ESTUDO TÉCNICO

N.º 05/2016

MDS

Metodologia de cômputo das taxas de pobreza e extrema pobreza das PNADs 1992 a 2014

SAGI

MINISTÉRIO DO DESENVOLVIMENTO SOCIAL E COMBATE À FOME

SECRETARIA DE AVALIAÇÃO E GESTÃO DA INFORMAÇÃO

Estudo Técnico
N.º 05/2016
Metodologia de cômputo das taxas de pobreza e extrema pobreza das PNADs 1992 a 2014.

Técnico responsável
Júlio César Gomes Fonseca
Caio Nogueira Gonçalves
Marcelo Lúcio Saboia Fonseca

Revisão
Paulo de Martino Jannuzzi
Marconi Fernandes de 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: *Análise Discriminante; estimativas PNAD; série histórica PNAD.*

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

Este estudo técnico apresenta os resultados e aplicação da metodologia de Análise Discriminante para identificação de pessoas com perfil de extrema pobreza (EP) e pobreza (PO) nos grupos de domicílios cuja declaração de rendimento domiciliar *per capita* sejam iguais a zero (SR) ou sem declaração (SD) dos microdados da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 1992 a 2014, com vistas a analisar e obter maior precisão nos dados para as populações pobres e extremamente pobres.

1. Introdução

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 não ocorrer apenas nos anos de realização do Censo Demográfico. O objetivo da Pesquisa é oferecer tratamento sistemático para características gerais da população, de caráter permanente, como educação, rendimento, habitação e trabalho; e características de natureza variável, como migração, fecundidade, saúde, nupcialidade, nutrição, e outros temas que são incluídos de acordo com a necessidade.

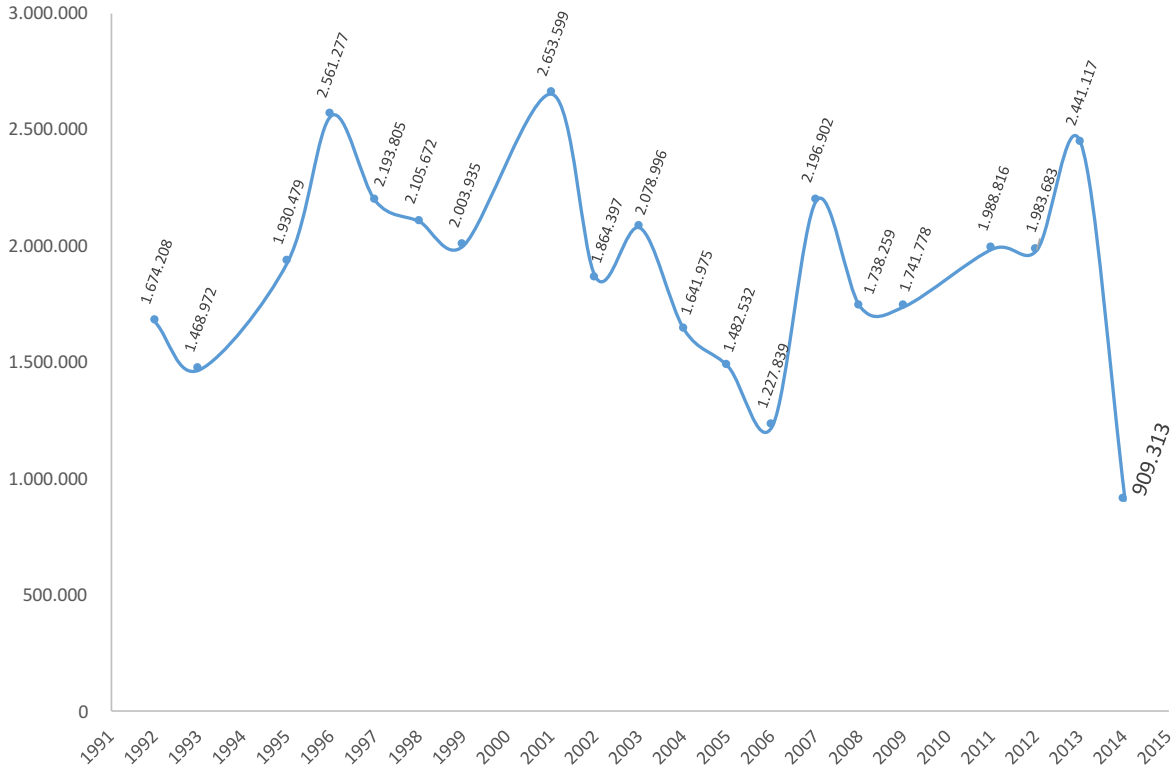
No presente estudo serão apresentados aprimoramentos realizados na metodologia de análise de discriminante publicada no Estudo Técnico 15/2014, onde à época lançou-se olhar no público com perfil de extrema pobreza nos grupos de domicílios cuja declaração de rendimento domiciliar *per capita* eram iguais a zero, também denominados *Sem Rendimento (SR)* e os grupos de domicílios informados como *Sem Declaração (SD)*. Os principais aprimoramentos destacados são: 1) inclusão de mais variáveis no modelo para melhor qualificar cenários específicos; 2) tratamento de não resposta em todas variáveis utilizadas; 3) análise de período completo da série de 1992 à 2014; 4) incorporação da estimativa das taxas de pobreza no período. Tais aprimoramentos, consideram assim outras melhorias, tais como as destacadas no Estudo Técnico 04/2016, onde são abordadas as questões de reponderação de pesos da PNAD dos anos 90. Desta forma, o estudo se divide em três seções principais: motivação, metodologia aplicada e considerações finais.

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. Essa linha de pesquisa engloba a análise crítica de séries históricas de pesquisas basilares para a construção de políticas públicas, entre PNAD realizada pelo IBGE.

Nesse âmbito, um dos métodos de pesquisa obteve resultados interessantes como demonstraram os Estudos Técnicos nº 15/2014 e nº 24/2012, quando se constatou uma superestimação dos índices de pobreza, segundo os métodos habituais de diversos pesquisadores e institutos de pesquisa. Estudos realizados pela SAGI/MDS sobre os segmentos tratados nestes estudos revelam comportamentos distintos na participação dos SR e dos SD, no estudo das estimativas de rendimento domiciliar *per capita* na PNAD nos últimos anos. Neste sentido, observa-se que a série dos microdados de 1992 a 2014, apesar dos picos apresentados em meados de 1995 e 2001, apresentou uma quebra de comportamento no quantitativo dos SR a partir de 2007 (Gráfico 1), onde se observa um aumento de 79,1% em relação a 2006, culminando na inversão do comportamento da série no período subsequente, a partir de 2008.

