PMKT
REVISTA BRASILEIRA DE PESQUISAS DE MARKETING, OPINIÃO E MÍDIA
BRAZILIAN JOURNAL OF MARKETING, OPINION, AND MEDIA RESEARCH

**PMKT – Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia**
**ISSN**: 1983-9456 (Impressa)
**ISSN**: 2317-0123 (On-line)
**Editor:** Fauze Najib Mattar
**Sistema de avaliação:** Triple Blind Review
**Idiomas:** Português e Inglês
**Publicação:** ABEP – Associação Brasileira de Empresas de Pesquisa

# Segmentação da População Brasileira[1]

## *Brazilian Population Segmentation*

Submissão: 28/mar./2014 - Aprovação: 8/abr./2014

**Rodrigo Otávio de Araújo Ribeiro**
Doutor e Mestre em Engenharia de Produção pela Universidade Federal Fluminense - UFF. Bacharel em Estatística pela Escola Nacional de Ciências Estatísticas - ENCE/IBGE. Especialista na aplicação de modelos estatísticos em grandes bases de dados. Diretor de Inteligência de Marketing no IBOPE DTM.
**E-mail:** rodrigo.ribeiro@ibopedtm.com
**Endereço profissional:** IBOPE DTM - Rua Voluntários da Pátria - 89 - sala 803 - 22270-000 - Botafogo - Rio de Janeiro/RJ – Brasil.

**Bruna Suzzara Bueno de Miranda**
Pós-graduada em Inteligência Competitiva pelo IBOPE Educação. Bacharel em Estatística pela Universidade Estadual de Campinas - UNICAMP. Elabora desenhos amostrais de estudos quantitativos de mercado, mídia e opinião pública e análises descritivas, multivariadas e de regressão. Coordenadora de Estatística no IBOPE Inteligência.
**E-mail:** bruna.suzzara@ibopeinteligencia.com

**Mariana Pereira Nunes**
Mestre em Administração pela Fundação Getúlio Vargas – FGV-RJ. Bacharel em Estatística pela Escola Nacional de Ciências Estatísticas - ENCE/IBGE. Especialista na aplicação de modelos estatísticos em grandes bases de dados. Coordenadora de Inteligência de Marketing no IBOPE DTM.
**E-mail**: mariana.nunes@ibopedtm.com

**Livia Gomes Cruz**
Pós-graduada em Opinião Pública e Pesquisa de Mercado pela Fundação Escola de Sociologia e Política de São Paulo - FESP/SP. Bacharel em Estatística pela Universidade Estadual Paulista - UNESP. Executa análises descritivas, multivariadas e regressões para estudos quantitativos de mercado e opinião pública. Estatística no IBOPE Inteligência.
**E-mail:** livia.cruz@ibopeinteligencia.com

[1] Este foi um dos trabalhos apresentados no 6º Congresso Brasileiro de Pesquisa - Mercado, Opinião e Mídia da ABEP (realizado em 24 e 25 de março de 2014), transformado em artigo por seu(s) autor(es), submetido à PMKT e aprovado para publicação.

## RESUMO

Este estudo teve como objetivo realizar uma segmentação da população brasileira com base nos dados do IBGE (2010) no Censo 2010. A metodologia elaborada consiste na aplicação de algoritmos de *cluster* de forma sequencial, visando à obtenção de agrupamentos relevantes. As soluções obtidas foram submetidas a um rigoroso processo de validação que, dentre outras atividades, utiliza métodos de classificação para averiguar a robustez dos agrupamentos obtidos. A disponibilização dos dados do Censo georreferenciados foi uma das motivações deste estudo e permitiu a caracterização de áreas geográficas a partir da predominância de determinados segmentos. Dentre as aplicações práticas do trabalho, tem-se: possibilidade de melhor caracterizar o mercado consumidor brasileiro e realizar pesquisas de opinião com amostras representativas para cada um dos segmentos identificados, entre outras.

## PALAVRAS-CHAVE:

Segmentação, Censo 2010, georreferenciamento, *two step cluster*.

## *ABSTRACT*

*This study aims to present a segmentation of the Brazilian population that was conducted Based on 2010 Census data from the IBGE (2010). The methodology developed consists of applying various clustering algorithms sequentially, in order to obtain relevant groupings. The solutions obtained were subjected to a rigorous validation process, which uses a classification method to assess the robustness of the clusters obtained, among other activities. The availability of georeferenced data from Census was one of the motivations of this study. It allowed the characterization of geographic areas from the predominance of certain segments. Some examples of possible practical applications of this work are: better characterize the Brazilian consumer market, conduct surveys with representative samples for each segment, among others.*

## *KEYWORDS:*

*Segmentation, 2010 Census, georeference, two step cluster.*

# 1. INTRODUÇÃO

A crescente necessidade dos profissionais de marketing brasileiros quanto ao entendimento de seus clientes constituiu o grande estímulo para a elaboração deste trabalho. Com base em uma metodologia inovadora, serão mostradas as diferentes características dos setores censitários da população brasileira quanto às suas condições sociodemográficas.

Os dados fornecidos pelo Instituto Brasileiro de Geografia e Estatística - IBGE contém grande riqueza de detalhes que permitem diversos tipos de análises estatísticas. As informações do Censo 2010 são um exemplo disso. Esta pesquisa forneceu um vasto conjunto de dados que possibilitaram uma caracterização profunda da população e dos domicílios brasileiros. Segundo o IBGE (2010), o menor nível de granularidade geográfica disponibilizado pelo IBGE é o setor censitário, com informações:

> [...] que compreendem sexo, idade, situação do domicílio, emigração internacional, ocorrência de óbitos, cor ou raça, registro de nascimento, alfabetização e rendimento, para a totalidade da população, bem como informações sobre composição e características dos domicílios. (IBGE, 2010).

Essas informações, pela primeira vez no Brasil, estão georreferenciadas, permitindo um mapeamento espacial das informações disponibilizadas pelo IBGE (2010) por meio do Censo 2010.

Os objetivos principais do presente artigo foram: segmentar a população brasileira pela caracterização dos municípios brasileiros e seus setores censitários utilizando dados sociodemográficos do Censo 2010 e mapear áreas de concentração desses segmentos.

Os resultados deste trabalho poderão servir de *insight* não apenas para pesquisas de marketing, mas também para pesquisas com foco na política e no comportamento da sociedade em geral. Será possível identificar o perfil dos eleitores de determinada zona eleitoral, conhecer as condições de moradia e o grau de instrução dos telespectadores de um programa de televisão específico, com um detalhamento mais profundo.

Futuramente será possível complementar esta caracterização com informações oriundas de outras pesquisas ou estudos. Os segmentos georreferenciados facilitarão a associação dos resultados encontrados neste trabalho com pesquisas ou estudos futuros.

Nos Estados Unidos, desde 1976, a Claritas (Nielsen) utiliza dados do censo para segmentar a população norte-americana. Esta segmentação é utilizada por empresas, políticos, ONGs, entre outras entidades/pessoas com os mais diversos objetivos.

# 2. REFERENCIAL TEÓRICO

## 2.1 IMPORTÂNCIA DO ESTUDO

As análises realizadas tiveram como objetivo segmentar a população brasileira e caracterizá-la segundo diversas variáveis disponíveis em dados oficiais, principalmente do Censo 2010.

Segundo Churchill Jr. e Peter (2000), a segmentação é entendida como uma forma de separar um mercado em grupos de compradores potenciais sendo esses indivíduos semelhantes segundo suas necessidades e desejos, percepções de valores ou comportamentos de compra.

