Textos para Discussão
Diretoria de Pesquisas
n.61

O Tratamento das Informações da PNAD Contínua

Fernanda Karine Ruiz Colenghi Baptista
Elizabeth Belo Hypólito
Felipe Quintas Conde

IBGE
Instituto Brasileiro de Geografia e Estatística

Presidente da República
Jair Messias Bolsonaro

Ministro da Economia
Paulo Roberto Nunes Guedes

Chefe da Assessoria Especial de Estudos Econômicos
Rogério Boueri Miranda

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Eduardo Luiz G. Rios Neto

Diretora-Executiva
Marise Maria Ferreira

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Cimar Azeredo Pereira

Diretoria de Geociências
Claudio Stenner

Diretoria de Tecnologia da Informação
Carlos Renato Pereira Cotovio

Centro de Documentação e Disseminação de Informações
Carmen Danielle Lins Mendes Macedo

Escola Nacional de Ciências Estatísticas
Maysa Sacramento de Magalhães

UNIDADE RESPONSÁVEL

Diretoria de Pesquisas

Coordenação de Pesquisas por Amostra de Domicílios
Adriana Araujo Beringuy

Ministério da Economia
Instituto Brasileiro de Geografia e Estatística - IBGE
Diretoria de Pesquisas
Coordenação de Pesquisas por Amostra de Domicílios

Textos para Discussão
Diretoria de Pesquisas
número 61

O tratamento das informações da PNAD Contínua

Fernanda Karine Ruiz Colenghi Baptista
Elizabeth Belo Hypólito
Felipe Quintas Conde

IBGE
Rio de Janeiro
2022

Instituto Brasileiro de Geografia e Estatística - IBGE

Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro, RJ - Brasil

ISSN 1518-675X **Textos para discussão - Diretoria de Pesquisas**

Divulga estudos desenvolvidos por técnicos do IBGE e/ou de outras instituições, bem como resultantes de consultorias e traduções consideradas relevantes para o Instituto. A série Textos para discussão está subdividida por unidade organizacional e os textos são de responsabilidade de cada área específica.

ISBN 978-85-240-4537-0

© IBGE. 2022

Impressão

IBGE/Centro de Documentação e Disseminação de Informações - CDDI, em 2022.

Capa

IBGE/Centro de Documentação e Disseminação de Informações – CDDI

Ficha catalográfica elaborada pela Gerência de Biblioteca e Acervos Especiais do IBGE.

Baptista, Fernanda Karine Ruiz Colenghi

O tratamento das informações da PNAD Contínua / Fernanda Karine Ruiz Colenghi Baptista, Elizabeth Belo Hypólito, Felipe Quintas Conde. - Rio de Janeiro : IBGE, 2022.

25 p. - (Textos para discussão. Diretoria de Pesquisas, ISSN 1518-675X, n. 61)

ISBN 978-85-240-4537-0

1. Pesquisa Nacional por Amostra de Domicílios Contínua. 2. Erros. 3. Erros de medida. 4. Erros de não resposta. I. Hypólito, Elizabeth Belo. II. Conde, Felipe Quintas. III. IBGE. Coordenação de Pesquisas por Amostra de Domicílios. IV. Título. V. Série.

CDU 314.6(81)

Sumário

Convenções

-	Dado numérico igual a zero não resultante de arredondamento;
..	Não se aplica dado numérico;
...	Dado numérico não disponível;
x	Dado numérico omitido a fim de evitar a individualização da informação;
0; 0,0; 0,00	Dado numérico igual a zero resultante de arredondamento de um dado numérico originalmente positivo; e
-0; -0,0; -0,00	Dado numérico igual a zero resultante de arredondamento de um dado numérico originalmente negativo.

Apresentação

A Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) produzida pelo Instituto Brasileiro de Geografia e Estatística (IBGE) é realizada por amostragem probabilística e tem como principal objetivo captar informações sobre a força de trabalho e suas variações no curto, médio e longo prazos.

Adicionalmente, a pesquisa capta informações sobre outros temas, tais como educação, habitação, outras formas de trabalho (trabalho para o próprio consumo ou uso, trabalho voluntário, afazeres domésticos e cuidados de pessoas), trabalho infantil, acesso à televisão e à Internet e posse de telefone móvel celular para uso pessoal, entre outros que são relevantes para o estudo do desenvolvimento socioeconômico do Brasil.

Desde a coleta dos dados até a produção dos microdados, o acompanhamento da PNAD Contínua é entendido como um processo composto por diversas etapas, cujas realizações estão sujeitas, naturalmente, à ocorrência de erros de dois tipos. Erros amostrais são aqueles relacionados ao planejamento amostral e, portanto, decorrentes do fato da pesquisa investigar uma amostra e não toda a população. Já os erros ocorridos nas demais etapas do processo, como na elaboração do questionário, na atualização do cadastro, na coleta ou no processamento das informações, por exemplo, são chamados de erros não amostrais.

O objetivo deste texto é abordar os processos utilizados pelo IBGE para lidar com dois tipos de erros não amostrais específicos: o erro de não resposta e o erro de medida. Serão detalhadas as ações preventivas adotadas para evitar tais erros; as etapas de crítica e imputação aplicadas para tratar os erros que permanecem nos dados após as ações preventivas e, por fim, o tratamento dado às variáveis de rendimento coletadas pela pesquisa.

Cimar Azeredo Pereira
Diretor de Pesquisas

1. Introdução

A Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) foi implantada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), em todo o território nacional, em janeiro de 2012. É realizada por amostragem probabilística e tem como principal objetivo captar informações sobre a força de trabalho e suas variações no curto, médio e longo prazos.

Adicionalmente, a pesquisa capta informações sobre outros temas, tais como educação, habitação, outras formas de trabalho (trabalho para o próprio consumo ou uso, trabalho voluntário, afazeres domésticos e cuidados de pessoas), trabalho infantil, acesso à televisão e à Internet e posse de telefone móvel celular para uso pessoal, entre outros que são relevantes para o estudo do desenvolvimento socioeconômico do país.

Os indicadores relacionados à força de trabalho são divulgados mensalmente para Brasil por meio de trimestres móveis; trimestralmente desagregados para Brasil, Grandes Regiões, Unidades da Federação, Regiões Metropolitanas que incluam o município da capital e capitais e, anualmente são divulgados indicadores complementares, com a mesma desagregação geográfica daqueles divulgados trimestralmente.

Os demais temas da pesquisa, em geral, são divulgados anualmente. Porém, o intervalo de tempo entre duas edições consecutivas para o mesmo tema pode ser reduzido ou ampliado de acordo com a variação dos seus indicadores ao longo do tempo.

A PNAD Contínua é parte de um Sistema Integrado de Pesquisas Domiciliares (SIPD), um modelo no qual as diversas etapas de produção de pesquisas são conduzidas de forma associada. Atualmente, esse sistema é composto por quatro pesquisas: PNAD Contínua, PNAD Covid19, Pesquisa Nacional de Saúde (PNS) e Pesquisa de Orçamentos Familiares (POF). A amostra de cada uma delas é um subconjunto de uma amostra mestra do SIPD. Além disso, os conceitos, os processos de coleta, de crítica e de imputação dos dados, por exemplo, são harmonizados entre essas pesquisas, sempre respeitando as particularidades de cada uma delas.

