Instituto Brasileiro de Geografia e Estatística

Diretoria de Pesquisas
Coordenação de Métodos e Qualidade
Gerência Técnica do Censo Demográfico

ESTUDOS E TRATAMENTO DA VARIÁVEL RENDIMENTO NO CENSO DEMOGRÁFICO 2010

Março de 2012

Introdução

No processo de Crítica e Imputação do Censo Demográfico 2010 as variáveis de rendimento passaram por um processo inicial de crítica utilizando o Sistema CANCEIS, que detectava as inconsistências e as tratava através de imputação obtida por valores de doadores. No caso dos questionários da Amostra, estas variáveis foram comparadas com aquelas existentes no tema Trabalho. Após esse tratamento surgiu a necessidade de uma segunda etapa de tratamento, para algumas pessoas cujo valor do rendimento se mostrou fora dos padrões esperados e que foram transformados em ignorado e imputados também pelo CANCEIS.

Para essa segunda etapa foram analisados, em paralelo, tanto os dados de rendimento do Universo como os da Amostra, sendo que os resultados de rendimento antes divulgados eram preliminares, por não terem sido submetidos a todos os processos de crítica e imputação.

1 – Pessoas com rendimento total de R$1,00

No conjunto Universo o grupo de pessoas com rendimento total inferior a R$20,00 era composto por 171.613 pessoas, o que representava 0,2% da população que declarou rendimento. Foi perceptível o comportamento discrepante em torno do valor R$1,00, quando comparado com valores mais próximos e com outros valores atrativos. Existiam 30.188 pessoas com rendimento R$1,00, o que representa 17,6% das pessoas com rendimento abaixo de R$ 20,00.

Observou-se uma concentração em algumas Unidades da Federação, como São Paulo. Nesta Unidade da Federação existiam 15.684 pessoas com rendimento R$1,00, o que representava 52,0% dos casos. Tais concentrações levantaram a suspeita da ocorrência de algum erro sistemático.

1.1 – Metodologia para detecção

Verificou-se que frequentemente em um mesmo domicílio mais de uma pessoa tinha rendimento de R$ 1,00. Acredita-se que o valor R$1,00 possa ter sido registrado pelo Recenseador para casos onde não foi possível obter o valor, devido à dificuldade de captação desta variável e à característica do quesito, que restringia a possibilidade de deixá-lo em branco.

Optou-se, então, por ignorar os rendimentos das pessoas residentes em domicílios particulares permanentes cujo domicílio tinha pelo menos dois moradores com rendimento de R$1,00 e proceder a imputação para esses registros utilizando o Sistema CANCEIS. Seguindo este critério, foram identificados 16.521 registros de pessoas (7.040 domicílios) para a imputação no conjunto Universo de acordo com as Tabelas 1 e 2.

Tabela 1 – Número de pessoas de 10 anos ou mais de idade com rendimento R$1,00 residentes em domicílios particulares permanentes apontadas para imputação no conjunto Universo, por Unidades da Federação, 2010.

Unidade da Federação	Pessoas com rendimento R$1,00
Brasil	16.521
Rondônia	37
Acre	19
Amazonas	66
Roraiama	11
Pará	286
Amapá	9
Tocantins	35
Maranhão	218
Piauí	64
Ceará	118
Rio Grande do Norte	31
Paraíba	106
Pernambuco	273
Alagoas	44
Sergipe	94
Bahia	488
Minas Gerais	1.077
Espírito Santo	123
Rio de Janeiro	1.706
São Paulo	9.452
Paraná	418
Santa Catarina	278
Rio Grande do Sul	820
Mato Grosso do Sul	22
Mato Grosso	45
Goiás	203
Distrito Federal	478

Fonte:IBGE, Diretoria de Pesquisas.

Tabela 2 – Número de domicílios particulares permanentes apontados para imputação no conjunto Universo, por número de moradores com rendimento R$ 1,00, segundo o total de moradores no domicílio. Brasil, 2010.