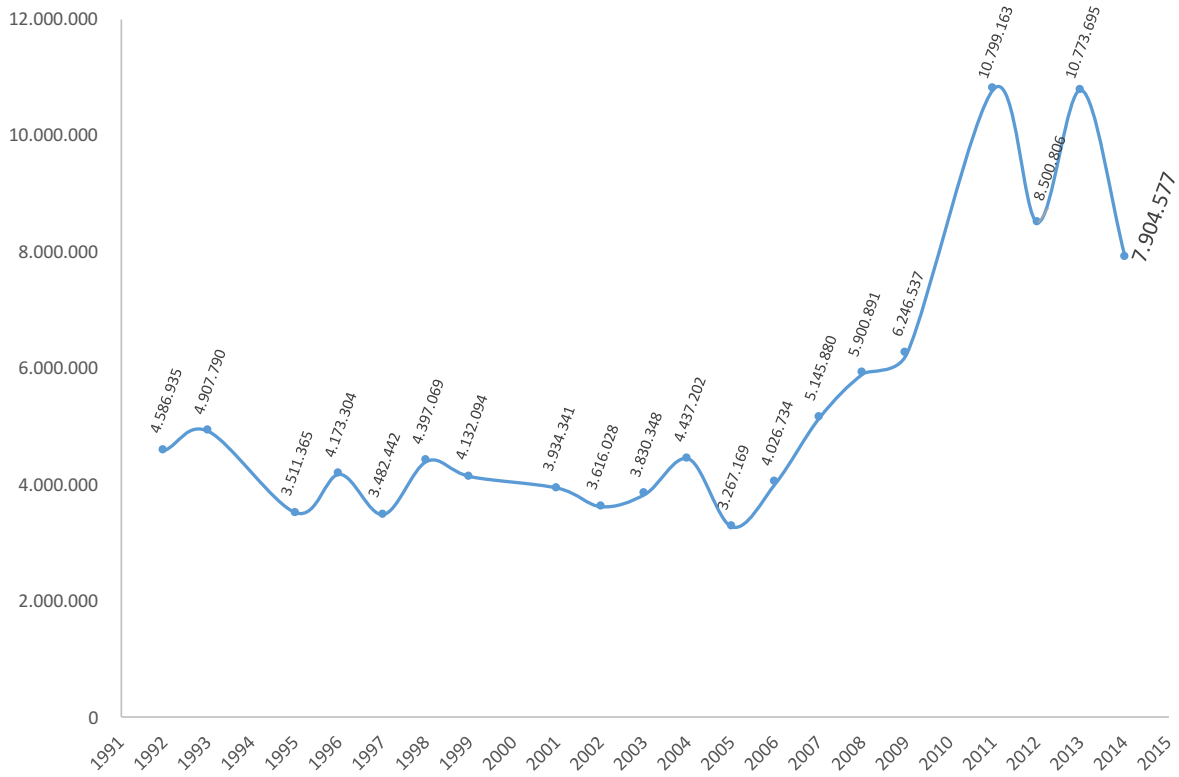
Gráfico 1 - Evolução do quantitativo de indivíduos sem rendimentos no rendimento domiciliar *per capita* – Brasil, 1992-2014



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

Já o grupo dos SD apresentou um padrão de comportamento mais expressivo a partir de 2007, quando assume uma tendência crescente (Gráfico 2), com expressivo aumento em 2011, onde se observou um quantitativo de 6,1 milhões de indivíduos em 2009, se elevar a 10,6 milhões em 2011, representando um aumento de 72,4% entre estes anos.

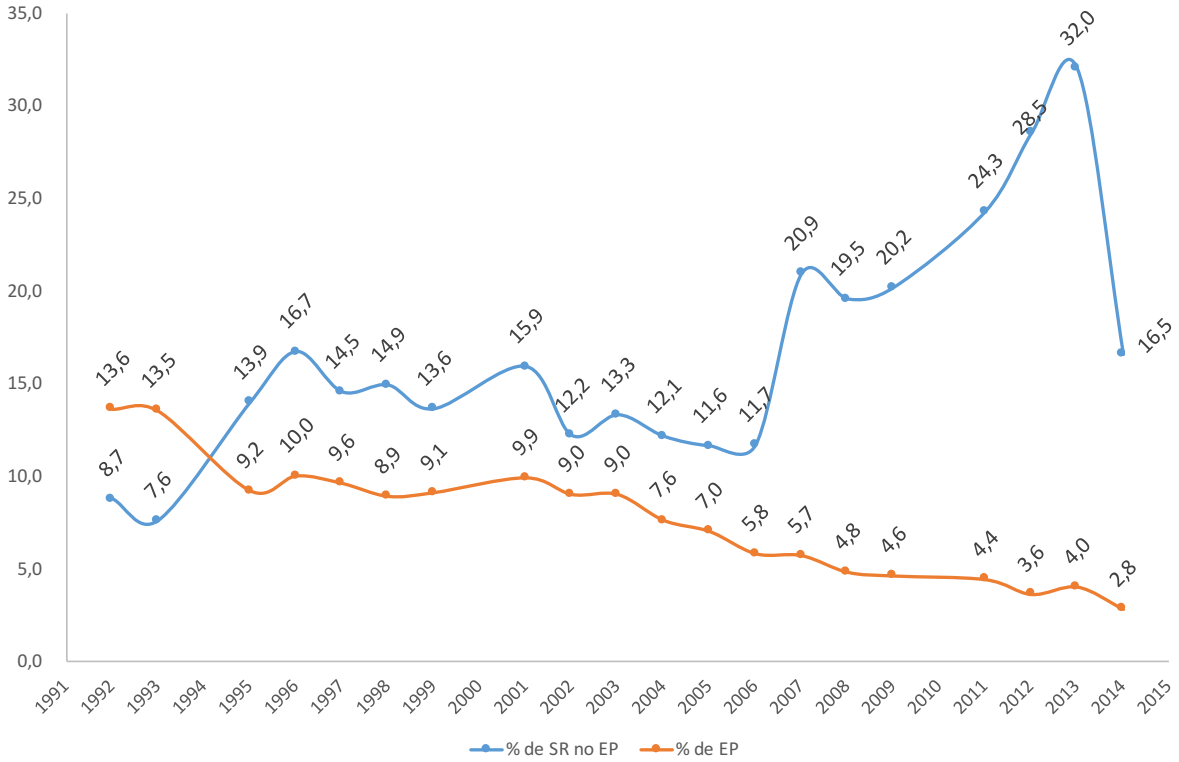
Gráfico 2 - Evolução do quantitativo de indivíduos sem declaração no rendimento domiciliar *per capita* – Brasil, 1992-2014



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

Tais cenários, conjuntamente apontam razões pelas quais no cômputo da taxa de extrema pobreza, o impacto da parcela dos rendimentos declarados como zero chega a 32,0% no ano de 2013 (Gráfico 3). Estes dados revelam o alto grau de superestimação das taxas calculadas utilizando-se pesquisas como a PNAD, na avaliação de vulnerabilidades no país, elevando assim os índices de pobreza e principalmente os de extrema pobreza. Desconsiderados os SD do cômputo das taxas de EP e PO, de forma a não serem contabilizados no denominador do cálculo, a avaliação do impacto dos SR apresentado no gráfico acima, revela uma mudança de patamar a partir de 2007, ano este onde o quantitativo de indivíduos que se auto declararam com rendimento domiciliar *per capita* iguais a zero, passaram de 1.227.839 declarados em 2006 para 2.196.902 em 2007, representando um aumento de 78,92%.

Gráfico 3 - Evolução do percentual dos sem rendimento na extrema pobreza no rendimento domiciliar *per capita* – Brasil, 1992-2014