A Amostra Mestra do SIPD é conglomerada em dois estágios, com estratificação das unidades primárias de amostragem (UPA), as quais são definidas como um setor censitário ou um conjunto de setores censitários com ao menos 60 domicílios particulares permanentes. A estratificação das UPAs ocorreu em quatro etapas sequenciais.

A primeira delas foi a estratificação por divisões administrativas importantes para a divulgação dos resultados, quais sejam: município da capital; demais municípios pertencentes à Região Metropolitana (RM) ou à Região Integrada de Desenvolvimento Econômico (RIDE); municípios pertencentes ao colar ou expansão metropolitana ou a outra RM, municípios pertencentes à RIDE que estejam situados em outra Unidade da Federação (UF); e demais municípios da UF.

A segunda etapa de estratificação foi baseada em informações de meso e microrregiões, além de outros conhecimentos fornecidos pela Diretoria de Geociências do IBGE. A terceira etapa foi a estratificação por situação, onde o objetivo foi agrupar as UPAs

em dois estratos segundo a situação dos seus domicílios, urbana e rural. Por fim, ocorreu uma etapa de estratificação estatística que buscou agrupar UPAs de forma a minimizar a variância do estimador do total da característica de interesse. Dentro de cada estrato final foram selecionadas UPAs com probabilidade proporcional ao tamanho. Assim, ao fim, 15.096 UPAs compõem a amostra mestra do SIPD.

A amostra trimestral da PNAD Contínua é composta por todas as UPAs da amostra mestra, sendo 14 domicílios selecionados aleatoriamente em cada uma delas, totalizando cerca de 211 mil domicílios. Cada domicílio selecionado é investigado cinco vezes, uma vez a cada trimestre. Assim, por exemplo, um domicílio que tem sua primeira entrevista realizada em fevereiro de um determinado ano, também é entrevistado em maio, agosto, novembro e, por fim, em fevereiro do ano seguinte.

A cada trimestre, a pesquisa tem cerca de 20% dos domicílios em cada uma das cinco entrevistas. Isso garante que, entre dois trimestres consecutivos haja uma sobreposição de 80% dos domicílios e, entre o mesmo trimestre de anos consecutivos, a sobreposição seja de 20% dos domicílios.

De janeiro de 2012 até março de 2020, as entrevistas da PNAD Contínua foram realizadas exclusivamente pelo modo de coleta CAPI (*computer-assisted personal interviewing*), ou seja, com contato face a face entre o entrevistador e o informante, por meio de questionário eletrônico implementado no Dispositivo Móvel de Coleta (DMC), um um *smartphone* com algumas funções bloqueadas.

De abril de 2020 até julho de 2021, com a necessidade de isolamento social devido à pandemia da Covid19, as entrevistas da PNAD Contínua passaram a ser coletadas via CATI (*computer-assisted telephone interviewing*), isto é, por telefone, com o auxílio do mesmo dispositivo eletrônico usado anteriormente.

A partir de agosto de 2021, com a flexibilização do isolamento e gradativo retorno presencial com equipamentos de segurança, a coleta está sendo feita por CAPI e CATI. Uma pergunta foi adicionada ao questionário com a informação do tipo de coleta, de forma a subsidiar estudos futuros.

Maiores informações sobre a PNAD Contínua estão disponíveis em suas notas metodológicas (IBGE, 2014), notas técnicas da pesquisa (IBGE, 2020) e no texto sobre o planejamento da amostra mestra do SIPD (FREITAS; ANTONACI, 2014).

Sendo a PNAD Contínua um processo de pesquisa, ela é composta por diversas etapas, cujas realizações estão sujeitas, naturalmente, à ocorrência de erros. Erros amostrais são erros relacionados ao planejamento amostral e, portanto, decorrentes do fato da pesquisa utilizar uma amostra e não a população total objeto de estudo. O patamar de erro amostral aceitável para fornecer as informações desejadas é levado em conta durante o planejamento amostral. No caso da PNAD Contínua, a pesquisa foi desenhada objetivando coeficiente de variação máximo de 15% para a taxa de desocupação das pessoas de 14 anos ou mais de idade, por Unidade da Federação. Os erros amostrais obtidos na prática, ou seja, após a realização da pesquisa, são calculados e reportados rotineiramente.

Os erros ocorridos nas demais etapas do processo, como no desenho do questionário, na atualização do cadastro, na coleta ou no processamento das informações, por exemplo, são chamados de erros não amostrais. Estes erros podem ter diversas origens, tais como:

- Mudança do tipo de espécie do domicílio entre o período de realização da listagem das unidades elegíveis para a pesquisa e o período de coleta da pesquisa. Nesta situação, domicílios que seriam elegíveis para seleção na pesquisa deixam de fazer parte do escopo do público-alvo determinado, como, por exemplo, um domicílio ocupado no momento de listagem passou a ser um domicílio vago no momento da coleta;
- Uma unidade selecionada se recusa a participar;
- Um entrevistador não segue os procedimentos de coleta e influencia a resposta;
- Um respondente fornece uma informação equivocada;
- Ocorrer falhas no sistema de captura e transmissão dos dados ou no processamento, entre muitas outras possibilidades.

Para Biemer e Lyberg (2003), erros não amostrais podem ser agrupados em cinco categorias: erro de especificação, erro de cadastro, erro de não resposta, erro de medida e erro de processamento.

Estabelecer limites máximos para os erros não amostrais durante o planejamento ou mesmo calculá-los após a realização da pesquisa não é uma tarefa simples uma vez que depende da existência de bases de dados livres de erros para comparação com os resultados obtidos. Diante desse obstáculo, o IBGE, de forma a garantir a qualidade da PNAD Contínua, investe grandes esforços na prevenção e no tratamento desses erros.

O objetivo do presente texto é abordar os processos utilizados pelo IBGE para lidar com dois tipos de erros não amostrais específicos: o erro de não resposta e o erro de medida. A seção 2 aborda os conceitos de erro e as ações preventivas adotadas pelo IBGE; a seção 3 apresenta os processos de crítica e imputação utilizados para tratar os erros que permanecem nos dados após todas as ações preventivas; por fim, a seção 4 detalha o tratamento dado às variáveis de rendimento coletadas pela pesquisa.

2. Erros de não resposta e de medida

A não resposta ocorre quando não é possível medir a informação desejada para a unidade selecionada. Pode ser de dois tipos: não resposta de unidade, quando nenhuma informação do questionário é obtida; e de item, quando a entrevista é realizada, porém um ou mais itens, ou até mesmo partes inteiras do questionário ficam em branco. O erro de não resposta ocorre quando a estimativa produzida com base nos respondentes difere da que seria produzida usando a amostra completa (Groves et al 2009).

A não resposta de unidade costuma ocorrer por falha de contato com a unidade quando o acesso está temporariamente interrompido, o(s) morador(es) está(ão) ausente(s) no período estabelecido para a coleta etc., recusa da unidade em participar da pesquisa ou impossibilidade do respondente fornecer as informações (barreiras linguísticas, doença física ou mental, etc.). A não resposta de item ocorre, por exemplo, quando o informante não sabe ou se recusa a responder um determinado item; o entrevistador salta a pergunta; a informação está fora dos limites estabelecidos para a resposta ou não atende determinadas regras da pesquisa e, portanto, é apagada durante o processamento; ou a resposta é perdida durante o processamento.

