EFEITOS DE VIÉS DE NÃO RESPOSTA NA EVOLUÇÃO DAS TAXAS DE OCUPAÇÃO E DE FORMALIZAÇÃO DA PNAD CONTÍNUA EM 2020 E 2021

Carlos Henrique Corseuil[1]

Gabriela Padilha[2]

Felipe Russo[3]

Resumo

As restrições de interação social causadas pela pandemia do novo Coronavírus obrigaram o Instituto Brasileiro de Geografia e Estatística (IBGE) a mudar seu método de realização de entrevistas em um momento crítico no mercado de trabalho brasileiro. Esse artigo visa primeiramente analisar o efeito dessa mudança no viés de não resposta proveniente de mudanças na composição da amostra da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua). Em seguida, propomos um método de correção para esse viés, que leva em conta tanto características observáveis como não observáveis para simular um contrafactual após o início da pandemia. Nós estimamos que a queda nas taxas de ocupação seria de 0,3 a 0,8 pontos percentuais (p.p.) a mais na ausência do viés. Para a taxa de formalização, o 2º trimestre de 2020 apresentaria uma queda 0,5 p.p. maior que a observada.

Palavras-Chave: viés de não resposta, população ocupada, formalização

Abstract

The restrictions on social interaction caused by the new Coronavirus pandemic forced the Brazilian Institute of Geography and Statistics (IBGE) to change its method of conducting interviews at a critical moment for the Brazilian job market. This article will first analyze the effect of this change in the non-response bias resulting from changes in the sample composition of the Continuous National Household Sample Survey (PNAD Contínua). We then propose a correction method for this bias that takes into account both observable and unobservable characteristics to simulate a counterfactual after the onset of the pandemic. We estimate that the decline in occupancy rates would be 0.3 to 0.8 percentage points (p.p.) greater in the absence of bias. For the formalization rate, the 2nd quarter of 2020 would show a drop of 0.5 p.p. greater than observed.

Keywords: non-response bias, employed population, formalization

Indicação de área ANPEC: Área 13 - Economia do Trabalho

Código JEL: C83 E24 E27

[1] Técnico de planejamento e pesquisa na Diretoria de Estudos e Políticas Sociais (Disoc) do Ipea.
[2] Assistente de pesquisa na Disoc/Ipea.
[3] Assistente de pesquisa na Disoc/Ipea.

1. Introdução

Em março de 2020, devido à pandemia de Covid-19 e seguindo as orientações do Ministério da Saúde, o Instituto Brasileiro de Geografia e Estatística (IBGE) interrompeu as entrevistas presenciais para a PNAD Contínua e passou a realizá-las através do telefone (IBGE, 2020a). Apesar da mudança na forma de coleta, o instituto conseguiu manter a divulgação dos dados da pesquisa sem grande perda de confiabilidade. Isso se manteve até julho de 2021, quando parte das entrevistas voltou a ser realizada presencialmente. Entretanto, segundo o próprio IBGE (2020b), o fato das entrevistas passarem a ser feitas pelo telefone resultou em uma maior proporção de entrevistas não realizadas.

Uma possível consequência desse fato é a geração de um "viés de não resposta" em indicadores computados a partir da amostra de respondentes da pesquisa, já que a redução da amostra de indivíduos entrevistados pode afetar com mais ou menos intensidade determinados grupos de indivíduos. Isso quer dizer que aqueles que seriam entrevistados, mas não foram, podem diferir em diversos aspectos daqueles que efetivamente foram entrevistados. Por exemplo, essa nova composição da amostra poderia acabar sobrerrepresentando grupos da população com maior ou menor propensão a ocupar postos de trabalho; afetando diretamente a taxa de emprego calculada com base na pesquisa (CORSEUIL e RUSSO, 2021).

Vale dizer que esse fenômeno não é restrito ao Brasil, haja visto que o problema da não resposta em pesquisas domiciliares nos meses que sucederam a chegada da pandemia foi percebido em diversos países[4]. Vale destacar as contribuições de Rothbaum e Bee (2021) e Dutz *et al* (2021), que analisam essa questão para os EUA e a Noruega respectivamente, propondo métodos para corrigir alguns indicadores afetados pelo abrupto aumento do viés de não resposta.

Para além da mudança na composição da amostra mencionada acima, o viés de não resposta também pode ser afetado por eventuais ajustes implementados no fator de ponderação (doravante peso) dos indivíduos respondentes. Por exemplo, se indivíduos respondentes passam a ter seu respectivo peso majorado para compensar o menor número de respostas, é de se esperar que um eventual viés de não resposta seja igualmente majorado. Nessa versão do artigo iremos focar apenas no componente de mudança da composição da amostra.

Como a PNAD Contínua é a principal fonte de informação sobre o mercado de trabalho brasileiro, acreditamos ser importante inferir em que medida alguns dos principais indicadores divulgados a partir dessa pesquisa podem ter tido sua evolução afetada por algum tipo de viés de não resposta.

O restante desse texto está organizado da seguinte forma. Na segunda seção, computamos a redução do número de entrevistados a partir do 2º trimestre de 2020 e apontamos que tal redução tende a ser maior para a parte da amostra composta por indivíduos que seriam entrevistados pela primeira vez no respectivo trimestre.

[4] A preocupação com a disseminação desse fenômeno em diversos países motivou a CEPAL a sugerir medidas para tentar minimizar a influência do viés de não resposta. O documento pode ser acessado em: https://www.cepal.org/en/publications/45553-recommendations-eliminating-selection-bias-household-surveys-during-coronavirus.

Na terceira seção, mostramos que a queda no número de entrevistados é generalizada no plano regional, mas que há uma heterogeneidade marcante numa partição do território nacional em 77 áreas consideradas. As quedas tendem a ser mais intensas em municípios que não são capitais nem parte de regiões metropolitanas.

Na quarta seção, examinamos indícios de viés de não resposta a partir de 2020, em particular, se a diminuição da quantidade de entrevistas alterou a composição da amostra em relação às características individuais.

Na quinta seção, continuamos a analisar indícios de viés de não resposta, investigando se há relação entre as variações nas taxas de ocupação e formalidade e a redução no número de entrevistas. Nesse caso procuramos contemplar tanto a parte do viés derivada da eventual mudança na composição da amostra.

