ANÁLISE DE AGRUPAMENTOS APLICADA AO ESTUDO DA POBREZA

Mariana Ferreira Peixoto dos Santos
Escola Nacional de Ciências Estatísticas
maryfps@gmail.com

José André de Moura Brito
Escola Nacional de Ciências Estatísticas
jose.m.brito@ibge.gov.br

Resumo

A pobreza é um problema complexo e multidimensional, com origem no acúmulo das riquezas por poucos. O primeiro e principal Objetivo de Desenvolvimento Sustentável é a extinção da pobreza em todas as suas formas e em todas as partes do mundo. Porém, a primeira grande barreira à implementação dessa meta, pode-se dizer, consiste justamente na estimação inicial do número de pessoas que são pobres. Uma das formas de lidar com essa problemática consiste na utilização de uma metodologia de predição de que uma família esteja em situação de pobreza. Mais especificamente, tratar a mensuração da pobreza de forma multidimensional, considerando mais dimensões além da renda monetária e definição de linha da pobreza. O pressuposto aqui assumido é o de que a utilização de determinado conjunto de variáveis possibilite o agrupamento de domicílios similares**,** de forma a distinguir os grupos e identificá-los em distintos grupos de "pobreza". Sendo assim, este artigo busca aplicar e avaliar os métodos hierárquicos e não hierárquicos de análise de agrupamentos, considerando a identificação e a classificação dos domicílios em distintos grupos de "pobreza" com base em um conjunto específico de variáveis em um estudo de caso com as seguintes unidades da federação: Distrito Federal, Roraima e Acre.

Palavras-Chaves: Pobreza; Análise de Agrupamentos; Multidimensionalidade.

Abstract

Poverty is a complex and multidimensional issue, which has its origin in the accumulation of wealth by few. The first and main objective of the Sustainable Development Goals is the extinction of poverty in all of its forms and all over the world. However, one can say the first big barrier to this goal implementation, is precisely the initial estimation of the number of poor people. One way of handling this issue consists in using a prediction methodology that will determine that a family is in state of poverty. More specifically, to address the poverty measurement in a multidimensional way, considering more dimensions than just the money income and poverty line definition. Our assumption is that by using a set of variables it will allow the grouping of similar households, so we can distinguish groups and identify them in different 'poverty' groups. Thus, this paper aims to apply and evaluate hierarchical and non-hierarchical methods of grouping, considering the identification and classification of households in different 'poverty' groups, based on a specific set of variables in a case study with the following federation units: Distrito Federal, Roraima and Acre.

Keyword: Poverty; Cluster Analysis; Multidimensionality.

1. INTRODUÇÃO

A pobreza é um problema complexo e multidimensional, com origem no acúmulo das riquezas por poucos. Constitui-se, também como um problema de ordem política, em especial de decisão política de combatê-la ou não. Os desafios concernentes à erradicação da pobreza e da fome, à maior equidade na distribuição da renda e ao desenvolvimento de recursos humanos são temas de alta relevância e que têm sido discutidos e considerados em toda parte do mundo, em vários fóruns. Neste sentido, o combate à pobreza é uma responsabilidade conjunta de todos os países e de todos os governantes (BRASIL, 2012).

É importante salientar que o primeiro e principal Objetivo de Desenvolvimento do Milênio é a erradicação da extrema pobreza e da fome no mundo. Para realizar esse Objetivo, a Cúpula do Milênio estabeleceu duas metas a serem alcançadas até 2015, a saber: a primeira é a redução do nível de incidência da pobreza extrema na população mundial à metade do observado em 1990; a segunda é a redução, à metade, da proporção de pessoas que passam fome. Para suceder os Objetivos de Desenvolvimento do Milênio, surgem os Objetivos de Desenvolvimento Sustentável, cujo primeiro objetivo diz respeito à extinção da pobreza em todas as suas formas e em todas as partes do mundo. Porém, a primeira grande barreira à implementação dessa meta, pode-se dizer, consiste justamente na estimação inicial do número de pessoas que são pobres. A medida da pobreza adotada depende da sua conceituação. Se enfocada de forma unidimensional, considera o aspecto renda monetária, via uma linha de pobreza. Se for de forma multidimensional, há índices que contemplam mais dimensões além da renda monetária e definição de linha da pobreza.

Segundo Sen (1976), as medidas de pobreza requerem dois passos: a identificação do grupo de pessoas consideradas pobres e que estão inseridas na população geral, sendo a principal ferramenta a definição da linha de pobreza; e a agregação, onde é construído um índice de pobreza, ou seja, deve-se determinar os indivíduos que são considerados pobres (por exemplo, os que estão abaixo da linha de pobreza), e os que não satisfazem a esse critério.

Para uma mensuração unidimensional no Brasil, utiliza-se informação da renda monetária coletada através de algumas pesquisas que incluem em seus questionários a declaração desse atributo pelo indivíduo. As principais pesquisas e bases de dados sociodemográficos de referência do Brasil, onde essa informação é encontrada, são os Censos Demográficos e as Pesquisas Nacionais por Amostra de Domicílios (PNADs), ambos do Instituto Brasileiro de Geografia e Estatística (IBGE). Porém, a periodicidade de realização e captação nas pesquisas pode gerar um problema concernente, especificamente, à renda dos indivíduos pesquisados.