Total de Moradores	Moradores com rendimento R$1,00									
	2	3	4	5	6	7	8	9	12	Total
2	2.453	0	0	0	0	0	0	0	0	2.453
3	1.332	463	0	0	0	0	0	0	0	1.795
4	1.015	336	203	0	0	0	0	0	0	1.554
5	380	184	105	73	0	0	0	0	0	742
6	123	71	39	25	21	0	0	0	0	279
7	61	26	12	13	8	3	0	0	0	123
8	20	4	13	6	2	3	5	0	0	53
9	11	5	3	4	1	1	0	2	0	27
10	2	2	2	0	3	0	0	0	0	9
11	2	1	0	0	0	0	0	0	0	3
12	0	0	0	1	0	0	0	0	1	2
Total	5.399	1.092	377	122	35	7	5	2	1	7.040

Fonte: IBGE, Diretoria de Pesquisas.

Para os dados da Amostra foi adotado o mesmo procedimento. Foram imputados 451 registros, o que representa aproximadamente 4.500 pessoas após a expansão.

2 – Pessoas com rendimento atípico (*outliers*)

Após uma série de análises exploratórias, constatou-se que havia alguns casos em que o valor do rendimento declarado não era condizente com características do morador, de seu domicílio ou, até mesmo, da localização da residência (em alguns casos estes valores extremos eram capazes de distorcer a distribuição de rendimentos do setor censitário e até mesmo do município, no caso de municípios pequenos), indicando um possível erro de preenchimento por parte do Recenseador. Com o objetivo de minimizar este tipo de ocorrência, foi desenvolvida uma metodologia para detecção destes valores extremos com maior probabilidade de serem erro (*outliers*).

2.1 – Metodologia para detecção

Para detecção dos *outliers* foi adotado um critério em dois estágios. O primeiro estágio é constituído do modelo de regressão com objetivo de detectar os possíveis *outliers* com base na análise dos resíduos. O logaritmo natural do rendimento total das pessoas foi ajustado em função de um conjunto de variáveis explicativas e seus resíduos analisados posteriormente.

No conjunto Universo foi ajustado um modelo para cada Unidade da Federação e situação do setor usando o seguinte conjunto de variáveis explicativas:

v4002: Tipo de espécie

v0201: Este domicílio é

v0202: Quantos banheiros de uso exclusivo dos moradores existem no domicílio

v0204: O esgoto ou sanitário é lançado (jogado) em

v0205: A forma de abastecimento de água utilizada neste domicílio é

v0206: O lixo deste domicílio é

v0207: Existe energia elétrica no domicílio

v0208: Existe medidor ou relógio no domicílio

v0401: Quantas pessoas moravam neste domicílio em 31 de Julho de 2010

v0502: Qual a relação de parentesco ou convivência com a pessoa responsável pelo domicílio

v0601: Sexo

v6033: Variável auxiliar de idade calculada

v0604: A sua cor ou raça é

v0611: Sabe ler e escrever

decil_uf: Variável que indica a categoria do décimo na Unidade da Federação do rendimento domiciliar *per capita* mediano do setor onde está o domicílio

tamanho_mun: Variável categórica que indicava o tamanho do município em termos populacionais (era composto por 8 faixas de população).

Na Amostra foi feito um ajuste semelhante utilizando as seguintes variáveis explicativas:

v4002: Tipo de espécie

v0201: Este domicílio é

v0202: O material predominante nas paredes externas é:

v0203: Quantos cômodos existem neste domicílio

v0205: Quantos banheiros de uso exclusivo dos moradores existem no domicílio

v0207: O esgoto ou sanitário é lançado (jogado) em

v0208: A forma de abastecimento de água utilizada neste domicílio é

v0215: Neste domicílio existe máquina de lavar roupa

v0217: Neste domicílio existe telefone celular

v0218: Neste domicílio existe telefone fixo

v0219: Neste domicílio existe microcomputador

v0222: Neste domicílio existe automóvel para uso particular

v0401: Quantas pessoas moravam neste domicílio em 31 de Julho de 2010

v0502: Qual é a relação de parentesco ou de convivência com a pessoa responsável pelo domicílio

v0601: Sexo

v6033: Variável auxiliar da idade calculada

v0606: A sua cor ou raça é

v6027: Sabe ler e escrever

v6028: Frequenta escola ou creche

v6400: Nível de instrução

v0637: Vive em companhia de cônjuge ou companheiro(a)

v0640: Qual é o estado civil

v0645: Quantos trabalhos tinha

v0648: Nesse trabalho era

v0650: Era contribuinte de instituto de previdência oficial em algum trabalho que tinha na semana de 25 a 31 de julho de 2010

v0651: No trabalho principal, qual era o rendimento bruto (ou a retirada) mensal que ganhava habitualmente em julho de 2010

decil_uf: Variável que indica a categoria do décimo na Unidade da Federação do rendimento domiciliar *per capita* mediano do setor onde está o domicílio

tamanho_mun: Variável categórica que indicava o tamanho do município em termos populacionais (era composto por 8 faixas de população).

