3
Método

Este estudo tem como base dados secundários da Pesquisa de Orçamentos Familiares do IBGE, realizada em dois períodos, 2002-2003 e 2008-2009, sendo estas as duas edições mais recentes da pesquisa. De forma geral, pode ser caracterizado como sendo de natureza descritiva.

3.1. Pesquisa de Orçamentos Familiares (POF)

A Pesquisa de Orçamentos Familiares – POF é realizada pelo IBGE (Instituto Brasileiro de Geografia e Estatística), tendo como propósito fornecer informações sobre o orçamento doméstico, sobre as condições de vida das famílias, inclusive percepções subjetivas sobre a qualidade de vida, e estudar o perfil nutricional da população. De forma mais específica, esta pesquisa busca principalmente "(...) mensurar as estruturas de consumo, dos gastos, dos rendimentos e parte da variação patrimonial das famílias. Possibilita traçar, portanto, um perfil das condições de vida da população brasileira a partir da análise de seus orçamentos domésticos" (POF, 2010, p.17).

Além da estrutura orçamentária, várias outras características do domicílio e da família também são investigadas, ampliando o potencial de utilização dos resultados.

Devido a sua alta abrangência, seus resultados contribuem com informações para subsidiar políticas públicas, a fim de melhorar as condições de vida da população, como políticas no campo da nutrição, orientação alimentar, saúde e moradia, dentre outras. Para as empresas privadas, a pesquisa é útil na definição de estratégias de investimento em que o conhecimento do perfil e da demanda por bens de serviços seja determinante.

A POF 2008/2009 é a quinta pesquisa feita pelo IBGE sobre orçamento familiar. A primeira pesquisa realizada foi o Estudo Nacional de Despesa Familiar – ENDEF 1974-1975, que abrangia todo o territorial nacional, à exceção das áreas rurais das Regiões Norte e Centro-Oeste. As duas seguintes, POF 1987-1988 e POF 1995-1996, foram feitas apenas nas Regiões Metropolitanas

de Belém, Fortaleza, Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo, Curitiba e Porto Alegre, além do município de Goiânia e do Distrito Federal. As últimas duas POFs, 2002-2003 e 2008-2009, foram de abrangência nacional, incluindo as áreas urbana e rural. Além disso, elas investigaram novas informações como as condições de vida da população a partir do consumo e as aquisições não monetárias.

A POF é uma pesquisa realizada por amostragem, em que se procura investigar os domicílios particulares permanentes. A unidade de consumo, identificada dentro do domicílio, é a unidade básica da pesquisa, que compreende o morador ou os moradores que compartilham a mesma fonte de alimentação ou as despesas com moradia.

O conceito de família utilizado nesta pesquisa, atendendo às especificações internacionais, se refere a "(...) pessoas ligadas por laços de parentesco, dependência doméstica ou normas de convivência, sem referência explícita ao consumo ou despesas" (POF, 2010, p. 19).

O período de coleta da POF 2008/2009 teve inicio no dia 19 de maio de 2008, e término no dia 19 de maio de 2009, já a POF 2002/2003, iniciou-se em julho de 2003 e se prolongou até junho de 2002.

Em função da grande diversidade de itens adquiridos com valores diferentes e frequências de aquisição diferentes, definiu-se quatro períodos de referência, com 7, 30, e 90 dias, e 12 meses, com o objetivo de aumentar a capacidade do respondente em informar as despesas com maior precisão. Para as informações sobre rendimentos considerou-se o período de um ano.

Em função dos vários períodos de referência das informações e do longo período de coleta de dados, foi fixada uma data de referência para compilação, análise e apresentação dos resultados. Na POF 2008/2009 esta data foi 15 de janeiro de 2009, já na POF 2002/2003, esta data foi 15 de janeiro de 2003.

Tanto para a POF 2002-2203 quanto para a 2008-2009, o plano amostral adotado foi o conglomerado em dois estágios, com estratificações geográficas e estatísticas das unidades primárias de amostragem (setores censitários). Dentro de cada setor, foram selecionadas as unidades secundárias de amostragem, que são os domicílios particulares permanentes. Assim, a amostra final da POF 2002-2003 foi de 44.248 domicílios e em 2008-2009 foi de 59.548.

