3
Método

Este estudo se caracteriza como sendo de natureza descritiva, baseado em dados secundários, provenientes das duas últimas edições da Pesquisa de Orçamentos Familiares do IBGE, realizadas nos anos de 2002/2003 e 2008/2009.

3.1. Pesquisa de Orçamentos Familiares (POF)

A Pesquisa de Orçamentos Familiares é uma das mais abrangentes pesquisas realizadas pelo IBGE, que visa principalmente "...mensurar as estruturas de consumo, dos gastos, dos rendimentos e parte da variação patrimonial das famílias" (POF, 2010 p. 17). Através dela, é possível traçar um perfil das condições de vida da população brasileira, com base na análise de seus orçamentos familiares. A POF é uma pesquisa domiciliar, feita por amostragem, e tem como unidade básica de pesquisa a unidade de consumo, identificada dentro dos domicílios particulares permanentes investigados. Entende-se como unidade de consumo "...um único morador ou um conjunto de moradores que compartilham da mesma fonte de alimentação ou compartilham as despesas com moradia..." (POF, 2010 p. 19). Neste estudo, os termos "família" e "domicílio" são utilizados para representar o conceito de unidade de consumo.

As variáveis investigadas englobam: características do domicílio, características das pessoas, despesas, rendimentos e condição de vida. As despesas, tópico principal da POF, são divididas em monetárias e não monetárias, e abertas nas seguintes categorias: alimentação, habitação, vestuário, transporte, higiene e cuidados pessoais, assistência a saúde, educação, recreação e cultura, fumo, serviços pessoais e despesas diversas.

O período de coleta dos dados foi de 12 meses. O período de referência das informações de despesas variou conforme a frequência de aquisição e o valor gasto: 7 dias, 30 dias, 90 dias e 12 meses. Para os rendimentos, o período de referência é de 12 meses. Em função do longo período de coleta e dos

diferentes períodos de referência das informações, foi estabelecida uma data de referência para a compilação, análise e apresentação dos resultados. No caso da POF 2002/2003, essa data é 15 de janeiro de 2003, e para a POF 2008/2009, é 15 de janeiro de 2009.

A POF 2002/2003 e a POF 2008/2009 investigaram, respectivamente, 48.470 e 55.970 domicílios em todo o Brasil.

3.2. Amostra

Como o objetivo desse estudo é analisar os orçamentos de consumidores de baixa renda, utilizou-se a renda familiar como critério para delimitação desse segmento. Serão consideradas como de baixa renda famílias com rendimento mensal entre 1 e 3 salários mínimos. Esse critério foi definido com base na literatura (GROSSI, MOTTA E HOR-MEYLL, 2008; SILVA E PARENTE, 2007; PONCHIO E ARANHA, 2009) e na análise da distribuição da renda das famílias pesquisadas, que mostra que nesta faixa não há grande diferença nos rendimentos, que apresentam aumento suave. Para que a renda não influencie na composição do orçamento familiar, decidiu-se por usar uma faixa de renda estreita.

Para o período de 2002/2003, o salário mínimo considerado é de R$240,00 (valor atualizado para abril de 2003) e, para o período de 2008/2009, o valor é de R$465,00 (valor atualizado para fevereiro de 2009). Assim, as faixas de renda consideradas nos dois períodos são R$240,00 a R$720,00 e R$465,00 a R$1.395, respectivamente.

O conceito de renda bruta total segue a definição da POF (2010). O rendimento familiar total corresponde aos rendimentos monetários somados aos rendimentos não monetários de todos os componentes da família. Os rendimentos monetários compreendem todos os tipos de ganhos monetários, que são: rendimento do trabalho (rendimento do empregado, rendimento do empregador e conta-própria), transferências (aposentadoria, pensão, programas sociais, pensão alimentícia, mesada, doação), rendimento de aluguel e outras rendas provenientes de vendas eventuais. Os rendimentos não monetários incluem tudo que é obtido através de doação, retirada do negócio, troca, produção própria, pesca e caça.

Serão consideradas, neste estudo, apenas as despesas monetárias, aquelas feitas através de pagamento, à vista ou a prazo. Despesas não

monetárias não serão consideradas, principalmente porque incluem o aluguel estimado (para famílias que residem em imóveis cuja condição de ocupação é diferente de alugado), que faz com que os gastos com habitação sejam inflacionados.

