UNIVERSIDADE FEDERAL DE ALFENAS
UNIFAL-MG

LARISSA GONÇALVES SOUZA

Agrupamento dos municípios do Sul/Sudoeste de Minas Gerais em relação ao envelhecimento populacional

ALFENAS - MG
2017

LARISSA GONÇALVES SOUZA

Agrupamento dos municípios do Sul/Sudoeste de Minas Gerais em relação ao envelhecimento populacional

Dissertação apresentada à Universidade Federal de Alfenas, como parte dos requisitos para obtenção do título de Mestre em Estatística Aplicada e Biometria.
Área de concentração: Estatística Aplicada e Biometria.
Linha de pesquisa: Modelagem Estatística e Estatística Computacional.

Orientadora: Dra. Patrícia de Siqueira Ramos
Coorientador: Dr. Lincoln Frias

ALFENAS - MG
2017

RESUMO

Nas últimas décadas, o Brasil tem experimentado transformações em seu regime demográfico, que conduzem ao envelhecimento populacional. Esse fenômeno não ocorre de maneira uniforme em todas as regiões do país, produzindo diferenciais demográficos. Nesse sentido, o objetivo deste trabalho é agrupar os municípios da mesorregião Sul/Sudoeste de Minas Gerais em relação ao envelhecimento populacional. Especificamente, identificar os grupos de municípios mais envelhecidos e menos envelhecidos com base em indicadores demográficos e o método não hierárquico das k-médias. As variáveis utilizadas foram: esperança de vida ao nascer, taxa de fecundidade total, mortalidade infantil, mortalidade até 5 anos de idade, razão de dependência, probabilidade de sobrevivência até 40 anos, probabilidade de sobrevivência até 60 anos e taxa de envelhecimento. Esses dados são provenientes do Censo Demográfico de 2010 do IBGE, consultados por meio do Atlas do Desenvolvimento Humano no Brasil. A linguagem R foi usada para implementar as análises por meio do programa RStudio. O método das k-médias propôs uma divisão dos municípios em quatro grupos. O G1 é o menos envelhecido e é composto por 17 municípios, o grupo G2, formado por 46 municípios, é considerado o mais envelhecido em relação aos demais. Nos grupos G3 e G4 estão 36 e 47 municípios, respectivamente, e são caracterizados por assumirem posições intermediárias entre os dois primeiros (G3 menos envelhecido que G4). Portanto, aproximadamente 64% dos municípios foram classificados nos grupos considerados mais envelhecidos (G2 e G4), o que corresponde a 93 dos 146 municípios.

Palavras-chave: Análise de agrupamento. Classificação. Transição demográfica.

LISTA DE TABELAS

Tabela 1 – Média das variáveis demográficas para o Brasil e para a mesorregião Sul/Sudoeste de Minas Gerais, 2010. 24

Tabela 2 – Siglas e descrições das variáveis demográficas do Atlas do Desenvolvimento Humano no Brasil. 45

Tabela 3 – Resumo estatístico das variáveis demográficas da mesorregião Sul Sudoeste de Minas Gerais, 2010. 52

Tabela 4 – Municípios com os melhores indicadores demográficos da mesorregião Sul/Sudoeste de Minas Gerais, 2010. 54

Tabela 5 – Municípios com os piores indicadores demográficos da mesorregião Sul/Sudoeste de Minas Gerais, 2010. 55

Tabela 6 – Resumo estatístico das variáveis demográficas dos grupos obtidos pelo método das k-médias. 62

Tabela 7 – Resumo estatístico das variáveis população (pop) e rendimento médio dos ocupados (renocup) dos grupos obtidos pelo método das k-médias. 66

Tabela 8 – Municípios da mesorregião Sul/Sudoeste de Minas Gerais. 76

Tabela 9 – Municípios da mesorregião Sul/Sudoeste de Minas Gerais classificados no grupo 1 (G1) pelo método das k-médias. 77