A coleta da pesquisa é feita por meio de sete questionários, assim descritos:

POF 1 – Questionário de características do domicílio e dos moradores

POF 2 – Questionário de aquisição coletiva

POF 3 – Caderneta de aquisição coletiva

POF 4 – Questionário de aquisição individual

POF 5 – Questionário de trabalho e rendimento individual

POF 6 – Avaliação das condições de vida

POF 7 – Bloco de consumo alimentar pessoal

3.2. Amostra

Com o objetivo de estudar o segmento de baixa renda da região metropolitana de Recife, escolheu-se a variável renda como critério de definição desse segmento. Para o presente estudo, considerou-se como pertencentes à classe de baixa renda as famílias com renda familiar mensal entre 1 e 3 salários mínimos. Para a pesquisa de 2002-2003, este intervalo é de R$240,00 até R$720,00, considerando o salário mínimo com data de referência abril de 2003 (R$240,00). Já para a pesquisa de 2008-2009, este intervalo vai de R$465,00 até R$1.395,00, considerando o salário mínimo com data de referência fevereiro de 2009 (R$465,00).

Este corte salarial, entre 1 e 3 salários mínimos, tem como objetivo minimizar a influência da renda nos padrões de consumo das famílias. O fato de selecionarmos apenas as famílias da região metropolitana de Recife segue este mesmo raciocínio, buscando minimizar o efeito geográfico no orçamento familiar.

O conceito utilizado como renda familiar nesta pesquisa seguirá a definição de renda bruta total da POF, sendo a soma dos rendimentos brutos monetários dos componentes das famílias, acrescidos de seus rendimentos não monetários. Como rendimento monetário, entende-se todo ganho monetário, exceto variação patrimonial, recebido nos 12 meses anteriores à coleta das informações. Este rendimento é proveniente de: rendimento do trabalho, rendimento do empregado, rendimento do empregador e conta própria, aposentadoria, pensão, previdência privada, entre outros. Já o rendimento não monetário engloba rendimentos obtidos por meio de doação, troca, produção própria, caça, pesca e coletado (POF, 2010).

No que se refere às despesas, considerou-se neste estudo apenas as despesas monetárias, aquelas pagas com dinheiro, cheque ou cartão de crédito, à vista ou a prazo. As despesas não monetárias foram desconsideradas, pois incluem o aluguel estimado, ou seja, para as famílias que moram em imóveis

próprios e por isso não gastam com aluguel é estimado um valor para este gasto, o que inflaciona este tipo de despesa.

O tamanho final da amostra, após exclusão de *outliers* e *missing values*, foi de 156 famílias na POF 2002-2003 e 422 famílias na POF 2008-2009. Vale ressaltar que esta amostra é representativa para região metropolitana do Recife, podendo os resultados encontrados neste estudo serem generalizados para a toda a população.

3.3. Análise Multivariada

A técnica estatística de análise multivariada consiste em analisar simultaneamente um conjunto de variáveis que caracterizam os indivíduos ou objetos de um conjunto de dados.

Os objetivos principais desta análise são: a redução dos dados, transformando-os em uma forma mais simples e facilitando a interpretação dos mesmos, porém sem perder as informações importantes; criação de grupos entre variáveis ou objetos similares, através de medidas de similaridade de algumas características; análise das variáveis, se são independentes ou dependentes de outras e, finalmente, determinação do relacionamento entre as variáveis a fim de prever os valores para as variáveis de interesse (JOHNSON e WICHERN, 2007; HAIR *et al*, 2005).

Esta ferramenta estatística é utilizada em diferentes áreas de aplicação como, por exemplo: Marketing, Administração, Sociologia, Psicologia, Medicina e Meteorologia, com o intuito de estudar dados complexos que envolvam mais de uma variável.

A análise multivariada abrange diversas técnicas para análise de dados, como análise fatorial, regressão múltipla, análise conjunta, análise discriminante, análise de cluster, entre outras. Neste trabalho, será utilizada como metodologia a análise de cluster, com o objetivo de agrupar as famílias de baixa renda da Região Metropolitana de Recife em subgrupos, onde todas as famílias de um grupo sejam similares entre si (JOHNSON e WICHERN, 2007; HAIR *et al*, 2005).

3.3.1. Análise de Cluster

A análise de cluster tem como objetivo principal agrupar os objetos com base na similaridade de algumas de suas características. Os grupos devem ser

homogêneos dentro de si, e heterogêneos entre si (HAIR et al, 2005). Nesta análise cabe ao pesquisador analisar se os grupos formados são bons ou não.