A captação de dados em períodos específicos do ano pode levar à situação de encontrarmos indivíduos categorizados como "sem rendimento" nas bases de dados, os quais, tomando por base um critério de classificação de pobreza e/ou extrema pobreza baseado apenas na variável renda, seriam assim categorizados - ainda que, na realidade, não o fossem, mas estivessem apenas momentaneamente sem renda. Além disso, não é difícil perceber que uma categorização baseada apenas na renda tende a subestimar o número real de pobres e extremamente pobres, já que os indivíduos passíveis de classificação nesta situação eventualmente poderiam não declarar suas respectivas rendas. Essa situação adquire contornos particularmente dramáticos na perspectiva de planejamento na área de políticas sociais e, especialmente, na estruturação ou fortalecimento de programas para mitigação da extrema pobreza.

Uma das formas de lidar com essa problemática consiste na utilização de uma metodologia de predição de que uma família esteja em situação de pobreza. Mais especificamente, tratar a mensuração da pobreza de forma multidimensional, considerando mais dimensões além da renda monetária e definição de linha da pobreza. Portanto, o pressuposto aqui assumido é o de que a utilização de determinado conjunto de variáveis

possibilite o agrupamento de domicílios similares**,** de forma a distinguir os grupos e identificá-los em distintos grupos de "pobreza".

Desta forma, o objetivo deste artigo é aplicar e avaliar os métodos hierárquicos e não hierárquicos de análise de agrupamentos, considerando a identificação e a classificação dos domicílios em distintos grupos de "pobreza" com base em um conjunto específico de variáveis que foram utilizadas em um estudo de caso com as seguintes unidades da federação: Distrito Federal, Roraima e Acre. Para tratar dessas questões, este artigo está dividido em três partes. A próxima seção apresenta o material e métodos utilizados na realização desse trabalho, incluindo uma revisão teórica dos métodos de agrupamento e uma descrição do planejamento da análise. O objetivo é fornecer ao leitor um ponto de partida para aprofundar seus conhecimentos sobre essa técnica e familiarizá-lo com a terminologia utilizada**,** bem como sistematizar os estágios que devem ser seguidos quanto ao estudo e ao tratamento desta problemática. A segunda seção traz os resultados relativos à aplicação em questão, apresentando os métodos utilizados na análise de agrupamentos. Por fim, a terceira parte apresenta as considerações finais do artigo e seus futuros desdobramentos.

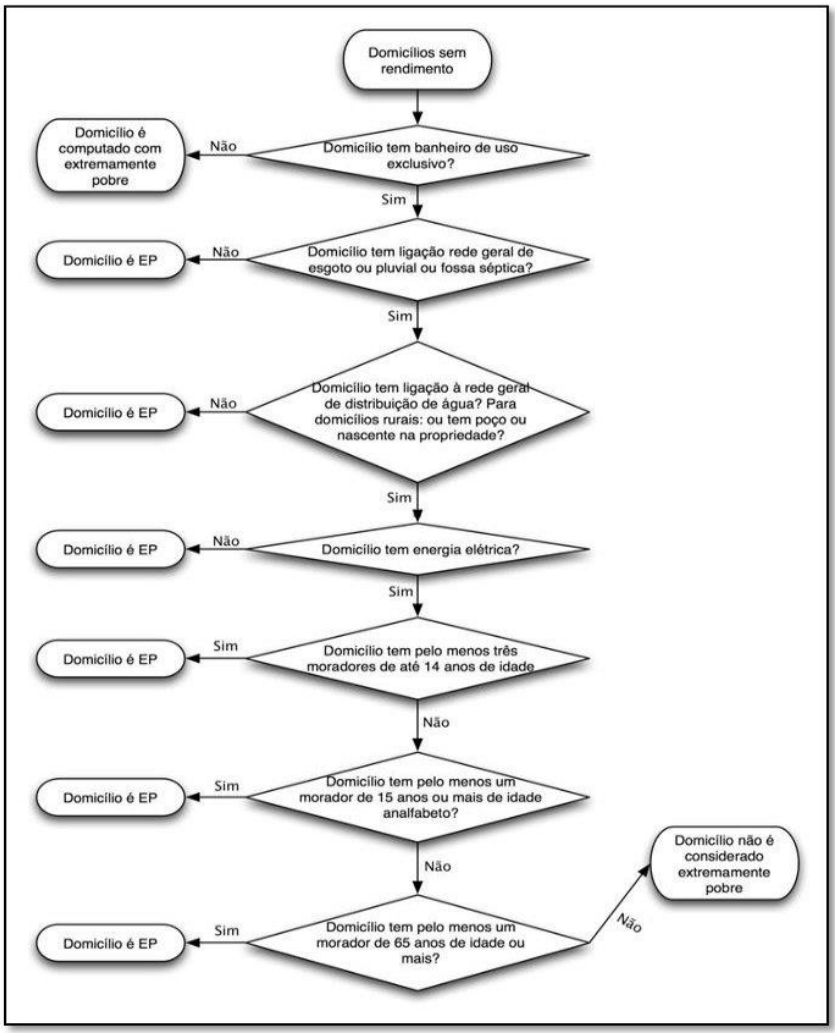
2. MATERIAL E MÉTODOS

Neste trabalho, de forma a estabelecer a classificação dos domicílios em distintos grupos de pobreza, utilizando a técnica de análise de agrupamentos, foi considerado um conjunto de variáveis e indicadores componentes da categoria de extrema pobreza descritos adiante. Mais especificamente, tomou-se por base central de análise os domicílios das unidades da federação do estudo em questão, com renda *per capita* que varia de R$ 0,00 a R$ 140,00, sendo o objetivo final a geração de grupos correlatos a essa faixa de renda, capazes de discriminar os domicílios em diferentes subconjuntos ditos extremamente pobres e pobres e identificar, dentre esses domicílios, a existência de um grupo de domicílios com renda declarada "Sem rendimento", porém com características distintas dos grupos identificados como pobres.