A análise de *clusters* é a principal técnica estatística aplicada neste trabalho, sua aplicação está relacionada ao agrupamento de objetos em grupos visando a maior homogeneidade possível dentro de um mesmo grupo e separabilidade em relação aos demais. Contudo, um conjunto de árvores de decisão por *BAGGING* (também chamado de *Random Forest*) foi utilizado para verificar a consistência dos agrupamentos encontrados por meio da análise de *cluster*.

## 2.2 O QUE PERMITIU O ESTUDO

Uma série de fatores em conjunto foi crucial para o desenvolvimento deste trabalho. Dentre eles destacam-se a crescente disponibilidade e busca por informações, a realização do Censo 2010 e a maior inovação trazida com ele: o georreferenciamento dos dados.

## 2.3 CRESCENTE NECESSIDADE E BUSCA POR INFORMAÇÃO

Nas duas últimas décadas o mundo vivenciou um aumento considerável da quantidade de dados armazenada. Como consequência deste volume surgiu a necessidade de se desenvolver formas de extrair e analisar as informações provenientes dessas grandes massas de dados.

Existem muitas técnicas analíticas capazes de extrair informações pertinentes destas grandes bases de dados. Algumas já consagradas tais como análise estatística multivariada e outras em constante desenvolvimento, baseadas em programação matemática, computação evolutiva e redes neurais. A junção de todo este ferramental analítico aplicado à extração de informações em grandes bases de dados pode ser denominado como *Big Data Analytics*.

Contudo, com a multiplicação das fontes de informações e suas mais variadas temáticas, o potencial de algumas fontes de informação pode passar despercebido no universo empresarial. Este é o caso do censo realizado pelo IBGE, muito utilizado para planejamento de políticas públicas, mas pouco explorado, em todo seu potencial, na geração de informações pertinentes sob a ótica do mercado consumidor brasileiro.

Em constante evolução metodológica por décadas, o censo realizado em 2010 passou a contar com informações georreferenciadas. Isto significa que seus dados podem ser associados a uma localização geoespacial.

Este foi o estímulo principal para o desenvolvimento deste estudo, pois, segundo Weiss (1988), "[...] o lugar em que se vive determina a maneira como se vive. Conhecendo onde e como as pessoas vivem, fica mais fácil atender às suas necessidades, com mais chance de sucesso nessa empreitada". Dias (2004), também acredita que o lugar onde se vive em todos os momentos (trabalho, residência, férias etc.) impacta diretamente o comportamento de compras das pessoas.

Portanto, em 2013, quando todas as informações coletadas no Censo 2010 já tinham sido disponibilizadas, observou-se a oportunidade de analisar profundamente esses dados com o objetivo entender o comportamento, estilo de vida e particularidades da população brasileira, sendo capaz de gerar novos insumos para pessoas e organizações das mais diversas áreas de atuação, que tiverem interesse de conhecer a população brasileira.

## 2.4 DADOS DO CENSO 2010

Para desenvolver esse trabalho, foram utilizados dados já coletados pelo IBGE (2010), no Censo 2010, abrangendo todo o território nacional. A principal fonte de informação utilizada foram as bases de microdados do Censo 2010:

> Os **microdados** consistem no menor nível de desagregação dos dados de uma pesquisa, retratando, sob a forma de códigos numéricos, o conteúdo dos questionários, preservado o sigilo estatístico com vistas a não individualização das informações [...] possibilitando aos usuários especializados [...] a leitura dos dados, o cruzamento em diferentes agregações geográficas, e a elaboração de múltiplas tabulações segundo sua perspectiva pessoal de interesse. (IBGE, 2010).

Existem dois tipos de microdados. Os microdados do questionário reduzido (dados do Universo) e do questionário completo (dados da Amostra). Os microdados do Universo contêm informações de domicílios e de pessoas incluindo resultados de rendimento que foi pesquisado para todas as pessoas de dez anos ou mais de idade. Estas informações são disponibilizadas no menor nível de que desagregação são os setores censitários.

Segundo o IBGE (2010) "um setor censitário é formado por uma área contínua, contida integralmente em uma área rural ou urbana, que respeita as divisões político-administrativas do Brasil."

Os microdados contêm diversas características individuais e domiciliares. Ao contrário da base do Censo, que possui leitura por setores censitários, essa base tem como menor unidade de desagregação a área de ponderação e esses dados não permitem que nenhum respondente do Censo 2010 tenha suas informações divulgadas.

> Define-se área de ponderação como sendo uma unidade geográfica, formada por um agrupamento mutuamente exclusivo de setores censitários contíguos [dentro de um único município], para a aplicação dos procedimentos de calibração dos pesos de forma a produzir estimativas com as informações conhecidas para a população como um todo. (IBGE, 2010).

Os temas pesquisados para o universo compreendem sexo, idade, situação do domicílio, emigração internacional, ocorrência de óbitos, cor ou raça, registro de nascimento, alfabetização e rendimento, bem como informações sobre composição e características dos domicílios.

## 2.5 GEORREFERENCIAMENTO DOS DADOS DO CENSO 2010

No Censo Demográfico de 2010 – Notas técnicas/Base territorial, o IBGE (2010) contou com diversas inovações e a principal delas foi com a base territorial:

> **Base territorial** é a denominação dada ao sistema integrado de mapas, cadastros e banco de dados, construído segundo a metodologia própria para dar organização e sustentação espacial às atividades de planejamento operacional, coleta e apuração de dados e divulgação de resultados do Censo Demográfico. [...] A base territorial foi elaborada de forma a integrar a representação espacial das áreas urbana e rural do Território Nacional em um ambiente de banco de dados geoespaciais, utilizando insumos e modernos recursos de tecnologia da informação. (IBGE, 2010).

A criação da base territorial e da utilização de técnicas geoespaciais tornou possível mapear os setores censitários. Isso permitiu que a caracterização, associada à localização de tais setores compusesse a segmentação da população brasileira.

## 2.6 INFLUÊNCIA DA DISPONIBILIDADE DOS DADOS NA METODOLOGIA ADOTADA

Devido à variedade de dados analisados, surgiu um fator que foi determinante para a escolha da metodologia utilizada: a granularidade dos dados, com algumas informações no nível de município, outras de área de ponderação ou setor censitário. Foi necessário fasear o processo de análise de forma que o máximo de informação pudesse ser analisado sem comprometer a consistência dos dados nem a interpretação dos resultados.

## 2.7 CLARITAS (NIELSEN)

A Claritas, empresa norte-americana, desenvolveu um sistema de segmentação dos consumidores norte-americanos que é vastamente utilizado desde a década de 1990. Nesta segmentação, a Claritas (Nielsen) criou 14 grupos de consumidores com base em dados sociodemográficos e esses grupos foram divididos em 66 segmentos com base em informações de consumo, estilo de vida, entre outras caracterizações. A partir dessas informações a empresa montou um sistema que permite que as pessoas consultem as características de cada segmento, assim como onde esses indivíduos estão localizados (CEP). Como se trata de uma ferramenta dinâmica, fácil de ser manipulada e rica de informações, serve como fonte (<www.claritas.com>).

## 3. METODOLOGIA

A clusterização é um método estatístico utilizado para o agrupamento de unidades analíticas em determinado conjunto de dados. Cada unidade é agrupada de acordo com a similaridade existente entre suas características. Neste trabalho, foram realizados agrupamentos em dois momentos distintos. Primeiramente foram agrupados os municípios brasileiros, já em um segundo momento, os setores censitários dentro de cada município.

A validação dos segmentos obtidos por meio da análise de *cluster* foi realizada pela aplicação de um algoritmo de *Random Forests* (BREIMAN, 2001) que buscou identificar os segmentos com base nas variáveis utilizadas na etapa de clusterização. A escolha deste método deve-se ao desempenho superior quanto a classificação de diferentes informações, tendo *performance* consistente em bases pequenas e grandes.