Já o erro de medida acontece quando a informação coletada difere da que deveria ter sido coletada. De acordo com Biemer e Lyberg (2003), pode ocorrer devido ao questionário, ao modo de coleta, ao entrevistador, ao respondente, ao sistema de informação (fonte utilizada pelo respondente para recuperar a informação solicitada) ou ao ambiente em que ocorre a entrevista, de forma individual ou interrelacionada. Por exemplo, a redação da pergunta pode ser complexa e levar a interpretações e respostas equivocadas sobre a mesma, o entrevistador pode influenciar a resposta ou mesmo falsificá-la, o respondente pode mentir ou chutar a informação, entre outras possibilidades.

Para reduzir a não resposta de unidade, o IBGE envia cartas de apresentação da PNAD Contínua que destacam os objetivos e as formas de utilização das informações coletadas, assim como a garantia do sigilo (para maiores informações, ver IBGE, 2018); investe em treinamento de abordagem para seus entrevistadores; disponibiliza canal para a confirmação da identidade do entrevistador[1]; faz mais de uma tentativa de contato, inclusive, em alguns casos, com alteração de entrevistador; entre outros esforços. A não resposta remanescente é tratada por meio de ajuste nos pesos básicos da pesquisa. Maiores informações sobre esse procedimento podem ser obtidas em Freitas e Antonaci (2014).

Em geral, a melhor forma de lidar com a não resposta de item e com o erro de medida é por meio de medidas preventivas relacionadas ao desenho do questionário (BIEMER; LYBERG, 2003). Ao longo do planejamento da PNAD Contínua, o IBGE realizou 15 fóruns com usuários[2], nos quais foram debatidos os objetivos e as perguntas do questionário da pesquisa. Também foram realizados dois testes de questionário em campo. O primeiro deles ocorreu em 2008 na Bahia, Rio de Janeiro, São Paulo e Rio Grande do Sul e teve como objetivo central a validação do questionário de trabalho e o aplicativo de coleta com base em uma pequena amostra selecionada de forma intencional. O segundo teste, em 2009, foi realizado no Pará, Pernambuco, São Paulo, Rio Grande do Sul e Distrito Federal e, além do questionário e aplicativo de coleta, testou a viabilidade do espalhamento da amostra da pesquisa e o esquema de rotação. Após a implantação da pesquisa, novos ajustes no questionário têm sido implementados sempre que necessários.

Outras medidas importantes para reduzir a não resposta de item e o erro de medida são a diminuição da sensação de ameaça à privacidade por meio da garantia de sigilo e o

.....

[1] https://respondendo.ibge.gov.br/verifique-a-identidade-do-entrevistador.html

[2] https://www.ibge.gov.br/arquivo/projetos/sipd/default.php

treinamento do entrevistador para esclarecimento de eventuais dúvidas. Como mencionado anteriormente, o IBGE investe continuamente em ambas as medidas.

Além disso, a PNAD Contínua adota diversos mecanismos para identificar erros de não resposta e de medida durante a coleta dos dados como a criação de relatórios de acompanhamento de campo, a inserção de regras de consistência no aplicativo de coleta e no sistema de codificação. Apesar de todos os esforços preventivos, a presença de erros não amostrais é inevitável na realização de pesquisas. Por conta disso, após o fechamento da coleta e durante o processamento centralizado dos dados, são realizadas diversas etapas de verificação e ajustes desses erros.

3. Tratamento dos erros de não resposta e de medida

Os dados da PNAD Contínua passam por etapas de crítica, ou seja, processo para detectar erros em dados estatísticos (Chambers, 2006) e de imputação, procedimento no qual se insere um valor para um item específico no qual a resposta está ausente ou é inutilizável (ONU, 2000). Esses processos permitem aumentar a qualidade do produto final e produzir um banco de dados completo.

A crítica de dados da PNAD Contínua identifica a não resposta e dois tipos de erros de medida: 1) respostas inválidas, ou seja, fora dos valores possíveis para a pergunta, e 2) respostas inconsistentes, isto é, que não atendem a regras específicas, previamente estabelecidas. Esses erros, quando detectados ainda no campo, seja no momento da coleta ou da codificação das atividades, podem ser corrigidos pela equipe de coleta, mediante verificação com o informante. Quando identificados após o fechamento da coleta são tratados com imputação.

A imputação na PNAD Contínua ocorre, preferencialmente, de forma determinística. Nesse caso, informações auxiliares são utilizadas para estabelecer, com base em regras preestabelecidas, um único valor possível a ser atribuído ao registro com não resposta ou erro de medida. Essa etapa é executada com base no pacote estatístico no SAS *Enterprise Guide* [3].

Na ausência de informações que permitam a imputação determinística, a PNAD Contínua adota a imputação probabilística. Nesse caso, a imputação é parte de um processo aleatório e, portanto, se repetida, pode gerar valores diferentes. O método escolhido para a pesquisa foi o vizinho mais próximo (NIM, do inglês *Nearest-neighbour Imputation Methodology*), implantado no *Canadian Census Edit and Imputation System* (CANCEIS), um pacote de crítica e imputação desenvolvido pelo *Statistics Canada*.

O NIM é um método de imputação hot-deck, ou seja, em que o valor com erro (receptor) é substituído pelo valor de um registro doador pertencente à mesma pesquisa, tão semelhante quanto possível ao receptor e selecionado dentre aqueles que não

.....

[3] https://support.sas.com/en/software/enterprise-guide-support.html.

violaram nenhuma das regras de crítica previamente estabelecidas.

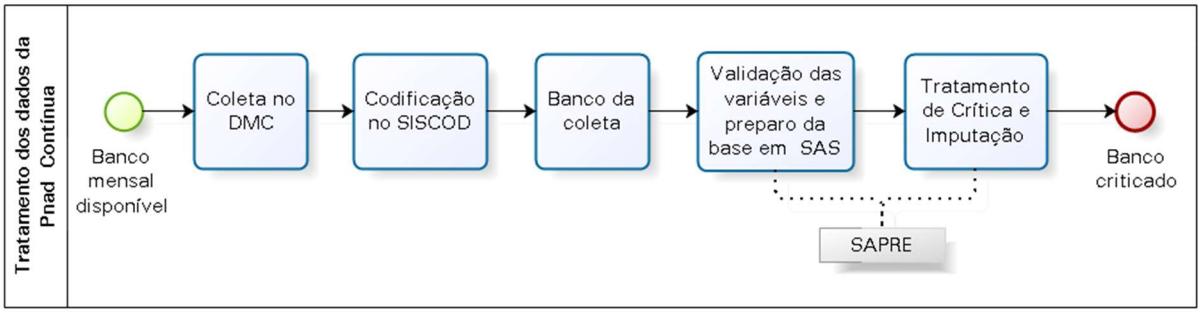
No NIM, a similaridade é medida com base em uma função de distância calculada para variáveis auxiliares disponíveis para os registros doadores e receptores. Assim, esses registros devem possuir o maior número possível de variáveis auxiliares qualitativas com valores iguais e as variáveis auxiliares numéricas com valores próximos. A importância de cada variável nesse processo é indicada por um peso definido pela equipe envolvida na pesquisa.

Para cada registro receptor, primeiramente são identificados os potenciais doadores, ou seja, os registros que foram aprovados pelas regras de crítica e, dentre eles, os vizinhos mais próximos. A preferência é dada para os registros que estão geograficamente perto do receptor, de acordo com a ordenação do banco de dados da pesquisa. Então, para cada vizinho mais próximo, são identificadas ações de imputação que, se implementadas, permitiriam ao receptor não violar nenhuma das regras de crítica. Por fim, uma ação que impute o menor número possível de variáveis é selecionada aleatoriamente (STATISTICS CANADA, 2007).