Os indicadores correlatos à faixa de renda, utilizados neste estudo, capazes de discriminar os domicílios em diferentes subconjuntos ditos extremamente pobres**,** foram extraídos dos microdados da amostra do Censo Demográfico 2010 (dados de uso público), com base na definição e aplicação de um conjunto de restrições aos domicílios particulares permanentes[1] sem rendimento definidos pelo IBGE, a partir de discussões técnicas com o Ministério do Desenvolvimento Social e Combate à Fome (MDS). Essa metodologia prevê que, dentre uma série de variáveis, caso um domicílio seja classificado em pelo menos uma das condições estipuladas, será considerado extremamente pobre. A figura a seguir apresenta a carta de variáveis e o fluxo de categorização de extrema pobreza passível de utilização nesse filtro.

[1] Domicílio particular permanente é aquele construído para servir, exclusivamente, à habitação e, na data de referência, tinha a finalidade de servir de moradia a uma ou mais pessoas (IBGE, 2010).

Figura 1 – Variáveis e fluxo de cálculo utilizados no filtro para estimação da extrema pobreza no Censo Demográfico 2010.



As variáveis definidas a partir do conjunto de restrições acima são de natureza nominal dicotômica (1=*sim* ou 0=*não)* e foram utilizadas como base para avaliar a dissimilaridade entre os domicílios considerados neste estudo. Mais especificamente, essas dissimilaridades (distâncias) são utilizadas com dado de entrada para os métodos de agrupamento hierárquico e não hierárquico utilizados neste estudo. Para realizar essas análises foi utilizado o *software* R 3.0.3.

2.1 ANÁLISE DE AGRUPAMENTOS

Análise de Agrupamentos ou, como também denominada, Análise de *Cluster* é uma técnica de análise multivariada que agrega um conjunto de métodos que têm como principal objetivo agrupar objetos a partir de suas características (variáveis ou atributos) (HAIR et al, 2005). Para tanto, devem ser incluídas apenas as variáveis que caracterizem os objetos que serão agrupados e se relacionem especificamente aos objetivos da análise, pois a análise de cluster não diferencia entre variáveis relevantes e irrelevantes, sendo necessário que essa inclusão seja teoricamente orientada. Em linhas gerais, os grupos resultantes da aplicação desses métodos devem exibir um alto grau de homogeneidade interna (*within-cluster*) e alta heterogeneidade externa (*between-cluster*). Formalmente, o problema de agrupamento pode ser definido da seguinte maneira:

Dado um conjunto X formado por n objetos x_i, tal que $X = \{x_1, x_2, \dots, x_i, \dots, x_n\}$, com cada objeto tendo p atributos, ou seja, $x_i = (x_{i1}, \dots, x_{ip})$, deve-se agrupar esses n objetos em k grupos G_1, G_2, ..., G_k, de forma que sejam observadas as seguintes restrições:

i. $G_1 \cup G_2 \cup \dots \cup G_k = X$

ii. (ii) $G_i \cap G_j = \emptyset, \quad i, j = 1, \dots, k (i < j)$

iii. (iii) $|G_i| \geq 1, \quad i = 1, \dots, k$

onde:

• (i) indica que a união dos grupos corresponde ao conjunto *X*.

• (ii) indica que um objeto pertence a exatamente um grupo.

• (iii) garante que cada grupo tem pelo menos um objeto.

Com os objetivos definidos e as variáveis selecionadas, deve-se medir a similaridade entre os objetos. Para distribuir os objetos com características similares nas variáveis em estudo em k grupos, podem ser consideradas uma das seguintes medidas de homogeneidade (distância entre os objetos): Distância Euclidiana, Distância de Manhattan, Distância Euclidiana ponderada, Distância de Minkowski, Distância de Mahalanobis e Distância de Jaccard. Estas distâncias, por sua vez, permitem quantificar o grau de dissimilaridade (d_{ij}) ou similaridade $(s_{ij} = 1 - d_{ij})$ entre dois objetos (i e j) quaisquer, considerando os seus p-atributos. A escolha da medida de distância depende do tipo de variável. Considerando esta última observação, a medida aqui empregada foi determinada pela tipologia das variáveis em estudo (binária assimétrica), ou seja, corresponde à medida de semelhança dada pelo Coeficiente de Jaccard (HAIR et al, 2005).

$$s_{ij} = \frac{a}{a + b + c}$$

em que

a – número de variáveis em que ambos os indivíduos tomam o valor 1;

b - número de variáveis em que *i* toma valor 1 e *j* toma valor 0

c - número de variáveis em que *i* toma valor 0 e *j* toma valor 1.