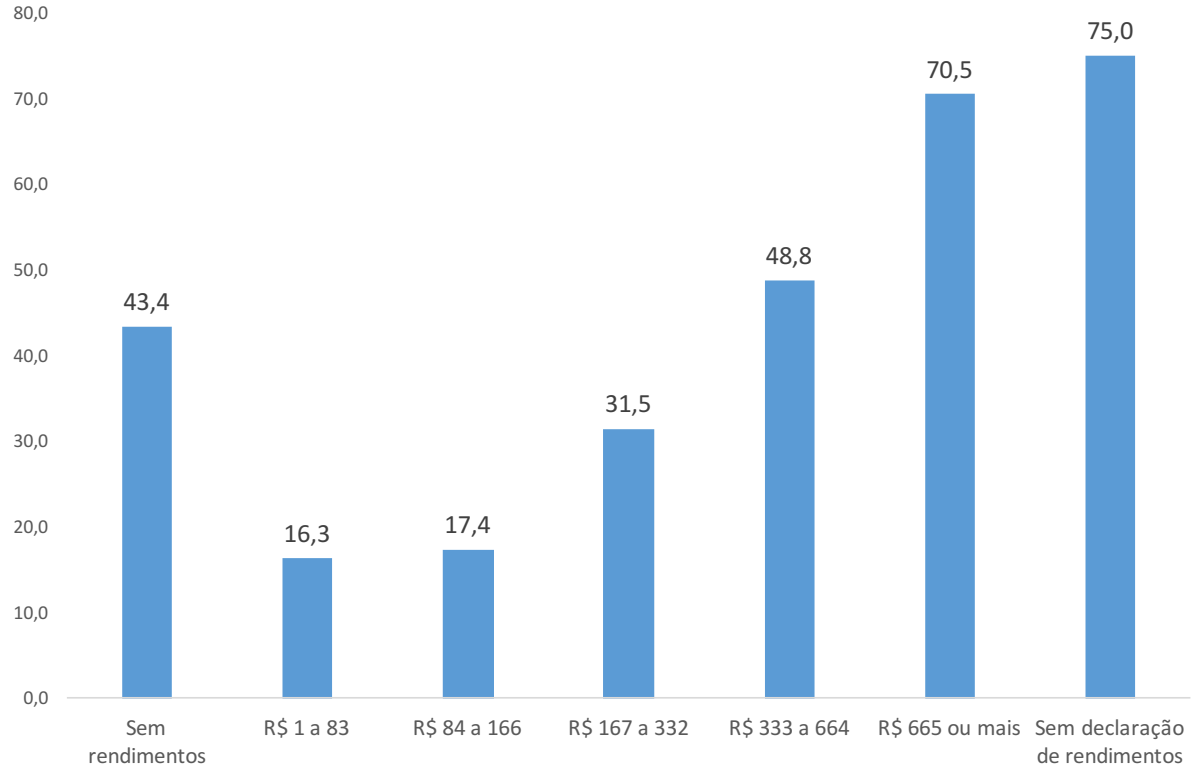


Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios (PNAD). Elaboração SAGI/MDS.

Dado este aumento, o impacto na taxa de extrema pobreza, causado por esta parcela, salta de 11,7% para 20,9%, os dados da pesquisa aparentam, de fato, apontar para algum tipo de procedimento que remeta à imputação de renda para subsequente avaliação de taxas que apresentam a vulnerabilidade sob diversos aspectos e/ou eixos temáticos – apesar dos grandes esforços feitos em 2014 com o intuito de aprimorar a coleta de dados.

Foi possível verificar mais evidências do problema exposto até então utilizando variáveis socioeconômicas, de modo que melhor representassem a exposição da população de baixa renda à situação de vulnerabilidade (Gráfico 4). A maior parte destas variáveis apresentou um alto percentual de subdeclaração de renda, superior aos níveis apurados nas faixas de renda que definem as linhas de pobreza discutidas pelos principais pesquisadores e instituições com publicações no âmbito da questão da vulnerabilidade social.

Gráfico 4 - Percentual dos domicílios que possuem máquina de lavar por faixas de rendimento domiciliar per capita – Brasil, 2014



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

2. Metodologia aplicada

Após trabalho investigativo sobre a aplicabilidade de um modelo estatístico preditivo, por se tratar de estimação da probabilidade de um indivíduo encontrar-se em estado de extrema pobreza ou pobreza, optou-se pela utilização de um método multivariado com caráter classificatório e não exatamente preditivo, em função da facilidade de uso e aplicabilidade no problema proposto. Assim, a metodologia utilizada para esse estudo balizou-se em técnicas provenientes do ramo da estatística que trata de análises multivariadas.

A técnica adotada foi a Análise de Discriminante (AD) por se tratar de método estritamente de investigação da classificação de grupos, segundo características determinísticas modelando separadamente os casos de extremamente pobres (EP), ou não, e de pobres (PO), ou não. Como a principal questão a ser investigada tratava da classificação dos indivíduos, cujo rendimento domiciliar *per capita* foi informado como "zero" ou não declarado, enquanto possibilidade de pertencerem, em verdade, aos grupos em função das características intrínsecas destes grupos, definidas pelas variáveis elencadas como determinísticas do grupo, definiu-se como variável de

interesse (variável resposta) a "presença", no caso dos extremamente pobres, no grupo cuja renda domiciliar *per capita* pertencesse ao intervalo de R$ 1,00 a R$ 70,00 (valores de junho de 2011[1]), e no caso dos pobres, no grupo cuja renda domiciliar *per capita* pertencesse ao intervalo de R$ 70,01 a R$ 140,00 (faixa de 2010), valores estes inflacionados conforme o ano da amostragem. Dessa forma, uma variável indicadora do tipo "dummy" foi definida como "1" para todos os indivíduos que pertencessem à faixa de renda do grupo acima citado, e "0" para os demais.

Quadro 1 - Variáveis utilizadas na Análise de Discriminante – Brasil, 1992-2014

Variável na PNAD	Descrição	Construção metodológica	Dummy utilizada
V8005	Idade do indivíduo/14 anos ou menos	0 – tem de 0 a 3 filhos 1 – tem mais de 3 filhos	V1_CRIANCA
V0230	Tem máquina de lavar roupa?	0 - não tem 1- tem	V2_MAQLAVAR
V2032	Tem carro ou motocicleta de uso pessoal?	0 - não tem 1- tem	V3_CARRO
V0228/V0229	Tem geladeira ou freezer?	0 - não tem 1- tem	V4_GELADEIRA
V0221/V0222 /V0223	Tem fogão de duas ou mais bocas? Tem fogão de uma boca?	0 - não tem 1- tem (elétrico ou gás)	V5_FOGAOBOM
V0211	Água canalizada em algum cômodo?	0 - não tem 1- tem	V6_AGUA
V0231/V0232	Microcomputador é utilizado para acessar a Internet?	0 - não tem 1- tem	V7_INTERNET
V4745	Nível de instrução da pessoa de referência	0 - não tem (Fundamental incompleto ou menos) 1- tem (Fundamental completo ou mais)	V8_ESCOLARIDADE
V2020	Tem telefone fixo convencional?	0 - não tem 1- tem	V9_TELFIXO
V0215/V0217	Tem banheiro ou sanitário no domicílio ou na propriedade com escoadouro? *	0 - não tem 1- tem	V10_BANHEIRO

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

[1] A linha de extrema pobreza é a de R$ 70 com referência ao mês de junho de 2011, mês de lançamento do Plano Brasil Sem Miséria. A linha foi atualizada pelo INPC nos meses de referência das PNADs para cômputo da extrema pobreza.

Utilizou-se o conjunto das 15 (quinze) variáveis abordadas nas análises de perfis, recodificando-as a fim de facilitar sua utilização e interpretação (Quadro 1).

Segundo as recomendações de Hair (2005), a base foi dividida na proporção 75% e 25% aplicando-se procedimento de amostra aleatória simples, selecionando-se registros conforme distribuição uniforme, delineando inicialmente uma variável cujo valor "1" representasse os EP e o valor "0" os NÃO EP, no modelo AD para classificação de extremamente pobres. Outra modelagem utilizou o mesmo grupamento para classificação de pobres, permitindo assim uma posterior validação cruzada entre os modelos gerados. A técnica permitiu o conhecimento das variáveis que mais se destacaram na discriminação dos grupos, a partir de testes estatísticos, como o lambda de Wilks, a correlação canônica, o qui-quadrado e o *eigenvalue*.

Com a seleção das variáveis discriminantes (explicativas) para formação dos grupos, o próximo passo foi a identificação das funções discriminantes. A função geral discriminante pode ser representada pela seguinte equação linear,

$$Z_n = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

onde Z_n é a variável dependente, α é o intercepto, X_i são as variáveis explicativas e β_i são os coeficientes discriminantes para cada variável explicativa.

Assim, os valores das variáveis explicativas de cada observação são inseridos nas funções de classificação e, consequentemente, um escore de classificação é calculado para cada grupo, para aquela observação.