Anteriormente à PNAD Contínua, o CANCEIS foi utilizado em outras pesquisas do IBGE, tais como o Censo Agropecuário de 2006, a Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2007 a 2015, o Censo Demográfico 2010 e a Pesquisa de Orçamentos Familiares (POF) 2008-2009. Posteriormente, foi utilizado na Pesquisa Nacional de Saúde (PNS) 2013 e 2019, assim como na POF 2017-2018.

A Figura 1 apresenta um resumo do processo de coleta e tratamento dos dados da PNAD Contínua. Maiores detalhes serão apresentados nas subseções 3.1, 3.2 e na seção 4, para o caso específico dos rendimentos de trabalho.

Figura 1: Tratamento dos dados da PNAD Contínua



Fonte: IBGE. PNAD Contínua. Agosto de 2020.

3.1. Crítica e correção realizada durante a coleta de dados

As primeiras críticas ocorrem ainda durante a entrada dos dados no DMC. As mesmas são classificadas como advertência ou erro. As críticas de advertência exigem que a informação seja confirmada. Por exemplo, quando uma pessoa de menos de 17 anos (V2009) possui carteira de trabalho assinada (V4029), aparece a seguinte mensagem na tela do DMC: " Menor de 17 anos com carteira assinada. Confirma? ". Caso o entrevistador ratifique a informação, a entrevista segue normalmente. Caso contrário, precisa efetuar a correção da informação.

As críticas de erro no DMC impedem a sequência da entrevista até que a informação seja revista. Por exemplo, se uma pessoa declara que não sabe ler e escrever (V3001) e diz que concluiu o curso Regular do Ensino Fundamental (V3009A e V3014), aparece uma mensagem de erro. Para prosseguir o fluxo da entrevista é necessário que o entrevistador corrija a resposta do quesito de nível de ensino ou do quesito que informa se a pessoa sabe ler e escrever, após verificação, *in loco*, com o morador. Outro exemplo de erro na entrada de dados no DMC que pode ocorrer é o registro de variáveis numéricas com valores fora da faixa preestabelecida para a variável. A idade deve variar entre 0 e 130 anos, as horas habitualmente trabalhadas na semana devem estar entre 1 e 120, entre outras.

Após a descarga dos dados do DMC para o *desktop* no sistema de armazenamento de dados em ORACLE, ocorrem outras duas etapas de crítica, uma no Sistema de Codificação (SISCOD), usado para as variáveis de ocupação e atividade dos questionários de trabalho das pessoas de 14 anos ou mais, trabalho voluntário e trabalho infantil, e outra no Sistema de Indicadores Gerenciais (SIGC), usado pelos supervisores e equipe de planejamento da pesquisa para acompanhamento da coleta.

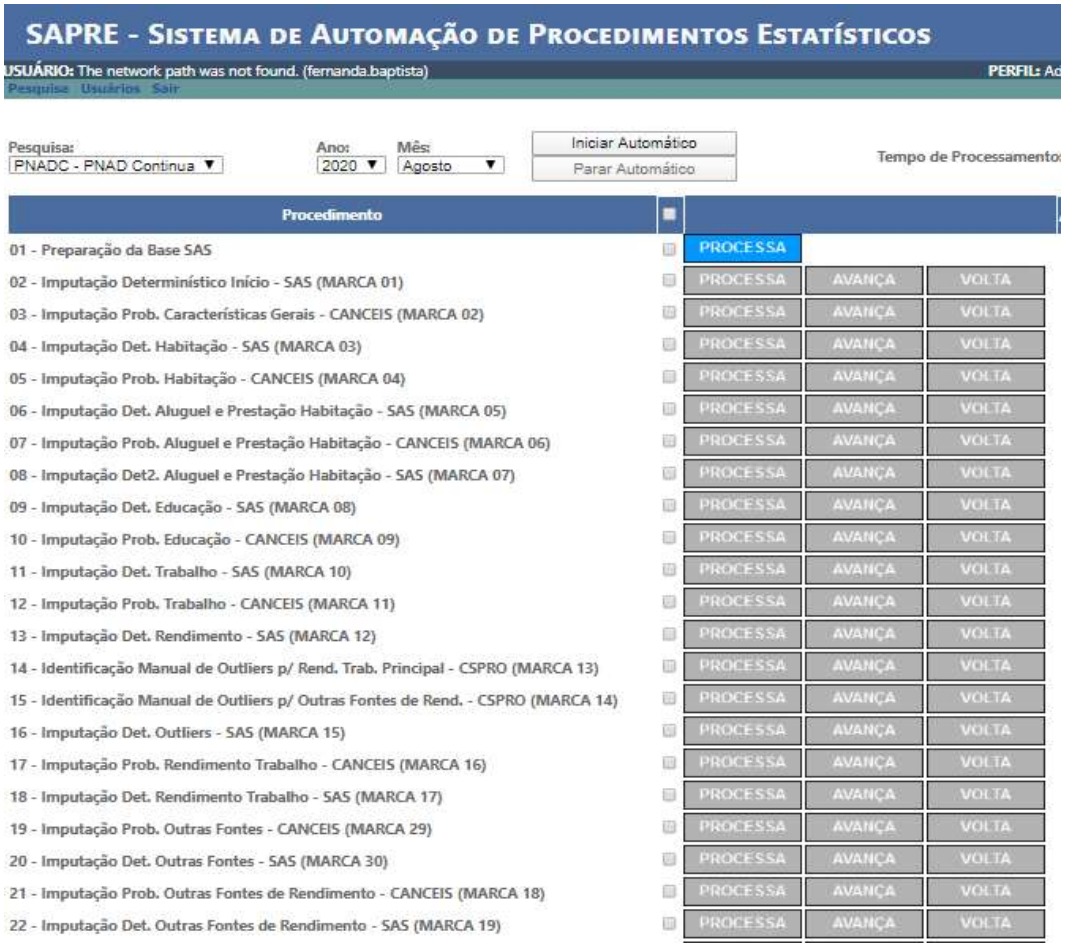
As críticas do SISCOD têm por objetivo detectar erros no conjunto de variáveis que estão relacionadas com as classes de atividades e com os grupos de ocupação. Por exemplo, são considerados erros os casos em que um advogado não possui nível superior completo ou que um vendedor ambulante não trabalha em via pública. Quando ocorrem, as informações precisam ser verificadas junto ao entrevistador ou ao informante e corrigidas. Caso o erro seja na descrição da ocupação ou da atividade, a correção pode ser feita no SISCOD. Caso contrário, é necessário retornar o questionário ao DMC, fazer o acerto, que pode gerar ou não mudança de fluxo das perguntas, e então transmitir novamente o questionário para, novamente, passar pelo SISCOD.

No SIGC são feitas críticas de valores de rendimentos ignorados ou extremos. Os registros com valores nestas situações são apresentados em um relatório e precisam ser confirmados com o entrevistador ou informante. Maiores detalhes sobre essa etapa serão apresentados na seção referente à crítica e imputação de rendimentos.