Uma vez calculada a similaridade, o próximo passo é decidir qual método de agrupamento deve ser aplicado. Os métodos utilizados neste estudo estão em duas categorias gerais: hierárquicos e não hierárquicos. Os procedimentos hierárquicos consistem em um agrupamento em classes que envolve uma combinação (ou divisão) dos objetos em agrupamento. Os dois tipos de procedimentos hierárquicos são, basicamente: os aglomerativos e os divisivos. Nos métodos aglomerativos cada objeto, a priori, é o seu próprio agrupamento. Estes objetos vão sendo combinados em um novo agregado, pelo critério de proximidade adotado, reduzindo, assim, o número de agrupamentos em uma unidade a cada passo, até que todos os objetos sejam reunidos em um grande agregado. Já os métodos divisivos prosseguem na direção oposta dos métodos aglomerativos, começando com um grande agregado contendo todos os objetos e, pela dissimilaridade entre si, esses objetos vão sendo separados e transformados em agrupamentos menores, até que cada objeto seja um agrupamento por si mesmo. O resultado dos dois procedimentos é a construção de uma hierarquia, ou estrutura em árvore (dendograma), que representa a formação dos agrupamentos. Segundo HAIR et al. (2005), uma característica importante dos procedimentos hierárquicos é que os resultados de um estágio anterior são sempre aninhados com os resultados de um estágio posterior, criando algo parecido com uma árvore.

Para cada uma das categorias foi escolhido e aplicado um método de agrupamento. O método hierárquico aglomerativo de Ligação Completa (*Complete Linkage*) foi escolhido para garantir que os objetos de um grupo preservem a máxima distância de outros grupos. Este método gera soluções mais compactas (HAIR et al, 2005). Para a realização da análise, foi definido o número de grupos, de acordo com o critério de parada, observando o gráfico que apresenta os 10 últimos passos sucessivos do processo, com a solução sendo definida quando

há uma súbita elevação, o que significa que causou um substancial decréscimo na similaridade.

Para o método não hierárquico foi escolhido o k-medóides (KAUFMAN; ROUSSEEUW, 1989), por trabalhar com atributos quantitativos e/ou qualitativos. Em uma fase posterior, os grupos produzidos foram validados utilizando o índice de *Silhueta*, com o propósito de verificar em que medida as variáveis de similaridade contribuíram para a formação dos grupos, bem como determinar o valor mais apropriado de algum parâmetro de um algoritmo aplicado correspondendo ao maior valor obtido.

3. RESULTADOS

A distribuição percentual das variáveis elencadas para os domicílios sem rendimento extremamente pobres e não extremamente pobres, assim como, para as demais faixas de renda, especialmente para a faixa de renda domiciliar per capita de R$ 0,01 a R$ 70,00 mensais é uma forma de aferir se o filtro de restrições produziu resultados coerentes.

Tabela 1 - Distribuição percentual dos domicílios particulares permanentes segundo variáveis do filtro de restrições por faixas de renda domiciliar *per capita* – Distrito Federal, 2010.

Variáveis do filtro de restrições		Faixas de renda domiciliar per capita		
		Sem rendimento	R$ 0,01 a R$ 70,00	R$ 70,01 a R$ 140,00
Existência de banheiro de uso exclusivo	Não	2,1%	2,9%	1,9%
	Sim	97,9%	97,1%	98,1%
Existência de rede geral de esgoto ou pluvial ou fossa séptica	Não	13,5%	25,7%	25,2%
	Sim	86,5%	74,3%	74,8%
Existência de ligação à rede geral de distribuição de água ou poço ou nascente na propriedade	Não	3,1%	5,4%	2,9%
	Sim	96,9%	94,6%	97,1%
Existência de energia elétrica	Não	0,4%	0,4%	0,6%
	Sim	99,6%	99,6%	99,4%
Existência de três moradores ou mais de até 14 anos de idade	Não	34,9%	48,6%	49,3%
	Sim	65,1%	51,4%	50,7%
Existência de morador de 15 anos ou mais analfabeto	Não	91,3%	80,1%	79,7%
	Sim	8,7%	19,9%	20,3%
Existência de morador de 65 anos ou mais	Não	92,4%	95,7%	93,8%
	Sim	7,6%	4,3%	6,2%

Fonte: IBGE. Microdados da amostra do Censo Demográfico 2010. Elaboração própria.

Sugere-se, pela Tabela 1, e como já escrito anteriormente, que os domicílios categorizados como "sem rendimento", tomando por base um critério de classificação de pobreza e/ou extrema pobreza fundamentado apenas na variável renda, podem estar assim classificados, porém, não representando a realidade e, sim, um período momentâneo sem renda, na percepção das variáveis selecionadas. Da mesma forma comportam-se os domicílios particulares permanentes dos estados de Roraima e do Acre, conforme suas distribuições apresentadas nas Tabelas 2 e 3.

Tabela 2 - Distribuição percentual dos domicílios particulares permanentes segundo variáveis do filtro de restrições por faixas de renda domiciliar *per capita* – Roraima, 2010.