A análise de cluster pode servir como ferramenta exploratória, quando pouco ou nada se sabe sobre o objeto estudado, ou como ferramenta confirmatória, quando se deseja confirmar alguma relação anteriormente verificada entre os objetos (HAIR et al, 2005).

A escolha das variáveis que irão compor os clusters deve ser feita de forma cuidadosa; variáveis que sejam irrelevantes na formação dos clusters não devem ser incluídas.

Aconselha-se que as variáveis sejam padronizadas, de forma a eliminar o efeito escala, assim todas as variáveis possuem o mesmo peso na formação do cluster. Segundo HAIR et al (2005), a padronização utilizada com mais frequencia é a z-scores, onde subtraímos o valor de cada variável pela média e a dividimos pelo desvio padrão desta mesma variável.

A amostra e os valores atípicos são dois pontos aos quais se deve prestar atenção ao fazer a análise de cluster. Como esta técnica não é de inferência estatística, o resultado final da análise só poderá ser generalizado para a população se a amostra for representativa, e este resultado será tão bom quanto a representatividade da amostra (HAIR *et al*, 2005). Já em relação aos *outliers*, estes devem ser identificados antes de ser iniciada a análise de cluster, pois influenciam em todos os métodos de agrupamento atípicos (JOHNSON e WICHERN, 2007; HAIR *et al*, 2005).

Ao escolher a medida de similaridade que será aplicada deve-se considerar três fatores: a natureza das variáveis, a escala de medida e o conhecimento sobre o problema (JOHNSON e WICHERN, 2007).

Para medirmos a similaridade entre os objetos, a medida mais utilizada é a medida de distância, sendo a distância euclidiana a mais comum, de acordo com HAIR et al (2005):

- **Distância Euclidiana**

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

Para a realização do agrupamento, deve-se escolher o algoritmo que será utilizado. Existem dois tipos de algoritmos de agrupamento: algoritmos hierárquicos e algoritmos não hierárquicos.

O procedimento hierárquico pode ser aglomerativo ou divisivo. No metodo aglomerativo, cada objeto constitui um grupo distinto, no próximo passo os dois objetos mais semelhantes são combinados em um novo agrupamento. Este método se repete diminuindo o número de agrupamentos em uma unidade em cada passo. Já no método divisivo, inicia-se com um único grupo contendo todos os objetos, onde a cada etapa as observações mais diferentes entre si são separadas e transformadas em agrupamento menores. Este processo se repete até que cada observação seja um agrupamento diferente. Os cinco algoritmos mais utilizados para desenvolver agregados são (HAIR et al, 2005):

- Ligação individual – Encontra os dois objetos separados pela menor distância e os coloca no mesmo grupo.
- Ligação completa – É baseado na distância máxima entre dois grupos.
- Ligação média – O critério de agrupamento é a distância média de todos os indivíduos de um grupo aos demais indivíduos do outro grupo.
- Método de Ward – A distância entre dois agrupamentos é a soma dos quadrados entre os dois agrupamentos, feitas sobre todas as variáveis.
- Método centróide – A distância entre dois agrupamentos é a distância entre seus centróides.

O dendograma apresenta os resultados da análise de cluster pelo método hierárquico. (JOHNSON e WICHERN, 2007; HAIR et al, 2005).

O método não hierárquico é utilizado para agrupar os objetos em um número preestabelecido de clusters. Uma vantagem deste método sobre o método hierárquico é a maior velocidade de processamento, o que permite que se trabalhe com maior número de objetos (JOHNSON e WICHERN, 2007).

O principal método não hierárquico é o k-means, cujo processo se desenvolve em 3 etapas:

1- Os indivíduos são divididos em k clusters iniciais
2- Calcula-se a distância de cada objeto para o centróide de cada um dos grupos e designa os objetos para o grupo cujo centróide é o mais próximo, recalculando o centróide.

3- Repete-se o passo 2 até que nenhum objeto mude de cluster.

Uma desvantagem deste método é a necessidade de algum prévio conhecimento dos dados, pois é obrigatório definir previamente o número de clusters.

3.4. Aplicação da Análise de Cluster

Antes de aplicar a análise de cluster na amostra selecionada da POF, foi feita uma análise exploratória do banco de dados com o intuito de avaliar as variáveis a serem utilizadas na formação e caracterização dos clusters, de identificar valores *outliers* e, também, verificar a coerência das informações.