Tabela 10 – Municípios da mesorregião Sul/Sudoeste de Minas Gerais classificados no grupo 2 (G2) pelo método das k-médias. 77

Tabela 11 – Municípios da mesorregião Sul/Sudoeste de Minas Gerais classificados no grupo 3 (G3) pelo método das k-médias. 78

Tabela 12 – Municípios da mesorregião Sul/Sudoeste de Minas Gerais classificados no grupo 4 (G4) pelo método das k-médias. 78

LISTA DE FIGURAS

Figura 1 – Pirâmides etárias absolutas do Brasil, 1980-2050. 16
Figura 2 – Razão de dependência do Brasil, 1940 a 2050. 18
Figura 3 – Índice de envelhecimento do Brasil, 1950 a 2000. 19
Figura 4 – Evolução da proporção de idosos com 60 anos ou mais na população brasileira, 1991 a 2060. 20
Figura 5 – Mapa das mesorregiões de Minas Gerais. 22
Figura 6 – Mapa da mesorregião Sul/Sudoeste de Minas Gerais. 23
Figura 7 – Diagrama de dispersão de dados fictícios de idade e renda de várias pessoas. 29
Figura 8 – Terminologia utilizada na descrição de dendrogramas. 32
Figura 9 – Ilustração de corte no dendrograma com 23 observações. 40
Figura 10 – Ilustração de corte no dendrograma com 20 observações. 41
Figura 11 – Fluxograma da metodologia aplicada no trabalho. 52
Figura 12 – Correlações entre as variáveis (antes e após a retirada de variáveis). . . 56
Figura 13 – Dispersão dos municípios em função dos escores dos componentes principais. 58
Figura 14 – Dendrograma pelo método de Ward e distância de Mahalanobis. 58
Figura 15 – Dispersão dos municípios em função dos escores dos componentes principais dos quatro grupos obtidos pelo método das k-médias. 59
Figura 16 – Mapa dos municípios da mesorregião Sul/Sudoeste de Minas Gerais classificados em quatro grupos pelo método das k-médias. 60
Figura 17 – *Boxplots* dos grupos da SSM de acordo com as variáveis: esperança de vida ao nascer (espvida), taxa de fecundidade total (tft), mortalidade infantil (mort1), razão de dependência (rd), probabilidade de sobrevivência até 60 anos (sobre60) e taxa de envelhecimento (t_env) 61
Figura 18 – *Boxplots* dos grupos da SSM de acordo com as variáveis população (pop) e rendimento médio dos ocupados (renocup). 66

SUMÁRIO

1 **INTRODUÇÃO** 12
2 **REVISÃO DE LITERATURA** 14
2.1 TRANSIÇÃO DEMOGRÁFICA BRASILEIRA 14
2.2 ENVELHECIMENTO POPULACIONAL 19
2.3 CARACTERIZAÇÃO DA MESORREGIÃO SUL/SUDOESTE DE MINAS GERAIS 21
2.4 ESTATÍSTICA MULTIVARIADA E ANÁLISE DE AGRUPAMENTO 25
2.4.1 Distâncias 29
2.4.2 Técnicas hierárquicas aglomerativas 32
2.4.3 Técnicas não hierárquicas: k-médias 37
2.4.4 Número de grupos 39
3 **DADOS E METODOLOGIA** 44
3.1 DADOS 44
3.1.1 Relação das variáveis demográficas com o envelhecimento populacional 48
3.2 METODOLOGIA 49
4 **RESULTADOS E DISCUSSÃO** 52
4.1 ANÁLISE DESCRITIVA DAS VARIÁVEIS 52
4.2 AGRUPAMENTOS 57
5 **CONSIDERAÇÕES FINAIS** 70
REFERÊNCIAS 72
ANEXOS 76

1 INTRODUÇÃO

Nas últimas décadas, o Brasil tem passado de forma gradual e progressiva a apresentar uma nova configuração de seu regime demográfico, caracterizada pelo envelhecimento de sua população (CAMARANO, 2014). De forma geral, antes do início do processo, todo o país possuía um perfil demográfico com muitas mortes e muitos nascimentos, resultando em um baixo crescimento vegetativo e população predominantemente jovem. Iniciada a transição demográfica, primeiro a mortalidade e, em seguida, a fecundidade declinam, acarretando um alto crescimento populacional. Posteriormente, na sua etapa final, o crescimento é lento novamente, mas agora movendo-se para um cenário de baixa fecundidade, aumento da longevidade e população envelhecida (LEE, 2003).

