

ESTUDO TÉCNICO
N.º 06/2016

**Abordagens para imputação
de rendimento na PNAD:
método e resultados**

MDS

SAGI

MINISTÉRIO DO DESENVOLVIMENTO SOCIAL E COMBATE À FOME

SECRETARIA DE AVALIAÇÃO E GESTÃO DA INFORMAÇÃO

Estudo Técnico

Nº 06/2016

Abordagens para imputação de rendimento na PNAD: método e resultados.

Técnico responsável

Júlio César Gomes Fonseca

Caio Nogueira Gonçalves

Revisão

Paulo de Martino Jannuzzi

Marconi Sousa

Estudos Técnicos SAGI é uma publicação da Secretaria de Avaliação e Gestão da Informação (SAGI) criada para sistematizar notas técnicas, estudos exploratórios, produtos e manuais técnicos, relatórios de consultoria e reflexões analíticas produzidas na secretaria, que tratam de temas de interesse específico do Ministério de Desenvolvimento Social e Combate à Fome (MDS) para subsidiar, direta ou indiretamente, o ciclo de diagnóstico, formulação, monitoramento e avaliação das suas políticas, programas e ações.

O principal público a que se destinam os Estudos são os técnicos e gestores das políticas e programas do MDS na esfera federal, estadual e municipal. Nesta perspectiva, são textos técnico-científicos aplicados com escopo e dimensão adequados à sua apropriação ao Ciclo de Políticas, caracterizando-se pela objetividade, foco específico e tempestividade de sua produção.

Futuramente, podem vir a se transformar em artigos para publicação no Cadernos de Estudos, Revista Brasileira de Monitoramento e Avaliação (RBMA) ou outra revista técnica-científica, para alcançar públicos mais abrangentes.

Palavras-chave: *Imputação de rendimento; estimativas PNAD; série histórica.*

Unidade Responsável

Secretaria de Avaliação e Gestão da Informação

Esplanada dos Ministérios | Bloco A | Sala 307

CEP: 70.054-906 Brasília | DF

Fone: 61 2030-1501 | Fax: 2030-1529

www.mds.gov.br/sagi

Secretário de Avaliação e Gestão da Informação

Paulo de Martino Jannuzzi

Secretária Adjunta

Paula Montagner

APRESENTAÇÃO

O presente estudo técnico apresenta o resultado de dois métodos de imputação de rendimento (Hot Deck) criados com o intuito de fornecer uma alternativa para o cômputo de estimativas que envolvam o rendimento domiciliar *per capita* dos domicílios da Pesquisa Nacional por Amostra de Domicílios (PNAD). Os métodos foram desenvolvidos sobre diferentes abordagens e em diferentes softwares, embora tenham se utilizado de conceitos semelhantes para criação de seus critérios de imputação de registros. O primeiro fixa-se inicialmente, nos resultados de classificação de domicílios com perfil de extrema pobreza e pobreza, obtidos após os estudos com os modelos de Análise de Discriminante (Estudos Técnicos nº 15/2014 e nº 05/2016), mantendo intactas as taxas de extrema pobreza e pobreza obtidos com o uso da técnica. O segundo não faz quaisquer restrições, embora utilize o conjunto de variáveis usado nos modelos de Análise de Discriminante para estabelecer padrões de comparação e posterior estimação das taxas de extrema pobreza e pobreza.

1. Introdução

O processo de imputação estatística é uma das técnicas disponíveis para a resolução do problema da não-resposta que pode ocorrer por dificuldade de estabelecer contato com o informante, ou por recusa do entrevistado em oferecer as informações solicitadas. Segundo Cortez et al (2010), rendimento é reconhecidamente uma variável suscetível à não resposta, por ser considerada como informação confidencial ou sensível.

O procedimento padrão de excluir registros que apresentem uma variável não respondida do banco de dados implica em perda de informações, o que em alguns casos resulta em uma imputação pela média, gerando distorção pois se está supondo que a relação entre as variáveis dos respondentes e dos não-respondentes é a mesma. Entre as técnicas de imputação existentes que preservam a relação entre as variáveis do banco de dados podemos citar: modelagem estatística, Hot Deck, vizinho (ou registro) mais próximo, etc. (CORTEZ et al, 2010).

A abordagem Hot Deck foi escolhida neste estudo, por simplicidade e facilidade de aplicação, a qual consiste na substituição dos valores sem preenchimento, baseando-se na escolha de outros registros presentes à base, que preencham parâmetros pré-definidos de similitude. Como ambas as técnicas de imputação apresentadas neste estudo tratam de Hot Deck,

doravante neste texto tratamos a título de distinção de métodos, como “Hot Deck AD” o método de imputação que tem em sua premissa a restrição em sua seleção de registros dada pelo resultado do modelo de Análise Discriminante, e como “Hot Deck Full” o método de imputação sem restrições.

A Pesquisa Nacional por Amostra de Domicílios (PNAD) produzida pelo Instituto Brasileiro de Geografia e Estatística (IBGE) consiste em uma pesquisa por amostragem de domicílios de periodicidade anual, programada para ocorrer apenas no período entre os Censos Demográficos. Embora seu objetivo seja oferecer tratamento sistemático para características gerais da população, de caráter permanente, como educação, rendimento, habitação e trabalho, entre outros; a dinâmica de coleta de suas informações não contempla retorno periódico¹ aos domicílios entrevistados, de forma que eventuais questionamentos que surjam sobre informações declaradas de um determinado domicílio só podem ser avaliadas caso haja um eventual retorno do pesquisador ao domicílio, a título de revisão das informações, e no ano específico da pesquisa, descartando quaisquer variações provenientes de anos posteriores, caso o domicílio seja selecionado na amostra de anos seguintes.

1.1. Motivação

Uma das linhas de atuação do Departamento de Monitoramento da Secretaria de Avaliação e Gestão da Informação (DM/SAGI) é o monitoramento analítico de indicadores sociais, em especial aqueles afetos à pobreza e extrema pobreza, por ser o público prioritário dos programas desenvolvidos pelo MDS. Nesse âmbito, a SAGI/MDS tem desenvolvido estudos que abordam a problemática e implicações da sub declaração de rendimentos na PNAD.

Neste estudo são apresentados dois métodos de imputação de dados para tratar de temas de forte impacto na estimação das taxas de extrema pobreza e pobreza, com base nas informações captadas pela PNAD, a saber: i) o rendimento *per capita* dos domicílios que nas variáveis de rendimento não possuem informação, também conhecidos como “*Sem Declaração*” (SD); ii) o rendimento *per capita* dos domicílios que nas variáveis de rendimento estão preenchidas com zero, também conhecidos como “*Sem Rendimento*” (SR).

¹ Com a divulgação da PNAD Contínua os problemas desta natureza poderão ser melhor avaliados, visto que o mesmo domicílio aparecerá na amostra em no mínimo três trimestres consecutivos, permitindo assim avaliar variações nas informações prestadas e gerando uma melhoria no processo de coleta de dados.

Mencionados em diversos Estudos Técnicos da SAGI [ETEC 16/2012, 24/2012, 08/2014, 15/2014, 17/2014, 24/2014] e também em Textos para Discussão do Instituto de Pesquisa Econômica Aplicada (IPEA) [TD 788], o tema até então foi tratado pela SAGI sob diversas abordagens que envolviam o cômputo das taxas mencionadas, no âmbito da *classificação* de domicílios SD e SR, sugerindo no máximo, diferentes formas de classificar os domicílios sem abarcar o mérito da imputação de rendimentos como alternativa para o cômputo das taxas de extrema pobreza e pobreza. Assim, após muitos estudos o que se concluiu é que havia a necessidade de se ter uma alternativa que permitisse não apenas a classificação de domicílios com perfis tão *sui generis*, como também uma solução que fosse pautável na interlocução com outros organismos que conduzem estudos sobre as taxas mencionadas, em um aspecto mais abrangente e não apenas de abrangência nacional, permitindo assim comparabilidade na avaliação de medidas da vulnerabilidade social brasileira, com outros contextos que necessitam da especificação de um valor monetário para seu cômputo.

Assim, o trabalho discorre subdivido em partes que vão desde a concepção de conceitos utilizados na criação dos métodos de imputação propostos, até a avaliação estatística dos resultados dos mesmos, proporcionando um estudo de acurácia na aplicação de ambos. Para tanto, o texto se divide em três partes principais: Introdução, Metodologia, Resultados e Conclusões Finais.

2. Metodologia

Esta seção apresenta cada um dos métodos propostos, considerando inicialmente o método que utiliza o modelo de Análise de Discriminante para aprimoramento e classificação dos perfis de SR e SD apresentados pelo Estudo Técnico SAGI n. 05/2016. Ambos os métodos desconsideram membros do domicílio que não compõe uma relação familiar com o responsável pelo domicílio. Segundo Medeiros e Osório (2001, p.11), há alguns arranjos domiciliares em que se encontram mais de uma pessoa nas categorias “agregado”, “pensionista”, “empregado doméstico” e “parente do empregado doméstico”, nestes o sistema de classificação da microlinhagem familiar pode ignorar as relações de parentesco de pessoas que reconhecem parentesco entre si, pois não representam mais que 1% dos casos.

O primeiro método de imputação, foi desenvolvido no pacote estatístico R e tem como resultado, estimativas de rendimento domiciliar *per capita* que ao serem utilizadas para o

cômputo das taxas de extrema pobreza e pobreza, não alteram os resultados das taxas apresentadas anteriormente. O Estudo Técnico SAGI 15/2014 apresentou estimativas apenas para as taxas de extrema pobreza para o período de 2004 à 2013. Posteriormente, os estudos avançaram para além do período, perfazendo a série de 1992 à 2014 e estimando também as taxas de pobreza. Os resultados destes esforços foram apresentados no Estudo Técnico SAGI 05/2016. Dessa forma, o método empregado neste estudo, busca balizar o que já se conhecia sobre as taxas de pobreza e extrema pobreza, permitindo seu uso em demais aplicações que se utilizam do conceito de pobreza monetária, expressa em unidades de rendimento domiciliar *per capita*, como base de cálculo, utilizando alguma linha de pobreza já conhecida:

“Isso porque mesmo quando a linha de pobreza é definida empiricamente, por um método que prometa “objetividade”, muitas das decisões necessárias para obtê-la são de natureza política, e sob o manto de cientificidade desses métodos os pesquisadores podem tomar decisões que deveriam ser objeto de discussão.” (OSORIO, R. G.; SOARES, S. S. D.; SOUZA, P. H. G. F., 2011, Pg. 34)

Em resumo, o primeiro método imputa valores na variável V4742² da PNAD, somente para os domicílios previamente identificados como SD ou SR, mantendo sua classificação original dada pelo modelo de Análise Discriminante, ou seja, seus resultados congelam as taxas de extrema pobreza e pobreza iguais às apresentadas pela ETEC 05/2016.

O segundo método de imputação não tem um limitador para as taxas estimadas ou qualquer restrição inicial, servindo assim como uma validação para o primeiro método. Seu desenvolvimento foi realizado utilizando-se a linguagem de programação *Python*, conjuntamente com as técnicas de indexação de bases de dados utilizadas pelo Departamento de Monitoramento da SAGI. Embora este método tenha alguns pontos em comum com o primeiro método em relação a alguns critérios de seleção de candidatos para imputação e variáveis a serem consideradas na escolha dos critérios, não utiliza do mesmo processo de seleção.