3.2. Crítica e imputação realizada após o fechamento da coleta

O banco de dados da PNAD Contínua é fechado mensalmente, ou seja, após a coleta e codificação dos dados de cada um dos meses de referência da pesquisa, o banco mensal é bloqueado para as Unidades Estaduais (UEs) e liberado para o processamento no Sistema de Automação de Procedimentos Estatísticos (SAPRE). O SAPRE encadeia, de forma simples e direta, diversas etapas de crítica e de imputação dos itens com não resposta, inválidos ou inconsistentes, ocorridas no CSPro, SAS Enterprise Guide e CANCEIS.

Figura 2: Etapas do processamento centralizado no SAPRE, agosto de 2020



Fonte: IBGE. PNAD Contínua. Agosto de 2020.

A Figura 2 ilustra as etapas do SAPRE para o processamento centralizado da base de dados de agosto de 2020. Conforme apresentado, a crítica e a imputação nesse sistema ocorrem sequencialmente para os quesitos de características gerais do domicílio e dos moradores, habitação, gastos com aluguel e prestação do domicílio, educação, trabalho de pessoas de 14 anos ou mais de idade, rendimentos de trabalho, rendimentos de outras fontes, trabalho de pessoas de 5 a 13 anos de idade, rendimentos do trabalho das pessoas de 5 a 13 anos de idade, tecnologias de informação e imputação

determinística final. Outros módulos temáticos, à medida que são incorporados à pesquisa, passam a compor novas etapas no SAPRE. Para cada uma dessas partes da PNAD Contínua, as variáveis, sejam elas quantitativas ou qualitativas, passam por tratamentos determinísticos e probabilísticos.

No SAS, são conferidas diversas regras de crítica e, havendo violação de alguma delas, assim como disponibilidade de informações auxiliares, os registros com falha são imputados deterministicamente. Em último caso, ou seja, quando não é possível fazer imputação determinística, algumas variáveis da parte de características gerais, como sexo, idade ou condição no domicílio, são imputadas por procedimento aleatório. Existem regras no SAS que se baseiam na imputação aleatória dentro de um intervalo predefinido de valores. A seguir são relacionados alguns exemplos de erros que compõem as regras de crítica da pesquisa.

- Conferência de parâmetros iniciais do sistema, como, por exemplo, os relacionados às datas de referência.
- Validação das entrevistas realizadas, identificando a existência de situações como a atribuição de entrevista realizada a um domicílio sem morador registrado.
- Consistência de informações de características gerais dos moradores (condição no domicílio, sexo, idade, cor ou raça), tais como: incompatibilidade de idade entre pais e filhos; obrigatoriedade de existir pessoa responsável pelo domicílio na variável de condição no domicílio.
- Conferência da variável de número de pessoas no domicílio (V2001) com a quantidade de pessoas registradas e validadas na lista de moradores. No caso de divergência, a variável V2001 é corrigida deterministicamente para que as informações sejam iguais.
- Consistência entre as variáveis indicadoras de recebimento de rendimento em dinheiro e das variáveis de valor recebido. Por exemplo, se há registro de valor em dinheiro do rendimento habitual de trabalho (V403312), mas, por alguma falha, a variável que indica o recebimento desse tipo de rendimento (V40331) está em branco, esta é imputada com o código 1 (recebia/fazia normalmente nesse trabalho rendimento/retirada em dinheiro). Em outro caso, um informante declara que recebeu rendimento de Benefício Assistencial de Prestação Continuada (BPC-LOAS), porém devido a algum erro o valor do rendimento não foi computado na variável correspondente (V5001A2). A variável V5001A2 é imputada pelo valor do salário mínimo vigente no mês de referência, uma vez que, pela lei, esse é o valor do benefício.

Após a etapa de imputação determinística, é realizada a imputação probabilística no CANCEIS. Para cada um dos temas, são definidas variáveis auxiliares para o cálculo da similaridade entre receptores e doadores, pertencentes aos temas anteriormente imputados, conforme a ordem definida para o SAPRE. As variáveis auxiliares, que compõem a matriz para o cálculo de distância, também pertencem ao tema que está sendo imputado.

É importante destacar que para cada variável imputável da pesquisa há uma variável de marca que registra se seus valores sofreram ou não modificação em relação

ao banco de dados gerado pelo fechamento da coleta. Essas marcas são atualizadas ao longo das etapas do SAPRE. Por exemplo, a variável V3001 (sabe ler e escrever?) tem suas alterações registradas na variável de marca MV3001, que recebe valor “1” quando há modificação em alguma etapa do processamento e “0” (zero) caso contrário.

Adicionalmente, o SAPRE produz uma série de relatórios e gráficos sobre o processo de crítica e imputação, permitindo verificar possíveis erros no processamento ou na coleta. Por exemplo:

- Relatórios de percentuais de imputação para cada uma das variáveis da pesquisa por Brasil, Grandes Regiões e UFs;
- Série histórica da imputação dos últimos 12 meses para cada variável;
- Gráficos com as variáveis mais imputadas e seu comparativo com os últimos 12 meses.

Além do tratamento dos dados, o sistema também permite fazer algumas conferências como, por exemplo, das UPAs ou da adequação dos formatos das variáveis; calcular o percentual de entrevistas realizadas e a quantidade de pessoas na amostra por UF; assim como gerar relatórios longitudinais de supervisão. Os relatórios longitudinais mostram situações com possíveis erros de coleta e são enviados para as Unidades Estaduais com os códigos de identificação do domicílio e morador em formato de planilha para controle e treinamento dos entrevistadores. Dentre tais relatórios, a título de exemplo, existe o de empregados do setor privado ou trabalhadores domésticos com carteira assinada e que trabalhavam mais de 40 horas semanais com rendimento habitual do trabalho inferior ao salário mínimo vigente.

Por fim, ainda com auxílio do SAPRE, são gerados os bancos finais mensais, trimestrais (acumulando os três últimos meses) e anuais (sempre após o processamento do mês de dezembro) da pesquisa; são calculados os pesos ajustados pela não resposta com pós-estratificação; e criadas as variáveis derivadas da pesquisa. O último processo é o carregamento dos bancos no ORACLE.

4. Detalhamento do processo de crítica e imputação dos rendimentos

4.1. Etapa em campo

No SIGC são feitos relatórios de valores de rendimentos ignorados ou extremos. Os limites de valores extremos são obtidos pelos dados das variáveis do banco da coleta do ano anterior de primeira e quinta entrevistas. A base é filtrada, mantendo apenas os 5% mais ricos, isto é, as pessoas que possuem rendimento de trabalho e de pensão e

aposentadoria acima do percentil 95, e calcula-se a média destes rendimentos por UF para o município da capital e resto da UF. O resultado desta média é utilizado como limite.

As variáveis de rendimentos de outras fontes que possuem limites estabelecidos em lei, como seguro desemprego e bolsa família, têm seus limites fixados de acordo com a legislação. Já as demais variáveis de rendimentos de outras fontes, por ter uma base pequena de respondentes que recebem, os limites são únicos para uma mesma grande região (pensão alimentícia, doação, rendimento de aluguel ou arrendamento ou outro programa social) ou para o Brasil (poupança, aplicação ou outros rendimentos).

Os registros com valores nas situações acima descritas precisam ser confirmados com o entrevistador ou informante. Contudo, não há um impedimento no sistema, como, por exemplo, bloqueio do valor informado, caso a confirmação não seja feita.