Após a função discriminante ser definida, foi calculado o escore discriminante da variável dependente (Z) para cada observação. Segundo Hair et al. (2005), o escore de corte é o critério em relação ao qual o escore discriminante de cada objeto é comparado para determinar em qual grupo o objeto deve ser classificado. Para os grupos de mesma dimensão amostral (tamanho), o cálculo do escore de corte é:

$$f = \frac{\overline{d_1} + \overline{d_2}}{2}$$

onde $\overline{d_1}$ e $\overline{d_2}$ representam as médias das funções discriminantes (centróides) nos grupos de "0" e "1", respectivamente. Para os grupos de tamanhos diferentes, têm-se:

$$f = \frac{n_1\overline{d_1} + n_2\overline{d_2}}{n_1 + n_2}$$

onde n_1 e n_2 são os tamanhos dos grupos "0" e "1", respectivamente. Sharma (1996) destaca que o valor do corte selecionado é aquele que minimiza o número de classificação incorreta.

Verificou-se que o modelo criado não apresentava "uma boa performance", em razão da baixíssima sensibilidade e alta especificidade verificadas. A solução foi modificar a forma de escolha do corte, utilizando a estatística de Kolmogorov-Smirnov (KS), que é definida como a diferença absoluta máxima das funções de distribuição acumulada (F) entre cada um dos grupos para os quais foram geradas estimativas de classificação. Assim, entre os EP e NÃO EP, obteve-se a estatística:

$$KS = max|F_{m.EP}(a) - F_{n.EP}(a)|$$

E entre PO e NÃO PO:

$$KS = max|F_{m.PO}(b) - F_{n.PO}(b)|$$

Onde *m* é o número registros atribuídos aos "0" e *n* o de "1".

Verificou-se, portanto, que a estimativa KS apresentava uma boa separação entre os grupos dicotômicos, ao alterar o ponto de corte onde as probabilidades obtidas pela função discriminante, associadas ao evento de "sucesso", deveriam de fato classificar um SR/SD como, respectivamente, um EP ou PO[2].

Utilizou-se o indicador de qualidade do modelo, conhecido por Lift. O Lift acumulado indica quantas vezes, em um determinado nível de rejeição, a seleção de pontuação do modelo é

[2] Classificações estas que foram avaliadas de maneira independente para cada modelo.

melhor que a seleção aleatória. Na prática, o Lift é computado correspondendo aos decis das famílias de melhor score, ou seja, a base é dividida em 10 partes (decil), calculando-se os indicadores.

$$Lift(a) = \frac{F_{m.EP}(a)}{F_{n.POP}(a)}$$

$$Lift(b) = \frac{F_{m.PO}(b)}{F_{n.POP}(b)}$$

Após calculado este indicador considerou-se como medida final para avaliação da qualidade do modelo estimado o valor Lift superior a 1,5, valor este que segundo especialistas em avaliação de carteiras de risco[3], combinado com o valor KS superior a 40%, produzem bons resultados em aplicações que utilizam técnicas de Data Mining. Dessa forma, estabeleceu-se um patamar mínimo para se definir um ponto de corte aceitável para o modelo criado. Em seguida, utilizamos a "Matriz de Confusão", representada pelo Quadro 2, para verificar sua eficácia. Neste quadro, a diagonal principal, representada pelos elementos "VP" e "VN", indica os quantitativos para os quais a classificação final está correta. Assim, os elementos pertencendo à diagonal secundária, representada pelos elementos "FP" e "FN", denotam as falhas de classificação.

Quadro 2 - Matriz de Confusão

Matriz de Confusão		**Modelo Empírico**		
		Positivo	*Negativo*	
Modelo Estimado	*Positivo*	Verdadeiros Positivos (VP)	Falsos Positivos (FP)	***Valor de Predição Positiva*** VP / (VP+FP)
	Negativo	Falsos Negativos (FN)	Verdadeiros Negativos (VN)	***Valor de Predição Negativa*** VN / (VN+FN)
		Sensibilidade VP / (VP+FN)	***Especificidade*** VN / (FP+VN)	**Acurácia** (VP+VN) / (VP+FP+FN+VN)

No intuito de utilizar os principais conceitos utilizados na avaliação de testes diagnósticos, citados por Martinez e Louzada-Neto (2000), avaliamos as principais medidas de acurácia

[3] Esta técnica é muito utilizada por analistas de carteiras de risco em bancos e seguradoras.

calculáveis por meio da matriz de confusão: *sensibilidade*, *especificidade, valor de predição positiva, valor de predição negativa* e *acurácia*.

Assim, o conceito de "positivo" observado na matriz de confusão, implica, na modelagem ou na observação empírica, que indivíduos com perfil de extrema pobreza, "*são classificado como extremamente pobre*" e vice-versa para o conceito "negativo", também na modelagem de indivíduos com perfil de pobreza. Para esclarecer melhor o ganho na utilização de tais medidas, destacamos abaixo os conceitos de cada uma delas, segundo sua utilização no contexto da modelagem das taxas de extrema pobreza e pobreza.

Sensibilidade: probabilidade do modelo estimado classificar o indivíduo como "positivo", dado que o modelo empírico o considera "positivo".

Especificidade: probabilidade do modelo estimado classificar o indivíduo como "negativo", dado que o modelo empírico o considera "negativo".

Valor de Predição Positiva: probabilidade do indivíduo ser "positivo", segundo o modelo empírico, dado que o modelo estimado o classifica como positivo.

Valor de Predição Negativa: probabilidade do indivíduo ser "negativo", segundo o modelo empírico, dado que o modelo estimado o classifica como negativo.

Acurácia: percentual global de precisão de acertos do modelo estimado em relação ao modelo empírico. Definido assim, observa-se que a *sensibilidade* e a *especificidade* são componentes da acurácia.

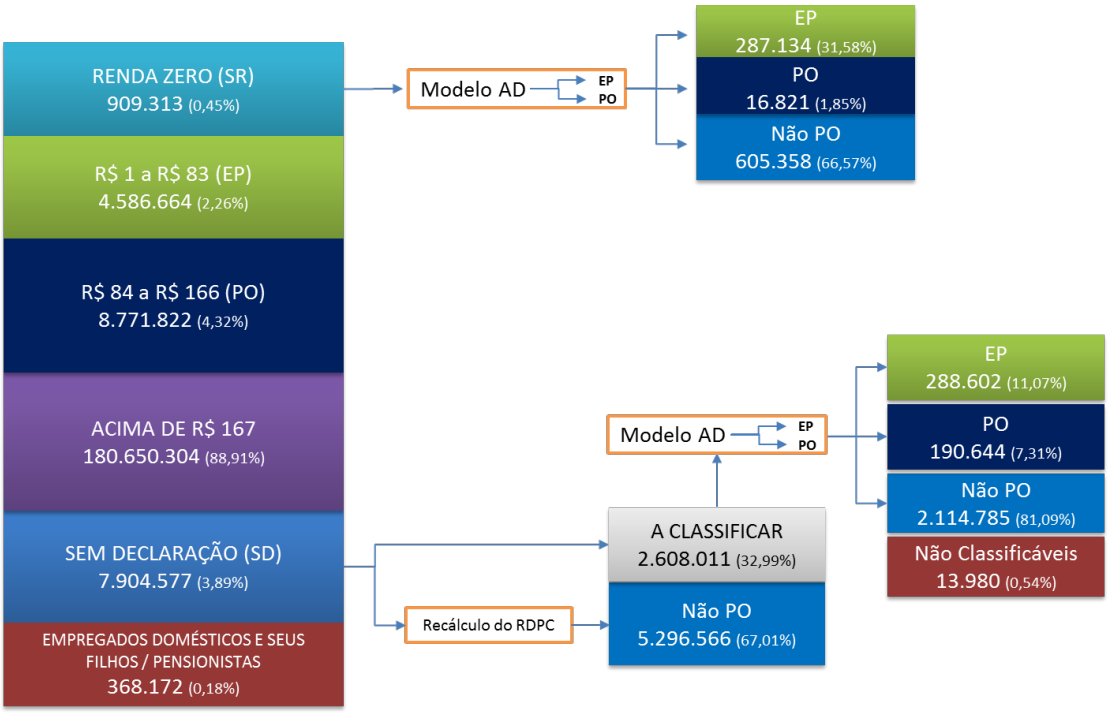
2.1. Ajuste do modelo

Dentre os diversos procedimentos investigados pela SAGI/MDS na busca pela solução do problema apresentado, além da Análise de Discriminante, um recálculo[4] da renda computável entre os SD, produziu resultados bastante promissores no cômputo das taxas de extrema pobreza e pobreza. Dessa forma, o modelo estimado foi aplicado aos SD após o cálculo de renda

[4] Procedimento este utilizado para avaliar entre as variáveis de rendimentos declarados por cada indivíduo, os que continham informação.

computável, procedimento este que leva em consideração o recálculo da renda domiciliar *per capita* para todos os indivíduos que apresentaram pelo menos um rendimento recebido nas variáveis que captam as transferências de renda obtidas do governo federal. No Quadro 3 apresenta-se de forma sumária os números obtidos com a aplicação da técnica, discutidos em detalhes a seguir.