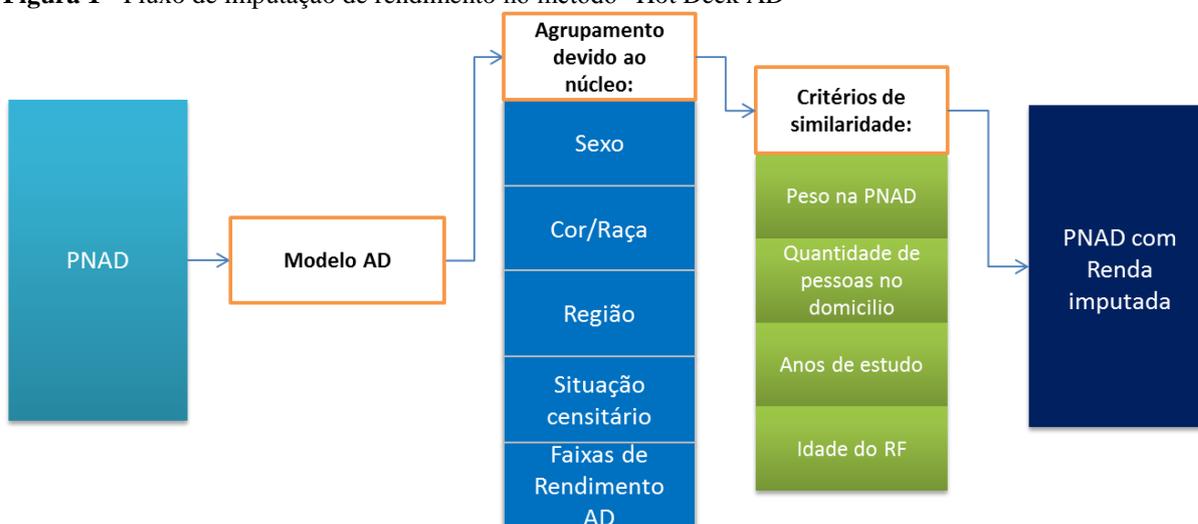
² Esta variável foi disponibilizada pelo IBGE a partir do ano de 2004, e armazena o rendimento domiciliar *per capita* (exclusive pensionistas, empregados domésticos, parentes dos empregados domésticos).

2.1. Hot Deck AD

O método descrito nesta seção, utilizou o pacote estatístico R para a aplicação do Hot Deck, baseando-se no modelo de Análise Discriminante e nos procedimentos utilizados na ETEC 05/2016 onde foi criada a variável EP_PO, com o intuito de classificar os indivíduos em três categorias: *i*) extremamente pobres (EP), *ii*) pobres (PO) e *iii*) não pobres (NP).

Para aplicação do método, a variável EP_PO foi adicionada a outras variáveis formando assim “núcleos”³ onde candidatos a fornecer a informação do rendimento domiciliar *per capita*, vulgo “doadores” e seus respectivos pares “receptores”, formaram grupos mantendo fixo para todos os seus elementos, o conjunto de informações relativas a sexo, cor/raça, região, setor censitário e EP_PO. Dentro de cada um destes grupos são aplicados critérios de similaridade (Figura 1) para definir o doador mais semelhante, considerando a ordem decrescente dos critérios de similaridade, com posterior escolha do registro doador obedecendo ao valor do rendimento *per capita* mais próxima, sem que ocorra repetição⁴ do registro selecionado como doador.

Figura 1 - Fluxo de imputação de rendimento no método “Hot Deck AD”



Fonte: Departamento de Monitoramento, SAGI.

É necessário mencionar que devido ao limitado tamanho de alguns grupos do “núcleo”, o setor censitário em sua definição foi dividido em apenas rural e urbano, a variável de cor/raça

³ Mais adiante no texto, é apresentada uma definição semelhante, a menos das faixas de rendimento. Vale ressaltar que na definição adiante o “setor censitário”, aqui citado, é denominado “urbano_urbano”.

⁴ Foi utilizado para este método um comando da biblioteca “VIM” disponibilizada no repositório do pacote estatístico R, detalhada em <https://cran.r-project.org/web/packages/VIM/VIM.pdf>.

agrupou brancos e amarelos na mesma categoria, assim como, pretos e pardos. Indígenas e não declarados tiveram raça/cor desconsiderados no núcleo de seus indivíduos.

2.2. Hot Deck Full

O método de imputação descrito nesta seção considera como candidatos para imputação os domicílios identificados como SR e SD. Uma particularidade deste método é o fato de utilizar seleção *sem reposição* para os SR e *com reposição* para os SD. A razão para se realizar seleção sem reposição nos SR advém do fato de que para este grupo de domicílios, a imputação considera a possibilidade de algum deles ser de fato um domicílio sem rendimento domiciliar *per capita*.

O fluxo de imputação utilizado na construção do método constituiu-se em três etapas:

1. Indexação⁵ das bases da PNAD para as quais se desejava realizar a imputação, segundo sua classificação inicial (SR, SD e VLR⁶);
2. Leitura dos arquivos indexados, estabelecendo para cada domicílio (SR e SD) as marcações de imputação, segundo os critérios apresentados adiante no texto, gerando uma log⁷ com todas informações sobre o registro que doou informação e o registro que recebeu a mesma;
3. Leitura das bases indexadas que identificam o domicílio, recompondo o arquivo PNAD e acrescentando a variável de RPDC imputada.

Diferentemente dos algoritmos que fazem uma única leitura no arquivo no qual ações serão executadas, o método desenvolvido primeiramente lê a base PNAD e armazena as informações necessárias para em uma segunda rodada realizar de fato a imputação. A compreensão do processo inicial de leitura das informações da PNAD e geração das bases

⁵ Para cada base da PNAD são gerados dois arquivos de índices, um com uma chave composta pelas variáveis que identificam o domicílio (UF + V0102 + V0103) e outro com uma chave composta pelas variáveis que identificam o que doravante no texto convencionou-se chamar “Núcleo”, definido pela concatenação das variáveis *região, sexo, urbano rural e raça-cor*. Estas bases indexadas possuem uma estrutura interna bastante flexível, permitindo o armazenamento agregado de algumas informações do domicílio.

⁶ Definiu-se como “VLR” todos os domicílios cujo rendimento familiar *per capita* (RDPC) não fossem nem SR, nem SD, ou seja, RDPC > 0 e RDPC < 999999999999, onde “999999999999” é a marca utilizada pelo IBGE, para domicílios que não informaram rendimentos em alguma das variáveis consideradas no cômputo do RDPC.

⁷ Arquivo contendo toda a movimentação feita durante o processo de imputação. Este arquivo contém um histórico completo dos critérios que foram utilizados na imputação do registro. Durante o processo de geração deste arquivo, os arquivos indexados também são atualizados com informações de endereçamento dos registros que doaram/receberam informações.

indexadas, passa pela definição de alguns conceitos que são utilizados ao longo do texto, para entendimento do fluxo de imputação. Assim, abaixo apresenta-se o conceito dos principais grupos (Figura 2) utilizados:

Núcleo: conjunto formado pelas variáveis identificadoras de *região, sexo, localidade do setor censitário (urbano/rural) e raça-cor*, utilizadas como base principal da imputação, ou seja, para cada domicílio pesquisado nas bases indexadas (SR e SD) estas informações dos indivíduos são utilizadas como chaves de pareamento⁸, fazendo com que o par de domicílios constituído pelo domicílio candidato a ceder a informação rendimento e o domicílio candidato a receber a informação de rendimento tenham minimamente estas informações semelhantes.

Padrão Socioeconômico: conjunto formado pelas variáveis utilizadas na ETEC 15/2014 para analisar o perfil dos domicílios declarados SR e SD. Embora naquele estudo técnico as variáveis tenham sido utilizadas para estimar um padrão de classificação que melhor se adequasse aos domicílios de ambos os grupos, sua utilização no presente estudo fundamenta-se no mesmo pilar de classificação, porém, em sentido mais amplo e em conjunto com outras variáveis que ajudam a caracterizar não apenas o perfil socioeconômico, mas também a proximidade territorial dos domicílios que fornecem informações.

⁸ Pareamento no contexto tratado neste estudo diz respeito ao processo de encontrar domicílios pertencentes às bases indexadas com perfil de SR e SD, cujo conteúdo do “núcleo” são idênticos, nas bases indexadas com perfil de VLR.

Figura 2 - Grupos utilizados no fluxo de imputação de rendimento



Fonte: Departamento de Monitoramento, SAGL.

A estrutura de dados⁹ adotada nas duas bases indexadas citadas anteriormente, contém elementos que ao longo do processo foram utilizados para agrupar domicílios cujo perfil socioeconômico e proximidade territorial fossem semelhantes, assim como os níveis de similaridade entre eles. Dessa forma, estas estruturas (Figura 3) foram internamente divididas em duas camadas a saber:

Camadas de Agregação: conjunto identificado¹⁰ de todas as combinações de “núcleo” e “padrão socioeconômico” existentes na base da PNAD, armazenados no arquivo indexado cuja chave é composta pelas variáveis que identificam o domicílio (UF + V0102 + V0103).

⁹ As bases indexadas utilizam as estruturas de dados de arquivos do tipo Berkley DB (<https://docs.python.org/2/library/bsddb.html>), também conhecidos como “dicionários” de dados (estrutura de dados composta por pares do tipo “chave” e “valor”, onde valor pode ser inclusive uma outra estrutura de dicionário).

¹⁰ Cada registro da PNAD é identificável no arquivo indexado que armazena as camadas de agregação, de forma que dado a identificação do domicílio (UF + V0102 + V0103), resgata-se a informação completa do entrevistado.

Camadas de Similaridade: conjunto de todas as combinações observadas nos domicílios declarados SR, SD e VLR, agrupadas pelo “núcleo”¹¹, segundo os seguintes elementos:

- Idade do responsável familiar;
- Nível de escolaridade do responsável familiar;
- Peso amostral do responsável familiar;
- Padrão socioeconômico (concatenação das respostas das variáveis utilizadas na ETEC 15/2014);
- Quantidade de moradores no domicílio.

Figura 3 - Camadas das estruturas de dados utilizadas no fluxo de imputação



Fonte: Departamento de Monitoramento, SAGI.

Compreendidos os conceitos adotados, a seguir são detalhadas as etapas do fluxo de imputação, considerando cada um dos conceitos apresentados acima.