Serão consideradas apenas famílias que residem na Região Metropolitana do Rio de Janeiro, para diminuir o impacto do fator geográfico no orçamento e, ao mesmo tempo, permitir uma análise mais ampla da baixa renda.

Para chegar à amostra final, foram excluídas famílias que apresentavam *outliers* ou valores faltantes nos dados. A amostra final analisada neste estudo foi de 187 famílias na POF 2002/2003 e 355 famílias na POF 2008/2009, que, ao serem expandidas, representam 800.833 e 1.203.864 famílias, respectivamente. É importante ressaltar que esta amostra é representativa para a Região Metropolitana do Rio de Janeiro, o que significa dizer que os resultados encontrados neste estudo poderão ser generalizados para toda a população representada.

3.3. Análise de Cluster

A análise de cluster, também chamada de análise de agrupamentos ou de conglomerados, é uma técnica da análise multivariada, cujo objetivo é identificar grupos naturais de objetos, com base na similaridade de algumas de suas características, tendo esses grupos alta homogeneidade interna e alta heterogeneidade externa (HAIR *et al*, 2005). Diferentemente dos métodos de classificação, na análise de cluster não há nenhum pressuposto sobre o número de grupos e a estrutura de cada grupo, cabendo ao próprio pesquisador o papel de identificar se o agrupamento é bom ou ruim (JOHNSON e WICHERN, 2007).

Ao realizar um agrupamento, o pesquisador pode ter dois objetivos: exploratório, quando não há nenhum pressuposto sobre os grupos a serem formados; ou confirmatório, quando se quer confirmar uma relação já identificada entre os objetos (HAIR *et al*, 2005).

A seleção das variáveis que serão incluídas na formação dos clusters deve ser cuidadosa. Deve-se inserir somente as variáveis baseadas em algum argumento, pois nesta técnica não é possível identificar as variáveis que são irrelevantes na criação dos grupos, mas que influenciam o resultado final. Além disso, é recomendável que se faça padronização das variáveis, para que se elimine o efeito de escala.

Existem várias formas de padronização das variáveis, sendo a mais comum a padronização por *z-scores*. Esta forma de padronização é feita subtraindo-se de cada variável a média e dividindo-se pelo desvio padrão (HAIR *et al*, 2005).

Um ponto importante na análise de cluster é a escolha da medida de similaridade que será utilizada no agrupamento. Três aspectos devem ser considerados na escolha da medida: a natureza das variáveis, a escala de medida e o conhecimento sobre o problema (JOHNSON e WICHERN, 2007). As duas medidas de similaridade mais usadas na análise de agrupamentos são medidas de distância e medidas de associação (HAIR *et al*, 2005).

Medidas de distância

As medidas de distância requerem dados métricos e é o método mais usado para medir a similaridade. Maiores distâncias significam menor similaridade e, portanto é uma medida de dissimilaridade (HAIR *et al*, 2005). A medida de distância mais usada é a distância euclidiana:

- **Distância Euclidiana**

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Medidas de associação

As medidas de associação são usadas quando os dados são não métricos. Para calcular a medida de similaridade, é introduzida uma variável binária, que representa a presença ou ausência de uma certa característica. Após isso, é escolhido o coeficiente de similaridade que será usado para a construção da matriz de similaridade (JOHNSON e WICHERN, 2007).

Como não é viável examinar todos os grupos possíveis que podem ser formados a partir de um banco de dados, foram desenvolvidos alguns algoritmos com o objetivo de agilizar o processo de agrupamento. Os algoritmos foram separados em dois tipos: algoritmos hierárquicos e algoritmos não hierárquicos (JOHNSON e WICHERN, 2007).

3.3.1. Método Hierárquico

Os algoritmos hierárquicos são divididos em aglomerativos e divisivos.

No método aglomerativo, considera-se, inicialmente, cada objeto distinto de todos os outros, ou seja, o número de clusters inicial é igual ao número de objetos. Em seguida, os objetos que tem menor distância, ou maior semelhança, vão sendo agrupados gradativamente, até ser formado apenas um grupo que contenha todos os objetos.