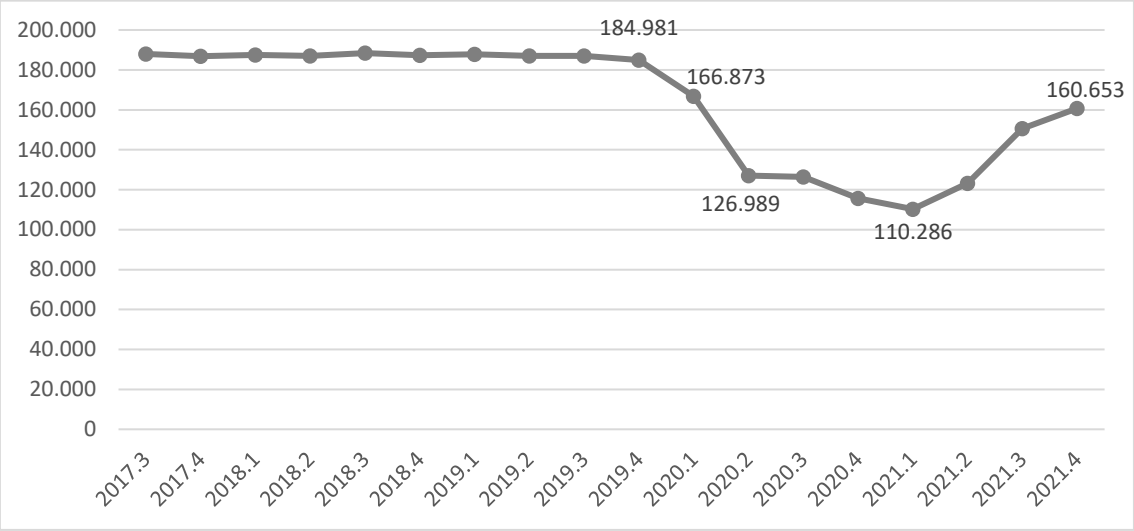
A nossa contribuição principal está na sexta e sétima seção, em que descrevemos e mostramos os resultados, respectivamente, de uma simulação que pretende identificar o efeito desse componente do viés de não resposta. Por fim, sumarizamos nossas conclusões na sétima seção.

2. Análise da evolução do número de entrevistas da PNAD Contínua

O primeiro passo para avaliar um possível viés de seleção da amostra é demarcar quais foram os momentos de maior variação no número de entrevistas durante a pandemia, e identificar se essa queda foi uniforme em relação à ordem das entrevistas. O gráfico 1 mostra a evolução da quantidade de domicílios entrevistados por trimestre na PNAD Contínua, no período do 3º trimestre de 2017 até o 4º trimestre de 2021. É possível observar que de 2017 até o último trimestre de 2019, o número de domicílios entrevistados permanecia relativamente estável, se mantendo entre 184 a 188 mil domicílios por trimestre. No 1º trimestre de 2020, com a transição para a coleta por telefone nas entrevistas de março, a amostra de domicílios caiu para 167 mil.

O impacto dessa mudança se tornou ainda maior no 2º trimestre de 2020, no qual apenas 127 mil domicílios foram entrevistados – uma redução de 32%, quando comparado ao mesmo trimestre do ano anterior. O número de domicílios entrevistados por trimestre continuou em queda até o 1º trimestre de 2021 (110.286), mas voltou a crescer nos trimestres seguintes, atingindo 160 mil domicílios no 4º trimestre de 2021 – porém, ainda sem conseguir recuperar o nível obtido no 4º trimestre de 2019.

Gráfico 1 – Total de domicílios entrevistados na PNAD Contínua, por trimestre (2017.3-2021.4)



Fonte: PNAD Contínua/IBGE.

O gráfico 2 mostra o mesmo indicador do gráfico anterior, desagregado por número de ordem da entrevista. Nota-se que a queda na quantidade de domicílios entrevistados ocorrida entre os primeiros trimestres de 2020 e de 2021 aconteceu em todos os cinco grupos de entrevistas; todavia, de forma desigual, sendo o grupo das primeiras entrevistas o mais afetado, posteriormente, o das segundas entrevistas, depois, o das terceiras, e assim seguindo nessa ordem.

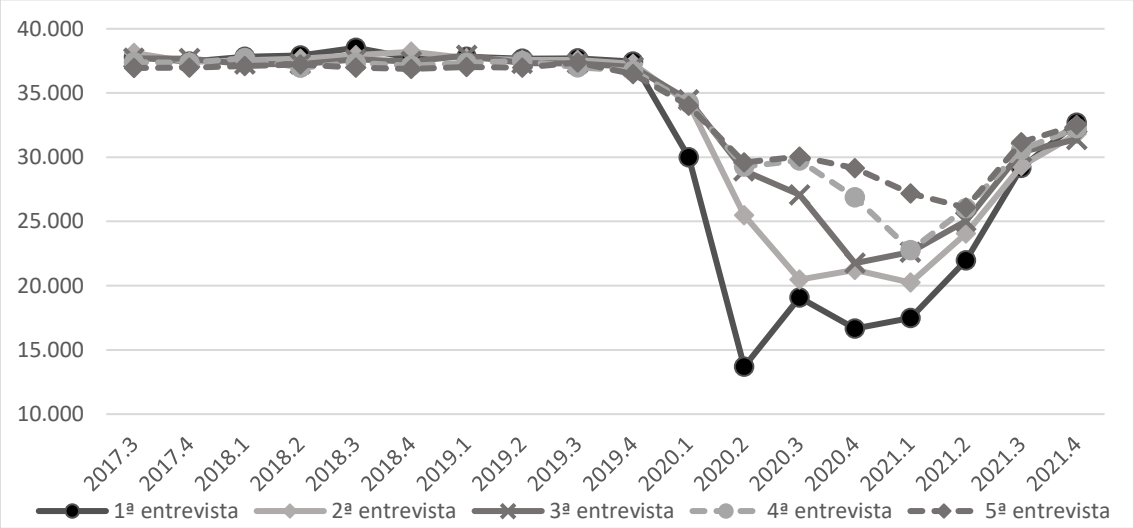
O gráfico também revela um efeito curioso: no 2º trimestre de 2020, a queda mais acentuada no número de domicílios entrevistados pela pesquisa se concentrou naqueles entrevistados pela primeira vez. No 3º trimestre de 2020, a amostra dos domicílios de primeira entrevista volta a subir um pouco; porém, é o grupo de domicílios entrevistados pela segunda vez que passa a cair. Esse efeito parece se alastrar para os três trimestres seguintes; por exemplo, o grupo de domicílios de terceira entrevista no 4º trimestre de 2020 também sofre uma redução considerável, mas com intensidade cada vez menor, até não ser tão visível nos 3º e 4º trimestres de 2021.

Esses resultados podem ser explicados pela dificuldade enfrentada pelo IBGE de conseguir contatar pelo telefone indivíduos que seriam entrevistados pela primeira vez. Como a amostra de domicílios do IBGE é construída a partir do Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE), é possível que esse cadastro não tivesse informações atualizadas de telefones, que eventualmente eram checadas (ou incluídas) na ocasião da primeira entrevista. Dessa forma, o 2º trimestre de 2020 teria sido o mais crítico para contatar indivíduos a serem entrevistados pela primeira vez. Esse mesmo grupo seria entrevistado pela segunda vez no 3º trimestre de 2020 e, portanto, também teria o tamanho da amostra comprometido[5] (CORSEUIL e RUSSO, 2021).