## 3.1 ANÁLISE DE *CLUSTER*

Existe uma grande variedade de algoritmos de clusterização, contudo não há um método ideal. A escolha do método depende do tipo de dados disponíveis e do objetivo do estudo (HAIR et al. 2005). Em geral, quando se realiza uma análise de *cluster*, aplicam-se diversos algoritmos de *cluster* e aquele que produzir os agrupamentos mais adequados é utilizado na clusterização final.

Os algoritmos que realizam a clusterização são denominados não supervisionados, pois não existe *a priori* definição do número de grupos que deverá ser encontrado. Estes algoritmos são divididos em duas categorias: hierárquicos e não hierárquicos.

A primeira categoria é composta por algoritmos que realizam sucessivas divisões (aglomerativos) do conjunto de dados com base em uma matriz de dissimilaridade, sendo as soluções finais de agrupamento definidas com base na análise de um gráfico denominado dendograma. A grande

vantagem está na flexibilidade quanto à escolha da medida de distância a ser considerada e à visualização da estrutura das divisões, contudo, possuem muito tempo de processamento.

A segunda categoria, é composta por algoritmos partitivos, que buscam particionar os objetos do conjunto, visando a otimização de algum critério predefinido. Como desvantagem, não há visualização da estrutura de agrupamento.

## 3.2 ALGORITMO PAM

O algoritmo *Partitioning Around Medoids* - PAM foi apresentado por Kaufman e Rousseeuw em 1987. Trata-se de um dos clássicos da literatura para pequenos conjuntos de dados, sendo muito parecido ao *k-means*. Contudo, ao invés de eleger protótipos para posição de centralidade dentro de um *cluster*, passa a eleger elementos existentes do conjunto de dados avaliado.

A vantagem deste método está na maior robustez na presença de *outliers*. Possui como objetivo selecionar *k* elementos denominados medóides. Após sua escolha, é realizada uma varredura no banco para criação dos *clusters*, agrupando os objetos ao medóide mais próximo (BRITO W.; SEMAAN; BRITO J., 2011).

## 3.3 ALGORITMO CLARA

Devido à complexidade computacional para aplicação do algoritmo PAM em grandes bases de dados, foi desenvolvido o algoritmo *Clustering Large Applications* - CLARA, que pode ser aplicado em bases de dados de grandes dimensões (KAUFMAN; ROUSSEEUW, 1990). Ao invés de determinar os *k*-medóides considerando toda a base de dados, o algoritmo CLARA seleciona *x* amostras compostas por *y* objetos da base de dados e aplica o algoritmo PAM em cada uma delas. Após a definição dos medóides, cada objeto que não pertence à amostra é alocado ao grupo com o medóide mais próximo.

## 3.4 ALGORITMO *TWO STEP CLUSTER*

O *Two-Step Cluster*, procedimento de clusterização utilizado neste trabalho, se baseia na utilização conjunta de um algoritmo não hierárquico com um hierárquico. A vantagem na junção destes dois tipos de algoritmos está na possibilidade de obtenção de resultados precisos com um menor tempo de processamento computacional. Sendo possível a análise mais precisa das soluções finais pela visualização de sua estrutura de agrupamento.

Primeiramente foi aplicado o método não hierárquico (CLARA) considerando uma grande quantidade de agrupamentos (*clusters*). No segundo passo, com base na solução dos medóides dos *clusters* obtidos pelo CLARA foi aplicado o método hierárquico com aglomerações realizadas pelo método de Ward.

O método de Ward é um algoritmo aglomerativo em que as partições formadas minimizam a perda associada a cada agrupamento. Todas as possíveis uniões entre pares de *clusters* são consideradas e aqueles que apresentam a mínima perda de informação (definida pela soma de quadrados), são selecionados para o agrupamento.

### *3.5 RANDOM FOREST*

O *Random Forest* é um método de aprendizado conjunto para a classificação desenvolvido por Breiman (2001). Este método é considerado uma boa alternativa para resolução de problemas de classificação em que há necessidade de melhor adequação ao processo tratado. Ele é construído com base na técnica de treinamento para coleções de classificadores instáveis, o *bootstrap AGGregatING - BAGGING*, que consiste em usar vários modelos distintos em conjunto com sua diversidade criada sob amostras aleatórias criadas por *bootstrap*.

Desta forma, o algoritmo constrói um grande número de árvores de decisão que realizam previsões individuais, que serão utilizadas conjuntamente na classificação de determinado elemento dentro de um conjunto dados. A alocação é definida por meio do "voto" de determinada árvore classificadora a determinada solução (COSTA, 2012).

A grande vantagem do método está na precisão de suas estimativas em bases de dados de diferentes tamanhos. Umas das principais desvantagens é a perda de interpretação do modelo, o que não acontece no método de árvore de decisão convencional.

## 4. ANÁLISE DE DADOS

### 4.1 PROCEDIMENTO

A obtenção dos resultados finais foi feita considerando duas etapas distintas. A primeira delas consiste na clusterização dos municípios brasileiros. Já na segunda etapa, foram obtidos os agrupamentos de setores censitários conforme suas características sociodemográficas. A Figura 1 ilustra a estratégia adotada.

| **ETAPA 1: MUNICÍPIOS (CIDADES)** | | **ETAPA 2: SETORES CENSITÁRIOS** |
|---|---|---|
| ***Cluster* 1** | | ***Cluster* 1.1** |
| ***Cluster* 2** | | ***Cluster* 1.2** |
| ***Cluster* 3** | | ***Cluster* 1.3** |
| ***Cluster* 4** | | ***Cluster* 1.4** |
| ***Cluster* 5** | | ***Cluster* 1.5** |
| ***Cluster* 6** | | ***Cluster* 1.6** |
| ***Cluster* 7** | | ... |
| ***Cluster* 8** | | |

**FIGURA 1**
Estratégia de clusterização.

É possível verificar, pela Figura 1, que a solução de *clusters* de setores censitários é dependente da solução obtida na primeira etapa. Desta forma, os resultados obtidos tiveram maior consistência.

Na primeira etapa, a divisão dos municípios é feita com base nas informações de PIB, População, IDH, Desigualdade (Gini) e localização em região metropolitana. A primeira quebra ocorre utilizando a informação e região metropolitana.

Em um segundo passo, os municípios de regiões metropolitanas (RMs) e não metropolitanas (RNMs) são agrupados de acordo com sua representatividade econômica, para cada um dos grupos. Por último foi utilizado o algoritmo *two-step cluster* para divisão e o *Random Forest* para classificação, considerando as informações restantes.

Na segunda etapa, foram utilizadas informações sociodemográficas do Censo 2010. Dentre elas se destacam:

- Renda do responsável pelo domicílio.
- Quantidade de pessoas vivendo no mesmo domicílio.
- Faixa etária dos componentes do domicílio.
- Raça.
- Tipo de domicílio (casa ou apartamento).
- Condição do domicílio (próprio ou alugado).

Estas informações foram recodificadas para a modelagem estatística, conforme a opinião de especialistas. A primeira quebra utilizada foi pela informação de renda. Foi aplicado o algoritmo *two-step cluster* para divisão e *Random Forest* para classificação em cada grupo de setores pertencentes a determinada combinação de *cluster* município e categoria de renda de cada setor.

Ao todo, foram encontrados 33 agrupamentos de setores censitários. A aplicação deste conhecimento será ilustrada por meio de exemplos apresentados na seção: "Caracterização do Brasil (setores censitários)".

## 4.2 CARACTERIZAÇÃO DOS MUNICÍPIOS

No total, foram identificados oito *clusters* de municípios (cidades) com características bastante diferentes entre si, conforme pode ser visto na Tabela 1.