Para controle das confirmações, há no sistema variáveis auxiliares com inicial "R" que indicam se o rendimento extremo foi confirmado pela rede de coleta no SIGC. A variável R403312, por exemplo, indica se houve a confirmação com o entrevistador ou informante sobre o rendimento habitual do trabalho principal em dinheiro alertado pelo sistema. Tais variáveis são utilizadas nas etapas posteriores de crítica no processamento do SAPRE.

4.2. Primeira etapa de imputação determinística

O primeiro procedimento no SAPRE para a imputação dos rendimentos após o fechamento do banco de dados é uma imputação determinística no SAS. Nesta etapa, são realizados três tipos de correções, os quais são listados abaixo:

- Valores de rendimento em dinheiro do trabalho de funcionários públicos e militares que estejam entre 0,85 e 1 salário mínimo são ajustados para 1 salário mínimo.
- Rendimentos em produtos e mercadorias recebidos em ocupações para as quais esse tipo de rendimento não é válido (atividades não compreendidas nos ramos da agricultura, silvicultura, pecuária, extração vegetal, pesca e piscicultura) são acrescentados aos rendimentos em dinheiro.
- Rendimentos do trabalho secundário de trabalhadores domésticos que declararam a mesma atividade nos dois trabalhos e que deveriam ter apenas um trabalho[4] são somados aos rendimentos do trabalho principal.

4.3. Inspeção visual de rendimentos extremos

A inspeção visual acontece mediante um critério de identificação de dados extremos (altos ou baixos) observados na distribuição estatística empírica de uma variável. Também pode consistir em dados não representativos de um conjunto de observações, como, por exemplo, valores de rendimentos habituais e efetivos inferiores a R$10,00, exceto

.....

[4] A orientação desta definição de único trabalho para trabalhadores domésticos que exercem a mesma atividade, está descrita no subitem 12.9 do manual básico da entrevista disponível em: https://biblioteca.ibge.gov.br/visualizacao/instrumentos_de_coleta/doc5361.pdf

rendimentos efetivos com valores iguais a zero. Os critérios usados para definir valores extremos nesta etapa baseiam-se em estatísticas de ordem, calculadas em determinados estratos, detalhados a seguir.

Para separar os registros em grupos homogêneos em relação aos valores de rendimento foi usado um método que utiliza Árvores de Regressão (BREIMAN ET AL, 1984) para variáveis de rendimento de trabalho. Em linhas gerais, esta técnica consiste em um método de estratificação que utiliza variáveis explicativas, ou seja, um conjunto de características das pessoas respondentes da PNAD Contínua, para classificar os registros em grupos homogêneos. Para tal procedimento, foi utilizada a função RPART do software R[5].

A formação da árvore se dá através de partições binárias, sempre distribuindo os indivíduos em dois grupos mutuamente exclusivos, que são chamados de nós. O grupo inicial que contém todos os indivíduos é chamado de nó raiz e os estratos finais, de nós terminais. Foram construídas árvores de regressão com os valores de rendimentos das variáveis dependentes do modelo para os dados de cada um dos meses de janeiro a dezembro de 2012 visando à definição dos estratos. Para as variáveis de rendimento de trabalho principal, foram definidos 7 estratos para cada UF, os quais consideraram como variáveis explicativas os anos de estudo, a posição na ocupação, a contribuição para previdência e os grupamentos de atividade.

Para as variáveis de rendimento de trabalho secundário, foram definidos 3 estratos para cada UF, os quais consideraram como variáveis explicativas os anos de estudo e a posição na ocupação. Para as variáveis de rendimento de outros trabalhos não foi factível fazer estratificação devido ao número reduzido de registros com valores nas variáveis. Para as variáveis de rendimentos de outras fontes também não houve estratificação de grupos em relação aos valores de rendimento. Por se referirem a valores monetários, para as variáveis de aluguel e prestação de domicílio foram definidos estratos de Unidade da Federação e Grandes Regiões, respectivamente.

O critério adotado para detecção de rendimentos extremos suspeitos em cada estrato usa o método da razão que considera a distância interquartílica (IQR = Q3 – Q1), o qual é utilizado para detecção de rendimentos e despesas extremos na POF (IBGE, 1991). Este método é definido da seguinte forma:

Se, $Q = \frac{X_{(i)} - X_{(i-1)}}{Q_3 - Q_1} > q$, então $X_{(i)}$ é rendimento extremo, onde:

$X_{(i)}$ é o i - ésimo maior valor;

$X_{(i-1)}$ é o $(i-1)$- ésimo maior valor;

$Q_3 - Q_1$ é a distância entre o 3o e o 1o quartil;

.....

[5] RPART - Recursive Partitioning. É uma função do software R que trabalha tanto com árvores de regressão quanto de classificação. O R é um software livre e pode ser obtido através do endereço http://www.R-project.org.

q é o valor do critério do corte.

Cabe observar que:

a) Todos os valores de $X_{(i)}$> $X_{(i)}$ também são considerados extremos;

b) Esse teste é feito na cauda superior da distribuição quando $X_{(i)}$> P95, onde P95 é o percentil de ordem 95;

c) Na cauda inferior o procedimento é similar, neste caso o numerador é tomado em valor absoluto, $|X_{(i)} - X_{(i-1)}|$, e o teste é feito para $X_{(i)}$< P5, onde P5 é o percentil de 5a ordem.

d) Os limites de q são definidos em 1,5 e 9, respectivamente, para detecção dos extremos inferior e superior (IBGE, 1991).

De forma independente ao método da razão, todos os rendimentos menores que R$10 ou maiores que R$ 50.000 são separados para serem avaliados na etapa seguinte. Este critério não é utilizado nas variáveis de rendimentos de outras fontes que possuem limites definidos em lei.

O método da razão usando a distância interquartílica é aplicado dentro de cada estrato. Quando não há estratificação, este método é aplicado considerando todos os registros como um estrato único.

Em seguida, a identificação de valores extremos verdadeiros ou incoerentes é realizada no CSPro. São 2 aplicações separadas: para rendimento do trabalho principal e rendimentos de outras fontes. Cada registro é avaliado individualmente por um técnico com base em outras variáveis do questionário e na entrevista anterior e se o valor extremo for considerado verdadeiro não será feita nenhuma marcação. A inspeção é feita mesmo que tenha ocorrido confirmação do valor extremo pela rede de coleta no SIGC, detalhado na subseção 4.1.

Se for considerado incoerente, será feita uma marcação no registro para ser imputado deterministicamente no SAS com base em informações do próprio registro ou de entrevista anterior, se atender as exigências que serão detalhadas na subseção 4.4, ou probabilisticamente, no CANCEIS, cujo procedimento será explicado na subseção 4.5.

Para rendimentos do trabalho secundário e de outros trabalhos, optou-se por adotar o método de razão interquartílica com cercas automáticas, isto é, os valores detectados são automaticamente imputados e não passam pela inspeção visual. Para valores de prestação e aluguel não há uma aplicação no CSPro devido ao número reduzido de registros. Quando aparecem valores extremos, estes são avaliados no SAS pelo técnico que faz o processamento do SAPRE na etapa 6 (ver Figura 2), antes da imputação probabilística dos valores de aluguel e prestação. Oportunamente, esses valores são submetidos a confirmação através do contato com o respondente, realizado pelas redes de coleta.

4.4. Segunda etapa de imputação determinística

Após a inspeção dos valores extremos, algumas regras de crítica e imputação são executadas no SAS de modo determinístico. Dentre estas, merecem destaque as críticas longitudinal e aquela que utiliza variável auxiliar.