Variáveis do filtro de restrições		Faixas de renda domiciliar per capita		
		Sem rendimento	R$ 0,01 a R$ 70,00	R$ 70,01 a R$ 140,00
Existência de banheiro de uso exclusivo	Não	51,4%	46,6%	27,7%
	Sim	48,6%	53,4%	72,3%
Existência de rede geral de esgoto ou pluvial ou fossa séptica	Não	75,0%	78,4%	72,3%
	Sim	25,0%	21,6%	27,7%
Existência de ligação à rede geral de distribuição de água ou poço ou nascente na propriedade	Não	38,0%	21,4%	14,1%
	Sim	62,0%	78,6%	85,9%
Existência de energia elétrica	Não	43,5%	29,8%	13,7%
	Sim	56,5%	70,2%	86,3%
Existência de três moradores ou mais de até 14 anos de idade	Não	76,3%	48,3%	51,9%
	Sim	23,7%	51,7%	48,1%
Existência de morador de 15 anos ou mais analfabeto	Não	57,9%	68,4%	71,4%
	Sim	42,1%	31,6%	28,6%
Existência de morador de 65 anos ou mais	Não	91,8%	96,3%	91,7%
	Sim	8,2%	3,7%	8,3%

Fonte: IBGE. Microdados da amostra do Censo Demográfico 2010. Elaboração própria.

Tabela 3 - Distribuição percentual dos domicílios particulares permanentes segundo variáveis do filtro de restrições por faixas de renda domiciliar *per capita* – Acre, 2010.

Variáveis do filtro de restrições		Faixas de renda domiciliar per capita		
		Sem rendimento	R$ 0,01 a R$ 70,00	R$ 70,01 a R$ 140,00
Existência de banheiro de uso exclusivo	Não	62,1%	77,0%	61,3%
	Sim	37,9%	23,0%	38,7%
Existência de rede geral de esgoto ou pluvial ou fossa séptica	Não	78,6%	90,8%	83,1%
	Sim	21,4%	9,2%	16,9%
Existência de ligação à rede geral de distribuição de água ou poço ou nascente na propriedade	Não	42,2%	46,1%	36,4%
	Sim	57,8%	53,9%	63,6%
Existência de energia elétrica	Não	25,9%	29,1%	11,9%
	Sim	74,1%	70,9%	88,1%
Existência de três moradores ou mais de até 14 anos de idade	Não	80,5%	44,0%	54,3%
	Sim	19,5%	56,0%	45,7%
Existência de morador de 15 anos ou mais analfabeto	Não	63,1%	44,7%	56,0%
	Sim	36,9%	55,3%	44,0%
Existência de morador de 65 anos ou mais	Não	94,1%	97,7%	93,6%
	Sim	5,9%	2,3%	6,4%

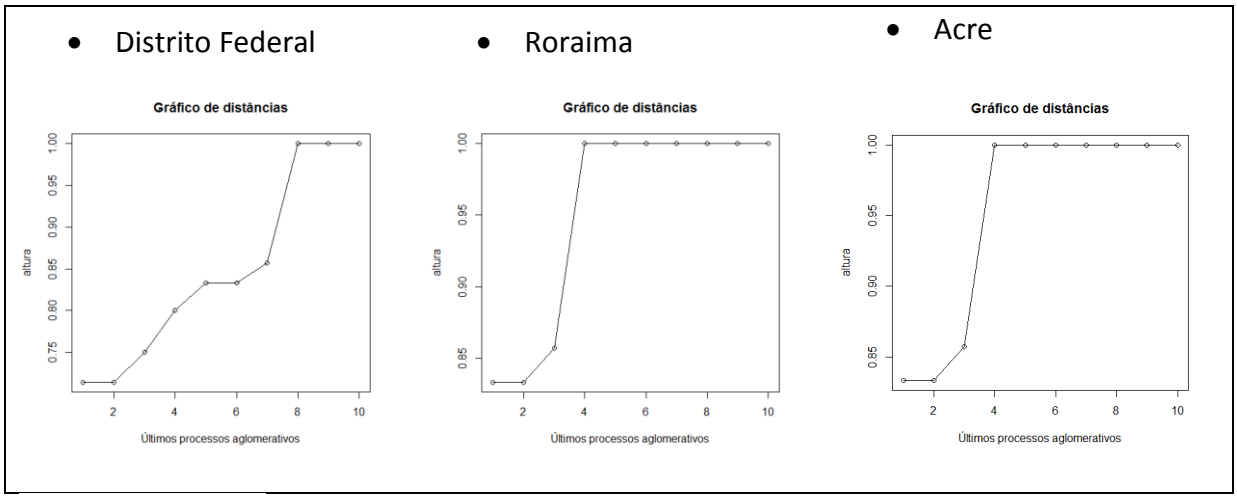
Fonte: IBGE. Microdados da amostra do Censo Demográfico 2010. Elaboração própria.

Para a definição dos agrupamentos com base nas variáveis do filtro, foram utilizados algoritmos em duas categorias gerais: hierárquicos e não hierárquicos. Serão descritos abaixo os resultados para cada método.

3.1 APLICAÇÃO DO MÉTODO HIERÁRQUICO

Para o método hierárquico, foi utilizado o método aglomerativo de Ligação Completa (Complete Linkage). Esse método utiliza a matriz de dissimilaridades D, mais especificamente, busca nessa matriz os dois objetos "menos semelhantes" entre si (com a maior dissimilaridade). Todos os objetos do grupo são ligados a qualquer outro a certa distância máxima ou por similaridade mínima. Isso é feito até todos os objetos pertencerem um único grupo. Ou seja, no caso deste trabalho, os domicílios são alocados a um único grupo inicial, mediante aplicação deste método. Para realizar uma separação em grupos hierárquicos é necessário fazer cortes, porém não existe um procedimento-padrão objetivo de seleção. Neste estudo, foi utilizado o critério de parada, observando a Figura 2, que apresenta os gráficos dos 10 últimos passos sucessivos do processo de agrupamento para o Distrito Federal e os estados de Roraima e do Acre.