**TABELA 1**
Estatísticas dos *Clusters* de Município.

| GRUPO DE MUNICÍPIOS | MUNICÍPIOS | | SETORES CENSITÁRIOS | | POPULAÇÃO | | | PIB | | PIB PER CAPTA (X 1000) | GINI MÉDIO | IDHM MÉDIO Médio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | QUANT. | % | QUANT. | % | QUANT. | % | MÉDIA | QUANTIDADE | % | | | |
| 1. Grandes metrópoles brasileiras | 73 | 1,3% | 89.898 | 29,7% | 61.599.712 | 32,5% | 843.832 | 1.787.274.654 | 47,4% | 29,01 | 0,55 | 0,78 |
| 2. Cidades de RMs com alto PIB e porte médio | 59 | 1,1% | 16.836 | 5,6% | 11.557.975 | 6,1% | 195.898 | 295.074.117 | 7,8% | 25,53 | 0,46 | 0,73 |
| 3. Cidades de RMs com baixo PIB e menor qualidade de vida | 300 | 5,4% | 20.444 | 6,7% | 13.234.213 | 7,0% | 44.114 | 142.569.979 | 3,8% | 10,77 | 0,50 | 0,66 |
| 4. Cidades pequenas nas regiões metropolitanas | 229 | 4,1% | 5.302 | 1,7% | 2.699.107 | 1,4% | 11.786 | 46.758.675 | 1,2% | 17,32 | 0,42 | 0,71 |
| 5. Cidades de médio porte de RNM | 308 | 5,5% | 50.427 | 16,6% | 34.227.730 | 18,1% | 111.129 | 646.247.914 | 17,1% | 18,88 | 0,55 | 0,68 |
| 6. Cidade de RNM com melhor qualidade de vida | 673 | 12,1% | 42.692 | 14,1% | 24.488.618 | 12,9% | 36.387 | 531.327.214 | 14,1% | 21,70 | 0,48 | 0,72 |
| 7. Cidades pequenas e muito pobres de RNM | 1.676 | 30,1% | 46.045 | 15,2% | 27.039.680 | 14,3% | 16.133 | 145.057.043 | 3,8% | 5,36 | 0,54 | 0,59 |
| 8. Cidades RNM com baixo PIB e mais desenvolvidas | 2.247 | 40,4% | 31.534 | 10,4% | 14.620.314 | 7,7% | 6.507 | 175.775.266 | 4,7% | 12,02 | 0,46 | 0,68 |
| **TOTAL** | **5.565** | **100,0%** | **303.178** | **100,0%** | **189.467.349** | **100,0%** | **34.046** | **3.770.084.862** | **100,0%** | **19,90** | **0,49438095** | **0,66** |

O *cluster* relativo às grandes metrópoles brasileiras é composto por 73 municípios que produzem quase 50% de toda a riqueza do país. Os *clusters* 7 e 8, apresentam mais de 70% da quantidade de

municípios e têm uma representatividade do PIB somada inferior a 10%. Mesmo não sendo o foco deste trabalho, o questionamento em relação à quantidade total de municípios existentes no Brasil se torna inevitável.

Também é possível verificar que, de forma geral, a qualidade de vida é superior nas regiões metropolitanas comparada às regiões não metropolitanas. Existem dois *clusters* de municípios de regiões não metropolitanas de alta representatividade no PIB nacional, *clusters* 5 e 6, que, somados, representam mais de 30% do PIB.

Foi observado também que existe um grande número de cidades pequenas em regiões metropolitanas cuja representatividade econômica e populacional é mínima.

Vale ressaltar que, neste trabalho, não foram considerados os setores censitários localizados em favelas, assim como os que não possuem informação completa divulgada pelo IBGE, que não divulga informações de setores muito pequenos para evitar a identificação dos respondentes. A inclusão dos setores de favelas poderá ser realizada em um estudo futuro

Analisando o mapa do Brasil (Figura 2), é possível verificar a forte desigualdade econômica existente no país. A maior parte da riqueza ainda está concentrada nas cidades.

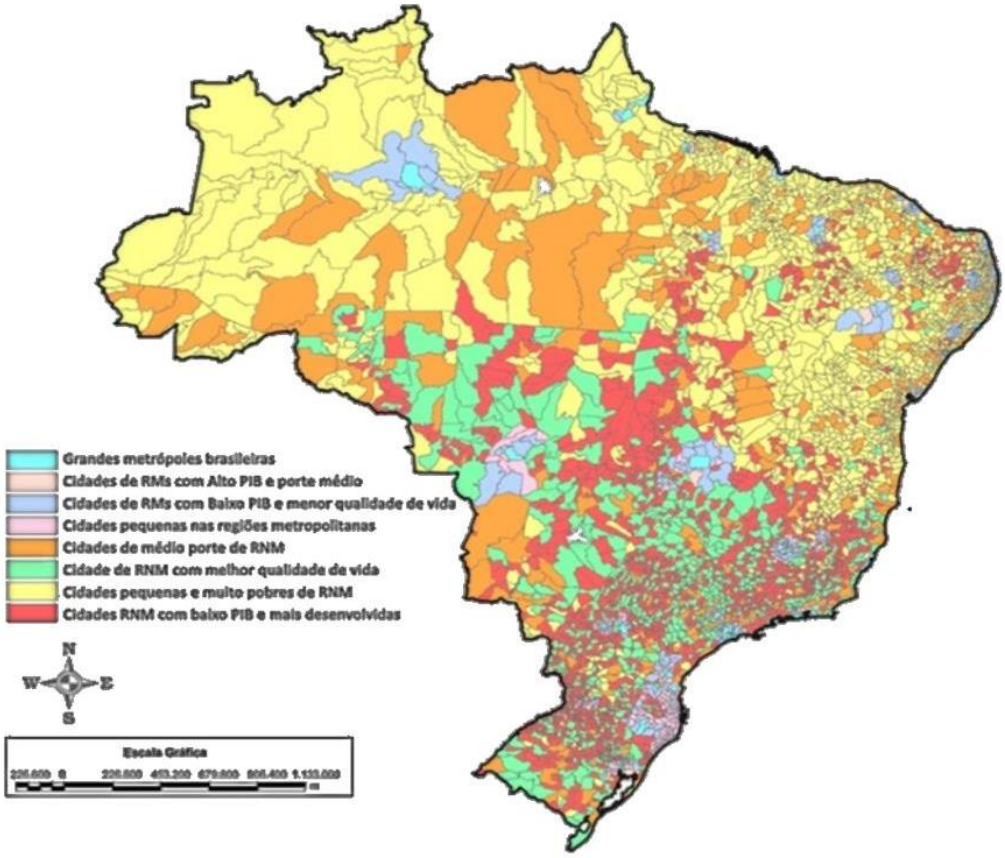


**FIGURA 2**
Mapeamento dos *clusters* de municípios.

## 4.3 CARACTERIZAÇÃO DO BRASIL (SETORES CENSITÁRIOS)

Após a aplicação dos filtros mencionados anteriormente, foi obtido um total de 287.425 setores censitários a serem clusterizados.

Um ponto interessante a ser ressaltado é a codificação das variáveis. A distribuição da população pelos setores censitários no Brasil apresenta comportamento bastante concentrado, conforme pode ser visto Figura 3.