[5] É importante destacar que houve um esforço do IBGE para construir um cadastro nacional de telefones visando contornar essa dificuldade do contato por telefone. Foi efetuada em 17 de abril de 2020 uma medida provisória do presidente da República, que permitia ao instituto ter acesso aos dados cadastrais dos clientes das companhias telefônicas. No entanto, em maio do mesmo ano, o STF impediu o IBGE de ter acesso a essas informações, obrigando-o a recorrer a outros meios para enriquecer seu cadastro com os telefones dos moradores dos domicílios potencialmente amostrados na PNAD Contínua (CORSEUIL e RUSSO, 2021).

Com o passar do tempo, o IBGE pode ter obtido mais alguns números de telefone, e assim conseguido contatar mais indivíduos da amostra prevista, fazendo com que a queda no número de entrevistas fosse um pouco atenuada nas entrevistas seguintes desse grupo, apesar de ainda relevante. Com a saída desse grupo mais afetado após sua quinta entrevista e a volta parcial das entrevistas presenciais em julho de 2021, vemos que o número de domicílios entrevistados pela PNAD Contínua volta a ficar mais equilibrado nos 3º e 4º trimestres de 2021.

Gráfico 2 – Total de domicílios entrevistados na PNAD Contínua, por trimestre e por número de ordem da entrevista (2017.3-2021.4)



Fonte: PNAD Contínua/IBGE.

Portanto, diante dos resultados, é possível destacar que em todo o período que foi adotada a coleta de entrevistas por telefone (1º trimestre de 2020 ao 1º trimestre de 2021), o tamanho da amostra da PNAD Contínua esteve em queda, só voltando a crescer com o retorno parcial das entrevistas presenciais, em julho de 2021. Destaca-se também que o aumento da proporção de entrevistas não realizadas no período foi composto principalmente por pessoas que seriam entrevistadas pela primeira vez, mas que não foram, e que esse efeito foi maior no 2º trimestre de 2020.

3. Possíveis determinantes para a queda no número de entrevistas:

Na seção anterior, destacamos a dificuldade enfrentada pelo IBGE de conseguir obter o telefone de indivíduos que seriam entrevistados pela primeira vez, especialmente no 2º trimestre de 2020, diante da mudança na forma de coleta da pesquisa – sendo assim um fator relevante para explicar a redução na amostra de entrevistados da PNAD Contínua.

Nessa seção procuraremos identificar alguns aspectos que podem ter influenciado a propensão de um domicílio ter respondido a pesquisa. Esses aspectos dizem respeito a probabilidade de obtenção do número de telefone do domicílio pelo IBGE, e a probabilidade de o telefone ser atendido por uma pessoa que responda ao IBGE. O primeiro aspecto a ser investigado diz respeito à localização do domicílio. Sabe-se que o IBGE recorreu às suas unidades regionais para tentar minimizar o problema da falta de

informação do número de telefone dos domicílios amostrados. Como essas unidades regionais estão localizadas nas capitais de cada unidade federativa (UF), supomos que o problema de não obter um número válido de telefone para o domicílio pode ser menor nas capitais e deve se tornar mais relevante à medida que a localização do domicílio se afasta da capital da sua respectiva UF.

O segundo aspecto que investigaremos diz respeito à composição demográfica do domicílio, mais especificamente a quantidade de adultos e de dependentes (crianças ou idosos). Supomos que uma vez tendo o número de telefone do domicílio, o IBGE pode ter mais dificuldade de estabelecer contato em um domicílio com apenas um adulto do que em domicílios com mais adultos e onde também há dependentes.

Portanto, essa seção pretende averiguar se a queda no número de entrevistas foi heterogênea no plano regional e/ou no âmbito da composição demográfica dos domicílios.

Podemos analisar mais profundamente o que ocorreu na dimensão regional através do painel 1 abaixo, que contém os gráficos 3 a 6. Todos são gráficos de dispersão, restritos à amostra de primeira entrevista, em que cada ponto representa um domínio de projeção da pesquisa. Os domínios de projeção são áreas geográficas para as quais o IBGE ajusta os pesos amostrais da PNAD Contínua de forma a coincidir com as projeções de população. Cada um deles pertence a uma UF e a um tipo de área, as quais podem ser: Capital; resto da Região Metropolitana (RM), excluindo a capital; resto da Região Integrada de Desenvolvimento Econômico (RIDE)[6], excluindo a capital; e resto da UF, excluindo a RM e a RIDE.

Em cada um dos gráficos o eixo Y representa o número de domicílios que foram entrevistados dentro de cada domínio de projeção no 2º, 3º ou 4º trimestre de 2019 ou no 1º trimestre de 2020, e o eixo X representa o mesmo indicador, para o mesmo trimestre do ano seguinte. Logo, a distância horizontal entre o ponto e a reta de 45º graus indica o quanto variou o número de entrevistas naquele domínio de projeção em um ano, contemplando um momento anterior e um momento posterior à chegada da pandemia.

Observando o painel, é possível destacar que em todos os gráficos os pontos se mantém concentrados no lado esquerdo da linha de 45º, mostrando que a redução do número de entrevistas do grupo de indivíduos a ser entrevistados pela primeira vez é algo disseminado no plano regional, atingindo todas as áreas que consideramos.