Quadro 3 - Distribuição da População da PNAD por Faixa de Renda – Brasil, 2014



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

É importante destacar que foram ajustados quatro modelos de Análise Discriminante, sendo dois para classificação dos indivíduos com perfil de extrema pobreza e dois para classificação dos indivíduos com perfil de pobreza. Ao final da modelagem as classificações foram combinadas segundo os números apresentados no Quadro 3. A aplicação do procedimento de recálculo do RDPC entre os 7,9 milhões de indivíduos declarados como SD na PNAD 2014 permitiu a identificar 5,3 milhões de indivíduos com rendimento domiciliar *per capita* acima do parâmetro de referência da pobreza. Outro fato importante diz respeito aos elementos "não classificáveis" de cada modelo. Embora o percentual observado destes elementos seja menor que 1% em todos os modelos ajustados, observou-se que entre os indivíduos declarados como SD e expostos ao ajuste após o procedimento de recálculo, 13.980 encontravam-se na categoria dos "não classificáveis". Entretanto, tantos os indivíduos declarados SD, como os declarados

como SR, não informaram qualquer valor em algumas das quinze variáveis utilizadas na modelagem, fazendo com que o modelo não se ajustasse para estes casos. Entre os declarados SD, dos 13.980 que se encontravam na categoria dos "não classificáveis" (Quadro 3), 73% não foram ajustados em função da ausência de informação em alguma das variáveis de predição, os demais eram de fato erros de classificação do modelo. Entre os declarados SR, observou-se que 7.948 não foram ajustados pelo mesmo motivo, porém, estes foram mantidos como "Sem Rendimento", contabilizando, portanto, no total de classificáveis como "EP", em ambos os modelos ajustados.

Aplicando o AD para definir se o indivíduo pertencia ou não ao grupo dos extremamente pobres (EP) observa-se, pelas variáveis preditoras e a função discriminante canônica padronizada (Quadro 4), que com exceção das variáveis V3_CARRO, para a qual só se tem informação na PNAD após o ano de 2008 e V7_INTERNET, para a qual só se tem informação na PNAD após o ano de 2001, as demais variáveis apresentaram ajuste sem distorções. Ainda no Quadro 4 é possível observar que o modelo não se ajustou em alguns anos para as variáveis V2_MAQLAVAR, V7_INTERNET e V9_TELFIXO.

Quadro 4 - Coeficientes das Variáveis Utilizadas na Análise de Discriminante do modelo para EP – Brasil, 1992-2014

	1992	1993	1995	1996	1997	1998	1999	2001	2002	2003	2004	2005	2006	2007	2008	2009	2011	2012	2013	2014
V1_CRIANCA	2,228	2,222	2,795	2,723	2,721	3,147	2,993	3,087	3,111	3,477	3,559	3,335	3,238	3,672	3,542	3,82	3,676	2,947	3,929	3.637
V2_MAQLAVAR	0,116	0,104	–	0,086	0,071	0,079	0,09	0,15	0,137	0,137	0,12	0,131	0,119	0,165	0,15	0,194	0,285	0,345	0,429	0.495
V3_CARRO	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0,25	0,313	0,353	0,371	0,368	0.395
V4_GELADEIRA	0,619	0,702	0,62	0,834	0,804	0,762	0,981	0,973	1,028	1,021	0,879	0,939	1,102	1,239	1,263	1,426	1,248	1,099	0,675	1.141
V5_FOGAOBOM	0,933	0,908	0,949	0,944	0,903	0,91	0,63	0,698	0,699	0,579	0,83	0,975	0,875	0,989	0,932	0,736	1,216	1,261	0,721	1.026
V6_AGUA	0,704	0,668	0,672	0,709	0,875	0,745	0,833	0,791	0,674	0,778	0,717	0,653	0,892	0,643	0,889	0,873	1,033	1,767	1,593	1.732
V7_INTERNET	–	–	–	–	–	–	–	-0,079	-0,085	-0,072	–	-0,091	-0,075	–	–	–	0,115	0,145	0,201	0.198
V8_ESCOLARIDADE	0,097	0,104	0,079	0,125	0,078	0,107	0,101	0,153	0,175	0,157	0,122	0,151	0,172	0,183	0,124	0,154	0,073	0,119	0,113	0.117
V9_TELFIXO	–	–	–	–	–	0,051	0,119	0,238	0,265	0,287	0,223	0,277	0,218	0,22	0,207	0,215	0,159	0,152	0,198	0.194
V10_BANHEIRO	0,19	0,184	0,161	0,157	0,173	0,176	0,176	0,147	0,259	0,149	0,129	0,195	0,203	0,137	0,234	0,209	0,434	0,336	0,387	0.320
(Constante)	-0,873	-0,856	-0,683	-0,727	-0,692	-0,698	-0,737	-0,734	-0,742	-0,723	-0,72	-0,704	-0,685	-0,723	-0,809	-0,847	-0,884	-0,903	-0,965	-0.965

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

Ou seja, no processo de estimação dos parâmetros do modelo, em alguma etapa da seleção de variáveis, a variável foi considerada de pouca significância. Em contrapartida, a variável

V1_CRIANCA cujo coeficientes apontavam correlações negativas no Estudo Técnico nº 15/2014, após o novo ajuste, passa a apresentar grande peso de decisão.

Quadro 5 - Coeficientes das Variáveis Utilizadas na Análise de Discriminante do modelo para PO – Brasil, 1992-2014

	1992	1993	1995	1996	1997	1998	1999	2001	2002	2003	2004	2005	2006	2007	2008	2009	2011	2012	2013	2014
V1_CRIANCA	1,903	1,963	2,318	2,315	2,288	2,589	2,47	2,649	2,662	2,789	2,996	2,902	3,393	3,602	3,748	4,146	4,101	4,497	4,527	4.674
V2_MAQLAVAR	0,332	0,362	0,206	0,285	0,258	0,293	0,299	0,379	0,384	0,391	0,408	0,364	0,302	0,346	0,256	0,28	0,341	0,348	0,437	0.502
V3_CARRO	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0,375	0,389	0,341	0,298	0,288	0.301
V4_GELADEIRA	0,813	0,889	0,936	0,955	0,903	0,866	0,864	0,753	0,813	0,779	0,721	0,842	0,914	1,002	0,975	0,997	0,879	0,773	0,654	0.858
V5_FOGAOBOM	0,357	0,293	0,435	0,345	0,341	0,245	0,24	0,196	0,242	0,3	0,411	0,482	0,495	0,611	0,607	0,613	0,893	1,06	0,852	0.945
V6_AGUA	0,552	0,501	0,646	0,674	0,867	0,772	0,761	0,725	0,704	0,653	0,714	0,65	0,83	0,723	0,873	0,828	1,113	1,392	1,437	1.599
V7_INTERNET	–	–	–	–	–	–	–	–	–	–	–	–	–	0,051	–	0,069	0,176	0,179	0,229	0.215
V8_ESCOLARIDADE	0,323	0,34	0,261	0,294	0,28	0,301	0,293	0,365	0,354	0,382	0,316	0,323	0,303	0,296	0,222	0,237	0,174	0,161	0,146	0.133
V9_TELFIXO	0,31	0,284	0,197	0,302	0,281	0,344	0,459	0,578	0,599	0,629	0,56	0,579	0,486	0,45	0,408	0,333	0,252	0,2	0,229	0.215
V10_BANHEIRO	0,404	0,39	0,346	0,364	0,383	0,398	0,411	0,316	0,36	0,321	0,27	0,32	0,242	0,246	0,297	0,274	0,447	0,394	0,371	0.351
(Constante)	-1,462	-1,448	-1,179	-1,23	-1,187	-1,195	-1,214	-1,227	-1,211	-1,229	-1,177	-1,171	-1,05	-1,06	-1,09	-1,092	-1,041	-0,945	-0,987	-0.966