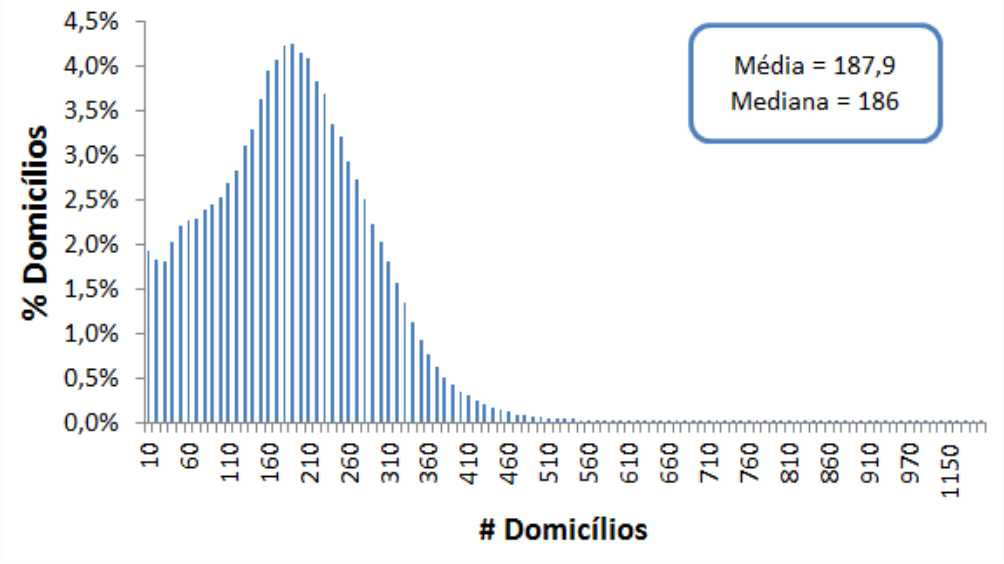


**FIGURA 3**
Distribuição da quantidade de domicílios por setor censitário.

A média de domicílios por setor é de 187,9 enquanto a mediana foi de 186, indicando que a distribuição apresenta simetria em relação às medidas centrais.

O coeficiente de variação encontrado foi de 0,52, o que indica uma variação inferior a um desvio-padrão. Este é um resultado importante, pois com base nele foi possível realizar padronizações de variáveis no nível de setores censitários.

Caso a variação fosse muito grande, algumas das transformações utilizadas não seriam possíveis, pois se estaria cometendo padronizações incoerentes.

Um ponto pertinente a ressaltar ao se trabalhar com informações ao nível de setor censitário está no fato que muito dificilmente um setor irá apresentar comportamentos extremos em relação aos outros quando se avaliam informações de sexo, idade ou quantidade de pessoas no domicílio.

A probabilidade de ocorrência de um setor que apresente 100% de homes ou mulheres ou apenas pessoas dentro de determinada faixa etária é baixíssima. Logo, variações nas médias aparentemente pequenas, já implicam em um perfil diferenciado.

Outro aspecto importante relacionado à baixa variabilidade desta distribuição é poder assumir, de maneira genérica que, quanto maior a quantidade de setores censitários, maior será quantidade de domicílios e, consequentemente, a população da localidade.

Para fins analíticos, os setores censitários foram categorizados conforme a distribuição etária de sua população. Na Figura 4, é possível perceber que foram identificados quatro tipos diferentes de padrão.

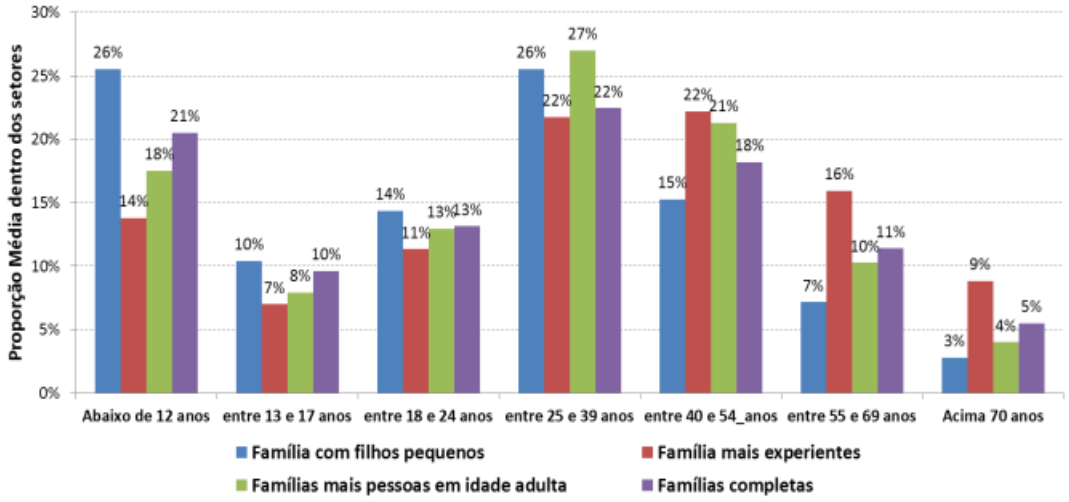


**FIGURA 4**
Distribuição da quantidade de domicílios por setor censitário.

O primeiro deles, chamado de família com filhos pequenos, apresentam mais de 50% de seus integrantes com idade inferior a 40 anos, sendo 26% com idade inferior a 12 anos. Por esse motivo, pode-se concluir que se trata de setores censitários em que provavelmente existe uma presença maior de crianças.

O segundo foi denominado de famílias mais experientes, em que a proporção de pessoas com mais de 55 anos é bem superior das demais distribuições. Este grupo apresenta, em média, 9% de sua população com mais de 70 anos.

O terceiro grupo apresenta maior proporção de pessoas com idades mais centrais, em média 48% têm entre 25 e 54 anos.

O último grupo é relativo a famílias completas, pois apresentou uma distribuição mais próxima a média geral, ainda com uma forte proporção de crianças.

Na Figura 5, é possível verificar a distribuição dos segmentos de setor censitário pelos grupos de municípios encontrados.

Conclui-se que, a maior quantidade de segmentos de setor está localizada nas grandes metrópoles brasileiras que concentram quase a metade da riqueza do país, 13 dos 33 segmentos de setor censitário se localizam nelas.

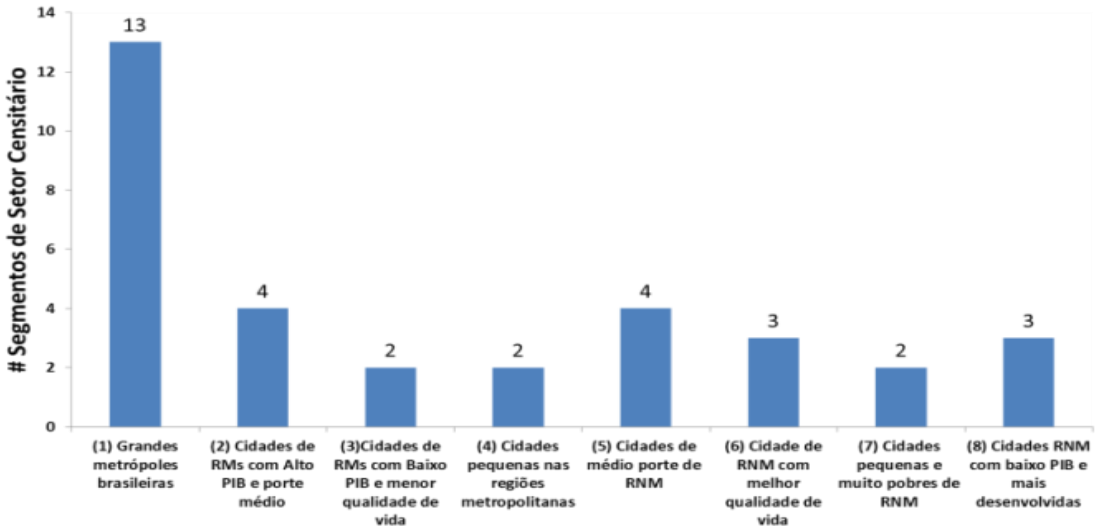


**FIGURA 5**
Quantidade de segmentos de setor censitário.

Os segmentos finais obtidos pelo estudo podem ser visualizados pela Tabela 2. Foi possível verificar que alguns segmentos acumularam uma grande quantidade de setores censitários. Um possível desenvolvimento futuro implicará na tentativa de detalhamento maior desses *clusters* visando à identificação de nichos de mercado específicos.