A imputação longitudinal ocorre quando, para uma determinada variável *y*, o valor com não resposta ou erro de medida do *i*-ésimo registro no trimestre *t*, y_{it}, é imputado pelo valor fornecido pelo mesmo registro na entrevista anterior, $y_{i(t-1)}$, realizada no trimestre *t-1*. Na PNAD Contínua ele é utilizado apenas para os rendimentos do trabalho principal e secundário. Além disso, no caso do rendimento habitual, as seguintes exigências devem ser satisfeitas:

- O domicílio deve estar entre a segunda e a quinta entrevista no trimestre t;
- O código da ocupação deve ser o mesmo nos dois trimestres; e
- O rendimento habitual do trimestre t-1 não pode ter sido imputado.

No caso do rendimento efetivo, as exigências são:

- O domicílio deve estar entre a segunda e a quinta entrevista no trimestre t;
- O mês de referência da entrevista, tanto no trimestre t como no trimestre t-1 deve ser diferente de dezembro;
- A posição na ocupação deve ser a mesma nos dois trimestres e igual a trabalhador doméstico, militar, empregado do setor público ou empregado do setor privado com carteira de trabalho assinada;
- O código de ocupação deve ser o mesmo nos dois trimestres; e
- O valor do rendimento efetivo no trimestre t-1 não pode ter sido imputado.

Quando essas condições não são satisfeitas, a pesquisa busca realizar a imputação dos rendimentos dos trabalhos principal e secundário por meio de uma variável auxiliar. Assim, o rendimento sem resposta ou com erro de medida para o i-ésimo registro no trimestre *t*, y_{it}, é imputado pela variável x_{it}.

Na PNAD Contínua, o valor do rendimento habitual é imputado pelo rendimento efetivo e vice-versa. Entretanto, existem condições para a aplicação dessa regra, conforme descrito, a seguir.

O rendimento habitual do trabalho principal (ou secundário) sem resposta ou com erro de medida é imputado pelo rendimento efetivo do trabalho principal (ou secundário) quando:

- O mês de referência da entrevista é diferente de dezembro;
- A posição na ocupação é diferente de conta própria e empregador; e
- O rendimento efetivo tem uma resposta válida e diferente de zero.

Já o rendimento efetivo do trabalho principal (ou secundário) é imputado pelo

rendimento habitual do trabalho principal (ou secundário) quando:

- O mês de referência da entrevista é diferente de dezembro;
- A posição na ocupação é diferente de conta própria e empregador; e
- O rendimento habitual tem uma resposta válida.

4.5. Imputação probabilística

Quando não é possível realizar a imputação determinística, como descrita na subseção 4.4, os rendimentos de trabalho da PNAD Contínua são imputados de forma probabilística, utilizando o método do vizinho mais próximo, no CANCEIS. A similaridade entre receptores e doadores é medida por meio das seguintes variáveis auxiliares: condição no domicílio, sexo, idade, posição na ocupação, anos de estudo, horas trabalhadas e contribuição para previdência em qualquer trabalho da semana de referência. Todas elas têm peso igual a 1 (um) no cálculo da distância.

Outras informações importantes para compreender e reproduzir o processo de imputação probabilística da PNAD Contínua são listadas abaixo.

- A imputação é realizada de forma conjunta para toda a população da pesquisa e não por subgrupos populacionais (classes de imputação). Assim, por exemplo, um registro pode ser imputado por um doador de outra UF.
- A similaridade entre registros não significa igualdade em todas as variáveis auxiliares. Por exemplo, o sexo do doador pode diferir do sexo do receptor.
- Um doador pode ser usado para mais de um receptor.
- Os pesos amostrais não são considerados na função de distância.

4.6. Terceira etapa de imputação determinística

Esporadicamente, pode acontecer falha na imputação probabilística por não haver doador com preenchimento nas variáveis de rendimentos do registro com falha. Essas situações são pouco prováveis e incomuns. Por exemplo, quando algum morador recebe rendimento de programas sociais, aposentadoria, pensão alimentícia e outros rendimentos, mas não declara o valor. Neste caso, pode ser que não tenha nenhum doador no banco da pesquisa desse mesmo mês com rendimentos em todas estas quatro variáveis dentro da distância que o CANCEIS procura por vizinhos. Dessa forma, não haverá um doador para este registro.

Quando isso acontece, uma nova etapa determinística em SAS faz a imputação pela mediana da variável. Para rendimentos do trabalho principal em dinheiro, a mediana é calculada por UF e posição na ocupação. Para o rendimento em dinheiro do trabalho secundário, a mediana é calculada por UF, sem distinção da posição na ocupação. Para as demais variáveis de rendimentos de trabalho e de rendimentos de outras fontes, considera-se o Brasil como estrato único.

Em geral, não há falhas na imputação probabilística. A título de exemplo, entre julho de 2018 a julho de 2020, houve falha em oito meses para as variáveis de rendimento de trabalho e falha em dois meses para rendimentos de outras fontes. Nestes casos, o número de falhas foi usualmente igual a um, com número máximo de falhas igual a dois em um único mês.

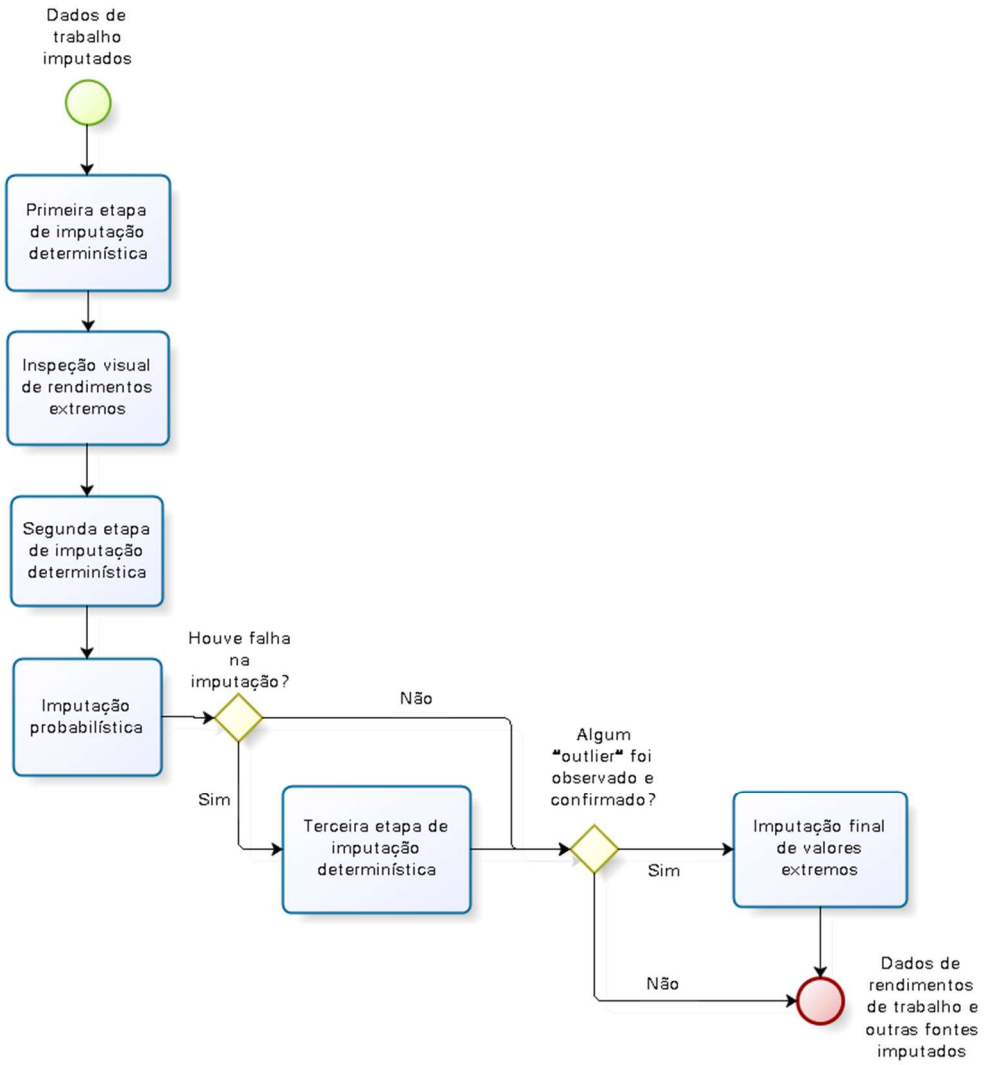
4.7. Imputação final de valores extremos