De forma contrária, o método divisivo inicialmente considera todos os objetos em um só grupo. A partir daí, os objetos são separados em subgrupos, de forma que os elementos de um grupo estejam distantes dos de outros grupos, ou seja, que sejam menos semelhantes. O processo termina quando o número de grupos for igual ao número de objetos (JOHNSON e WICHERN, 2007; HAIR *et al*, 2005).

Um enfoque maior é dado ao método aglomerativo, que é o mais empregado nos principais programas estatísticos. Este método tem cinco algoritmos mais usados: ligação simples, ligação completa, ligação média, método de Ward e método centróide (JOHNSON e WICHERN, 2007; HAIR *et al*, 2005).

- **Ligação simples**: considera a menor distância entre um grupo e outro. Este procedimento também é chamado de vizinho mais próximo.
- **Ligação completa**: considera a maior distância entre um grupo e outro. Também chamado de vizinho mais distante.
- **Ligação média**: considera a distância média entre todos os indivíduos de um grupo para os do outro grupo.
- **Método de Ward**: utiliza como distância a soma dos quadrados entre os dois agrupamentos, feita sobre todas as variáveis. Este método forma grupos de maneira a minimizar a soma interna de quadrados, o que equivale a buscar o mínimo desvio padrão entre os dados de cada grupo.
- **Método Centróide**: considera a distância entre os centróides dos agrupamentos, onde os centróides são os centros (médias) dos agrupamentos.

Os resultados da análise de cluster pelo método hierárquico geralmente são exibidos em um dendograma (JOHNSON e WICHERN, 2007; HAIR *et al*, 2005).

3.3.2. Método Não-Hierárquico

Os métodos não hierárquicos são utilizados para agrupar objetos em um número determinado de grupos. Esse método é recomendado quando se tem um número grande de observações, pois não necessita que a matriz de distâncias seja determinada e os dados não precisam ser armazenados durante o processo. O principal método não hierárquico é o método conhecido como K-means (JOHNSON e WICHERN, 2007).

O método K-means divide os objetos em K clusters e designa cada objeto para o cluster cujo centróide é o mais próximo. Resumidamente, esse método segue os seguintes passos:

1. Divide os objetos inicialmente em K clusters
2. Designa cada objeto para o cluster cujo centróide é o mais próximo (geralmente se usa a distância euclidiana) e recalcula o centróide de cada cluster que recebeu um novo objeto e do que perdeu o objeto.
3. Segue-se realizando o passo 2 até que nenhum objeto precise mais trocar de cluster.

Outra maneira de iniciar o método K-means é especificando K centróides iniciais. A partir daí, segue-se os passos 2 e 3.

Os métodos não hierárquicos tem como desvantagem o fato de ser necessário definir previamente o número de clusters.

Dois pontos merecem atenção especial na análise de cluster: valores atípicos e amostra. Todos os métodos de agrupamento são sensíveis a *outliers*, que interferem nos grupos distorcendo a verdadeira estrutura da população. Uma análise preliminar dos dados deve, portanto, ser feita, para identificar possíveis *outliers*. Os clusters finais também devem ser examinados com cuidado, para verificar a interferência desses valores atípicos (JOHNSON e WICHERN, 2007; HAIR *et al*, 2005).

A amostra, especialmente na análise de agrupamentos, é ponto crítico, porque a técnica não é de inferência estatística e, portanto, o resultado final da análise será tão bom quanto a representatividade da amostra. Assim, para que o resultado possa ser generalizado para a população, a amostra deve ser representativa (HAIR *et al*, 2005).

Para finalizar a análise de cluster, é preciso interpretar os grupos finais encontrados. Nesta etapa, o objetivo é examinar cada grupo formado, considerando-se as variáveis que participaram da sua formação. Após esse exame, designa-se um nome ou rótulo que descreva cada agrupamento (HAIR *et al*, 2005).

3.4. Aplicação da Análise de Cluster

Através de uma análise exploratória inicial do banco de dados para verificar a distribuição das variáveis, a coerência das informações e a existência de *outliers*, identificou-se que, aproximadamente, metade das famílias apresentava total de gastos superior ao total de rendimento. Para que essas famílias fossem mantidas no estudo, optou-se por fazer a análise de cluster com os percentuais de gastos de cada categoria de despesa monetária, em relação ao total de gastos.