A primeira etapa do fluxo de imputação lê a base de dados da PNAD, desconsiderando os registros referentes a pensionistas, empregados domésticos e seus filhos, e gera dois arquivos de índices, respectivamente “RDPC_PNAD_YYYY.db” e “NUCLEO_PNAD_YYYY.db”, onde “YYYY” representa o ano da PNAD considerada. O primeiro deles utiliza como chave de indexação a identificação do domicílio (UF + V0102 + V0103) e armazena além das principais

¹¹ Os dados desta camada são armazenados em arquivo indexado diferente da camada de agregação e são utilizados basicamente na segunda etapa do fluxo de imputação.

informações sobre o domicílio¹², os elementos da camada de agregação e o endereço¹³ de cada um dos registros do domicílio na base original. O segundo arquivo utiliza como chave de indexação as informações concatenadas do “núcleo” (região, sexo, urbano_rural¹⁴ e raça-cor¹⁵) e armazena, por tipo de perfil declarado do domicílio (SR, SD ou VLR), as informações da camada de similaridade definida anteriormente somente para os responsáveis familiares de cada domicílio. Assim, os arquivos possuem três informações em comum (“núcleo”, “padrão socioeconômico” e identificação do domicílio) que permitem o intercâmbio de informações armazenadas por cada um deles de maneira instantânea, uma vez que duas destas são chaves.

Criados os arquivos de indexação na primeira etapa do fluxo de imputação, a segunda etapa passa a ler cada um dos “núcleos” contidos no arquivo “NUCLEO_PNAD_YYYY.db”, tratando inicialmente aqueles cujo perfil domiciliar declarado sejam de SR¹⁶, buscando para eles pares semelhantes de mesmo núcleo, nos domicílios de perfil VLR e SR, aplicando-lhes filtros na camada de similaridade que permitam selecionar candidatos com o maior grau de similaridade segundo os seguintes critérios:

1. Presença na mesma UF;
2. Igualdade no número moradores do domicílio;
3. Igualdade no “padrão socioeconômico”;
4. Menor diferença etária entre os responsáveis familiares.

Caso sejam encontrados muitos candidatos, restringe a seleção calculando um novo nível de similaridade considerando os seguintes critérios:

1. Presença na mesma UF;

¹² As principais informações registradas são: número de pessoas no domicílio, valor da RDPC originalmente declarada para o domicílio, informações da camada de agregação, composta por “núcleo” e por “padrão socioeconômico”.

¹³ Utilizando-se de técnica já comumente utilizada no Departamento de Monitoramento na aplicação de pareamento de bases de dados, o “endereço” citado refere-se ao deslocamento de bytes necessário no arquivo em questão, para que se posicione o arquivo no registro que se deseja.

¹⁴ A definição de “urbano_rural” é formada pela variável V4728 (código da situação censitária), onde é considerado “urbano” os valores de 1 a 3 e “rural” os valores de 4 a 8.

¹⁵ Os valores utilizados desta variável neste trabalho agrupam as categorias “Branços” com “Amarelos” e “Pretos” com “Pardos”.

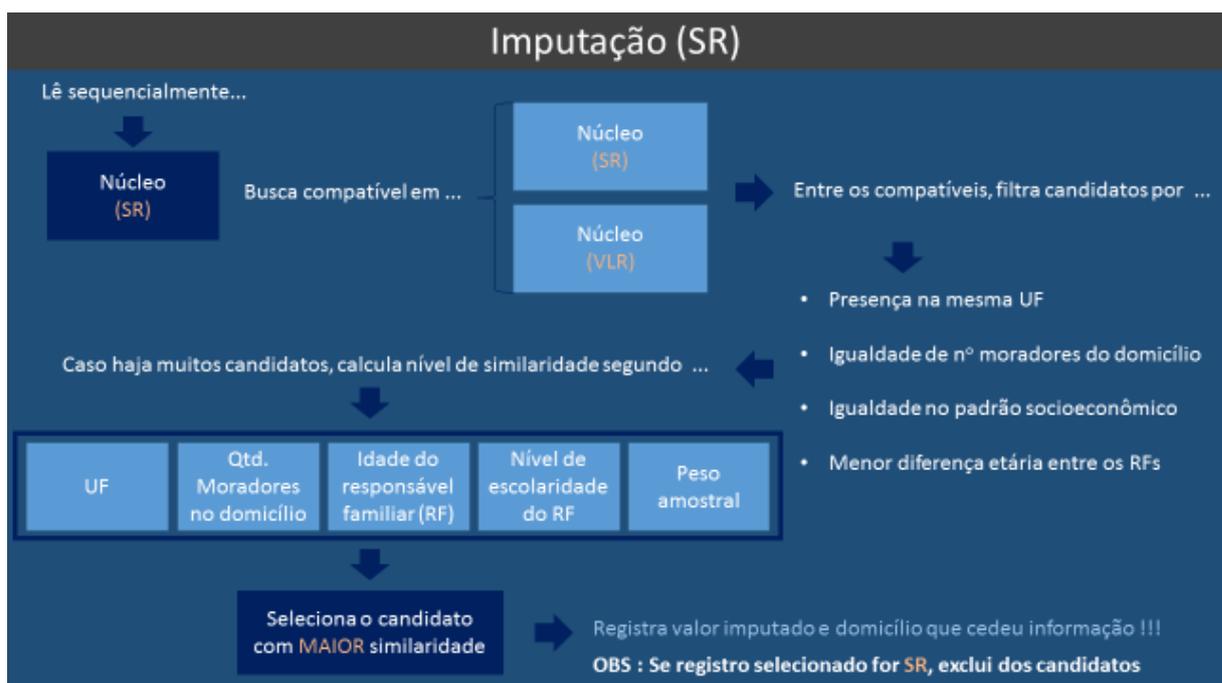
¹⁶ Por se tratar de procedimento que desconsidera a reposição de registros já selecionados no grupo dos domicílios de perfil declarado SR, o que conseqüentemente reduz o número de iterações, busca parear primeiramente os pertencentes a este perfil.

2. Maior proximidade no número de moradores do domicílio;
3. Menor diferença etária entre os responsáveis familiares
4. Semelhança no nível de escolaridade dos responsáveis familiares;
5. Menor diferença entre os pesos amostrais.

Se ainda houver mais de um candidato, ordena os níveis de similaridade calculados e seleciona aquele que possuir o maior grau de similaridade, registrando no arquivo “RDPC_PNAD_YYYY.db” o valor imputado e o identificador do domicílio escolhido para ceder a informação. A cada passo em que seleciona um candidato para tomar-lhe emprestado o valor do rendimento *per capita*, registra a movimentação em um arquivo de saída, utilizado para conferência das imputações realizadas, as informações da camada de similaridade utilizada para parear os registros.

Repete todo o procedimento, para o mesmo “núcleo”, tratando a seguir os domicílios cujo perfil domiciliar declarado sejam de SD, buscando para eles pares semelhantes de mesmo núcleo, nos domicílios apenas de perfil VLR, aplicando-lhes a mesma sequência de filtros anteriormente citada. A Figura 4 apresenta de forma simplificada a segunda etapa do fluxo de imputação, no caso particular dos domicílios com perfil de SR.

Figura 4 - Diagrama da segunda etapa do fluxo de imputação de rendimento, para os domicílios com perfil de SR



Fonte: Departamento de Monitoramento, SAGI.

A terceira e última etapa do fluxo de imputação gera um novo arquivo da PNAD, contendo a variável RDPC_IMP, que representa a nova variável V4742 (rendimento domiciliar *per capita*). Para isto, lê o arquivo original da PNAD e utilizando o identificador do domicílio (UF + V0102 + V0103) como chave de indexação, procura no arquivo “RDPC_PNAD_YYYY.db”, agora atualizado com as informações imputadas, a informação do rendimento *per capita* cedida, caso o domicílio em questão tenha perfil declarado de SR ou SD, caso contrário replica o valor original da V4742.

3. Resultados

Criados os métodos, passamos a validar seus resultados a partir do cálculo de vários indicadores relacionados com o rendimento domiciliar *per capita*. Antes mesmo de avaliar quaisquer indicadores, comparamos o rendimento médio imputado com o rendimento médio original em toda a série (Gráfico 1). Tomando alguns anos como exemplo (Tabela 1), nota-se que os rendimentos imputados são média 3% maiores em relação ao rendimento médio original nos anos de 1995 e 2004. Porém em 2014, a diferença média relativa observada cai para 2%, algo esperado já que o quantitativo de SR em 2014 reduz 63% em relação à 2013. Em suma, ainda que o método Hot Deck Full apresentado apresente resultados menores, o impacto desta imputação é muito pequeno na renda média.

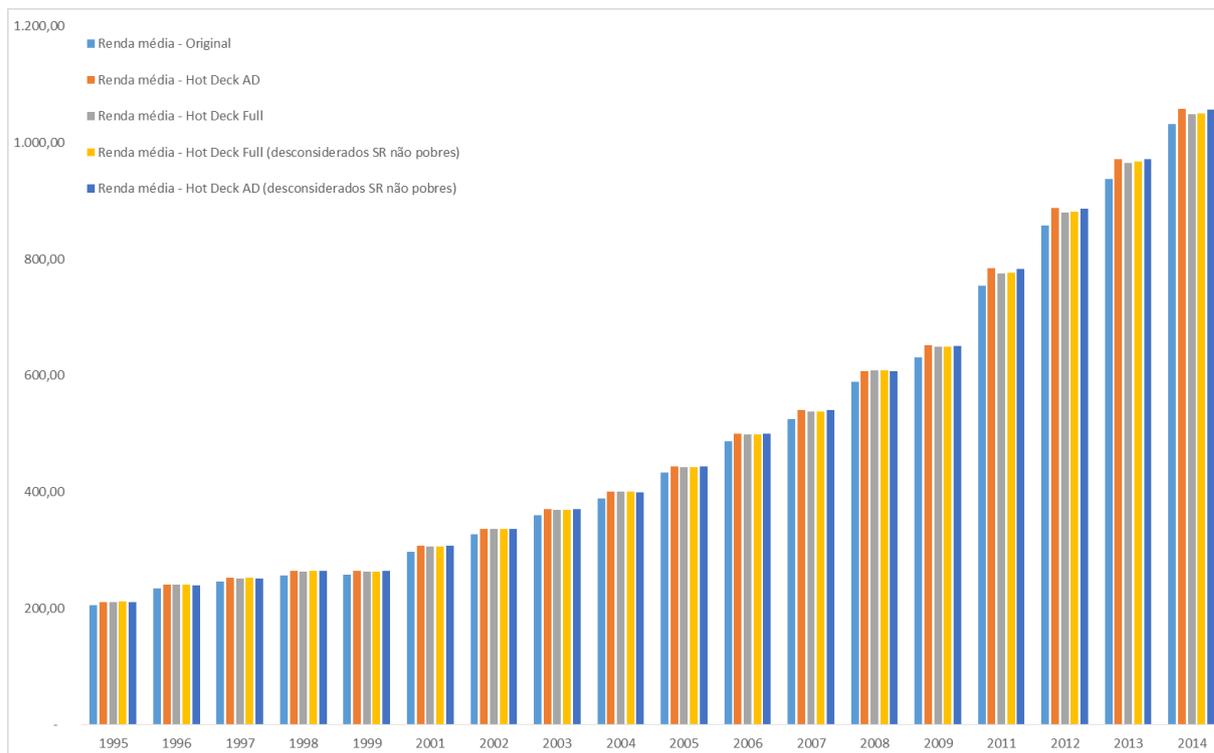
Tabela 1 – Diferença relativa do rendimento domiciliar *per capita* calculado pelos métodos de imputação em relação ao rendimento médio domiciliar *per capita* original - Brasil, 1995-2004-2014