Com base na distribuição dos resíduos do modelo calculou-se para cada Unidade da Federação o terceiro quartil (Q3) e o intervalo interquartílico (IIQ) destes resíduos por situação do setor. Surge então a primeira regra de seleção de *outliers*: o resíduo do rendimento total de um indivíduo deve ser superior a Q3 mais 6 (seis) vezes o IIQ.

No segundo estágio, aplicaram-se duas restrições para selecionar dentre os possíveis *outliers* apontados na etapa anterior aqueles que iriam passar pelo processo de imputação.

Inicialmente foi criada uma variável denominada **escore,** que consistia na soma de valores atribuídos às respostas de quesitos selecionados que apresentaram alta correlação com o rendimento domiciliar, atribuindo uma “nota” (escore) a cada um dos domicílios. Esta variável buscou medir a adequabilidade das condições do domicílio em relação a um conjunto de características básicas.

Como esperado, as distribuições do rendimento domiciliar *per capita* apresentaram diferenças entre as populações urbana e rural tanto na tendência quanto no conjunto de variáveis que apresentaram relação com o rendimento. Em decorrência, a variável escore foi construída de forma diferente para cada tipo de recorte: urbano e rural.

Para o recorte urbano foram selecionadas as seguintes variáveis:

- Condição de propriedade do domicílio
- Quantidade de banheiros
- Forma de abastecimento de água
- Forma de coleta de lixo
- Esgotamento sanitário
- Total de moradores no domicílio
- Alfabetização do responsável pelo domicílio

Para o recorte rural foram selecionadas as seguintes variáveis:

- Quantidade de banheiros
- Forma de abastecimento de água
- Forma de coleta de lixo
- Esgotamento sanitário
- Total de moradores no domicílio
- Disponibilidade de Energia elétrica
- Alfabetização do responsável pelo domicílio

Com o objetivo de sintetizar estas informações em um indicador, as variáveis mais relacionadas ao rendimento foram categorizadas em dois grupos:

0 - Inadequado

1 - Adequado

Como já mencionado, além da relação de variáveis ser diferente entre as populações urbana e rural, a dicotomização diferenciada também foi necessária para algumas respostas de variáveis comuns entre os dois grupos.

O escore de adequabilidade foi obtido então a partir da soma destas variáveis dicotomizadas (Di).

$$ESCORE = \sum Di$$

Recategorização das variáveis de adequabilidade de acordo com a situação do domicílio:

ESTE DOMICÍLIO É	URBANO	RURAL
Próprio de algum morador - já pago	1	-
Próprio de algum morador - ainda pagando	1	-
Alugado	1	-
Cedido por empregador	0	-
Cedido de outra forma	0	-
Outra condição	0	-

QUANTIDADE DE BANHEIROS	URBANO	RURAL
0	0	0
1 ou mais	1	1

ESGOTO	URBANO	RURAL
Rede geral de esgoto ou pluvial	1	1
Fossa séptica	1	1
Fossa rudimentar	0	0
Vala	0	0
Rio, lago ou mar	0	0
Outro	0	0

ABASTECIMENTO DE ÁGUA	URBANO	RURAL
Rede geral de distribuição	1	0
Poço ou nascente na propriedade	0	1
Poço ou nascente fora da propriedade	0	0
Carro-pipa	0	0
Água da chuva armazenada em cisterna	0	0
Água da chuva armazenada de outra forma	0	0
Rios, açudes, lagos e igarapés	0	0
Outra	0	0
Poço ou nascente na aldeia	0	0
Poço ou nascente fora da aldeia	0	0