Assim, dentre as variáveis presentes no banco de dados da POF, as que participaram da formação dos clusters foram as referentes às categorias de consumo do orçamento familiar, definidas pelo IBGE. A escolha dessas variáveis deu-se em função do objetivo de identificar e comparar perfis de consumo, com base no orçamento das famílias. Também baseou-se no estudo de Silva e Parente (2007). As variáveis utilizadas na formação dos clusters foram:

- **Alimentação**

Alimentação dentro e fora do domicílio

- **Habitação**

Aluguel, condomínio, serviços e taxas de energia elétrica, telefone fixo, celular, pacote de telefone, TV e Internet, gás doméstico, água e esgoto, manutenção do lar e pequenos reparos, serviços domésticos, artigos de limpeza, mobiliário e artigos do lar, eletrodomésticos.

- **Vestuário**

Roupas para homem, mulher e crianças, sapatos e apetrechos, jóias e bijuterias, tecidos e armarinhos.

- **Transporte**

Transporte urbano, combustível, manutenção e acessórios, aquisição de veículos, viagens, estacionamento, pedágio, óleo diesel, gás combustível e seguro obrigatório.

- **Higiene e cuidados pessoais**

Perfume, produtos para cabelo, sabonete, maquiagem, produtos para pele, lâmina de barbear, alicate e cortador de unha.

- **Assistência à saúde**

Remédios, planos de saúde, consulta médica, tratamento dentário, tratamento médico e ambulatorial, serviços de cirurgia, hospitalização, exames e material de tratamento.

- **Educação**

Mensalidade, despesas com cursos, livros didáticos, revistas técnicas, artigos escolares, uniforme escolar, matrícula e outras despesas com educação.

- **Recreação e cultura**

Brinquedos, jogos, celular, livros, revistas e periódicos não didáticos, recreações, esportes, instrumentos musicais, equipamentos esportivos, artigos de acampamento e demais despesas similares.

- **Serviços pessoais**

Cabeleireiro, manicuro e pedicuro, consertos de artigos pessoais, depilação, maquiagem, esteticista, e demais despesas similares.

- **Aumento do ativo**

Aquisição de imóveis, construção e melhoramento de imóveis, títulos de capitalização, títulos de clube, aquisição de terrenos para jazigo e investimentos direcionados para aumento do patrimônio em geral.

Além das dez variáveis descritas, outras quatro categorias de consumo não foram consideradas para a formação dos grupos, por apresentarem baixo percentual de resposta: fumo, despesas diversas, outras despesas correntes e diminuição do passivo.

Para a definição dos clusters, adotou-se como medida de similaridade a distância euclidiana quadrada. Além disso, as variáveis foram padronizadas em escores padrão (z-scores).

Inicialmente, o método utilizado na definição dos clusters foi o método hierárquico pois, como não se tem nenhuma informação prévia do número de grupos existentes dentro dessa classe, esse método é mais apropriado (JOHNSON e WICHERN, 2007; HAIR *et al*, 2005).

Dentro do método hierárquico, foram testados todos os algoritmos aglomerativos (ligação simples, ligação completa, ligação média, método centróide e método de Ward) e analisados os dendogramas. Os melhores resultados foram encontrados com o método de Ward, que forma grupos de maneira a minimizar a soma interna de quadrados, o que equivale a buscar o mínimo desvio padrão entre os dados de cada grupo, sendo o algoritmo mais completo.

Com base na análise do dendograma, identificou-se que os dados poderiam ser divididos em 3 a 5 clusters. Assim, a partir dos resultados encontrados com o método hierárquico, foi aplicado o método não hierárquico (K-means) para 3, 4 e 5 clusters.

Na análise de cluster, a definição do número ótimo de clusters depende, principalmente, do propósito do pesquisador. Sendo assim, após analisar cada uma das opções, observou-se que, para os propósitos deste estudo, os dados ficaram bem divididos em 4 clusters tanto para a POF de 2002/2003 quanto para de 2008/2009. Esta definição também levou em consideração o que foi observado na revisão de literatura (CASTILHOS E ROSSI, 2009; SILVA E PARENTE, 2007; MATTOSO E ROCHA, 2009; CHAUVEL E MATTOS 2008).

Vale ressaltar que os mesmos métodos e procedimentos foram adotados para as duas POF's para que fosse possível comparar os resultados encontrados nos dois períodos.