Nas próximas décadas é esperado que o envelhecimento populacional se acentue no Brasil. A velocidade com que esse fenômeno acontece nos países em desenvolvimento é considerada preocupante. Isso ocorre porque os países desenvolvidos iniciaram o processo muito antes e de maneira mais lenta (LIMA-COSTA; VERAS, 2003). Portanto, a maioria deles teve tempo pra se ajustar à nova realidade de um país de idosos. A França e a Suécia, por exemplo, levaram, respectivamente, 115 anos e 85 anos para a proporção de idosos, com 65 anos e mais de idade, aumentar de 7% para 14%. Nos países em desenvolvimento, por sua vez, o cenário é diferente. No Brasil, a população idosa dobrou sua proporção de 7% para 14% em apenas 21 anos (DOBRIANSKY; SUZMAN; HODES, 2007). Esses dados mostram que, ainda que todos os países estejam passando por profundas transformações que levam ao envelhecimento da população, o fenômeno não ocorre de forma homogênea. O processo se inicia em momentos, magnitude e velocidade diferentes. Isso pode ser observado não só entre países distintos, mas também dentro de um mesmo país, pois o processo é desigual em relação a suas regiões, estados e municípios.

A importância de estudar esse fenômeno por região encontra-se no fato de que, de forma geral, o envelhecimento populacional exige uma redefinição de políticas públicas direcionadas para esse segmento populacional. Em termos de políticas públicas de saúde, por exemplo, o idoso surge como prioridade (MARIN; PANES, 2015). Nesse contexto, é necessário conhecimento sobre todos os aspectos do problema de forma a caracterizar sua magnitude e características, que apoiarão as decisões sobre a alocação de recursos nesse novo cenário demográfico.

Atualmente, há razoável disponibilidade de dados demográficos de municípios do Bra-

sil, relacionados ao envelhecimento, principalmente através de pesquisas do Instituto Brasileiro de Geografia e Estatística (IBGE) (em especial, os Censos Demográficos e o Perfil dos Estados e dos Municípios Brasileiros). Porém, na maior parte das vezes, as variáveis disponíveis nesses bancos de dados são analisadas separadamente e uma visão geral pode ficar comprometida. Por isso, a análise multivariada é fundamental para a análise dos dados municipais, uma vez que a realidade dos municípios é multidimensional e muitas dessas variáveis estão intimamente relacionadas.

Nesse sentido, o objetivo deste trabalho é agrupar os municípios da mesorregião Sul/Sudoeste de Minas Gerais em relação ao processo de envelhecimento populacional. Especificamente, identificar os grupos de municípios mais envelhecidos e menos envelhecidos com base em indicadores demográficos e o método não hierárquico das k-médias. Dessa forma, os agrupamentos obtidos podem subsidiar a tomada de decisões sobre políticas públicas específicas para cada grupo de municípios.

2 REVISÃO DE LITERATURA

O objetivo dessa seção é apresentar a transição demográfica brasileira, o seu envelhecimento populacional e a caracterização da mesorregião Sul/Sudoeste de Minas Gerais. Por fim, são apresentados aspectos da estatística multivariada e da análise de agrupamento.