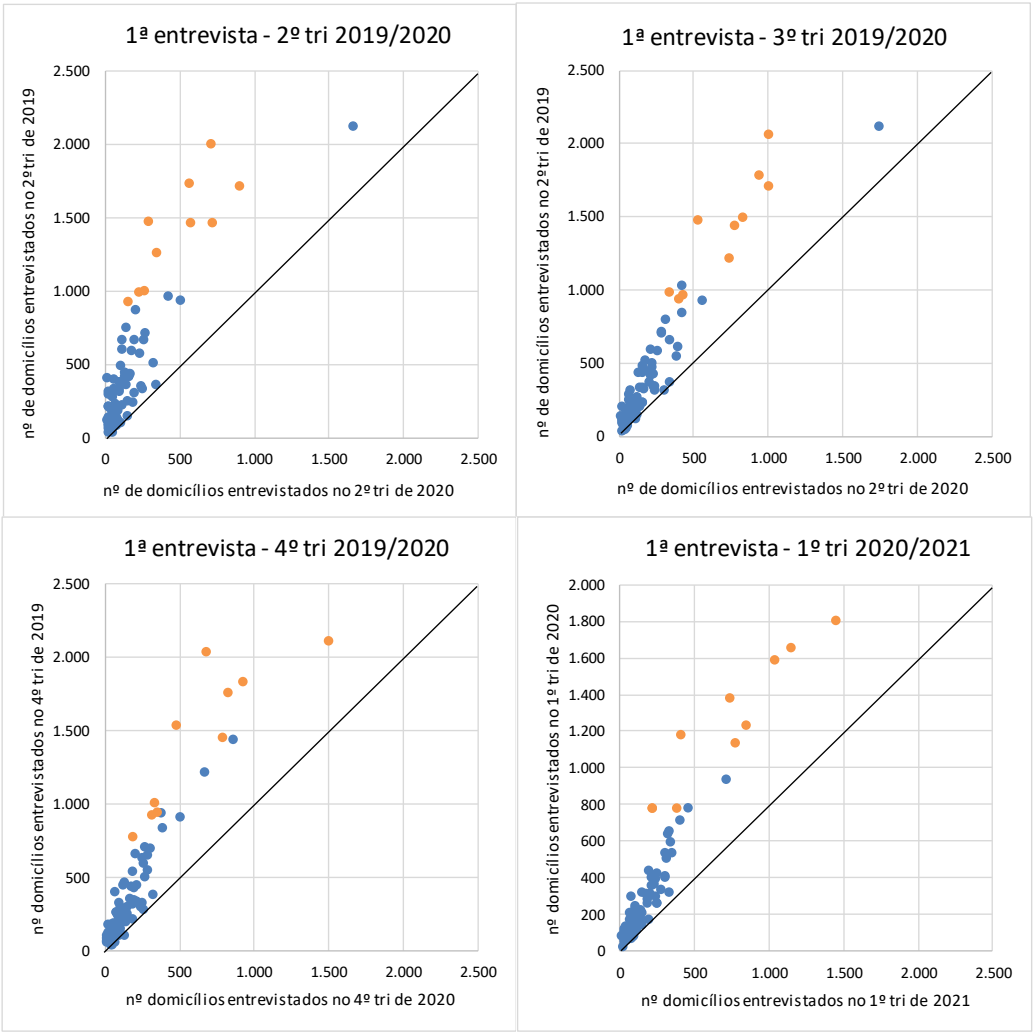
No entanto, o fator que mais chama atenção nesses quatro gráficos é o de que as distâncias horizontais de cada um dos pontos até a reta de 45º variam bastante em tamanho. Em outros termos, nota-se que para certos domínios de projeção, a diferença entre o número de domicílios que compõem a amostra de primeira entrevista de um trimestre e o número dos que compõem a do mesmo trimestre no ano seguinte é pequena; já para outros, essa diferença é muito grande. Portanto, é notável que a redução no número de entrevistas não se deu de maneira homogênea em relação aos domínios de projeção.

Os pontos em laranja dos gráficos 3, 4, 5 e 6 são os que possuem as maiores distâncias horizontais em relação a reta de 45º; ou seja, são os dez domínios de projeção com maior diferença entre a quantidade de domicílios entrevistados pela primeira vez no

[6] A RIDE é uma área análoga às regiões metropolitanas brasileiras, porém, situada em mais de uma unidade federativa.

2º, 3º, 4º trimestre de 2019 ou no 1º trimestre de 2020, e no mesmo trimestre do ano seguinte. O quadro 1 descreve com detalhes suas características.

Painel 1: Gráficos 3 a 6 - Número de domicílios entrevistados pela 1ª vez em um trimestre vs número de domicílios entrevistados pela 1ª vez no mesmo trimestre do ano seguinte



Fonte: PNAD Contínua/IBGE.

Do total de 77 domínios de projeção, apenas os 10 da primeira parte do quadro foram responsáveis por 39,5% da queda no número de domicílios de 1ª entrevista entre o 2º trimestre de 2019 e o 2º trimestre de 2020. Na comparação entre os terceiros trimestres de 2019 e 2020, essa contribuição foi de 39,2%; e para os quartos trimestres, de 38,6%. Para o 1º trimestre de 2021, a contribuição foi de 40,8%.

O quadro revela que há uma prevalência de domínios de projeção localizados em áreas restantes da UF, ou seja, que não são capitais e não fazem parte da região metropolitana ou da RIDE. Esse padrão é coerente com a nossa suposição de que o instituto tenha enfrentado uma dificuldade maior de obter os telefones de moradores de domicílios que se encontram nos domínios de projeção mais distante das capitais, resultando assim em uma maior queda no número de entrevistas nessas áreas.

Outro fato revelado pelo quadro 1, é que há poucas mudanças na lista de domínios de projeção de um trimestre para o outro; em sua maioria são os mesmos domínios de projeção que aparecem nos períodos analisados. Esse fato descarta a possibilidade da diminuição no número de entrevistas num dado domínio de projeção ser causado majoritariamente por um aspecto ocasional ou específico de um certo momento do tempo. O principal elemento que causou a redução de entrevistas para o grupo de indivíduos entrevistados pela primeira vez deve ser algo permanente; tal como a distância para a capital.

No que se refere às regiões do país, o quadro não aparenta mostrar nenhum padrão, reunindo domínios de projeção de diversas áreas do país. Contudo, ao analisar as UFs, o estado do Rio de Janeiro aparece como destaque, tendo contribuído com 9,7% e 9,3% da queda total no número de domicílios de 1ª entrevista entre o 2º trimestre de 2019 e 2020 e entre o 3º trimestre de 2019 e 2020, respectivamente. Essa contribuição diminui no 4º trimestre de 2020, registrando 6,2%; porém volta a aumentar para 9,0% no 1º trimestre de 2021. Também é a única UF do quadro na qual um domínio de projeção localizado em uma capital aparece entre as 10 áreas com maior diminuição no número de entrevistas.

Quadro 1 – Domínios de projeção com maior diferença anual entre a quantidade de domicílios de primeira entrevista dos trimestres analisados