DESTINO DO LIXO	URBANO	RURAL
Coletado diretamente por serviço de limpeza	1	1
Colocado em caçamba de serviço de limpeza	0	1
Queimado (na propriedade)	0	0
Enterrado (na propriedade)	0	0
Jogado em terreno baldio ou logradouro	0	0
Jogado em rio, lago ou mar	0	0
Tem outro destino	0	0

TOTAL MORADORES	URBANO	RURAL
ATÉ 4	1	1
5 OU MAIS	0	0

ENERGIA ELÉTRICA	URBANO	RURAL
Sim, de companhia distribuidora	-	1
Sim, de outras fontes	-	1
Não existe energia elétrica	-	0

RESPONSÁVEL-SABE LER E ESCREVER	URBANO	RURAL
Sim	1	1
Não	0	0

O objetivo foi construir um indicador que auxiliasse a minimizar os erros de detecção, ou seja, imputar rendimento onde houvesse um valor que não fosse *outlier.* Considerando-se a hipótese de que, à medida que se afasta da "inadequabilidade" (escore=0) para a "adequabilidade" (escore=7), um domicílio tem maiores chances de pertencer realmente ao conjunto dos domicílios com rendimento alto, não foram indicados para imputação os domicílios que pertenciam ao escore máximo, ou seja, escore=7.

Além do escore, calculou-se o rendimento domiciliar *per capita* mediano de cada setor por Unidade da Federação e situação do setor. Em seguida, para cada tipo de recorte foram usados os decis do rendimento *per capita* mediano para a construção de 10 categorias por Unidade da Federação e situação do setor. Este indicador também foi usado para minimizar o erro de seleção, pois não foram indicados para imputação os domicílios situados em setores cujo rendimento domiciliar *per capita era* superior ao último decil daquela Unidade da Federação.

Assim, os valores de rendimento identificados como outliers foram ignorados e submetidos ao procedimento de imputação utilizando o Sistema CANCEIS. Foram identificados para imputação todos os registros que atenderam simultaneamente às três restrições:

1 – O resíduo foi maior que o terceiro quartil (Q3) mais 6 vezes o intervalo (IIQ);

2 – Não residiam em setores cujo rendimento domiciliar *per capita* mediano era superior ao último decil da respectiva Unidade da Federação;

3 – Não residiam em domicílios com escore igual a 7.

Através destas restrições foram levadas para imputação 643 pessoas no conjunto Universo e 274 registros na Amostra (equivalente a 2.114 pessoas após a expansão). Os rendimentos destas pessoas estão distribuídos conforme as Tabelas 3 e 4.

Tabela 3 – Número de valores *outliers* detectados no conjunto Universo, por Unidades da Federação, 2010

Unidade da Federação	Frequência
Brasil	643
Rondônia	8
Acre	6
Amazonas	20
Roraima	2
Pará	18
Amapá	3
Tocantins	6
Maranhão	5
Piauí	3
Ceará	6
Rio Grande do Norte	3
Paraíba	3
Pernambuco	12
Bahia	29
Minas Gerais	80
Espírito Santo	13
Rio de Janeiro	44
São Paulo	133
Paraná	51
Santa Catarina	44
Rio Grande do Sul	66
Mato Grosso do Sul	19
Mato Grosso	27
Goiás	42

Fonte: IBGE, Diretoria de Pesquisas.

Tabela 4 – Número de valores *outliers* detectados na Amostra, por Unidades da Federação, 2010

Unidade da Federação	Frequência - Não Expandido	Frequência - Expandido
Brasil	274	2.114
Rondônia	4	29
Acre	2	20
Amazonas	2	20
Roraima	1	3
Pará	11	128
Tocantins	11	32
Maranhão	14	88
Piauí	3	9
Ceará	4	36
Rio Grande do Norte	2	6
Paraíba	9	26
Pernambuco	8	86
Alagoas	1	9
Bahia	18	167
Minas Gerais	31	260
Espírito Santo	7	44
Rio de Janeiro	10	145
São Paulo	47	494
Paraná	17	83
Santa Catarina	17	100
Rio Grande do Sul	18	93
Mato Grosso do Sul	9	75
Mato Grosso	13	103
Goiás	15	59

Fonte: IBGE, Diretoria de Pesquisas.