	1995	Dif. %	2004	Dif. %	2014	Dif. %
Renda média - Original	205,0	-	388,8	-	1.031,9	-
Renda média - Hot Deck AD	210,6	2,7	400,2	2,9	1.058,2	2,6
Renda média - Hot Deck Full	211,1	2,9	400,5	3,0	1.049,6	1,7
Renda média - Hot Deck Full (desconsiderados SR não pobres)	211,3	3,1	400,5	3,0	1.050,2	1,8
Renda média - Hot Deck AD (desconsiderados SR não pobres)	210,6	2,7	399,8	2,8	1.057,4	2,5

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

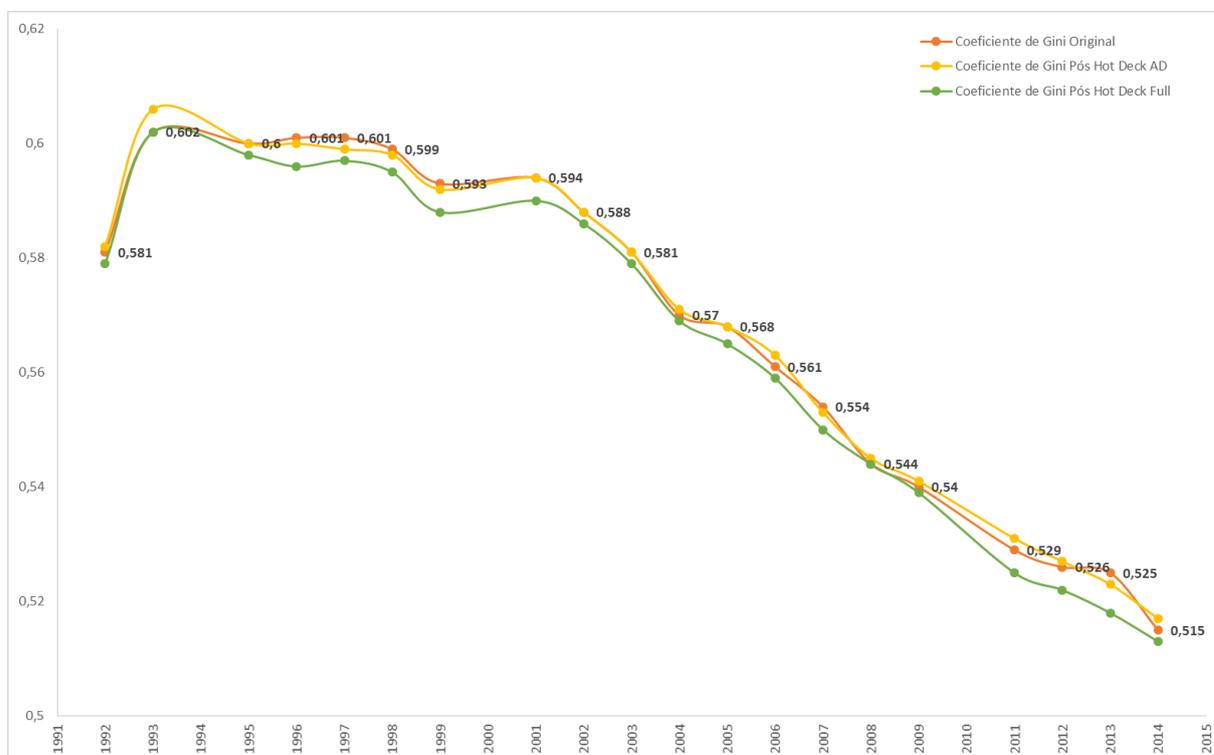
Avaliamos também o coeficiente de Gini (Gráfico 2), que de uma maneira geral consiste em um número entre 0 e 1, onde 0 corresponde à completa igualdade, no caso corrente seria dizer que todos os respondentes da PNAD possuem o mesmo rendimento domiciliar *per capita*, e 1 à completa desigualdade, o que significaria estar diante de um cenário onde o rendimento domiciliar *per capita* se concentraria em apenas um domicílio, tendo todos os demais, ausência do mesmo.

Gráfico 1 - Rendimento domiciliar *per capita* médio segundo dados originais da PNAD e dados pós-imputação (Hot Deck AD e Hot Deck Full) - Brasil, 1995-2014



Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Gráfico 2 - Coeficiente de Gini segundo dados originais da PNAD e dados pós-imputação (Hot Deck AD e Hot Deck Full) - Brasil, 1992-2014



Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Observa-se pelo Gráfico 2 que embora os resultados de ambos os métodos se aproximem muito do valor original do indicador¹⁷, o método Hot Deck Full produz os menores coeficientes ao longo de toda a série, ficando abaixo da série original em todos os anos. O método Hot Deck AD por sua vez, oscila em torno do valor original, apresentando um valor um pouco superior no ano de 1993. Embora os resultados observados no Gráfico 2 indiquem oscilações dos métodos de imputação apresentados no cálculo do coeficiente de Gini, o que se deve provavelmente em maior escala em razão da imputação de renda em domicílios com renda zero, para se ter uma avaliação mais precisa da significância estatística da desigualdade apontada por este indicador, seria importante conhecer a variância do mesmo (NEDER, 2004), possibilitando assim a análise dos respectivos intervalos de confiança. Neste estudo não tratamos deste tema, porém uma referência de como avaliar os cálculos se encontra em Azevedo (2006 apud Barros et al, vol. 1, p. 163).

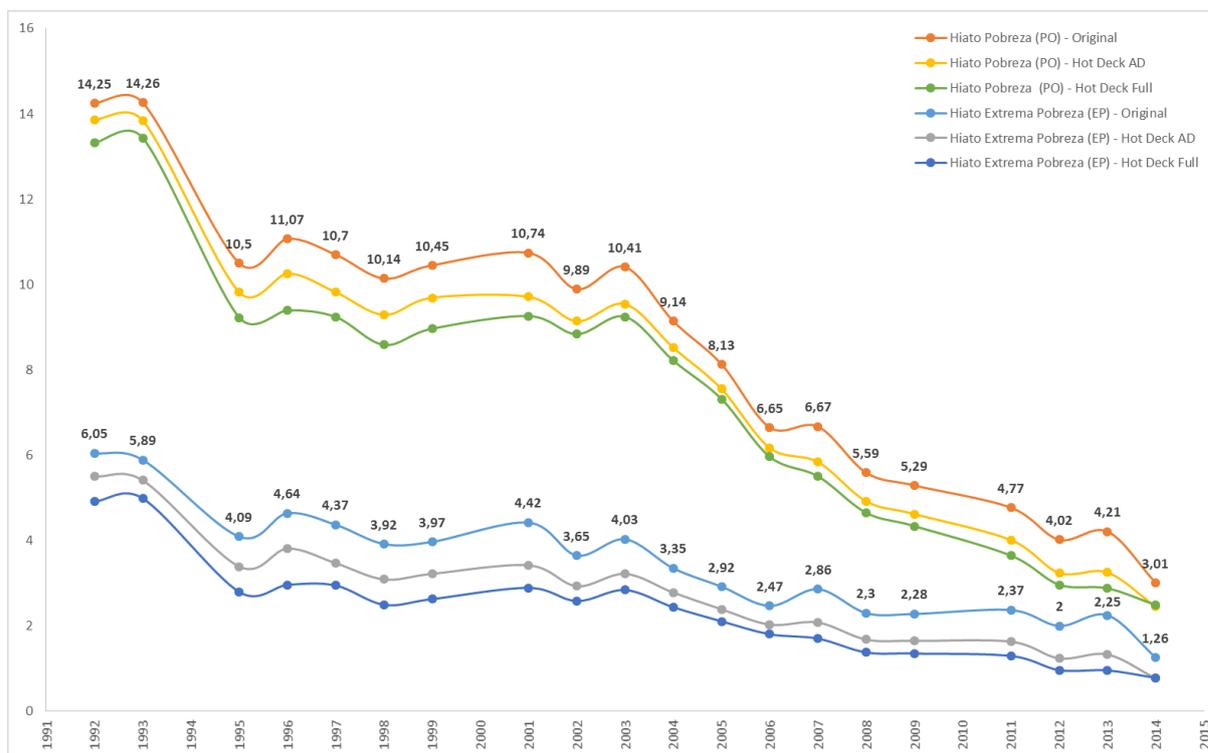
De maneira geral, o indicador com a imputação é consistente e apresenta comportamento semelhante, com leve impacto no coeficiente original, ficando evidente a necessidade de avaliação de outro indicador mais focado na população sensível. Assim, avaliamos também o hiato (Gráfico 3) e a severidade (Gráfico 4) da pobreza, onde é possível observar a tendência de queda da desigualdade ao longo dos anos e também o valor nominal dos indicadores originais.

O hiato é uma medida que representa a porcentagem da renda da população necessária para levar todos os pobres até a linha de pobreza. O hiato médio da linha de pobreza é calculado somando quanto falta para cada indivíduo atingir a linha de pobreza. Nesse aspecto há redução evidente dos valores devido ao aumento constante de indivíduos com perfil de SR e SD mencionado no Estudo Técnico nº 24/2012.

A severidade da pobreza mede a desigualdade da pobreza, de maneira que quanto mais indivíduos estiverem distantes da linha da pobreza, maior será o índice de severidade da pobreza. Evidenciando o ponto mencionado anteriormente o indicador de severidade é reduzido em até 27% no ajuste do indicador para a pobreza em 2014 e 55,3% para a extrema pobreza no mesmo período. Assim, o impacto do método fica mais evidente à medida em que produz uma redução significativa nos valores dos indicadores.

¹⁷ São apresentados no gráfico apenas os valores da série original.