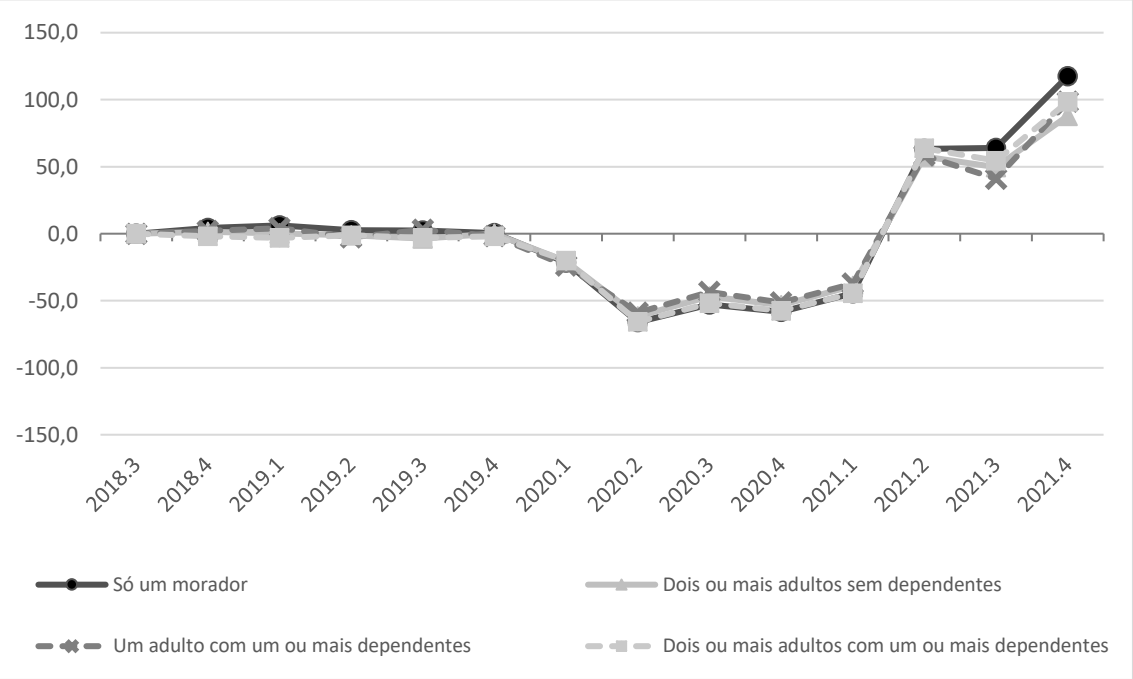
2º tri 2019/2020			3º tri 2019/2020			4º tri 2019/2020			1º tri 2020/2021		
UF	Tipo de área	Contribuição para a queda total no nº de domicílios de 1ª entrevista	UF	Tipo de área	Contribuição para a queda total no nº de domicílios de 1ª entrevista	UF	Tipo de área	Contribuição para a queda total no nº de domicílios de 1ª entrevista	UF	Tipo de área	Contribuição para a queda total no nº de domicílios de 1ª entrevista
SC	Resto da UF	5,4%	SC	Resto da UF	5,7%	SC	Resto da UF	6,5%	PR	Resto da UF	6,2%
PR	Resto da UF	5,0%	PR	Resto da UF	5,2%	PR	Resto da UF	5,1%	SP	Resto da UF	5,1%
MA	Resto da UF	5,0%	SP	Resto da UF	4,6%	SP	Resto da UF	4,5%	RJ	Capital	4,5%
CE	Resto da UF	3,9%	MA	Resto da UF	3,8%	MA	Resto da UF	4,3%	RJ	Resto da UF	4,5%
BA	Resto da UF	3,8%	RS	Resto da UF	3,7%	RJ	Capital	3,3%	MA	Resto da UF	4,4%
SP	Resto da UF	3,5%	BA	Resto da UF	3,7%	BA	Resto da UF	3,2%	SC	Resto da UF	4,0%
RJ	Resto da RM	3,3%	RJ	Capital	3,5%	MG	Resto da UF	3,0%	AL	Resto da UF	3,2%
RJ	Capital	3,3%	PE	Resto da UF	3,2%	AL	Resto da UF	2,9%	BA	Resto da UF	3,1%
RS	Resto da UF	3,2%	RJ	Resto da UF	2,9%	ES	Resto da UF	2,9%	RS	Resto da UF	2,9%
RJ	Resto da UF	3,1%	RJ	Resto da RM	2,9%	RJ	Resto da RM	2,9%	MG	Resto da UF	2,9%

Fonte: PNAD Contínua/IBGE.

Já em relação ao perfil dos domicílios, o gráfico 8 mostra a variação interanual no número de domicílios de primeira entrevista na PNAD Contínua, desagregada por cada uma das seguintes características demográficas: só um morador; dois ou mais adultos sem dependentes; um adulto com um ou mais dependentes; dois ou mais adultos com um ou mais dependentes. O conceito de dependentes utilizado foi o de pessoas com idade menor que 14 anos, ou maior ou igual a 80 anos.

Nota-se que, diferente da dimensão regional, a queda no número de primeiras entrevistas na PNAD Contínua parece ter sido homogênea no que se refere à composição demográfica dos domicílios.

Gráfico 8 – Variação interanual do número de domicílios entrevistados pela primeira vez na PNAD Contínua, por configuração demográfica do domicílio (2018.3 – 2021.4)
(Em %)



Fonte: PNAD Contínua/IBGE.

4. Evidências preliminares do viés de não resposta para as taxas de emprego e formalização

Na seção anterior, vimos que a redução no número de entrevistas com a mudança para coleta por telefone da PNAD Contínua se deu de maneira heterogênea na dimensão regional. Agora investigaremos se há indícios dessa redução ter afetado a evolução de indicadores do mercado de trabalho, em particular as taxas de ocupação (razão entre ocupados e PIA) e de formalização (razão entre empregos formais[7] e PIA).[8]

Buscando investigar indícios de influência de um viés de não resposta, o quadro 2 mostra as variações interanuais das taxas de ocupação e formalização (sem peso) para duas amostras: i) a amostra restrita às primeiras entrevistas, e ii) amostra restrita aos 10 domínios de projeção com a maior queda no número de primeiras entrevistas no 2º trimestre de 2020.

Analisando primeiramente as variações das taxas de ocupação, podemos notar que essas são bem maiores na amostra restrita aos 10 domínios de projeção que obtiveram maiores reduções do número de entrevistas no 2º trimestre de 2020. Nos trimestres analisados, a variação nesse grupo de domínios de projeção frequentemente passa de 10

[7] Empregados com carteira e trabalhadores conta-própria e empregadores que contribuem para Previdência.
[8] Como estamos interessados em analisar as mudanças na composição amostral da pesquisa em relação à determinantes da taxa de emprego e de emprego formal, em todos os dados dessa seção e nas seções seguintes, a amostra selecionada é restrita a entrevistados que compõem a PIA (população em idade ativa), isto é, com idade maior ou igual a 14 anos.

pontos de percentagem, com a única exceção sendo a variação anual da taxa de ocupação computada para o 4º trimestre de 2020.

O mais importante para a nossa análise é constatar que tais reduções anuais da taxa de ocupação são sempre maiores na amostra restrita aos 10 domínios de projeção que tiveram maiores reduções do número de entrevistas do que na amostra com todos os domicílios entrevistados pela primeira vez. Na verdade, as reduções nas taxas de ocupação na primeira amostra tendem a ser praticamente o dobro (ou até maior) que as reduções computadas para a segunda amostra; à exceção do 4º trimestre de 2020. Essa grande discrepância na evolução da taxa de ocupação entre amostras diferentemente afetadas pela redução do número de entrevistas sugere uma interferência de um viés de não resposta na evolução desse indicador.

Quadro 2 – Variação interanual das taxas amostrais de ocupação e formalização para diferentes amostras (2020.2 - 2021.1 – em p.p.)