Para fins analíticos, classificou-se a renda média do responsável pelo domicílio da seguinte maneira: renda muito baixa, valores abaixo de R$ 1.000,00; renda baixa, entre R$ 1.000,00 e R$ 2.000,00; renda média, entre R$ 2.000,00 e R$ 4.000,00; renda alta, entre R$ 4.000,00 e R$ 8.000,00; renda muito alta, valores acima de R$ 8.000,00.

Já em relação à quantidade média de pessoas por domicílio (ppd), assumiu-se como alta quando a média foi superior a 3; média, entre 2,5 e 3 e muito baixa, quando a média ficou abaixo de 2,5. A Tabela 2 contém estatísticas descritivas dos agrupamentos de setores considerando as informações de renda média, quantidade de domicílios e perfil das famílias com maior penetração nos segmentos.

Verificou-se uma grande diferença em relação à renda nos setores censitários. Os segmentos 11,12 e 13 localizados nas grandes metrópoles brasileiras são os únicos que apresentaram renda muito alta (acima de R$ 8.000,00). Nas demais localidades, a renda foi muito baixa, baixa ou média. Este resultado ressalta que o grande abismo existente entre ricos e pobres no Brasil ainda se mantém até os dias de hoje.

Geralmente, setores classificados como de famílias com filhos possuem maior quantidade de pessoas no domicílio e renda mais baixa. A inversão deste fenômeno ocorre apenas nos segmentos de renda muito alta, em que a renda é mais alta quando a quantidade de pessoas no domicílio é maior. Os *clusters* 3, 22, 30 e 31 foram os mais representativos em relação ao total de setores censitários. A título de ilustração, serão apresentados três exemplos práticos da utilidade do presente estudo.

**TABELA 2**
Estatísticas dos *Clusters* de Setores Censitários.

| CÓDIGO *CLUSTER* SETOR | CÓDIGO *CLUSTER* MUN. | DESCRIÇÃO | SETOR CENSITÁRIO | | MÉDIA PESSOAS DOMIC. | RENDA MÉDIA | % FAM. FILHOS PEQ. | % FAM. + EXPERI-ENTES | % FAM. + ADULTOS | % FAM. COMPL. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | QUANT. | % | | | | | | |
| 1 | 1 | Família com filhos pequenos e renda muito baixa | 19.641 | 6,8% | 3,46 | 756 | 57,4% | 8,2% | 16,4% | 18,0% |
| 2 | 1 | Família de renda baixa e filhos pequenos | 4.392 | 1,5% | 3,36 | 1.204 | 65,8% | 12,6% | 13,6% | 8,0% |
| 3 | 1 | Famílias mais experientes com renda baixa | 23.474 | 8,2% | 3,25 | 1.284 | 12,7% | 39,6% | 31,2% | 16,6% |
| 4 | 1 | Famílias mais experientes com renda baixa e domicílios maior densidade demográfica | 5.431 | 1,9% | 2,84 | 1.472 | 24,8% | 38,7% | 33,2% | 3,3% |
| 5 | 1 | Famílias mais experientes com renda média | 10.181 | 3,5% | 2,67 | 3.192 | 9,3% | 75,8% | 14,1% | 0,7% |
| 6 | 1 | Famílias mais experientes com renda média e alta quantidade de pessoas por domicílio | 5.682 | 2,0% | 3,22 | 2.792 | 8,1% | 62,1% | 26,7% | 3,1% |
| 7 | 1 | Famílias mais experientes com renda média e baixa quantidade de pessoas por domicílio | 1.756 | 0,6% | 2,39 | 3.032 | 17,5% | 33,5% | 46,6% | 2,3% |
| 8 | 1 | Famílias mais experientes com renda alta e alta quantidade de pessoas por domicílio | 784 | 0,3% | 3,06 | 5.908 | 9,9% | 65,4% | 23,7% | 0,9% |
| 9 | 1 | Famílias mais experientes com renda alta e média quantidade de pessoas por domicílio | 2.015 | 0,7% | 2,86 | 5.973 | 7,5% | 70,0% | 21,9% | 0,6% |
| 10 | 1 | Famílias mais experientes com renda alta e muito baixa quantidade de pessoas por domicílio | 1.446 | 0,5% | 2,33 | 5.951 | 0,0% | 100,0% | 0,0% | 0,0% |
| 11 | 1 | Famílias mais experientes com renda muito alta e alta quantidade de pessoas por domicílio | 1.156 | 0,4% | 3,27 | 10.164 | 5,6% | 66,8% | 26,7% | 0,9% |
| 12 | 1 | Famílias mais experientes, renda muito alta e média quantidade de pessoas por domicílio | 665 | 0,2% | 2,85 | 9.559 | 0,0% | 100,0% | 0,0% | 0,0% |
| 13 | 1 | Famílias mais experientes, renda muito alta e baixa quantidade de pessoas por domicílio | 673 | 0,2% | 2,40 | 9.014 | 1,5% | 95,5% | 3,0% | 0,0% |
| 14 | 2 | Família com filhos pequenos, renda muito baixa e alta quantidade de pessoas por domicílio | 2.634 | 0,9% | 3,55 | 656 | 94,5% | 1,7% | 1,2% | 2,5% |
| 15 | 2 | Família com filhos pequenos + completas e renda muito baixa | 4.573 | 1,6% | 3,30 | 716 | 34,9% | 9,6% | 15,7% | 39,8% |
| 16 | 2 | Famílias diversas, renda média e alta quantidade de pessoas por domicílio | 878 | 0,3% | 2,79 | 2.220 | 24,6% | 42,0% | 30,6% | 2,7% |
| 17 | 2 | Famílias diversas, renda baixa e alta quantidade de pessoas por domicílio | 7.929 | 2,8% | 3,20 | 1.330 | 26,9% | 26,1% | 28,4% | 18,6% |
| 18 | 3 | Família com filhos pequenos, renda muito baixa e alta quantidade de pessoas por domicílio | 12.645 | 4,4% | 3,53 | 641 | 56,9% | 10,2% | 6,0% | 26,9% |
| 19 | 3 | Famílias diversas, com renda baixa e alta quantidade de pessoas por domicílio | 7.034 | 2,4% | 3,17 | 1.431 | 24,6% | 33,4% | 24,7% | 17,3% |
| 20 | 4 | Famílias diversas, renda muito baixa e alta quantidade de pessoas por domicílio | 2.503 | 0,9% | 3,26 | 765 | 25,3% | 32,6% | 7,7% | 34,3% |
| 21 | 4 | Famílias diversas, renda baixa e alta quantidade de pessoas por domicílio | 2.782 | 1,0% | 3,14 | 1.322 | 19,1% | 38,0% | 21,2% | 21,7% |
| 22 | 5 | Família com filhos pequenos, renda muito baixa e alta quantidade de pessoas por domicílio | 25.613 | 8,9% | 3,61 | 635 | 57,7% | 10,0% | 5,5% | 26,7% |
| 23 | 5 | Famílias diversas, renda baixa e alta quantidade de pessoas por domicílio | 17.631 | 6,1% | 3,18 | 1.335 | 24,0% | 36,0% | 23,3% | 16,7% |
| 24 | 5 | Famílias mais experientes, renda média e média quantidade de pessoas por domicílio | 2.824 | 1,0% | 2,64 | 3.495 | 12,9% | 64,6% | 20,3% | 2,2% |
| 25 | 5 | Famílias mais experientes, renda média e alta quantidade de pessoas por domicílio | 3.154 | 1,1% | 3,10 | 3.301 | 9,8% | 63,8% | 22,7% | 3,6% |
| 26 | 6 | Famílias diversas, renda muito baixa e alta quantidade de pessoas por domicílio | 19.962 | 6,9% | 3,26 | 771 | 31,9% | 27,9% | 8,8% | 31,4% |
| 27 | 6 | Famílias diversas, renda baixa e alta quantidade de pessoas por domicílio | 19.225 | 6,7% | 3,10 | 1.330 | 19,4% | 43,7% | 21,1% | 15,8% |
| 28 | 6 | Famílias mais experientes, renda média e média quantidade de pessoas por domicílio | 3.261 | 1,1% | 2,85 | 2.872 | 6,4% | 73,3% | 17,7% | 2,7% |
| 29 | 7 | Fam. com filhos pequenos, + completas, renda muito baixa, alta quant. pessoas p/ domic. raças div. | 12.282 | 4,3% | 3,73 | 550 | 43,8% | 13,9% | 2,2% | 40,1% |
| 30 | 7 | Fam. com filhos peq., + completas, renda muito baixa, alta quant. pessoas p/ domic. raça = pardos | 33.681 | 11,7% | 3,84 | 466 | 61,1% | 7,0% | 1,1% | 30,8% |
| 31 | 8 | Famílias + experientes, + completas, renda muito baixa e alta quantidade de pessoas por domicílio | 23.532 | 8,2% | 3,22 | 726 | 19,0% | 41,1% | 5,5% | 34,4% |
| 32 | 8 | Famílias diversas, renda baixa e alta quantidade de pessoas por domicílio | 3.907 | 1,4% | 3,14 | 1.241 | 22,9% | 15,5% | 24,3% | 37,3% |
| 33 | 8 | Famílias experientes, renda baixa e média quantidade de pessoas por domicílio | 4.079 | 1,4% | 2,93 | 1.408 | 0,0% | 100,0% | 0,0% | 0,0% |