A distribuição dos valores dos rendimentos das pessoas apontadas para imputação pode ser vista nas Tabelas 5 e 6.

Tabela 5 – Número de valores *outliers* detectados no conjunto Universo, por faixas de valor do rendimento total. Brasil, 2010.

Faixa de Rendimento	Frequência
R$30.000,00 a R$39.999,00	6
R$40.000,00 a R$49.999,00	1
R$50.000,00 a R$99.999,00	48
R$100.000,00 a R$499.999,00	533
R$500.000,00 a R$999.998,00	55
Total	643

Fonte: IBGE, Diretoria de Pesquisas.

Tabela 6 – Número de valores *outliers* detectados na Amostra, por faixas de valor do rendimento total. Brasil, 2010

Faixa de Rendimento	Frequência
R$10.000,00 a R$14.999,00	2
R$15.000,00 a R$19.999,00	7
R$20.000,00 a R$29.999,00	3
R$30.000,00 a R$39.999,00	9
R$40.000,00 a R$49.999,00	3
R$50.000,00 a R$99.999,00	40
R$100.000,00 a R$499.999,00	193
R$500.000,00 a R$999.998,00	15
R$ 1.200.000,00	1
R$ 1.600.000,00	1
Total	274

Fonte: IBGE, Diretoria de Pesquisas.

3 – Rendimentos registrados como R$999.999,00

Em princípio, o valor máximo de rendimento passível de registro nos questionários Básico e da Amostra era de R$999.999,00. Portanto este deveria ser um valor válido, o que de fato ocorreu durante a coleta de dados.

O que se observou, porém, é que a quase totalidade dos casos onde estes registros ocorreram estava concentrada nos estados do Rio de Janeiro e São Paulo, mais especificamente nas capitais, o que indicava que o valor R$999.999,00 pode ter sido utilizado erroneamente para registrar um valor ignorado onde o rendimento não foi declarado, talvez por ser este um procedimento adotado em outras pesquisas do IBGE.

Observou-se, também, uma concentração nos valores R$99.999,00 e R$9.999,00 nestes mesmo estados, o que talvez tenha ocorrido por uma falha adicional de digitação ao

tentar registrar o valor R$999.999,00. Baseado nisso optou-se por indicar todos os registros de rendimento com valor R$999.999,00 para imputação (9.094 pessoas no conjunto Universo e nenhuma ocorrência no conjunto de questionários da Amostra). Os demais valores com dígitos repetidos ficaram sujeitos ao modelo de detecção de *outliers* descrito no item anterior.

4 – Estudo sobre os rendimentos de pessoas residentes em domicílio com rendimento zero

No Censo Demográfico 2000 os dados da Amostra apontaram a existência de 2.109.697 domicílios sem rendimento no Brasil, o que representava 4,7% do total de 45.023.421 domicílios (particulares permanentes ou improvisados).

Ao repetir-se a análise para o Censo Demográfico 2010 (mesmo antes de realizar os procedimentos de crítica), obteve-se o número de 2.467.188 domicílios sem rendimento para o conjunto Universo. Este número representava 4,3% dos 57.419.189 domicílios (particulares permanentes e improvisados). Quando analisados os dados expandidos da Amostra, verificou-se que 2.767.889 dos 57.326.393 domicílios apresentavam rendimento domiciliar zero. Este número correspondia a 4,8% dos domicílios.

Deve ser mencionado que a forma de coleta dos dados de rendimento nos dois censos foi diferente. No Censo 2000, os dados foram coletados em papel. No Censo 2010, a coleta foi realizada com um Computador de Mão (PDA) que continha regras mais restritivas de preenchimento, e dificultava a possibilidade de o Recenseador registrar que a pessoa possuía rendimento e não declarar o valor.

Analisando a distribuição de rendimentos proveniente da Pesquisa Nacional por Amostra de Domicílios – PNAD, edição de 2009, verifica-se que foram estimados 771.179 domicílios sem rendimento, representando 28,0% do total estimado através da Amostra do Censo 2010. Além disso, a PNAD aponta 1.812.311 domicílios “sem declaração” de rendimento.