Gráfico 3 - Hiato Médio da Extrema Pobreza (EP) e da Pobreza (PO), segundo dados originais da PNAD e dados pós-imputação (Hot Deck AD e Hot Deck Full) - Brasil, 1992-2014



Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Tabela 2 - Diferença relativa nos indicadores de desigualdade (Gini, Hiato, Severidade) calculados pelos métodos de imputação em relação ao cômputo original de cada indicador - Brasil, 1995-2004-2014

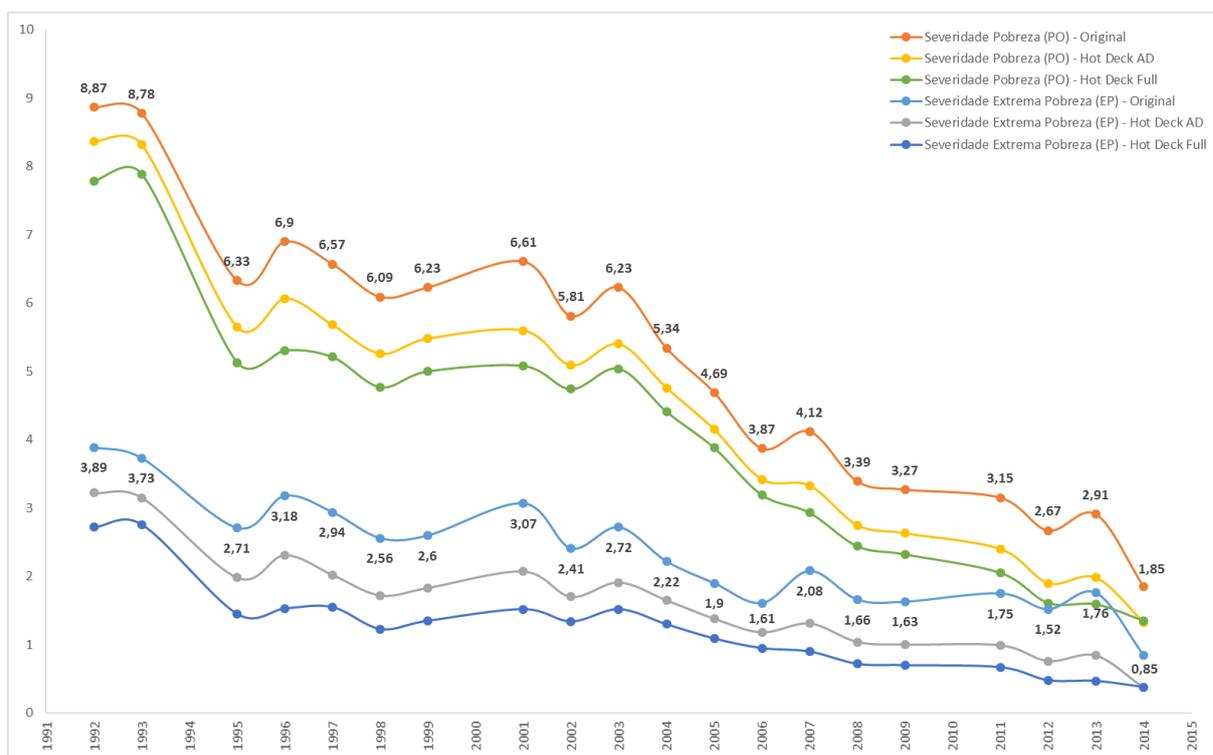
	1995	Dif. %	2004	Dif. %	2014	Dif. %
Coefficiente de Gini Original	0,600	-	0,570	-	0,515	-
Coeficiente de Gini Pós Hot Deck AD	0,600	0,00	0,571	0,18	0,517	0,39
Coeficiente de Gini Pós Hot Deck Full	0,598	-0,33	0,569	-0,18	0,513	-0,39
Hiato Extrema Pobreza (EP) - Original	4,090	-	3,350	-	1,260	-
Hiato Extrema Pobreza (EP) - Hot Deck AD	3,390	-17,11	2,780	-17,01	0,770	-38,89
Hiato Extrema Pobreza (EP) - Hot Deck Full	2,795	-31,65	2,437	-27,24	0,783	-37,88
Hiato Pobreza (PO) - Original	10,500	-	9,140	-	3,010	-
Hiato Pobreza (PO) - Hot Deck AD	9,820	-6,48	8,520	-6,78	2,450	-18,60
Hiato Pobreza (PO) - Hot Deck Full	9,226	-12,13	8,216	-10,11	2,490	-17,28
Severidade Extrema Pobreza (EP) - Original	2,710	-	2,220	-	0,850	-
Severidade Extrema Pobreza (EP) - Hot Deck AD	1,980	-26,94	1,650	-25,68	0,370	-56,47
Severidade Extrema Pobreza (EP) - Hot Deck Full	1,450	-46,49	1,300	-41,44	0,380	-55,29
Severidade Pobreza (PO) - Original	6,330	-	5,340	-	1,850	-
Severidade Pobreza (PO) - Hot Deck AD	5,650	-10,74	4,760	-10,86	1,330	-28,11
Severidade Pobreza (PO) - Hot Deck Full	5,130	-18,96	4,410	-17,42	1,350	-27,03

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

A sensibilidade de alguns dos indicadores (Tabela 2) fica evidente quando observa-se a variação em relação em cômputo original de cada um deles, observada após a aplicação dos diferentes métodos aplicados. Dessa forma observa-se que os diferentes métodos de imputação têm impacto maior no cômputo da severidade, indicador este mais sensível na

avaliação da desigualdade de rendimentos. A tabela também evidencia o perfil de queda, em relação ao valor original do cômputo dos diversos indicadores, com exceção do ano de 2014, fato este que corrobora a hipótese de ajuste na captação realizada naquele ano.

Gráfico 4 - Severidade da Extrema Pobreza (EP) e da Pobreza (PO), segundo dados originais da PNAD e dados pós-imputação (Hot Deck AD e Hot Deck Full) - Brasil, 1992-2014

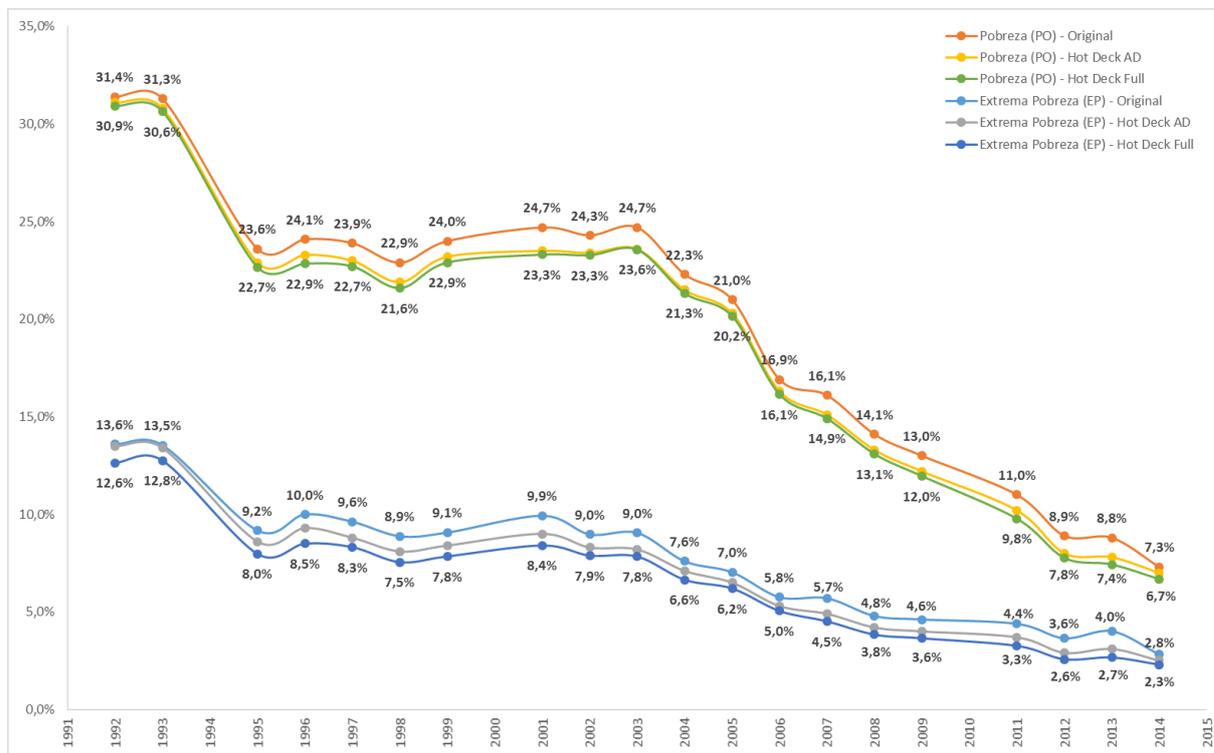


Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

De uma maneira geral o comportamento dos indicadores de hiato e severidade apresentados nos Gráficos 3 e 4 não aponta para divergências nos métodos utilizados, ainda que observados de maneira pontual. Conclusões semelhantes são obtidas ao analisar o Gráfico 5, onde observa-se que, de fato, o método Hot Deck Full conduz a uma ligeira diferença nas taxas de pobreza e extrema pobreza, com cifras menores do que o método Hot Deck AD. Assim, visualmente o que parece ser notório é o fato do método Hot Deck Full conduzir a taxas de pobreza mais próximas do método Hot Deck AD do que o faz com relação às taxas de extrema pobreza, apresentando uma variação maior no cômputo desta última (Tabela 3).

Esses resultados evidenciam que as taxas de pobreza e extrema pobreza calculadas diretamente pelos microdados da PNAD estão sobre-estimadas e que, o modelo de tratamento apresentado no ETEC 05/2016 - modelo AD para SR e SD - já corrigia boa parte dos problemas de captação de renda em campo.

Gráfico 5 - Percentual de Extrema Pobreza (EP) e da Pobreza (PO), segundo dados originais da PNAD e dados pós-imputação (Hot Deck AD e Hot Deck Full) - Brasil, 1992-2014



Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Tabela 3 - Diferença relativa nas taxas de extrema pobreza e pobreza calculadas pelos métodos de imputação em relação ao cômputo original - Brasil, 1995-2004-2014

	1995	Dif. %	2004	Dif. %	2014	Dif. %
Extrema Pobreza (EP) - Original	9,2	-	7,6	-	2,8	-
Extrema Pobreza (EP) - Hot Deck AD	8,6	-6,17	7,1	-6,48	2,5	-11,34
Extrema Pobreza (EP) - Hot Deck Full	8,0	-13,02	6,6	-12,46	2,3	-18,51
Pobreza (PO) - Original	23,6	-	22,3	-	7,3	-
Pobreza (PO) - Hot Deck AD	22,9	-2,97	21,5	-3,59	7,0	-4,11
Pobreza (PO) - Hot Deck Full	22,7	-1,03	21,3	-0,78	6,7	-4,54