### 4.3.1 EXEMPLO PRÁTICO 1: DIRECIONAMENTO DE PESQUISA

Uma agência de viagem, localizada na cidade São Paulo e voltada para clientes de alta renda, quer ampliar seu portfólio de destinos turísticos para pessoas da terceira idade. Por isso eles pretendem realizar uma pesquisa de opinião, para identificar os destinos e estilos de viagem mais promissores para este público.

Direcionando a pesquisa para o grupo "Famílias mais experientes, com renda muito alta e alta quantidade de pessoas por domicílio" foi possível mapear o público mais recomendado para ser abordado (Figura 6), não só na cidade, mas em toda a região metropolitana de São Paulo.

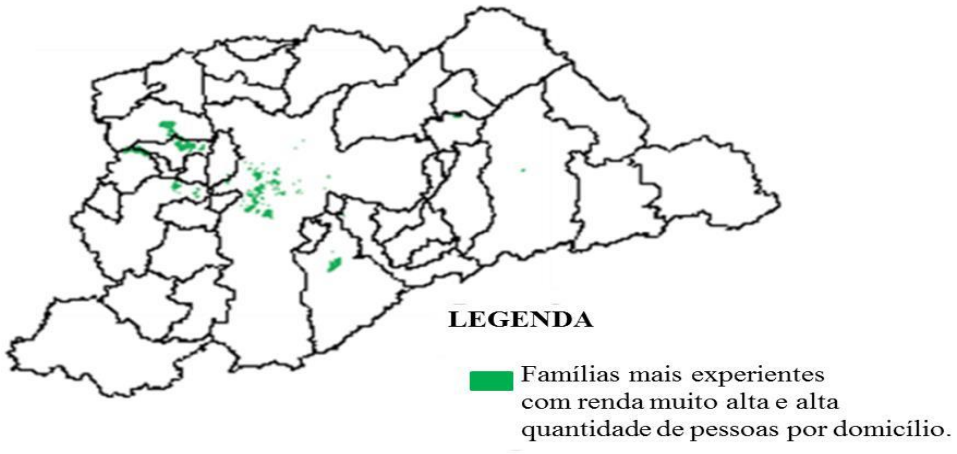


**FIGURA 6**
Mapa da Região Metropolitana do Estado de São Paulo.

Assim, as regiões com maior presença deste grupo são Pinheiros e Barueri, dado que estas pessoas representam 9% da população destas regiões, enquanto nas demais regiões (RMSP) este percentual é de 0,5%.

Isso permitirá minimizar o custo e otimizar a alocação dos entrevistadores, sem comprometer os resultados, uma vez que há uma maior chance de que as pessoas que participaram da pesquisa sejam, de fato, os consumidores finais deste novo portfólio.

### 4.3.2 EXEMPLO PRÁTICO 2: ABERTURA DE NOVAS UNIDADES DE NEGÓCIO

O Supermercado X quer abrir, em Porto Alegre, uma nova loja especializada em comida pré-pronta, de alta qualidade. Eles acreditam que exista um grupo de pessoas das classes A e B que trabalham fora e que não gostam, ou não tenham muito tempo para cozinhar em casa durante a semana.

Estas lojas atenderiam às necessidades destas pessoas, mas os donos do Supermercado X veem a distância das suas lojas como um limitador da sua estratégia. Por isso, precisavam mapear a localização do seu público-alvo para identificar o endereço mais estratégico para a abertura da sua nova unidade.

O grupo de "Famílias mais experientes, renda muito alta e baixa quantidade de pessoas por domicílio" (Figura 7), foi identificado como o público-alvo para a estratégia deste supermercado.

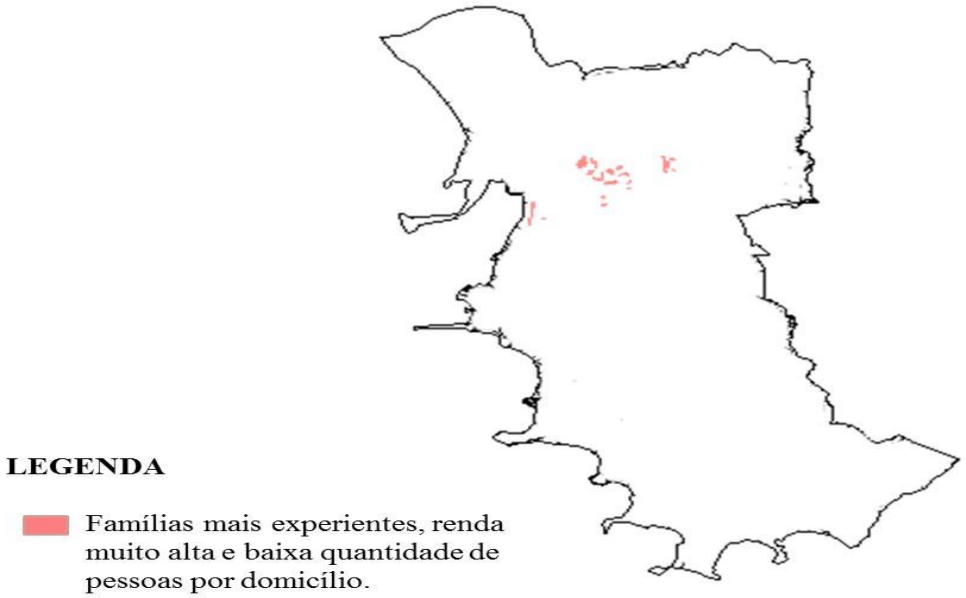


**FIGURA 7**
Mapa do município de Porto Alegre/RS.

Após analisar a dispersão deste grupo de pessoas em Porto Alegre, o supermercado escolheu a região do centro de Porto Alegre perto dos bairros Moinho de Vento, Bela Vista e Petrópolis, pois estão a uma distância próxima e o público-alvo representa 16% da população destes bairros.