Ano e trimestre	Variação da taxa de ocupação		Variação da taxa de formalização	
	1ª entrevista nas 10 áreas com maiores quedas no número de primeiras entrevistas em 2020.2	Amostra na 1ª entrevista	1ª entrevista nas 10 áreas com maiores quedas no número de primeiras entrevistas em 2020.2	Amostra na 1ª entrevista
2020.2	-14.47	-7.51	-5.83	-2.80
2020.3	-11.85	-6.28	-8.14	-2.74
2020.4	-5.79	-4.22	-1.97	-2.11
2021.1	-10.39	-4.01	-4.54	-2.13

Fonte: PNAD Contínua/IBGE.

O mesmo ocorre quando comparamos as evoluções da taxa de formalização entre essas amostras. Embora as quedas sejam menos pronunciadas em geral do que as registradas para a taxa de ocupação, a comparação entre as duas amostras também indica que as quedas na taxa de formalização são frequentemente muito maiores na amostra restrita aos 10 domínios de projeção que tiveram maiores reduções do número de entrevistas. Novamente o 4º trimestre de 2020 configura uma exceção a esse padrão. Logo, há indícios também que a evolução da taxa de formalização tenha sido afetada por viés de não resposta.

5. Simulando o efeito do viés de não resposta na evolução da taxa de ocupação e de formalização: metodologia

5.1. Antecedentes

A literatura em estatística e econometria já colecionava contribuições relevantes para estimar e/ou contornar o viés de não resposta em pesquisas domiciliares anteriores à pandemia. Grove (2006); Meyer *et al* (2015); Litle e Rubin (2019); e DiNardo *et al* (2021) são algumas das contribuições influentes nesse tema analisando dados pré-pandêmicos.

O aumento abrupto das não respostas em pesquisas domiciliares com a chegada da Covid-19 inspirou análises mais focadas nesse período. Consideramos que os trabalhos de Rothbaum e Bee (2021) e Dutz *et al* (2021) ilustram bem as duas principais opções metodológicas desenvolvidas para lidar com o viés de não resposta. Em ambas as opções, procura-se encontrar um cenário contrafactual que indique como seria a composição da amostra sem o efeito do viés de não resposta. A diferença entre os trabalhos é que o primeiro trabalho se atém a analisar a composição baseada nas características observáveis dos indivíduos; enquanto o segundo trabalho advoga que é necessário atentar também para alterações na composição da amostra no que diz respeito às características não observáveis. Dutz *et al* (2021) vão além e proveem evidências contundentes de que métodos para lidar com viés de não resposta restritos às características observáveis não conseguem identificar corretamente a influência desse problema nos dados dinamarqueses nos meses que sucedem o início da pandemia.

Motivados por essa constatação, propomos analisar a possível influência de um viés de não resposta na PNAD Contínua após a pandemia, levando em conta também eventuais alterações na composição dos indivíduos entrevistados quanto às características não observáveis. No entanto, não podemos utilizar o método proposto por Dutz *et al* (2021), pois o mesmo depende de haver uma distinção gerada aleatoriamente entre grupos de indivíduos, a qual se refere ao grau de incentivo para responder a pesquisa domiciliar em questão. De forma alternativa, nossa metodologia explora o fato dos domicílios serem entrevistados mais de uma vez na PNAD Contínua para trabalharmos com uma amostra idêntica em dois momentos do tempo.

5.2. Metodologia Proposta

Nessa seção apresentaremos uma estratégia para identificar um termo contrafactual que descreveria a evolução dos nossos indicadores de interesse num intervalo de um ano, livres de viés de não resposta.

Seja "t0" um momento pré-pandemia em que não houve variações abruptas no número de não respostas, e, portanto, a composição da PIA não foi afetada nem em aspectos observáveis nem em não observáveis. Seja "t1" um momento pós-pandemia em que a composição da PIA pode ter sido afetada (via aumento de não respostas) tanto em aspectos observáveis como em não observáveis.

Considere que a nossa variável de interesse, probabilidade de emprego (formal), pode ser modelada da seguinte forma:

$$y_{i,t} = X'_{i,t}.\beta_t + \alpha_i + \varepsilon_{i,t}$$

Em que X' descreve um vetor de características observáveis do indivíduo e/ou do domicílio onde mora; α denota características não observáveis fixas no tempo (e

possivelmente correlacionadas com X); ε representa características não observáveis que influenciam a probabilidade de emprego (formal); e t $\in$ {t0, t1}. Em cada um desses momentos pode-se representar a média da variável de interesse ($\bar{y}$) entre os indivíduos amostrados em "t" da seguinte forma:

$$\bar{y}_t = \bar{X}'_t.\beta_t + \bar{\alpha} + \bar{\varepsilon}_t$$

Logo, em um painel balanceado, a evolução temporal da média da nossa variável de interesse (evolução da taxa de ocupação ou de formalização) não é influenciada pelas características não observáveis e pode ser descrita como:

$$\bar{y}_{t1} - \bar{y}_{t0} = (\bar{X}'_{t1}.\beta_{t1} - \bar{X}'_{t0}.\beta_{t0}) \qquad (1)$$

Note que a expressão acima é também o ponto de partida de decomposições do tipo Oaxaca-Blinder. De fato, argumentaremos a seguir que a nossa proposta para obter a variação das taxas de ocupação e formalização livres de viés de não resposta corresponde a um dos termos desse tipo de decomposição. Em particular, propomos que deve ser estimado o seguinte termo que representa o contrafactual de interesse; denotado abaixo pelo sobrescrito "sv" para representar "sem viés":

$$(\bar{y}_{t1} - \bar{y}_{t0})^{sv} = \bar{X}'_{t0}.(\beta_{t1} - \beta_{t0}) \qquad (2)$$

A identificação do termo do lado direito da equação (2) passa a ser trivial se supormos a seguinte hipótese:

H1: Na ausência de problemas de viés de não resposta derivados da pandemia, a composição da PIA não deveria se alterar (em termos médios) entre t0 e t1, no que diz respeito à covariadas. Ou seja,

$$\bar{X}'_{t1} = \bar{X}'_{t0}$$

Sob a hipótese acima, a evolução contrafactual da média da nossa variável de interesse (evolução da taxa de ocupação ou de formalização) pode ser reescrita a partir de (1) como:

$$\bar{y}_{t1} - \bar{y}_{t0} = \bar{X}'_{t0}.(\beta_{t1} - \beta_{t0})$$

Note que o lado direito da equação acima traz o mesmo objeto representado do lado direito da equação (2); que vem a ser o nosso contrafactual de interesse. Para estimar esse objeto que representaria a evolução da taxa de emprego (formal) que prevaleceria na pandemia se não houvesse a queda drástica no número de entrevistas, basta estimar de forma consistente o termo $(\beta_{t1} - \beta_{t0})$, e multiplicá-lo pela composição da PIA que prevaleceria nesse cenário ($\bar{X}'_{t0}$).