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

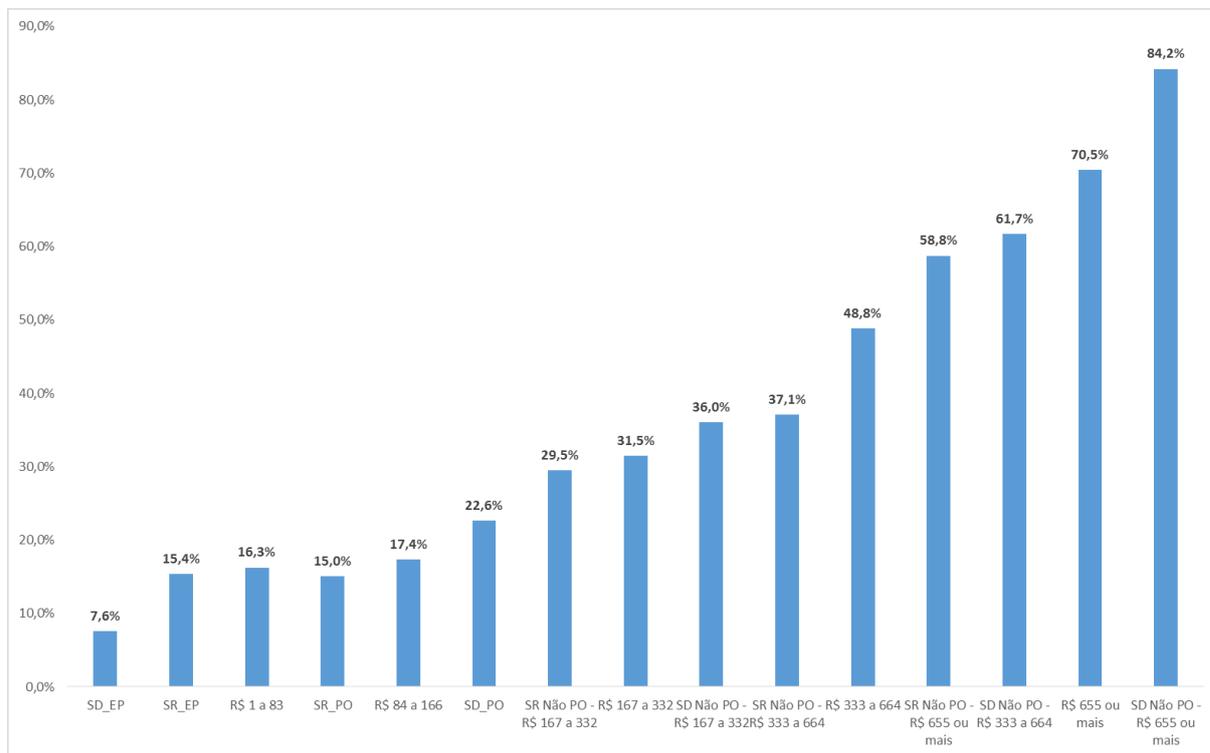
Os efeitos desta característica no cômputo da taxa de extrema pobreza podem ser constatados no Gráfico 6 quando se observa o perfil socioeconômico dos domicílios que possuem máquina de lavar, classificados como extremamente pobres após aplicação do método Hot Deck Full. Se comparados seus resultados com os apresentados no estudo técnico do modelo AD (Gráfico 7), onde o percentual dos domicílios originalmente sem rendimento classificados como extremamente pobres apontava para 3,0%, após a aplicação da imputação o método identificou mais domicílios com perfil de extrema pobreza, subindo para 15,4% no ano de 2014. Nos domicílios originalmente sem declaração de rendimentos, os dados do modelo AD apontavam para um percentual de 16,1% e pós aplicação do método caíram para 7,6%

(Gráfico 6). Os resultados para os domicílios originalmente sem rendimento classificados como pobres pelo modelo AD apontava para 2,6% em 2014, subindo para 22,6% após a aplicação do método Hot Deck Full. Estes resultados mostram que o método de imputação identifica mais perfis de domicílios sem rendimento do que perfis de domicílios sem declaração.

O impacto dos indivíduos sem rendimento domiciliar *per capita* fica ainda mais evidente no método Hot Deck Full quando observamos a evolução do percentual deste grupo na extrema pobreza (Gráfico 8). Com o modelo AD observou-se que em 1992 o percentual de participação dos SR na taxa de extrema pobreza era de 2,9%. Este mesmo percentual avaliado após o método Hot Deck Full cai para 1,7% no mesmo ano. Em 2013, onde se observava até então um ápice de 32% sob a ótica do cômputo original, cai para 11,6% com a aplicação do modelo AD e para 2,4% após a aplicação da imputação Hot Deck Full. Esta expressiva queda mostra o quanto o este último método tem impacto na classificação dos SR, podendo ser constatado pelo percentual de indivíduos que permanecem SR após a aplicação do método (Gráfico 9).

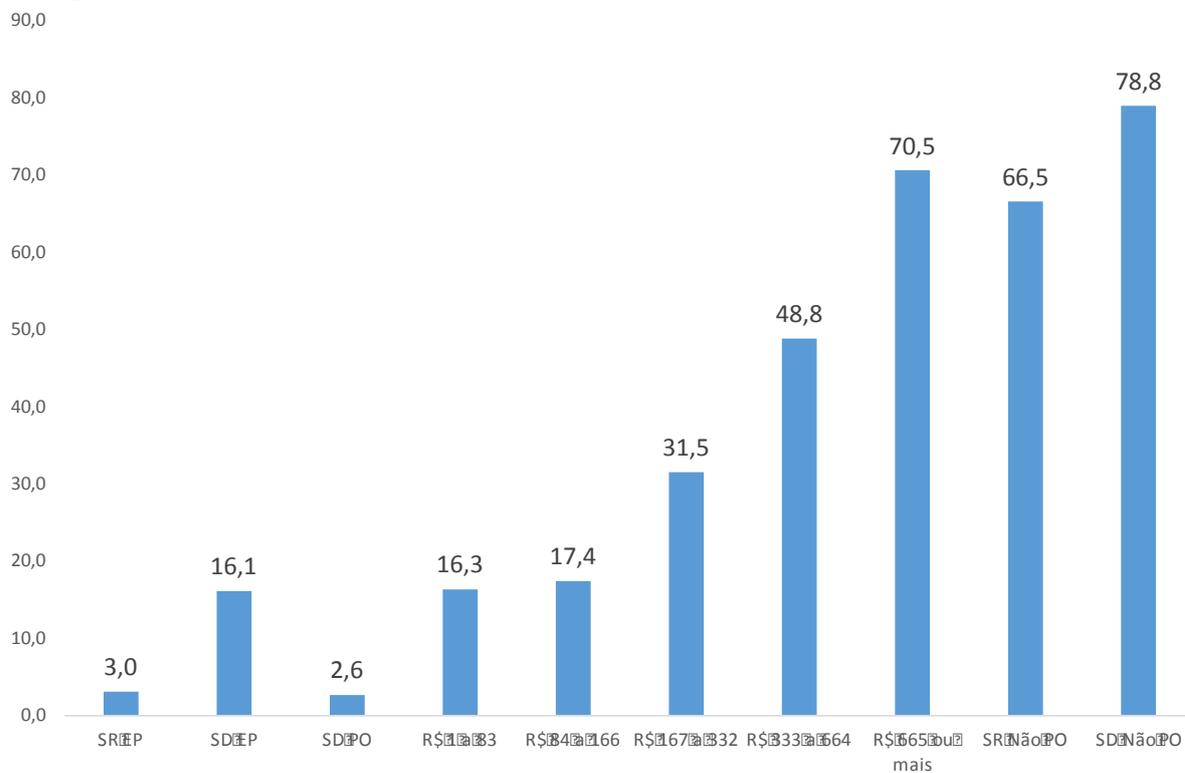
Com a finalidade de investigar a alta classificação dos domicílios sem rendimento domiciliar *per capita* observada nos resultados do método Hot Deck Full, levantou-se a hipótese de que o modelo estaria classificando um número maior de indivíduos com o perfil de SR caso os mesmos tivessem declarado recebimento de valores na variável V1273. Esta variável, segundo o dicionário da PNAD, capta além de transferências recebidas de programas sociais, juros de caderneta de poupança e de outras aplicações financeiras, dividendos, e outros rendimentos. Ou seja, o que se gostaria de avaliar é se o método Hot Deck Full estaria corrigindo parte da subdeclaração de recebimentos dos programas Bolsa Família (PBF) e Benefício de Prestação Continuada (BPC).

Gráfico 6 - Percentual dos domicílios que possuem máquina de lavar por faixas de rendimento domiciliar *per capita* pós imputação Hot Deck Full - Brasil, 2014



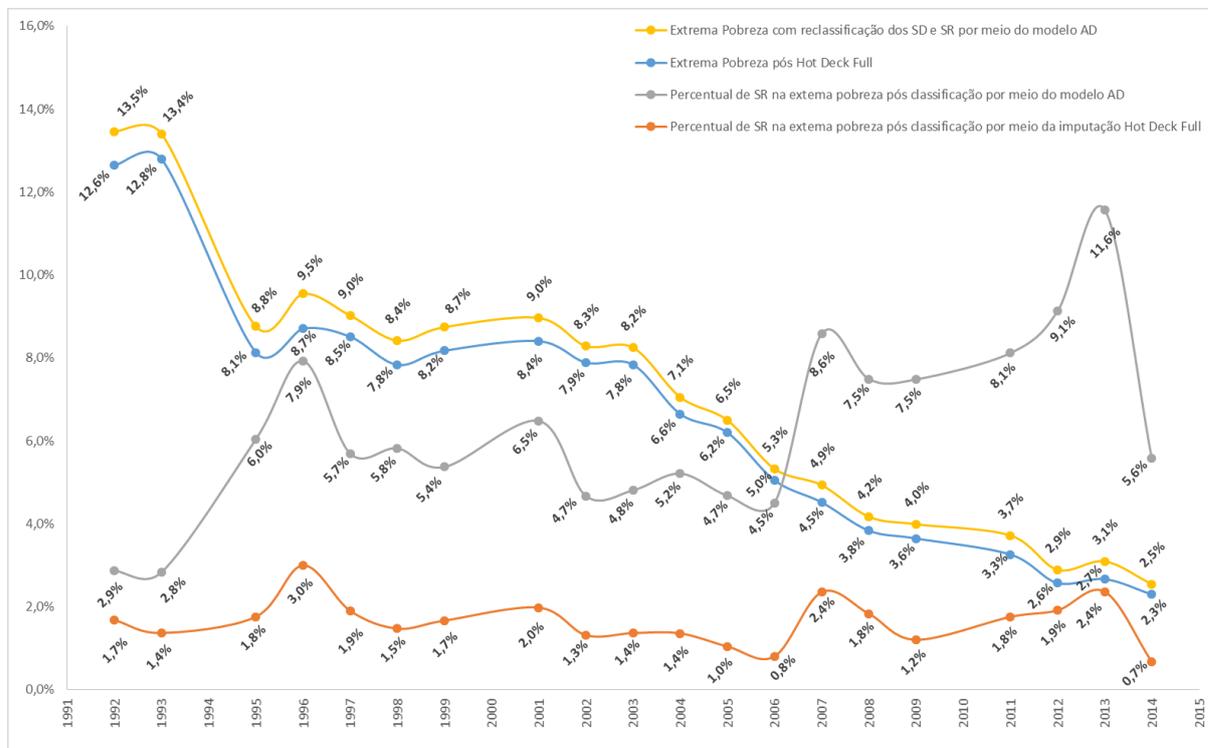
Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Gráfico 7 - Percentual dos domicílios que possuem máquina de lavar por faixas de rendimento domiciliar *per capita* pós AD – Brasil, 2014