Levando em consideração as diferenças de captação da informação sobre rendimento entre as duas pesquisas, pode-se observar que as demais classes de rendimento guardam similitude entre o Censo Demográfico de 2010 e PNAD 2009. Quando se realiza a soma dos domicílios sem rendimento com os domicílios “sem declaração” na PNAD, chega-se a uma estimativa de 2.583.490 domicílios, algo muito próximo do total de domicílios com rendimento zero encontrados no Censo Demográfico 2010. Acredita-se que este fato é um forte indício a contribuir para o raciocínio de que a opção “não tem rendimento” serviu no Censo, em um número significativo de situações, como alternativa para a categoria “ignorado”.

4.1 – Metodologia para detecção

Para avaliar a possibilidade de imputação de rendimento nestes domicílios, buscou-se uma metodologia de detecção que identificasse os casos de rendimento zero “suspeitos”. O ideal seria fazer tal crítica para as pessoas com rendimento zero, porém as perguntas

do questionário Básico (que definem o conteúdo do conjunto Universo), para os moradores dos domicílios, não são suficientemente correlacionadas com a de rendimento a ponto de servirem na aplicação de um modelo. Optou-se, então, por fazer a crítica no nível dos domicílios, buscando-se dentro do conjunto dos domicílios sem rendimento identificar aqueles suspeitos de conterem morador(es) com rendimento não declarado mas registrado como zero.

O aplicativo de coleta do Censo 2010 não tinha um código específico para o registro de rendimento ignorado. A forma encontrada para distinguir estes domicílios daqueles com rendimento efetivamente igual a zero foi buscar uma fonte alternativa, no caso a PNAD 2009. A abordagem adotada foi a de construir um modelo de regressão logístico, com base em dados da PNAD, que permitisse estimar a probabilidade de um domicílio ter rendimento não declarado, considerando-se para isso o conjunto dos domicílios particulares permanentes sem rendimento e sem declaração.

4.2 - O modelo logístico

Considere o conjunto dos domicílios particulares permanentes investigados na PNAD 2009, com rendimento zero ou não declarado. Defina-se então a variável:

$Y_i = 0$, se o rendimento total do i-ésimo domicílio investigado foi zero

$Y_i = 1$, se o rendimento total do i-ésimo domicílio investigado não foi declarado

$p_i = P(Y_i = 1)$: probabilidade do rendimento total do i-ésimo domicílio investigado ser não declarado

O modelo logístico acima citado pode ser descrito na forma:

$$\boldsymbol{log(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i,}$$

onde: $\varepsilon_i \sim N(0,\sigma^2)$ e é denominado o erro aleatório do modelo, que capta a parcela da variabilidade de Y_i não explicada pelas variáveis $X_1, X_2, \ldots, X_k$

De forma que pode-se estimar:

$$\boldsymbol{p_i = exp(b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki})},$$

onde b_j é uma estimativa de β_j, j=1,2,...,k.

Levando-se em conta a hipótese de que as variáveis explicativas adotadas no modelo da PNAD 2009 sejam conceitualmente compatíveis com as variáveis do Censo 2010, então podemos utilizar o modelo para estimar a probabilidade de cada domicílio com rendimento zero do Censo possuir um rendimento não declarado.

4.3 – Aplicação do modelo nos dados da Amostra

Através da similaridade com os dados do censo, foi criado um banco com dados da PNAD 2009 contendo 1.605 (67,0%) domicílios sem rendimento e 798 (33,0%) domicílios com rendimento não declarado.

Foram utilizadas as seguintes variáveis explicativas:

- Grupamento de Unidades da Federação - com base na similaridade da proporção de domicílios sem rendimento

- Número de dormitórios

- Número de moradores com 10 anos ou mais

- Tem automóvel (sim ou não)

- Há morador com 65 anos ou mais (sim ou não)

- Responsável sabe ler e escrever (sim ou não)

- Tem telefone fixo (sim ou não)

- Tem máquina de lavar (sim ou não)

- Nível de instrução mais elevado alcançado por algum dos moradores

Após o ajuste do modelo, aplicou-se o modelo ajustado nos próprios dados da PNAD 2009 e obteve-se a classificação “aleatória” (baseada nas probabilidades dadas pelo modelo) do domicílio em duas classes: com rendimento não declarado e sem rendimento. Neste modo, o modelo produziu 77,0% de classificações corretas.