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

O mesmo vale para o modelo AD aplicado para definirmos se o indivíduo pertence ou não ao grupo dos pobres (PO) (Quadro 5), que demonstra que apesar da variável V7_INTERNET ser coletada desde 2001 ela só se tornou relevante a partir de 2007.

Avaliando as principais medidas de acurácia do modelo aplicado no Grupo de Ajuste do Modelo (Amostra de 75%, como já mencionado), apresentadas pelo Quadro 6, observa-se, por exemplo no modelo ajustado para EP, que apesar de uma alta precisão global de acertos, a modelagem apresentou uma baixa sensibilidade. Resultado semelhante se reproduz com relação à especificidade quando o modelo é aplicado no Grupo de Validação (Amostra dos restantes 25%, não usado no ajuste do modelo).

Na busca de uma melhoria desta medida, de forma a conservar um alto valor de sensibilidade e especificidade, aplicou-se o Teste de Kolmogorov-Smirnov com o objetivo de determinar um novo ponto de corte que permitisse um ajuste no qual ambas as medidas fossem maiores.

Quadro 6 - Matriz de Confusão para EP antes do ajuste do corte – Brasil, 2014

Matriz de Confusão		Modelo Empírico		
		Positivo	Negativo	
Modelo Estimado	Positivo	267	1.849	*Valor de Predição Positiva* **12,6%**
	Negativo	1.269	85.561	*Valor de Predição Negativa* **98,6%**
		Sensibilidade **17,6%**	*Especificidade* **97,9%**	*Acurácia* **96,5%**

Quadro 7 - Matriz de Confusão para EP antes do ajuste do corte (Grupo Validação) – Brasil, 2014

Matriz de Confusão		Modelo Empírico		
		Positivo	Negativo	
Modelo Estimado	Positivo	483	14.706	*Valor de Predição Positiva* **3,2%**
	Negativo	38	14.442	*Valor de Predição Negativa* **99,7%**
		Sensibilidade **92,7%**	*Especificidade* **49,5%**	*Acurácia* **50,3%**

Tal como apresentado na seção anterior, o ponto de corte onde o valor do Lift é superior a 1,5 (Quadro 8) é dado pelo percentil 90, onde observa-se um Lift de 1,81. Assim, concluímos também pela matriz de confusão (Quadro 9) que este ponto de corte produz o melhor equilíbrio em sensibilidade e especificidade, ainda que o valor de acurácia de 71% seja menor que o apurado antes do ajuste do ponto de corte. Entretanto, observa-se um significativo aumento na sensibilidade do teste, com ligeira diminuição do valor de predição positivo (VPP).

Quadro 8 - Teste de Kolmogorov-Smirnov do modelo para EP (Grupo Validação) – Brasil, 2014

Decil	MIN	MAX	NALVO	ALVO	% NALVO	%ALVO	POP	%POP	ODDS	RATIO	% NALVO ACUM	%ALVO ACUM	KS	Taxa Alvo	Lift	Lift Acum
10%		0.0015	4,011	2	13.7%	0.4%	4,013	0.4%	0.0	(36.2)	14%	0%	13%	0%	0.03	0.03
20%	0.0015	0.0020	2,490	1	8.5%	0.2%	2,491	0.2%	0.0	(45.0)	22%	1%	22%	0%	0.02	0.03
30%	0.0020	0.0027	2,600	7	8.9%	1.3%	2,607	1.3%	0.1	(6.7)	31%	2%	29%	0%	0.15	0.06
40%	0.0027	0.0036	2,533	8	8.7%	1.5%	2,541	1.5%	0.2	(5.7)	40%	3%	36%	0%	0.18	0.09
50%	0.0036	0.0048	2,820	18	9.7%	3.4%	2,838	3.4%	0.4	(2.8)	50%	7%	43%	1%	0.36	0.14
60%	0.0048	0.0057	3,335	46	11.4%	8.7%	3,381	8.7%	0.8	(1.3)	61%	16%	45%	1%	0.77	0.26
70%	0.0057	0.0088	3,006	38	10.3%	7.2%	3,044	7.2%	0.7	(1.4)	71%	23%	49%	1%	0.70	0.32
80%	0.0088	0.0116	2,192	58	7.5%	11.0%	2,250	11.0%	1.5	1.5	79%	34%	45%	3%	1.45	0.43
90%	0.0116	0.0183	3,645	121	12.5%	23.0%	3,766	23.0%	1.8	1.8	91%	57%	35%	3%	1.81	0.63
100%	0.0183		2,541	228	8.7%	43.3%	2,769	43.3%	5.0	5.0	100%	100%	0%	8%	4.64	1.00
TOTAL			**29,173**	**527**	**100.0%**	**100.0%**	**29,700**	**100.0%**					**49%**	**1.77%**	**1.00**	

Quadro 9 - Matriz de Confusão para EP após o ajuste do corte (Grupo Validação) – Brasil, 2014

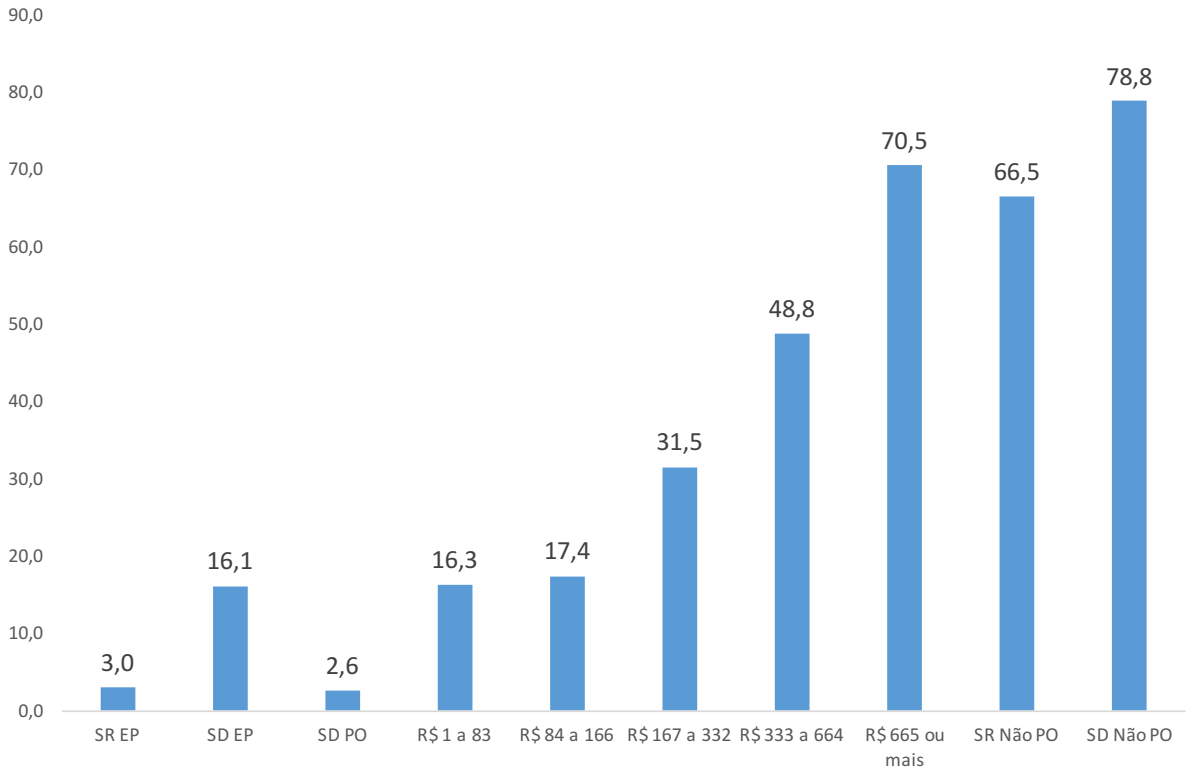
Matriz de Confusão		Modelo Empírico		
		Positivo	Negativo	
Modelo Estimado	Positivo	393	8.330	*Valor de Predição Positiva* **4,5%**
	Negativo	128	20.818	*Valor de Predição Negativa* **99,4%**
		Sensibilidade **75,4%**	*Especificidade* **71,4%**	*Acurácia* **71,5%**