A última etapa de tratamento dos rendimentos de trabalho consiste na identificação e imputação de rendimentos extremos que foram confirmados pela rede de coleta e considerados válidos na inspeção visual de valores extremos. No entanto, por possuírem valores muito elevados, elas podem gerar impactos artificiais nos indicadores de rendimento médio ou total, assim como na desigualdade.

Nesta etapa, são considerados como *outliers* os rendimentos superiores ao valor da média mais seis desvios-padrão dentro dos estratos definidos pelas Grandes Regiões. Esses valores são imputados deterministicamente pelo valor mais alto dentre aqueles que não ultrapassam o limite, ou seja, que não são considerados *outliers*. Maiores detalhes sobre esse método estão disponíveis no anexo 8 das notas técnicas da pesquisa (IBGE, 2020).

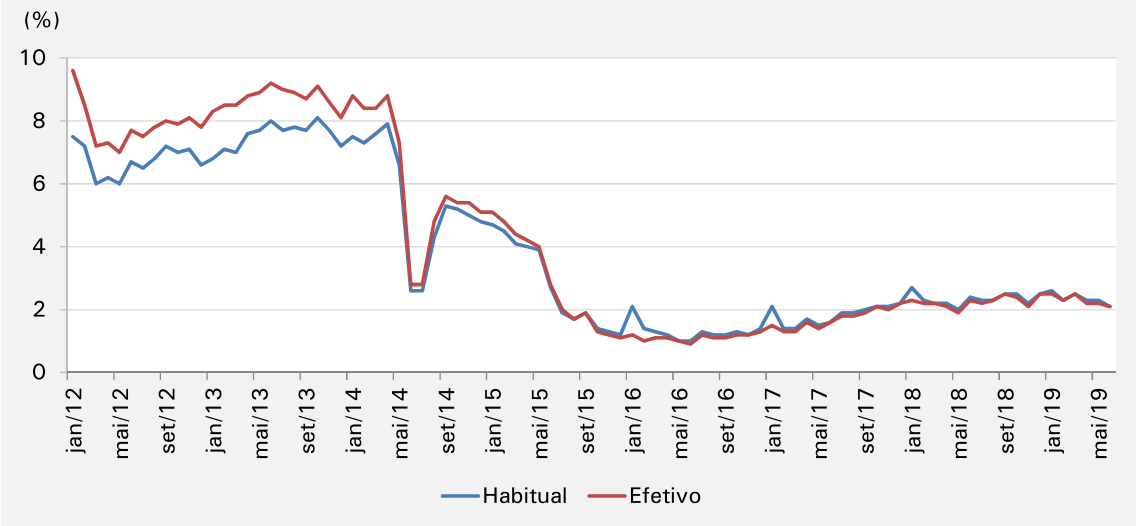
A Figura 3 apresenta o resumo das etapas de crítica e imputação dos rendimentos da PNAD Contínua, anteriormente descritas, exceto a 4.1 que não acontece no SAPRE. Para exemplificar o resultado final do processo, a Figura 4 apresenta as taxas de imputação dos rendimentos habitual e efetivo do trabalho principal desde o início da pesquisa, em 2012, até junho de 2019. Pode-se observar que, mesmo no período da implantação da Pesquisa, a taxa de imputação não superou 10%. A partir de 2015, essa taxa caiu consideravelmente e manteve-se abaixo de 3% até junho de 2019.

Figura 3: Processo de crítica e imputação de rendimentos



Fonte: IBGE. PNAD Contínua. Setembro 2020

Figura 4: Taxa de imputação das variáveis de rendimento do trabalho principal, por tipo de rendimento, Brasil, janeiro de 2012 a junho de 2019



Fonte: IBGE. PNAD Contínua. Janeiro de 2012 a junho de 2019

5. Referências

BIEMER, P. P.; LYBERG, L. E. Introduction to Survey Quality. New Jersey: John Wiley Sons, 2003.

BREIMAN, L. et al. Classification and regression trees. Belmont. Calif.: Wadsworth International Group, 358 p. (1984).

Chambers, R. L. (2006). Evaluation criteria for editing and imputation in EUREDIT. In Statistical Data Editing, Volume No. 3, Impact on Data Quality, U. E. S. Division (ed), 11. New York and Geneva: United Nations Statistical Commission, and United Nations Economic Commission for Europe.

FREITAS, M. P. S. de; ANTONACI, G. de A. Sistema Integrado de Pesquisas Domiciliares: Amostra mestra 2010 e amostra da PNAD Contínua. Texto para discussão n. 50. Rio de Janeiro, 2014.

GROVES, R. M. et al. Survey Methodology. New Jersey: John Wiley Sons, 2009. v. 2nd edition. 448 p.

IBGE. Pesquisa Nacional por Amostra de Domicílios Contínua. Notas técnicas. Versão 1.7. Diretoria de Pesquisas Coordenação de Trabalho e Rendimento. Rio de Janeiro, 2020. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101708_notas_tecnicas.pdf

IBGE. Confidencialidade no IBGE Procedimentos adotados na preservação do sigilo das informações individuais nas divulgações de resultados das operações estatísticas. Rio de Janeiro, 2018. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101636.pdf

IBGE. Pesquisa Nacional por Amostra de Domicílios Contínua. Notas metodológicas. Volume 1. Diretoria de Pesquisas. Coordenação de Trabalho e Rendimento. Rio de Janeiro, 2014. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/trabalho/9171-pesquisa-nacional-por-amostra-de-domicilios-continua-mensal.html?=&t=downloads

IBGE. Pesquisa de Orçamentos Familiares, Tratamento das Informações, v.2. Série Relatórios Metodológicos, v. 10. Rio de Janeiro, 1991.

ONU. Glossary of terms on statistical data editing. Conference of european statisticians methodological material. Genebra, 2000. Disponível em https://ec.europa.eu/eurostat/ramon/statmanuals/files/un_editing_glossary_2000.pdf

STATISTICS CANADA. Canceis: User's Guide. Versão 4.5. Ottawa, 2007

Se o assunto é **Brasil**,
procure o **IBGE**.

 /ibgecomunica  /ibgeoficial  /ibgeoficial  /ibgeoficial

www.ibge.gov.br 0800 721 8181

ISBN 978-85-240-4537-0