Figura 2 – Gráficos das distâncias dos últimos 10 processos aglomerativos para o Distrito Federal e os estados de Roraima e Acre.



Fonte: Autoria própria.

A solução é definida quando há uma súbita elevação, ou seja, que foi observado um substancial decréscimo na similaridade. Quando um grande incremento ocorre, é selecionada a solução anterior baseado na lógica de que a última combinação causou o substancial decréscimo na similaridade que, neste caso, leva a definição de 4, 8 e 8 grupos, respectivamente. A seguir, na Tabela 4, são apresentadas as frequências relativas de domicílios por grupo para cada atributo e o quantitativo de domicílios alocados em cada grupo definido para as análises do Distrito Federal.

Tabela 4 - Distribuição percentual dos domicílios particulares permanentes por grupos resultantes da análise de conglomerados segundo variáveis do filtro de restrições – Distrito Federal.

Variáveis do filtro de restrições		Grupos			
		1	2	3	4
TOTAL DE DOMICÍLIOS		**966**	**1.047**	**24**	**293**
Existência de banheiro de uso exclusivo	Não	0,0%	1,9%	100,0%	2,0%
Existência de rede geral de esgoto ou pluvial ou fossa séptica	Não	3,6%	33,4%	37,5%	19,5%
Existência de ligação à rede geral de distribuição de água ou poço ou nascente na propriedade	Não	5,2%	0,9%	29,2%	3,4%
Existência de energia elétrica	Não	0,2%	0,0%	8,3%	2,4%
Existência de três moradores ou mais de até 14 anos de idade	Sim	98,2%	16,7%	100,0%	70,0%
Existência de morador de 15 anos ou mais analfabeto	Sim	1,0%	2,7%	16,7%	100,0%
Existência de morador de 65 anos ou mais	Sim	7,9%	2,7%	4,2%	17,4%

Fonte: Autoria própria

Pode-se sugerir, para os grupos que apresentam percentual baixo para os atributos, que eles não fazem parte dos domicílios com renda per capita que varia de R$ 0,00 a R$ 140,00, pois eles quase não apresentam características indicadoras de pobreza. Levanta-se a questão se esses são os domicílios com renda declarada "sem rendimento", mas que se encaixam nos casos, por exemplo, em que a coleta da pesquisa foi realizada somente em período específico do ano, não captando, assim, a verdadeira renda, como é o caso do grupo 1, com 966 domicílios dentre os 2.330 analisados. Análise semelhante foi feita para os estados de Roraima e Acre. Os resultados são apresentados a seguir nas Tabelas 5 e 6.

Tabela 5 - Distribuição percentual dos domicílios particulares permanentes por grupos resultantes da análise de conglomerados segundo variáveis do filtro de restrições – Roraima.

Variáveis do filtro de restrições		Grupos							
		1	2	3	4	5	6	7	8
TOTAL DE DOMICÍLIOS		**645**	**582**	**728**	**1.209**	**204**	**602**	**231**	**492**
Existência de banheiro de uso exclusivo	Não	0,0%	0,0%	100,0%	88,6%	44,1%	0,0%	47,2%	4,7%
Existência de rede geral de esgoto ou pluvial ou fossa séptica	Não	61,7%	0,0%	94,4%	97,7%	70,1%	100,0%	86,1%	65,2%
Existência de ligação à rede geral de distribuição de água ou poço ou nascente na propriedade	Não	4,3%	0,0%	3,3%	70,1%	22,1%	0,0%	100,0%	10,6%
Existência de energia elétrica	Não	7,4%	0,0%	12,8%	100,0%	35,8%	0,0%	3,5%	2,6%
Existência de três moradores ou mais de até 14 anos de idade	Sim	100,0%	0,0%	52,2%	46,8%	4,9%	0,0%	38,1%	29,5%
Existência de morador de 15 anos ou mais analfabeto	Sim	5,9%	0,0%	27,9%	63,4%	66,7%	0,0%	3,9%	100,0%
Existência de morador de 65 anos ou mais	Sim	6,0%	0,0%	2,9%	3,1%	100,0%	0,0%	9,5%	0,0%

Fonte: Autoria própria

Tabela 6 - Distribuição percentual dos domicílios particulares permanentes por grupos resultantes da análise de conglomerados segundo variáveis do filtro de restrições – Acre.

Variáveis do filtro de restrições		Grupos							
		1	2	3	4	5	6	7	8
TOTAL DE DOMICÍLIOS		**748**	**701**	**1.186**	**1.034**	**370**	**2.262**	**377**	**179**
Existência de banheiro de uso exclusivo	Não	13,8%	93,3%	66,8%	100,0%	60,3%	71,4%	0,0%	59,2%
Existência de rede geral de esgoto ou pluvial ou fossa séptica	Não	100,0%	97,0%	87,8%	95,0%	64,3%	85,9%	0,0%	63,1%
Existência de ligação à rede geral de distribuição de água ou poço ou nascente na propriedade	Não	3,7%	58,8%	0,0%	0,0%	0,0%	100,0%	0,0%	65,4%
Existência de energia elétrica	Não	1,7%	100,0%	0,0%	0,0%	65,7%	20,2%	0,0%	34,1%
Existência de três moradores ou mais de até 14 anos de idade	Sim	29,8%	0,0%	48,0%	34,4%	100,0%	51,1%	0,0%	31,3%
Existência de morador de 15 anos ou mais analfabeto	Sim	14,6%	55,8%	100,0%	0,0%	48,6%	47,0%	0,0%	80,4%
Existência de morador de 65 anos ou mais	Sim	21,7%	0,0%	0,0%	0,0%	1,6%	0,0%	0,0%	100,0%