Comparativamente, os novos resultados apresentam um valor mais aceitável em relação ao ponto de corte utilizado pelo modelo original (Quadro 7), quando se avaliava uma baixa especificidade na classificação, corrigindo também a baixa sensibilidade apresentada no Grupo de Ajuste (Quadro 6).

2.2. Validação do modelo

Observando o comportamento do percentual de domicílios com posse de máquina de lavar por faixas de rendimento domiciliar *per capita* (Gráfico 5), tem-se mais uma evidência da importância da aplicação do modelo, visto que o percentual de domicílios com posse de máquina de lavar dentre os SR chegava a 43,4% (Gráfico 4). Após a reclassificação este percentual entre os SR cai para 3,0%, colocando-se abaixo do percentual apresentado pelos extremamente pobres com rendimento declarado. Adequação similar se observa entre os SD. Comportamento semelhante se pode observar com relação a outras variáveis.

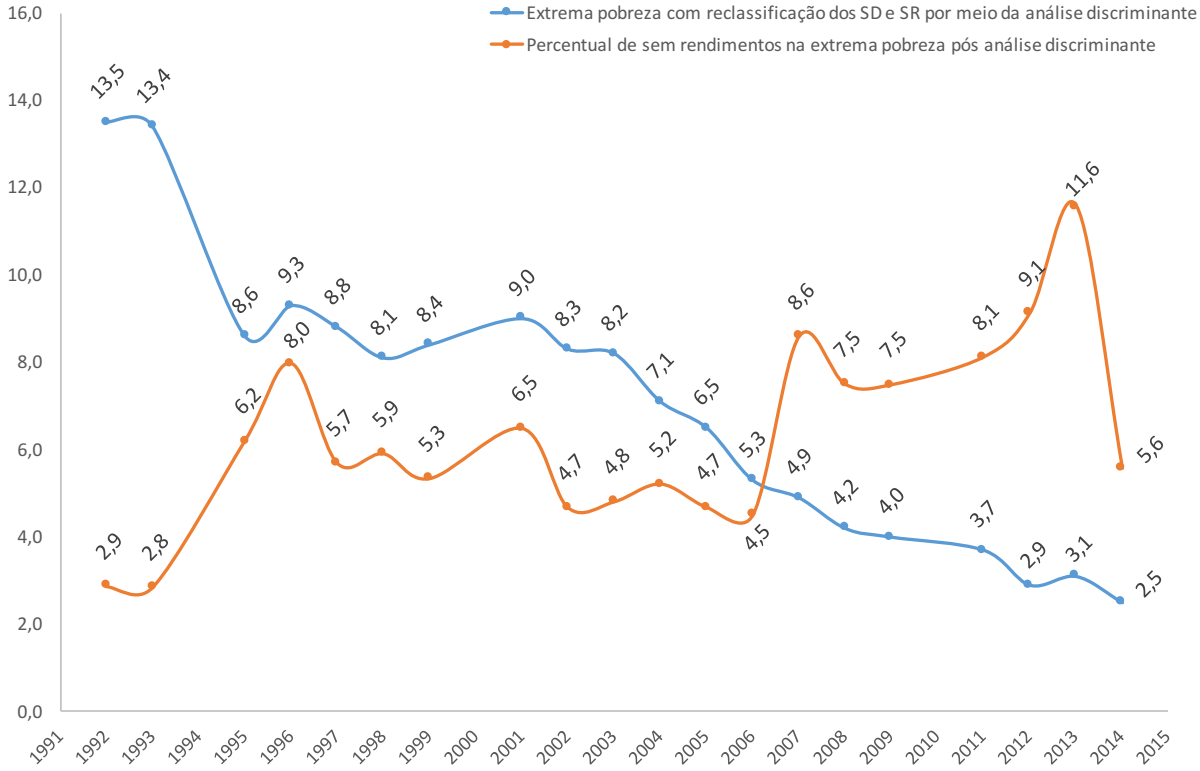
Gráfico 5 - Percentual dos domicílios que possuem máquina de lavar por faixas de rendimento domiciliar *per capita* pós AD – Brasil, 2014



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

Recalculada a série histórica temos que o impacto dos indivíduos sem rendimento domiciliar *per capita*, passou de 32,0% em 2013 para 11,6%, após a aplicação do modelo de análise discriminante. Em 2014, tínhamos 16,5% de SR entre os EP que após AD passaram a representar 5,6% (Gráfico 6).

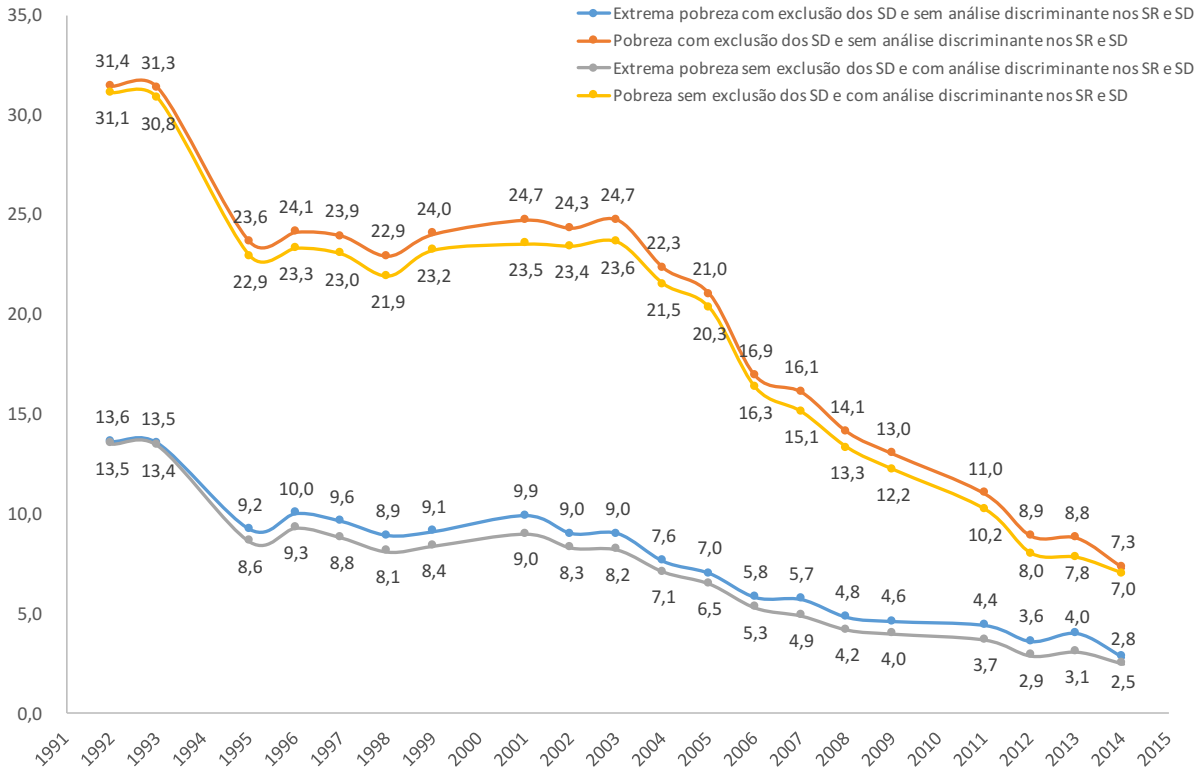
Gráfico 6 - Evolução do percentual dos sem rendimento na extrema pobreza no rendimento domiciliar *per capita* pós AD – Brasil, 1992-2014