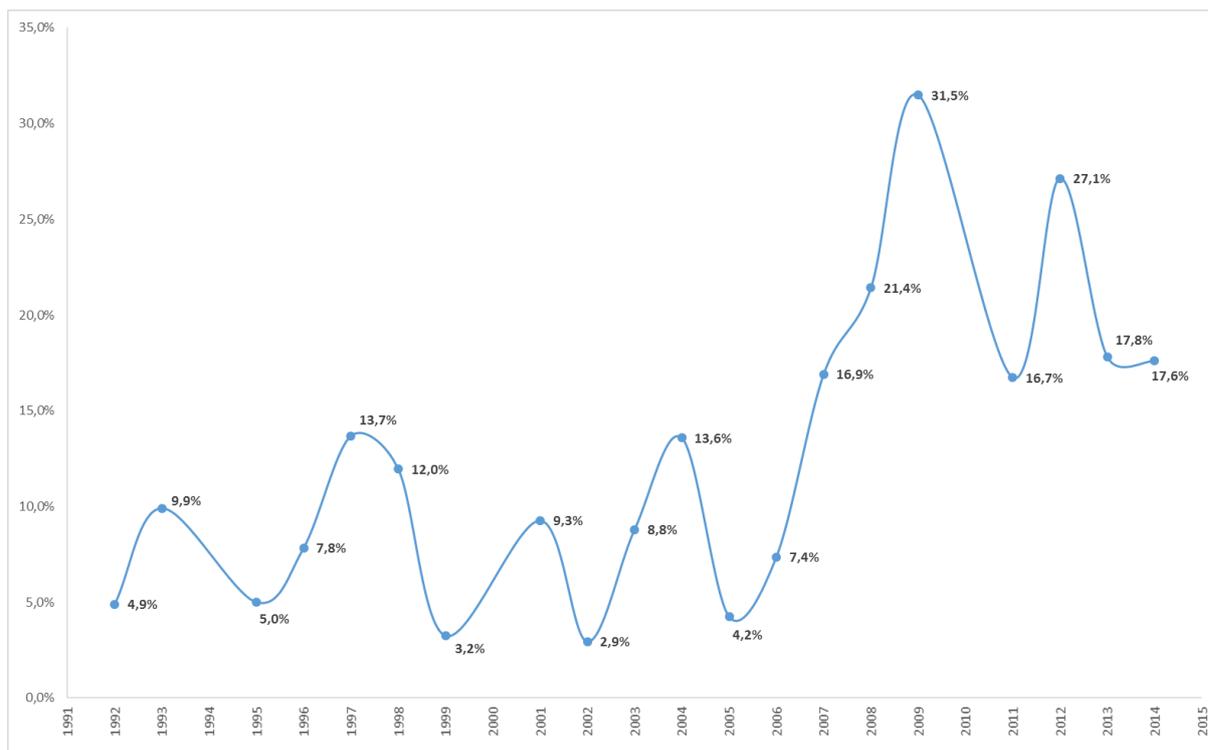
Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Gráfico 8 - Evolução do percentual dos Sem Rendimento na extrema pobreza pós imputação Hot Deck Full - Brasil, 1992-2014



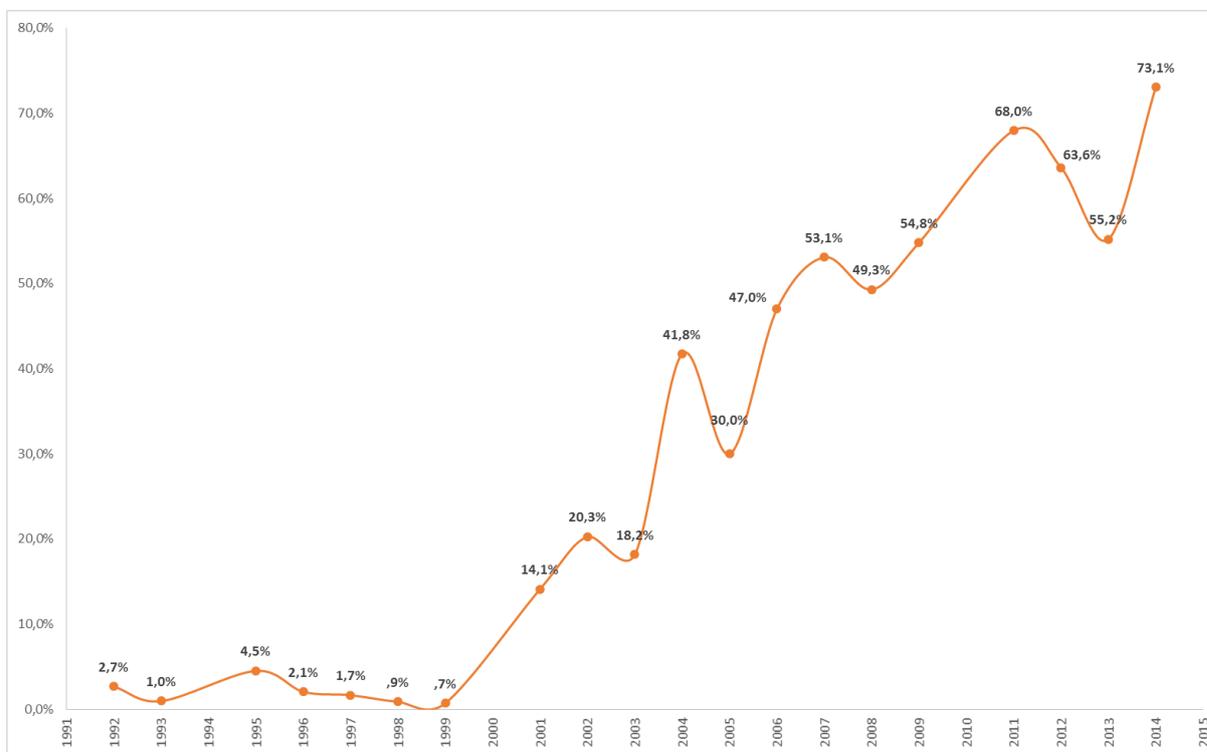
Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Gráfico 9 - Percentual dos Sem Rendimento que permaneceram sem rendimento domiciliar *per capita* pós imputação Hot Deck Full - Brasil, 1992-2014



Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Gráfico 10 - Percentual dos Sem Rendimento na taxa de pobreza que declararam receber rendimentos na V1273 pós imputação Hot Deck Full - Brasil, 1992-2014



Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Para avaliar a hipótese, observou-se entre os domicílios com perfil de SR após a aplicação do método Hot Deck Full, o percentual de indivíduos classificados como pobres que declararam receber algum rendimento na V1273. O Gráfico 10 mostra que a evolução percentual destes indivíduos é praticamente nula nos anos 1990 e começa a mudar drasticamente de comportamento nos anos 2000. Dentre os indivíduos com perfil de SR classificados como pobres, o percentual que declarou receber rendimentos na V1273 variou de 2,7% em 1992 a 73,1% em 2014¹⁸, seguindo a tendência de aumento do número potencial de beneficiários desses benefícios.

Há, pois, alguma evidência de que o método Hot Deck Full estaria conseguindo identificar, entre os SR, aqueles que não teriam declarado o valor dos benefícios do PBF ou BPC, mantendo-se ainda como pobres (pelo método). Ainda que a variável V1273 não fosse um critério de seleção para doares, o fato é que a aplicação do método acabou revelando que os doadores para esses indivíduos SR têm, em larga maioria, registro de rendimentos na variável V1273. Essa suposta correção da subdeclaração do recebimento de PBF e/ou BPC explicaria

¹⁸ Entre os extremamente pobres era de 2,2% em 1992, chegando a 76,2% em 2014.

por que as taxas de pobreza e extrema pobreza calculadas pelo método Hot Deck Full são um pouco menores que as obtidas pelo modelo AD.

Em outra linha de validação dos métodos, realizamos testes de comparação de médias entre os rendimentos domiciliares imputados por cada um dos métodos, em cada ano de edição da pesquisa, para cada uma das cinco regiões brasileiras, a fim de verificarmos qual o grau de significância estatística da diferença da distribuição empírica dos rendimentos imputados. Seriam estatisticamente diferentes a distribuição de renda domiciliar média *per capita*, computados pelos dois métodos?

Os resultados destes procedimentos são apresentados no anexo e mostram que a distribuição do rendimento domiciliar *per capita* após a aplicação dos mesmos é significativamente diferente em todas as regiões brasileiras com exceção da Região Norte. Nessa região, em 10 das 20 edições anuais da PNAD não se observa diferenças estatisticamente significativas. Naturalmente, esses resultados precisariam ser reavaliados considerando teste de significância que contemplasse o efeito do desenho amostral da Pnad. Como os intervalos de confiança da renda domiciliar aumentariam (pelo efeito do desenho complexo da amostra), as diferenças apontadas podem não se manter¹⁹. Mesmo que as diferenças entre as duas distribuições não se mantivessem, a aplicação do método Hot Deck Full se justificaria por todo seu impacto na regularização da trajetória dos indicadores de desigualdade e das taxas de pobreza, além da correção da subdeclaração de rendimentos e benefícios do PBF e BPC nas operações de coleta de campo da pesquisa.

4. Considerações Finais

Os dois métodos de imputação de rendimentos revelam a necessidade de se tratar registros sem declaração e sem rendimento na PNAD. A subdeclaração de rendimentos e benefícios e outras dificuldades na captação de dados sobre rendimentos na operação de campo da pesquisa afetam as estimativas de indicadores cruciais para a avaliação dos efeitos das políticas sociais em curso.

A abordagem aqui empregada focou-se na imputação de renda domiciliar *per capita*, não em todas as variáveis de rendimento captados na pesquisa (cerca de quinze tipos de rendimentos). Outras iniciativas metodológicas deveriam ser empreendidas, como tem sido feito pelo Departamento de Monitoramento da SAGI. Vale lembrar que, desde 2011, se tem produzido

¹⁹ Os resultados no anexo deste texto levam em consideram as diferenças entre os rendimentos domiciliares *per capita* da amostra, não ponderados portanto, para minimizar o efeito do plano amostral.

vários estudos técnicos que discutem essa questão, abordando-a seja no Censo Demográfico, seja na PNAD, oferecendo soluções técnicas cada vez mais consistentes. Inserem-se nesse esforço, inclusive, os estudos realizados para reponderação das PNADs dos anos 1990 e do cálculo dos intervalos de confiança para as taxas de pobreza e extrema pobreza considerando o desenho complexo da amostra da PNAD, como documentado, respectivamente, nos ETECs 04/2016 e 24/2014.

A proposta de imputação pelo método Hot Deck Full é, certamente, um avanço metodológico em relação à proposta anterior, assim como o modelo AD já foi uma inovação importante para resolver as inconsistências de captação de renda na PNAD, em particular após o ano de 2005 (com aumento do SR e SD). Os estudos técnicos realizados mostram que é um equívoco computar taxas de pobreza e extrema pobreza sem qualquer tratamento das variáveis de rendimento captadas na pesquisa. Os resultados da PNAD 2014 com relação à forte diminuição de registros com renda zero já deixam claro que havia problemas na captação de rendimentos nas edições anteriores da pesquisa e portanto, da necessidade de compatibilização de indicadores para análise longitudinal da pobreza, desigualdade e nível médio de rendimentos. Solução mais definitiva para as questões aqui tratadas deverá ser contemplada pela proposta de imputação de variáveis em nível individual adotada na apuração de resultados da PNAD Contínua.

Ademais, vale registrar que outras questões podem ser abordadas a partir dos resultados aqui apresentados, como o cálculo de intervalos de confiança para o coeficiente de Gini, do Hiato e Severidade. Os desafios das políticas sociais implicam na necessidade de aprimoramento das pesquisas e indicadores delas derivados. É nessa perspectiva que a SAGI se orienta.

Referências bibliográficas

AZEVEDO J.P. **Avaliando a Significância Estatística da Queda na Desigualdade no Brasil.** Desigualdade de Renda no Brasil: uma análise da queda recente. IPEA 2006, vol. 1, pp. 163-173.

BARROS, R.P.; CARVALHO, M.; FRANCO S. **Pobreza Multidimensional no Brasil.** Texto para Discussão IPEA n. 1227/2006.