Fonte: Autoria própria

Os grupos que apresentam percentual baixo para os atributos, que afiguram não fazer parte dos domicílios com renda per capita que varia de R$ 0,00 a R$ 140,00, por não apresentarem características indicadoras de pobreza no estado de Roraima são representados pelo grupo 6 e mais especificamente pelo grupo 2, onde totalizam 1.184 domicílios dos 4.693 analisados. E o grupo que afigura ter mais características de pobreza, de acordo com as variáveis para estimação da extrema pobreza acordadas pelo IBGE e MDS, é o 4. Portanto, diante deste conjunto de restrições, os domicílios alocados ao grupo 4 são identificados como extremamente pobres. No caso do estado do Acre, o grupo que se destaca por não conter as características de pobreza descritas é o grupo 7 com 377 domicílios dos 6.857 analisados.

Com um cruzamento dos domicílios entre os grupos alocados e a faixa de renda declarada, podem-se confirmar as constatações dos resultados analisados no que se refere aos grupos de domicílios com percentual baixo para os atributos indicadores de extrema pobreza. As Tabelas 7, 8 e 9 apresentam estes cruzamentos para o Distrito Federal e os estados de Roraima e Acre, respectivamente.

Tabela 7 - Distribuição dos domicílios particulares permanentes por faixa de rendimento domiciliar per capita segundo grupos resultantes da análise de conglomerados – Distrito Federal.

Grupos	Quantidade de domicílios			
	Faixa de rendimento domiciliar per capita em julho de 2010			
	Renda 0	De 1 a 70	De 71 a 140	Total
1	614	85	267	966
2	458	139	450	1.047
3	11	6	7	24
4	95	46	152	293
Total	1.178	276	876	2.330

Fonte: Autoria própria**.**

Tabela 8 - Distribuição dos domicílios particulares permanentes por faixa de rendimento domiciliar per capita segundo grupos resultantes da análise de conglomerados – Roraima.

Grupos	Quantidade de domicílios			
	Faixa de rendimento domiciliar per capita em julho de 2010			
	Renda 0	De 1 a 70	De 71 a 140	Total
1	67	230	348	645
2	328	127	127	582
3	203	308	217	728
4	756	319	134	1.209
5	121	19	64	204
6	238	164	200	602
7	86	67	78	231
8	139	154	199	492
Total	1.938	1.388	1.367	4.693

Fonte: Autoria própria.

Tabela 9 - Distribuição dos domicílios particulares permanentes por faixa de rendimento domiciliar per capita segundo grupos resultantes da análise de conglomerados – Acre.

Grupos	Quantidade de domicílios			
	Faixa de rendimento domiciliar per capita em julho de 2010			
	Renda 0	De 1 a 70	De 71 a 140	Total
1	263	141	344	748
2	375	188	138	701
3	253	379	554	1.186
4	302	285	447	1.034
5	109	130	131	370
6	707	751	804	2.262
7	226	51	100	377
8	78	25	76	179
Total	2.313	1.950	2.594	6.857

Fonte: Autoria própria.

Nota-se que, para os grupos citados: 1 (do Distrito Federal), 2 e 6 (do estado de Roraima) e 7 (do estado do Acre), que têm características semelhantes ao não conter indicadores de pobreza, a maioria dos domicílios tiveram declaração "sem rendimento" na pesquisa, poucos com declaração na faixa de renda de R$ 1,00 a R$ 70,00 e um aumento no quantitativo de domicílios com declaração na faixa de R$ 71,00 a R$ 140,00 (exceto pelo grupo 2 do estado de Roraima, onde os quantitativos das faixas de renda de R$ 1,00 a R$ 70,00 e de R$ 71,00 a R$ 140,00 são iguais). Essa análise possibilita a confirmação citada anteriormente de que a renda declarada "sem rendimento" dos domicílios alocados nesses grupos não seja a verdadeira renda, justamente por eles quase não apresentarem características indicadoras de pobreza.

3.2 APLICAÇÃO DO MÉTODO NÃO HIERÁRQUICO

Diferentemente dos métodos hierárquicos, os métodos não hierárquicos trabalham com a quantidade de grupos definida a priori. Em particular, neste trabalho, o método não

hierárquico utilizado foi dos k-medóides. Esse método trabalha com atributos quantitativos e/ou qualitativos e com o conceito de objeto representativo. Esse objeto é chamado de medóide e tem a melhor localização central em relação aos demais objetos (KAUFMAN; ROUSSEEUW, 1989).