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

Com o aprimoramento realizado, os resultados da série das taxas de EP e PO visualizadas lado a lado (Gráfico 7), apresentam outra evidência do ajuste do modelo, atribuindo maior estabilidade e comparabilidade na série.

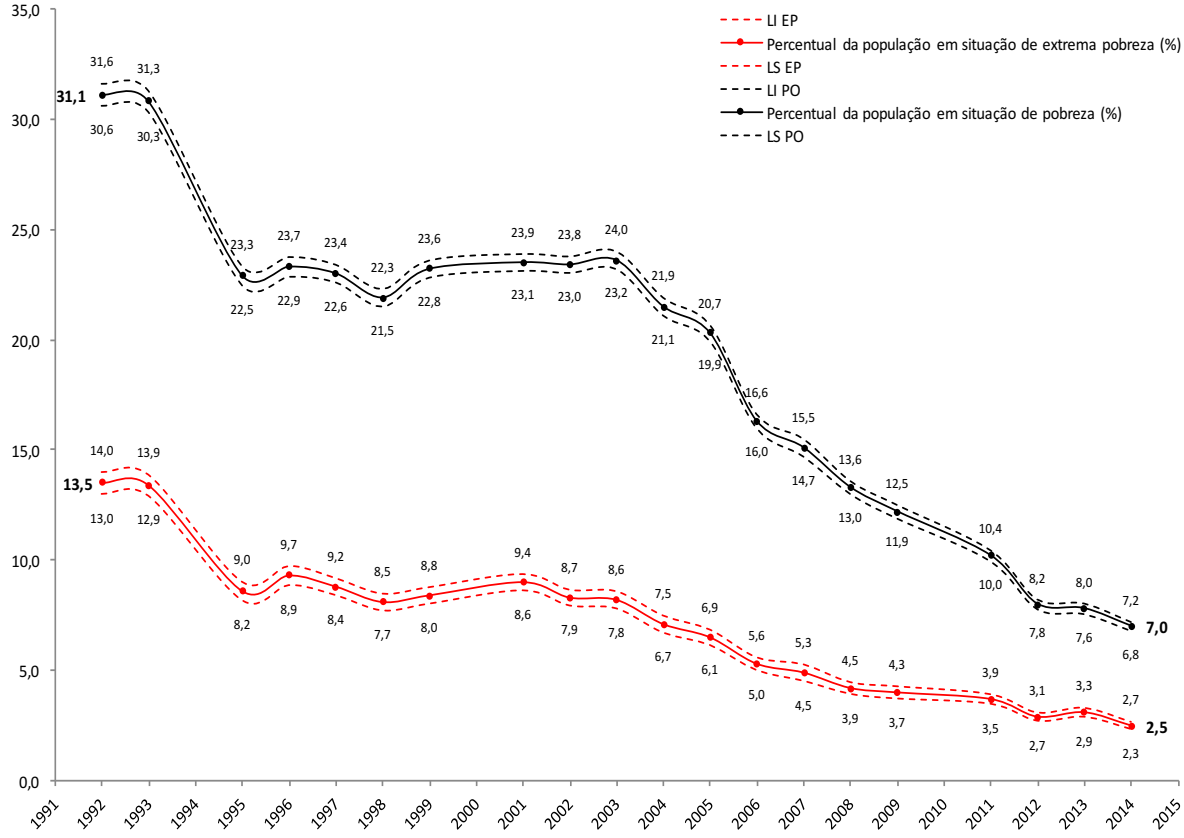
Gráfico 7 - Distribuição percentual da comparação pré e pós AD para EP e PO – Brasil, 1992-2014



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

Utilizando os resultados fornecidos pelo modelo, calculou-se também os intervalos de confiança de 95% (Gráfico 8) considerando-se o desenho complexo da amostra da PNAD (SILVA; PESSOA 2002, vol. 7, n.4, pp. 659-670), segundo o qual observou-se que desde 1995 o resultado do cálculo tradicional de 23,6% para PO encontra-se fora do intervalo construído, sendo este avaliado entre 22,5% e 23,3%. Para EP o resultado original de 9,2% também encontra-se fora do intervalo construído, avaliado entre 8,2% e 9,0%. Este fato é observado até 2014, onde o valor nominal de PO de 7,3% encontra-se fora do intervalo avaliado entre 6,8% e 7,2% e o valor nominal de EP em 2,8%, enquanto o intervalo avaliado estava entre 2,3% e 2,7%. Estes resultados corroboram a hipótese de que a não declaração de rendimentos avaliada conjuntamente sob todos os aspectos apresentados neste e em outros estudos, evidencia a necessidade de aprimoramentos na captação das informações.

Gráfico 8 - Distribuição percentual da comparação pós AD para EP e PO com Intervalos de Confiança de 95% – Brasil,1992-2014



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

3. Considerações Finais

Esforços vêm sendo empreendidos pela equipe técnica do DM/SAGI/MDS às diversas informações pertinentes ao tema de pobreza e extrema pobreza publicado pelos órgãos oficiais brasileiros. Esses esforços visam a obtenção de resultados que possam melhorar o cômputo das estimativas, de modo que os indicadores sociais fiquem cada vez mais próximos da realidade brasileira, propiciando de maneira qualificada, que as políticas públicas de combate e superação da pobreza e extrema pobreza no Brasil atinjam seus objetivos. Neste sentido, se enquadram os aprimoramentos realizados desde 2014, em particular as que se referem à aplicação do modelo AD à série histórica das PNADs de 1992 a 2014.

Referências bibliográficas

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.; LUCENA, F.F.A. **Erro amostral das Taxas de Extrema Pobreza na PNAD: procedimentos e estimativas para Brasil, Estados e Regiões Metropolitanas em 2013.** Estudo Técnico SAGI n. 24/2014.

FONSECA, J.C.G.; LUCENA, F.F.A.; OFUJI, A.I.; FONSECA, M.L.S.; CARVALHO, J.F. **Harmonização dos pesos das PNADs de 1992 a 1999: método e resultados.** Estudo Técnico SAGI n. 04/2016.

HAIR, J. F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. **Análise multivariada de dados**. Tradução: Adonai Schlup Sant'Anna e Anselmo Chaves Neto. 5. Ed. Porte Alegre: Bookman, 2005.

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.

MARTINEZ, E. Z.; LOUZADA-NETO, F. Metodologia estatística para testes diagnósticos e laboratoriais com respostas dicotomizadas. *Revista de Matemática e Estatística*, v. 18, p. 83 – 101, 2000.

SHARMA, S. Applied multivariate techniques. New York: John Wiley & Sons, 1996.

SILVA, Pedro Luis do Nascimento; PESSOA, Djalma Galvão Carneiro e LILA, Maurício Franca. Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. **Ciênc. saúde coletiva [online]**. 2002, vol.7, n.4, pp. 659-670 .

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.