CORTEZ, B.F.; MOREIRA, G.G.; FERNANDES, M.V.M. **Descrição e avaliação do processo de imputação nos quesitos de fecundidade da PNAD 2008.** In: XVII Encontro Nacional de Estudos Populacionais, 2010, Caxambu. XVII Encontro Nacional de Estudos Populacionais.

FONSECA, J.C.G.; BARBOSA, M.V.S. **Análise Discriminante no tratamento dos grupos de domicílios Sem Rendimento (SR) e Sem Declaração (SD).** Estudo Técnico SAGI n. 15/2014.

FONSECA, J.C.G.; GONÇALVES, C.N.; FONSECA, M.L.S. **Metodologia de cômputo das taxas de pobreza e extrema pobreza nas PNADs 1992 a 2014.** Estudo Técnico SAGI n. 05/2016.

JANNUZZI, P.M.; SOUZA, M.; VAZ, A.C.N.; FONSECA, J.C.G.; BARBOSA, M.V.S. **Dimensionamento da Extrema Pobreza no Brasil: aprimoramentos metodológicos e novas estimativas para 2001 a 2013.** Estudo Técnico SAGI n. 17/2014.

MARTIGNONI, E. **Estimativas da População em Extrema Pobreza da PNAD ajustadas aos Censos 2000 e 2010: método e resultados.** Estudo Técnico SAGI n. 23/2012.

MEDEIROS, M.; OSORIO, R.G. **Arranjos Domiciliares e Arranjos Nucleares no Brasil: Classificação e Evolução de 1977 a 1998.** Texto para Discussão IPEA n. 0788/2001.

NEDER, H.D. **Desenvolvimento de Metodologias Estatísticas Aplicadas aos dados das PNADs.** In: José Graziano da Silva. (Org.). III Seminário Novo Rural Brasileiro. Campinas: Embrapa, 2004.

OSORIO, R.G.; SOARES, S.S.D.; SOUZA, P.H.G.F. **Erradicar a Pobreza Extrema: Um Objetivo ao Alcance do Brasil.** Texto para Discussão IPEA n. 1619/2011.

RODRIGUES, N.M.C.; SOUZA, R.C. **Concatenação Estatística de Dados e Afiliação Estocástica.** Rio de Janeiro, 2009. 159p. Tese de Doutorado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

SOARES, S.S.D. **Metodologias para Estabelecer a Linha de Pobreza: Objetivas, Subjetivas, Relativas, Multidimensionais.** Texto para Discussão IPEA n. 1381/2009.

SOUZA, M.; VAZ, SANTOS, J.R.S. **Análise do filtro de restrições aplicado aos sem remuneração nos dados preliminares do Universo do Censo Demográfico 2010.** Estudo Técnico SAGI n. 16/2012.

SOUZA, M.; JANNUZZI, P.M. **Questões metodológicas acerca do dimensionamento da extrema pobreza no Brasil nos anos 2000.** Estudo Técnico SAGI n. 08/2014.

VAZ, A.C.N. Metodologias de estimação de população em extrema pobreza: um estudo dos "Sem Declaração" e dos "Sem Rendimento" na PNAD. Estudo Técnico SAGI n. 24/2012.

ANEXO

Tabela 4 - Testes de diferenças de médias entre os métodos HOT DECK Full e HOT DECK AD – Brasil - Região Norte, 1992-2014

Região NORTE				
Ano PNAD	Diferença entre os rendimentos domiciliares <i>per capita</i> médios	Intervalo de Confiança (95%)		p-valor
		Limite Inferior	Limite Superior	
1992	1.201,8533	-49,8355	2.453,5422	0,0598
1993	98,3861	62,6652	134,1071	0,0000
1995	-0,1467	-0,7173	0,4239	0,6144
1996	1,7005	1,2719	2,1291	0,0000
1997	1,0901	0,4814	1,6988	0,0004
1998	0,0356	-0,8317	0,9030	0,9359
1999	0,7512	0,1293	1,3732	0,0179
2001	-0,8773	-1,7010	-0,0536	0,0368
2002	1,7469	1,2077	2,2860	0,0000
2003	0,0819	-0,8473	1,0110	0,8629
2004	1,6252	0,7601	2,4903	0,0002
2005	-0,2670	-0,9002	0,3662	0,4086
2006	1,4557	0,5808	2,3306	0,0011
2007	1,2061	-0,1424	2,5545	0,0796
2008	1,3340	-0,1276	2,7956	0,0736
2009	-0,7739	-2,4566	0,9087	0,3673
2011	5,1679	2,7226	7,6133	0,0000
2012	1,7545	-0,1892	3,6981	0,0769
2013	4,2037	1,9453	6,4621	0,0003
2014	-0,8093	-3,6408	2,0222	0,5753

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Tabela 5 - Testes de diferenças de médias entre os métodos HOT DECK Full e HOT DECK AD – Brasil - Região Nordeste, 1992-2014

Região NORDESTE				
Ano PNAD	Diferença entre os rendimentos domiciliares <i>per capita</i> médios	Intervalo de Confiança (95%)		p-valor
		Limite Inferior	Limite Superior	
1992	639,4044	-739,9128	2.018,7217	0,3636
1993	-11,6064	-40,8955	17,6826	0,4373
1995	-0,8477	-1,2232	-0,4723	0,0000
1996	-0,4982	-1,1726	0,1761	0,1476
1997	-0,2329	-0,6979	0,2322	0,3263
1998	-0,4926	-0,9334	-0,0519	0,0285
1999	-0,4405	-0,8610	-0,0199	0,0401
2001	-1,2263	-1,7380	-0,7145	0,0000
2002	-1,3149	-1,8641	-0,7657	0,0000
2003	-0,9814	-1,5345	-0,4283	0,0005
2004	0,8969	0,3119	1,4819	0,0027
2005	-1,1912	-2,0181	-0,3644	0,0047
2006	-0,9582	-1,5140	-0,4023	0,0007
2007	-0,2264	-0,9610	0,5081	0,5457
2008	-2,0079	-2,8311	-1,1846	0,0000
2009	-1,6070	-2,3948	-0,8191	0,0001
2011	-3,5774	-4,9972	-2,1576	0,0000
2012	-5,8140	-10,7323	-0,8957	0,0205
2013	-1,7433	-4,1622	0,6757	0,1578
2014	-0,6924	-2,6660	1,2811	0,4917

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Tabela 6 - Testes de diferenças de médias entre os métodos HOT DECK Full e HOT DECK AD – Brasil - Região Sudeste, 1992-2014

Região SUDESTE				
Ano PNAD	Diferença entre os rendimentos domiciliares <i>per capita</i> médios	Intervalo de Confiança (95%)		p-valor
		Limite Inferior	Limite Superior	
1992	2.666,2253	1.213,6550	4.118,7956	0,0003
1993	-126,3617	-195,6985	-57,0248	0,0004
1995	1,5288	0,8719	2,1858	0,0000
1996	-0,1272	-0,6292	0,3749	0,6196
1997	-0,6314	-1,3370	0,0741	0,0794
1998	-1,4168	-2,2348	-0,5988	0,0007
1999	-2,7730	-3,4296	-2,1165	0,0000
2001	-3,0875	-4,0094	-2,1656	0,0000
2002	0,1210	-0,8116	1,0536	0,7993
2003	-0,6436	-1,8752	0,5880	0,3057
2004	0,4688	-0,9781	1,9157	0,5254
2005	-2,2843	-3,3989	-1,1697	0,0001
2006	-5,2941	-6,6843	-3,9040	0,0000
2007	-7,9241	-9,7463	-6,1019	0,0000
2008	0,1952	-2,0445	2,4350	0,8643
2009	-8,3784	-10,5145	-6,2423	0,0000
2011	-25,1271	-28,1256	-22,1285	0,0000
2012	-18,7509	-23,3139	-14,1880	0,0000
2013	-15,1831	-17,7894	-12,5767	0,0000
2014	-19,9671	-22,8697	-17,0644	0,0000

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Tabela 7 - Testes de diferenças de médias entre os métodos HOT DECK Full e HOT DECK AD – Brasil - Região Sul, 1992-2014

Região SUL				
Ano PNAD	Diferença entre os rendimentos domiciliares <i>per capita</i> médios	Intervalo de Confiança (95%)		p-valor
		Limite Inferior	Limite Superior	
1992	-4.559,9824	-7.026,3680	-2.093,5968	0,0003
1993	-138,1881	-207,3987	-68,9775	0,0001
1995	-1,4216	-2,2229	-0,6204	0,0005
1996	0,2881	-0,3625	0,9386	0,3855
1997	-0,5514	-1,4313	0,3284	0,2193
1998	-3,0677	-3,9328	-2,2026	0,0000
1999	-2,0166	-2,6718	-1,3614	0,0000
2001	-2,2098	-3,3769	-1,0428	0,0002
2002	-1,8841	-2,8684	-0,8998	0,0002
2003	-2,0997	-2,9159	-1,2835	0,0000
2004	1,3411	0,2315	2,4507	0,0178
2005	-3,3341	-4,3291	-2,3391	0,0000
2006	-4,0805	-5,9443	-2,2168	0,0000
2007	-3,9187	-5,3376	-2,4997	0,0000
2008	-7,0342	-8,9458	-5,1227	0,0000
2009	-3,8546	-5,7576	-1,9515	0,0001
2011	-7,6250	-10,1440	-5,1061	0,0000
2012	-11,4462	-14,2836	-8,6089	0,0000
2013	-15,9985	-19,3870	-12,6101	0,0000
2014	-6,8504	-9,9181	-3,7828	0,0000

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Tabela 8 - Testes de diferenças de médias entre os métodos HOT DECK Full e HOT DECK AD
– Brasil - Região Centro-Oeste, 1992-2014

Região CENTRO-OESTE				
Ano PNAD	Diferença entre os rendimentos domiciliares	Intervalo de Confiança (95%)		p-valor
		Limite Inferior	Limite Superior	
1992	-3.564,6968	-6.468,3770	-661,0167	0,0161
1993	-31,9501	-118,8011	54,9009	0,4709
1995	-0,9911	-1,9382	-0,0440	0,0403
1996	1,3118	0,3491	2,2744	0,0076
1997	0,4077	-0,5847	1,4000	0,4207
1998	2,6488	0,2091	5,0886	0,0333
1999	1,0864	-0,0877	2,2606	0,0697
2001	6,0313	4,3708	7,6919	0,0000
2002	1,8953	0,4805	3,3101	0,0087
2003	2,7738	1,1196	4,4280	0,0010
2004	3,9694	1,6114	6,3275	0,0010
2005	0,9566	-0,6477	2,5609	0,2425
2006	12,4527	8,8245	16,0808	0,0000
2007	15,1435	11,4582	18,8287	0,0000
2008	14,9800	10,7280	19,2321	0,0000
2009	19,5017	12,4530	26,5505	0,0000
2011	8,2625	4,0666	12,4585	0,0001
2012	41,7466	33,8991	49,5942	0,0000
2013	27,5805	18,6422	36,5188	0,0000
2014	11,8610	6,7143	17,0077	0,0000

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.