Definiu-se, inicialmente, o quantitativo de grupos de domicílios com base no número de grupos produzido pelo método hierárquico utilizado e descrito na seção anterior. Uma vez definidos os grupos a partir da aplicação do método dos k-medóides, utilizou-se o índice Silhueta, que está associado a um critério de validação relativo (NALDI, 2011). Esse índice mede a qualidade dos grupos com base na distância entre os objetos de um grupo e na distância dos objetos de um grupo ao grupo mais próximo. Em geral, calcula-se a média das silhuetas (n objetos) a avalia-se a qualidade do agrupamento a partir dessa média. Esse critério de validação foi utilizado não só para validar a quantidade de grupos pré-definidos, como para indicar o melhor quantitativo de grupos definidos para esse conjunto de dados. Por isso, foi calculado o índice de Silhueta não só para o valor pré-definido do número de grupos, como para valores próximos a ele. O índice ao maior valor indicou a melhor quantidade de grupos para alocação dos domicílios desse conjunto de dados.

Para os dados do Distrito Federal, o quantitativo de grupos foi mantido e obteve a distribuição de 1.095, 785, 258 e 192 domicílios para os grupos 1, 2, 3 e 4, respectivamente. O resultado para o índice Silhueta foi de 0,7012, indicando que foi encontrada uma estrutura substancial, isto é, uma solução de boa qualidade. No caso do estado de Roraima, o índice Silhueta foi calculado para 8 grupos, a priori, e, posteriormente, para 7 e 6 grupos. O valor do índice aumenta quando calculado para 7 grupos e diminui quando calculado para 6, optando-se por adotar, neste critério de agrupamento, 7 grupos distintos de pobreza. Sendo assim, foi obtida a seguinte distribuição: 576, 582, 626, 849, 593, 738, 423 e 306 domicílios para os grupos 1, 2, 3, 4, 5, 6 e 7, respectivamente. Para o estado do Acre, foi observada uma melhora significativa no valor do índice Silhueta à medida que aumentava o número de grupos. Foram testados de 6 a 10 grupos. Optou-se, arbitrariamente, por adotar os 8 grupos definidos na aplicação do modelo hierárquico, pois notou-se um critério de agrupamentos já bem definido. A distribuição obtida dos grupos 1, 2, 3, 4, 5, 6, 7 e 8 foi, respectivamente de 438, 1.144, 942, 611, 795, 1.614, 936 e 377 domicílios.

4. CONSIDERAÇÕES FINAIS

O estudo dos domicílios extremamente pobres lançam desafios de ordem metodológica, conceitual e teórica na busca da compreensão dos aspectos não relacionados estritamente à renda e capazes de diferenciar perfis socioeconômicos nas pesquisas domiciliares. Este estudo teve como finalidade aplicar os métodos hierárquicos e não hierárquicos da análise de agrupamentos para tentar classificar domicílios em distintos grupos de "pobreza" com base em um conjunto específico de variáveis.

Verificou-se que as variáveis de similaridade consideradas contribuíram para uma boa formação dos grupos. Elas proveem um significado que avalia a correspondência dos resultados com os propostos anteriormente pela teoria. Entretanto, os resultados obtidos, por serem limitados e restritos a poucas variáveis, não eliminam a necessidade de, em futuros estudos, ampliar o conjunto de variáveis, expandir o estudo para todas as unidades da federação e testar outros métodos de agrupamento. Acredita-se, assim, que este trabalho é apenas o início de uma linha de estudos necessários para o aperfeiçoamento continuado da pesquisa sobre estimação de pobreza.

5. REFERÊNCIAS BIBLIOGRÁFICAS

[1] BRASIL. Plano Brasil Sem Miséria. Disponível em: <http://www.brasilsemmiseria.gov.br/noticias/ultimos-artigos/2012/fevereiro/para-combater-a-pobreza-e-preciso-mobilizacao>. Acesso em: 04 fev. 2015.

[2] BRITES, M. et al. Medida Multidimensional de Pobreza: Um estudo da importância das dimensões. 1o Seminário de Jovens Pesquisadores em Economia e Desenvolvimento, p. 1–23, out. 2013.

[3] BRITO, J. A. de M. et al. Um algoritmo para o agrupamento baseado em K-Medoids. Revista brasileira de estatística, v. 71, n. 234, p. 75-100, 2010.

[4] CORRAR, L. J.; PAULO, E.; FILHO, J. M. D. Análise Multivariada para os Cursos de Administração, Ciências Contábeis e Economia. São Paulo: Editora Atlas S.A., 2014.

[5] HAIR, J. F. et al. Análise Multivariada de Dados. 5a edição ed.Porto Alegre: Bookman, 2005. p. 593

[6] KAUFMAN L.; ROUSSEEUW P.J. Finding Groups in Data – An Introduction to Cluster Analysis. Wiley-Interscience Publication, 1989.

[7] MINGOTI, S. A. Análise de Dados através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada. Belo Horizonte: Editora UFMG, 2005.

[8] MUNICIPAIS, IBGE Indicadores Sociais. uma análise dos resultados do universo do Censo Demográfico 2010. Estudos & Pesquisas: informações demográfica e socioeconômica, n. 28, 2011.

[9] NALDI, C. N. Técnicas de Combinação para Agrupamento Centralizado e Distribuído de Dados. Tese de Doutorado, USP - São Carlos, 2011.

[10] SEN, Amartya K. Poverty as a ordinal approach to measurement. Econometrica, v. 44, mar. 1976.

[11] SOUSA, M. F. de; SANTOS, J. R. S. Análise do filtro de restrições aplicados aos sem remuneração nos dados preliminares do Universo do Censo Demográfico 2010. Estudo técnico, Nº 16. Brasília, 2012.