O banco de dados possui uma extensa e complexa lista de variáveis que inclui desde características dos domicílios até avaliação das condições de vida. As variáveis que formarão os clusters são as referentes aos gastos nas categorias de despesa. Após a definição dos grupos, outras variáveis serão utilizadas para estudar a caracterização dos mesmos. Segue abaixo a definição das variáveis.

Variáveis usadas na formação dos clusters

- **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:** remédio, plano 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.

- **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.

Outras cinco categorias de despesa foram desconsideradas na formação dos clusters em função de apresentarem baixo percentual de resposta. São elas:

- **Serviços pessoais:** cabeleireiro, manicuro e pedicuro, consertos de artigos pessoais, depilação, maquiagem, esteticista, e demais despesas similares.

- **Despesas diversas:** jogos e apostas, comunicação, cerimônias e festas, serviços profissionais, despesas com imóveis de uso ocasional, reforma e manutenção de jazigo, alimentos e outros produtos para animais, etc.

- **Outras despesas:** impostos pagos, pensões, mesadas, doações, previdência privada, seguros, pagamento de asilo, indenização a terceiros e demais despesas de mesma natureza.

- **Fumo:** cigarros, charutos, fumo para cachimbo, fumo para cigarro e outros artigos para fumante.

- **Diminuição do passivo:** pagamentos de débitos, juros e seguros com empréstimos pessoais.

Variáveis usadas na caracterização dos clusters

Para a caracterização dos grupos, foram investigadas, alem das variáveis descritas acima, as seguintes variáveis:

- Renda bruta familiar mensal
- Número de moradores por domicílio
- Características demográficas do chefe do domicílio (sexo, cor/raça, idade, anos de estudo, nível de instrução)
- Posse de cartão de crédito e posse de cheque
- Posse de plano de saúde
- Benefício do Bolsa Família (somente para a 2008/2009)
- Quantidade e tipo de alimento consumido
- Tipo de domicílio
- Condição de ocupação do domicílio
- Percentual de gastos com despesas diversas, outras despesas, fumo, serviços pessoais e diminuição do passivo.
- Município de moradia (capital ou fora da capital)
- Avaliação geral da condição de vida
- Inadimplência
- Avaliação das condições de moradia
- Avaliação dos problemas do domicílio
- Avaliação dos serviços
- Situação financeira das famílias (gastam mais do que ganham no mês)

A variável “situação financeira”, não consta no questionário da POF/IBGE, a mesma foi construída a partir de dados sobre gastos e receitas das famílias. Caso a família tenha mais gastos do que rendimentos ela é considerada como endividada.

Na análise inicial identificou-se também que um número significante de famílias (61% em 2002-2003 e 21% em 2008-2009) apresentava um valor total de gastos acima do valor total de rendimentos. Em função disso, e também para ser possível comparar com o estudo feito anteriormente para a região metropolitana do Rio de Janeiro, optou-se por aplicar a análise de cluster no percentual de gastos de cada categoria de despesa monetária em relação ao total de gastos.

Para realizar a análise de cluster, as variáveis foram padronizadas em escores padrão (*z-scores*) e definiu-se como medida de similaridade a distância euclidiana quadrada, que é a mais usada neste tipo de análise.

Na definição do método a ser utilizado, como não se tinha nenhum conhecimento prévio do número de subgrupos existentes nesta população, optou-se por utilizar o método hierárquico, que é mais indicado nestes casos (JOHNSON e WICHERN, 2007; HAIR *et al*, 2005).

Ao aplicar o método hierárquico, todos os algoritmos aglomerativos descritos anteriormente foram testados: ligação simples, ligação completa, ligação média, método centróide e método de Ward. Após analisar os resultados e os dendogramas, verificou-se que o método de Ward foi o algoritmo que apresentou os melhores resultados. Este algoritmo busca o mínimo desvio padrão entre os dados de cada grupo sendo considerado o mais completo.

O dendograma final mostrou que os dados ficariam bem divididos em 4 ou 5 grupos. Como na análise de cluster a definição do número ótimo de clusters depende, principalmente, do propósito do pesquisador, após avaliar os resultados de cada uma das opções, optou-se por segmentar os dados em 4 grupos, tanto na POF 2002-2003 quanto na POF 2008-2009, pois essa classificação apresentou grupos homogêneos dentro de si e heterogêneos entre si.

Os métodos acima descritos foram aplicados de maneira igual nos dados das duas pesquisas para que fosse possível comparar os resultados dos dois períodos estudados.