Argumentaremos em seguida que tal estimativa pode vir da simples diferença de coeficientes estimados por MQO em cada um dos dois momentos considerados (t1 e t0), com a amostra restrita às observações disponíveis em ambos os momentos, ou seja, a partir de um painel balanceado.

A variável indicadora D_i vale 1 quando o indivíduo atende essa condição de ser observado em ambos os momentos. Por ser um conjunto selecionado de forma não aleatória teríamos a seguinte expressão para representar a estimativa obtida por MQO em cada um dos instantes considerados:

$$\hat{\beta}_t = \beta_t + \left[cov\left(X'_{i,t}; \alpha_i | D_i = 1\right) \Big/ var(X'_{i,t} | D_i = 1) \right]$$
$$+ \left[cov\left(X'_{i,t}; \varepsilon_{i,t} | D_i = 1\right) \Big/ var(X'_{i,t} | D_i = 1) \right]$$

Note que não estamos supondo que o estimador de MQO seja consistente no nosso contexto. Mas ainda assim essas estimativas vão ser úteis para chegarmos ao nosso objetivo. Para isso precisamos usar uma hipótese adicional.

H2: em um painel balanceado teremos que

$$cov\left(X'_{i,t0}; \varepsilon_{i,t0} | D_i = 1\right) = cov\left(X'_{i,t1}; \varepsilon_{i,t1} | D_i = 1\right)$$

e

$$cov\left(X'_{i,t0}; \alpha_i | D_i = 1\right) = cov\left(X'_{i,t1}; \alpha_i | D_i = 1\right)$$

A hipótese acima garante que o termo $(\beta_{t1} - \beta_{t0})$ pode ser estimado pelo termo $(\hat{\beta}_{t1} - \hat{\beta}_{t0})$; ou seja, pela diferença das estimativas obtidas em cada *cross section* por MQO.

Note que a variância no denominador da expressão acima não deve variar entre os dois momentos, em virtude de se tratar de variáveis que tendem a ser fixas no tempo e mensuradas para um conjunto fixo de indivíduos observado tanto em t0 como em t1.

6. Análise da composição amostral em relação às características observáveis

Nas duas seções anteriores, vimos que há indícios de que a redução no número de entrevistas tenha resultado em viés de não resposta para a evolução amostral das taxas de ocupação e formalização. Além disso, propomos um método para expurgar a influência desse viés nas evoluções dessas taxas, contemplando um viés oriundo de mudanças na composição da amostra, em relação a características observáveis e não observáveis dos indivíduos.

O objetivo dessa seção é analisar a evolução da composição da amostra de indivíduos entrevistados no que diz respeito a características observáveis, tanto antes como depois da chegada da pandemia. Essa análise tem dois propósitos. Em primeiro lugar, pretendemos checar se eventuais mudanças na composição das características observáveis ocorrem de forma mais pronunciada em áreas mais afetadas pela redução no número de entrevistas, tal como reportado para as quedas nas taxas de ocupação e formalização. Em outras palavras, estaríamos checando o quão plausível é a hipótese de que basta lidar com as características observáveis para contornar o problema do viés de não resposta. Essa análise será apresentada na primeira subseção dessa seção.

Em segundo lugar, pretendemos checar em que medida tal composição era estável antes da pandemia, tal como preconizado em uma das hipóteses da metodologia que propomos. Essa parte da investigação compõe a segunda subseção dessa seção.

6.1. A evolução da composição da amostra no pós-pandemia

Para avaliar a evolução da composição amostral, podemos calcular a variação interanual da proporção de pessoas com certas características individuais (de gênero, raça,

nível de escolaridade e faixa etária) no total de entrevistados da PNAD Contínua a cada trimestre. O quadro 3 mostra a média dessa variação para 2020 e 2021 para a amostra de indivíduos na primeira entrevista, e o subconjunto desses indivíduos nos 10 domínios de projeção com maiores reduções no número de entrevistados no 2º trimestre de 2020.

Quadro 3 – Variação interanual da composição da amostra e de grupos específicos por características individuais (taxa de variação média em 2020 e 2021 – em p.p.)

Proporção de características individuais	1ª entrevista (restrito as 10 áreas com maiores variações em 2020.2)		1ª entrevista	
	2020	2021	2020	2021
Mulheres	0,76	-0,70	0,60	-0,17
Pretos e pardos	-1,63	1,11	-0,55	-0,50
Até fundamental incompleto	0,10	-0,28	-1,34	0,13
Ensino superior completo	1,58	-1,15	1,05	-0,32
Entre 14 e 17 anos	-0,63	0,02	-0,18	-0,06
Entre 18 e 24 anos	-1,40	1,08	-0,76	0,04
Entre 25 e 44 anos	-2,37	-0,13	-1,56	-0,57
Entre 45 e 59 anos	-0,09	0,71	0,68	0,30
60 anos ou mais	4,48	-1,68	1,83	0,29

Fonte: PNAD Contínua/IBGE.

É possível observar dois fatos relevantes no quadro 3. Em primeiro lugar, as magnitudes de variação na composição das amostras consideradas são bem inferiores àquelas reportadas no quadro 2 para as taxas de ocupação e formalização das mesmas amostras. Isso já sugere uma influência reduzida de mudanças na composição relativas as características observáveis nas quedas bruscas nas taxas de ocupação e formalização no período pós-pandemia. Em segundo lugar, o quadro 2 revela diferenças bem limitadas, em geral inferiores a 1 ponto de percentagem, na comparação das mudanças na composição entre cada uma das duas amostras. As maiores diferenças nessa comparação dizem respeito às variações computadas para a parcela de pretos e pardos; e a parcela de idosos (60 anos ou mais).

6.2. A evolução da composição da amostra no pré-pandemia

O quadro 4 mostra a evolução da composição para 2018 e 2019, e para as mesmas características individuais consideradas no quadro 3. As evoluções são mostradas para a amostra limitada a indivíduos entrevistados pela primeira vez e para todos os indivíduos. O principal resultado para os nossos propósitos é que as variações reportadas em todas as dimensões consideradas são bem restritas, quase todas menores do que 1 ponto de percentagem em valor absoluto (sendo a única exceção a variação reportada para indivíduos com até o fundamental incompleto em 2019).

Quadro 4 – Variação interanual da composição da amostra e de grupos específicos por características individuais (taxa de variação média em 2018 e 2019 – em p.p.)

Proporção de características individuais	1ª entrevista		Amostra total	
	2018	2019	2018	2019
Mulheres	0,06	0,10	0,08	0,11
Pretos e pardos	0,13	0,46	0,20	0,24
Até fundamental incompleto	-0,97	-1,04	-0,90	-1,15
Ensino superior completo	0,88	0,40	0,76	0,62
Entre 14 e 17 anos	-0,46	-0,22	-0,50	-0,19
Entre 18 e 24 anos	-0,27	-0,22	-0,16	-0,35
Entre 25 e 44 anos	-0,27	-0,32	-0,29	-0,34
Entre 45 e 59 anos	0,19	0,23	0,13	0,28
60 anos ou mais	0,81	0,52	0,82	0,58

Fonte: PNAD Contínua/IBGE.

