demográfico 2010**: manual do recenseador. Rio de Janeiro: IBGE, 2010.

_____. **Pesquisa Nacional por Amostra de Domicílios Contínua**. Rio de Janeiro: IBGE, 2014. (Notas metodológicas).

_____. **Pesquisa Nacional por Amostra de Domicílios Contínua**: manual básico de entrevista. Rio de Janeiro: IBGE, 2016.

_____. **Pesquisa Nacional por Amostra de Domicílios Contínua**. Rio de Janeiro: IBGE, 2020a. (Notas técnicas, versão 1.7). Disponível em: <https://bit.ly/3CBgdTe>. Acesso em: 15 jul. 2020.

_____. **Pesquisa Nacional por Amostra de Domicílios (PNAD) Contínua**: informações referentes à coleta do mês de abril de 2020. Rio de Janeiro, 2020b. (Nota técnica). Disponível em: <https://bit.ly/3zqiRsO> . Acesso em: 15 jul. 2020.

JÄCKLE, A. Dependent interviewing: Effects on respondent burden and efficiency of data collection. **Journal of Official Statistics**, v. 24, n. 3, p. 441-430, 2008.

LYNN, P. *et al*. The effects of dependent interviewing on responses to questions on income sources. **Journal of Official Statistics**, v. 22, n. 3, p. 357-384, 2006.

MOORE, J. Self/proxy response status and survey response quality. **Journal of Official Statistics**, n. 4, p. 155-172, 1988.

OFFICE FOR NATIONAL STATISTICS. **Labour force survey user guide**: volume 1 – lfs background and methodology. [S.l.], 2007.

SALA, E.; LYNN, P. **Measuring change in employment characteristics**: the effects of dependent interviewing. United Kingdom: University of Essex, 2004. (Working Paper, n. 26).

STATISTICS CANADA. **Guide to the labour force survey 2018**. [S.l.], 2018.

THOMSEN, I.; VILLUND, O. Using register data to evaluate the effects of proxy interviews in the norwegian labour force survey. **Journal of Official Statistics**, 2011.

US CENSUS BUREAU – UNITED STATES CENSUS BUREAU. **Current population survey design and methodology**. [S.l.], 2006.

WOLFGANG, G.; BYRNE, R.; SPRATT, S. **Analysis of proxy data in the accuracy and coverage evaluation**. [S.l.], 2003.

Missão do Ipea

Aprimorar as políticas públicas essenciais ao desenvolvimento brasileiro por meio da produção e disseminação de conhecimentos e da assessoria ao Estado nas suas decisões estratégicas.

ISSN 1676-0883

ipea Instituto de Pesquisa Econômica Aplicada

MINISTÉRIO DA ECONOMIA

PÁTRIA AMADA BRASIL GOVERNO FEDERAL