Também foi testado um modelo com corte de probabilidade em 0,5. Neste caso, o modelo produziu 84,0% de classificações corretas.

O próximo passo foi aplicar o modelo aos dados da Amostra do Censo 2010. A Amostra apresenta 257.127 domicílios com rendimento zero, sendo um total expandido de 2.537.672. Primeiramente foi aplicado o critério de classificação aleatória. Este critério classificou 725.727 domicílios (28,6%) como sem declaração (67.131 registros, sem expansão). No nível de unidade da federação, as taxas de troca, domicílio sem rendimento para domicílio com rendimento ignorado, variam entre 4,4% em Mato Grosso do Sul e 40,9% em São Paulo. Nestes domicílios existem 2.177.632 moradores com 10 anos ou mais.

Em seguida, aplicou-se o modelo com o critério de classificação com ponto de corte em 0,5. Este critério classificou 571.754 domicílios como sem declaração (22,5%). No nível de Unidade da Federação, as taxas de troca, domicílio sem rendimento para domicílio com rendimento ignorado, variam entre 1,3% em Mato Grosso do Sul e 37,9% em São Paulo. Nestes domicílios existem 1.913.055 moradores com 10 anos ou mais.

4.4 – Aplicação do modelo nos dados do conjunto Universo

Para o ajuste do modelo para o conjunto Universo selecionou-se no banco de dados da PNAD 2009 1.606 (32,5%) domicílios sem rendimento e 3.331 (67,5%) domicílios com rendimento não declarado.

Foram utilizadas as seguintes variáveis explicativas:

- Grupamento de Unidades da Federação - com base na proporção de domicílios sem rendimento

- Situação do setor censitário (urbano ou rural)

- Há morador com 65 anos ou mais (sim ou não)

- Responsável sabe ler e escrever (sim ou não)

- Número de moradores com 10 anos ou mais

Novamente após o ajuste do modelo, aplicou-se o modelo ajustado nos dados da PNAD 2009 e foi obtida a classificação “aleatória” do domicílio em duas classes: com rendimento não declarado e sem rendimento. Neste modo, o modelo produziu 59,0% de classificações corretas. Quando aplicado o modelo com corte de probabilidade de 0,5 o modelo produziu 60,0% de classificações corretas.

Em seguida, aplicou-se o modelo ajustado aos dados do conjunto Universo do Censo 2010, que possui 2.168.753 domicílios com rendimento zero. Utilizado o critério de classificação aleatória, 1.037.197 domicílios foram classificados como sem declaração (47,8%). No nível de Unidade da Federação, as taxas de troca - domicílio sem rendimento para domicílio com o rendimento ignorado – variavam entre 22,0% no Espírito Santo e 66,0% no Pará.

Quando aplicado o modelo com o critério de classificação com ponto de corte em 0,5, foram classificados 975.467 domicílios como sem declaração (45,0%). No nível de Unidade da Federação, as taxas de troca - domicílio sem rendimento para domicílio com o rendimento ignorado - variavam entre 11,0% no Espírito Santo e 75,0% no Pará.

4.5 - Conclusão

Uma questão de ordem técnica e operacional conduziu à decisão de não se aplicar o modelo, ainda que os resultados obtidos do ajuste dos modelos fossem satisfatórios para a Amostra e para o Universo. Ao verificar nos dados da Amostra quem eram as pessoas residentes nos domicílios apontados para a imputação constatou-se que, em aproximadamente 90,0% dos casos, eram pessoas não-ocupadas. Isso implicaria em uma das duas seguintes possibilidades: 1 - Imputar toda a parte de trabalho e rendimento destas pessoas; ou 2 – Imputar somente um rendimento de outras fontes para estas pessoas. A opção mais factível seria a imputação apenas de um rendimento de outras fontes. A distribuição

do rendimento de outras fontes é naturalmente diferente da distribuição dos rendimentos de trabalho. Como não há esta separação nos dados do Universo, as distribuições de rendimento nos domicílios imputados da Amostra e do Universo seriam consideravelmente diferentes. Dadas estas restrições, optou-se por não realizar a imputação de rendimento nestes domicílios e manter o valor de rendimento zero.