Esse resultado suporta a suposição que recorremos na metodologia proposta de que antes da pandemia não havia variações significativas na composição da amostra. As diferenças entre a amostra de 1ª entrevista e a amostra geral também são muito pequenas, não passando de dois décimos de um ponto percentual em nenhum caso. Essa semelhança é importante para simulação que será apresentada na próxima seção.

7. Simulando o efeito do viés de não resposta na evolução da taxa de ocupação e de formalização: resultados

Essa seção apresenta os resultados da estimação da equação (2). Calculamos os coeficientes β_{t0} e β_{t1}a partir do painel balanceado da PNAD Contínua do 2º trimestre de 2019 ao 1º trimestre de 2021. Assim, por exemplo, para a simulação da taxa interanual do 2º trimestre de 2019 para o 2º trimestre de 2020 usamos os indivíduos que estavam na 1ª entrevista no primeiro período e na 5ª entrevista no segundo período.

Após a estimação dos coeficientes, multiplicamos estes pelas médias das características observáveis no período pré ($\bar{X}'_{t0}$). Notem que essa média é feita sobre toda a amostra no período pré, e por isso uma hipótese implícita em nosso cálculo é que o resto da amostra se comportaria ao longo do ano da mesma maneira que o subgrupo na 1ª entrevista. Os resultados do quadro 4 na seção anterior mostram que essa não é uma suposição forte, pelo menos no período anterior à pandemia.

Quadro 5 – Variação interanual da taxa de ocupação e formalização, simulado e observado (em p.p.)

	Variação interanual da taxa de ocupação		Variação interanual da taxa de formalização	
	Simulação (I)	Amostra PNAD Contínua (II)	Simulação (III)	Amostra PNAD Contínua (IV)
2º tri 2020/2019	-6.04	-5.66	-1.46	-0.88
3º tri 2020/2019	-7.49	-6.65	-2.11	-1.96
4º tri 2020/2019	-5.48	-5.01	-1.61	-1.79
1º tri 2021/2020	-4.79	-4.25	-1.61	-1.75

Notas: As colunas (I) e (III) apresentam os resultados da estimação da equação (2) para a variação anual da taxa de ocupação e taxa de formalização respectivamente. As colunas (II) e (IV) apresentam a variação anual dos valores observados na amostra publicada pelo IBGE.

Fonte: PNAD Contínua/IBGE

O quadro 5 mostra os resultados da estimação que descrevemos acima. Nossos valores simulados para a evolução da taxa de ocupação, que a princípio eliminam o viés proveniente de mudanças na composição da amostra, são consistentemente menores que o observado na amostra publicada do IBGE. Ou seja, as dificuldades provenientes da mudança para entrevistas telefônicas tiveram como efeito a suavização da queda da taxa de ocupação durante a pandemia. Comparando as colunas (I) e (II), essa diferença está entre 0,3 e 0,8 pontos percentuais, e corresponde a aproximadamente 645 mil trabalhadores.

As colunas (III) e (IV) realizam o mesmo exercício para a taxa de formalização. Na variação do 2º trimestre de 2019 e 2020, novamente o viés de mudança de composição agiu de forma a atenuar a queda do indicador; nossa variação simulada foi 0,59 p.p. mais negativa que a observada. Ao contrário da taxa de ocupação, a diferença entre os valores simulados e observados diminui no trimestre seguinte e muda de sinal a partir da comparação anual do 4º trimestre. Para os dois últimos trimestres de nossa série, a variação simulada foi pouco mais de 0,1 ponto percentual maior que a observada.

A conclusão de nossas simulações é que na ausência do viés de não resposta, as quedas da taxa de ocupação após o começo da pandemia seriam ainda maiores que as divulgadas até o 1º trimestre de 2021, pelo menos. Já para a taxa de formalização, o viés parece ser mais relevante no 2º trimestre de 2020 e sem ele novamente a queda do indicador seria mais forte que a divulgada.

Devemos destacar novamente que nossa correção visa apenas o viés proveniente de mudanças na composição da amostra devido a não resposta e por isso comparamos nossos resultados apenas com as estimativas sem peso dos microdados do IBGE. A partir de novembro de 2021, o IBGE passou a divulgar a PNAD Continua usando um novo método de reponderação, com o objetivo de mitigar o viés de não resposta proveniente da pandemia (IBGE, 2021a e 2021b). Uma extensão natural para esse artigo é explorar como nossa simulação interage com os novos fatores de ponderação publicados e os disponíveis em 2020 durante a publicação original dos resultados.

REFERÊNCIAS

CORSEUIL, C.H.L.; RUSSO, F. M. A redução no número de entrevistas na PNAD Contínua durante a pandemia e sua influência para a evolução do emprego formal. **Carta de Conjuntura:** nº 50, 1º tri. 2021.

DINARDO, J. *et al.* A Practical Proactive Proposal for Dealing with Attrition: Alternative Approaches and an Empirical Example. **Journal of Labor Economics**, v. 39, n. S2, p. S507-S541, 2021.

DUTZ, D. *et al.* **Selection in surveys**. National Bureau of Economic Research, 2021.

GROVES, R. M. Nonresponse rates and nonresponse bias in household surveys. **Public opinion quarterly**, v. 70, n. 5, p. 646-675, 2006.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua):** informações referentes à coleta do mês de abril de 2020. Rio de Janeiro: IBGE, 2020a. (Nota Técnica). Disponível em: https://bit.ly/3qsuF9Z.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua):** informações referentes à divulgação dos dados do 2º trimestre de 2020. Rio de Janeiro: IBGE, 2020b. (Nota Técnica). Disponível em: https://bit.ly/3ifiTwC.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua):** Sobre o processo de ponderação da PNAD Contínua. Rio de Janeiro: IBGE, 2021a. (Nota Técnica). Disponível em: https://bit.ly/3aMRpiw.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua):** Sobre a divulgação da Reponderação da PNAD Contínua em 2021. Rio de Janeiro: IBGE, 2021b. (Nota Técnica). Disponível em: https://bit.ly/3Pinpdl.

LITTLE, R. J. A.; RUBIN, D. B. **Statistical analysis with missing data**. John Wiley & Sons, 2019.

MEYER, B. D.; MOK, W. K. C.; SULLIVAN, J. X. Household surveys in crisis. **Journal of Economic Perspectives**, v. 29, n. 4, p. 199-226, 2015.

ROTHBAUM, J.; BEE, A. Coronavirus infects surveys, too: Survey nonresponse bias and the Coronavirus pandemic. **US Census Bureau**, 2021.