### 4.3.3 EXEMPLO PRÁTICO 3: PLANEJAMENTO SOCIAL

Imaginando que os governantes do Estado do Piauí desejam entender quais são as áreas com maiores dificuldades quanto à infraestrutura para construção de escolas. Este estudo permitirá verificar se existem regiões que concentram grupos de pessoas que mais se beneficiariam destas novas escolas, para direcionar os esforços.

Com este intuito, mapeou-se o grupo de "Famílias com filhos pequenos, mais completas, renda muito baixa, alta quantidade de pessoas por domicílio e raça = pardos" que são mais carentes de infraestrutura e que se beneficiariam muito desta iniciativa.

Contudo, como pode ser visto na Figura 8, 40% dos setores censitários do Estado do Piauí possuem esta característica e estes setores estão dispersos em 142 dos 224 municípios (63%) do estado.

Apesar das famílias de baixíssima renda e com muitos filhos estarem espalhados pelo Estado do Piauí, considerou-se mais crítica a condição dos habitantes dos municípios do extremo norte e extremo sul do estado. Isto porque, os demais setores, localizados em municípios das áreas mais centrais do estado, podem utilizar a infraestrutura de municípios vizinhos, mesmo como medida paliativa, ao passo que os municípios mais críticos não têm opções no entorno.

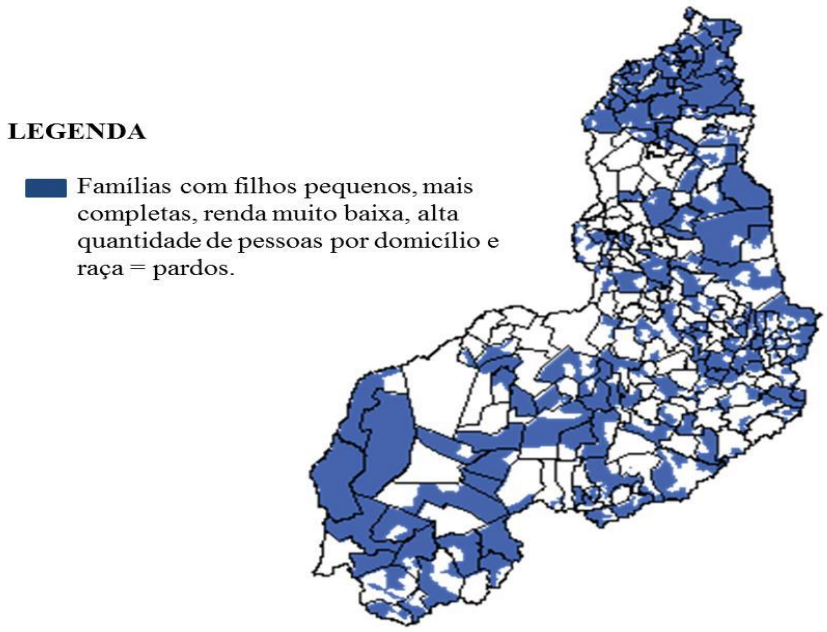


**FIGURA 8**
Mapa dos Municípios do Estado do Piauí.

Assim, conclui-se que é preciso começar a construção de escolas nas regiões norte e sul do Estado do Piauí e "caminhar" em direção ao centro do estado, até atender a toda a população carente.

## 5. CONCLUSÕES

Por meio da metodologia desenvolvida foi possível descobrir segmentos populacionais com características muito diferentes entre si.

Existem grandes diferenças quanto ao poder aquisitivo da população. Segmentos com renda média acima de R$ 8.000,00 aparecem apenas no segmento de município das grandes metrópoles brasileiras, composto por 73 cidades cuja representatividade econômica é próxima de 47,4% do PIB nacional.

De forma geral, quanto maior for a quantidade de pessoas no domicílio, menor será a renda. Contudo, esta generalização não é verdadeira quando são avaliados os segmentos de maior poder aquisitivo, em que a maior renda média foi do segmento que apresentou também a maior quantidade de pessoas por domicílio.

A maior parte da extensão territorial do é formada por cidades pequenas e muito pobres de regiões não metropolitanas, principalmente Norte e Nordeste, tendo segmentos populacionais com renda muito baixa (abaixo de R$ 1.000,00) ou baixa (entre R$ 1.000,00 e R$ 2.000,00) e composta, em sua maioria, por famílias com filhos pequenos.

## 6. LIMITAÇÕES DA PESQUISA E SUGESTÕES PARA NOVAS PESQUISAS

Foram sugeridos alguns exemplos simulando a utilização desta segmentação em problemas reais de empresas privadas e órgãos públicos. Contudo, trata-se apenas de uma pequena amostra do potencial de aplicação das informações obtidas neste trabalho.

Em desenvolvimentos futuros, deve ser feita a inserção dos setores censitários referentes às favelas, assim como a inclusão de informações atitudinais de pesquisas IBOPE (*Target Group Index* e Conectaí) visando melhor caracterização dos segmentos encontrados. Dentre os segmentos que tiveram representatividade populacional muito grande, será buscado um detalhamento maior.

## 6. REFERÊNCIAS BIBLIOGRÁFICAS

BREIMAN, L. *Random Forests*. Machine Learning. 45 (1): 5–32. DOI:10.1023/A: 1010933404324 2001.

BRITO, W. M.; SEMAAN, G. S.; BRITO, J. A. de M. *Um algoritmo genético para o problema dos k-médoides*. Fortaleza, 2011.

DIAS, Sérgio R. *Gestão de marketing*. 2. ed. São Paulo: Saraiva, 2004.

CHURCHILL JR., Gilbert A.; PETER, J. Paul. *Marketing*: criando valor para os clientes. São Paulo: Saraiva, 2000.

CLARITAS.COM. Disponível em: <http://www.claritas.com/MyBestSegments/Default.jsp>. Acessado em: 15 jan. 2014.

COSTA, H. S. da R. M. *Estudo comparativo de abordagens ao problema de débito de transações bancárias em contas com saldo insuficiente*. Dissertação (Mestrado em Engenharia Matemática) – Faculdade de Ciências da Universidade do Porto, Porto, 2012. [Orientador: Prof. Dr. Luis Torgo].

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. *Análise multivariada de dados*. Porto Alegre, Bookman, 2005.

INSTITUTO BRASILEIRO DE CIÊNCIAS ESTATÍSTICAS - IBGE. Censo Demográfico 2010 – Características da população e dos domicílios – Resultados do universo, Rio de Janeiro, 2011. Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/censo2010/caracteristicas_da_populacao/resultados_do_universo.pdf>. Acessado em: 5 fev. 2014.

KAUFMAN, L.; ROUSSEEUW, P. J. Finding Groups in Data. NewYork: John Wiley, 1990.

KAUFMAN, L.; ROUSSEEUW, P. J. *Clustering by means of medoids*. In: Statistical Data Analysis based on the L1 Norm, pp. 405-416. Amsterdam: Y. Dodge, 1987.

WARD, J. H. Hierarquical grouping to Optimize na objective Function. Disponível em: <http://iv.slis.indiana.edu/sw/data/ward.pdf>. Acessado em: 7 jan. 2014. *Journal of the American Statistical Association*. v. 58, n. 301, 1963, pp. 236-244.

WEISS, Michael J. *The clustering of America*. New York: Tilden, 1988.