2.1 TRANSIÇÃO DEMOGRÁFICA BRASILEIRA

O Brasil vivenciou grandes mudanças nas últimas décadas no que diz respeito à sua dinâmica demográfica. Em torno de 1960, a população crescia num ritmo acelerado, sendo um país jovem, enquanto que, cerca de cinquenta anos depois, vivencia uma desaceleração desse crescimento. A idade mediana era 18 anos em 1960 e passou para 27 anos em 2010. Isso é apenas uma indicação de fenômenos mais gerais que têm sido observados: uma brusca diminuição das taxas de fecundidade e mortalidade em todas as idades, envelhecimento da população, novos arranjos familiares se formando, modificações na magnitude e limites etários da população economicamente ativa, além de outras mudanças (VASCONCELOS; GOMES, 2012).

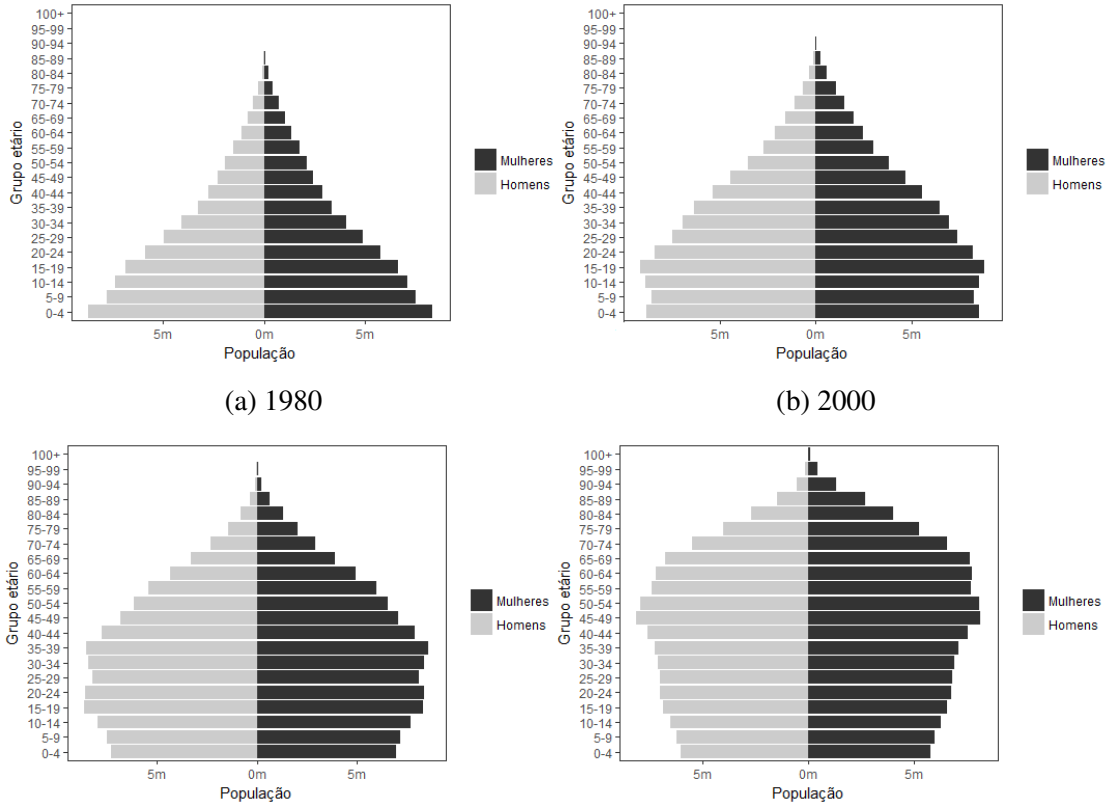
Até meados da década de 1940, o país se encontrava na "pré-transição demográfica", caracterizada por elevadas taxas de mortalidade e natalidade, o que resultava em um baixo crescimento vegetativo (diferença entre as taxas de natalidade e mortalidade). Nesse cenário, sua população era tipicamente jovem. Contudo, em virtude da evolução da medicina, urbanização, introdução dos antibióticos, melhoria nas condições sanitárias e difusão de novas tecnologias, o Brasil ingressou na primeira fase da transição, caracterizada pela diminuição dos níveis de mortalidade. Nesse período, de 1940 a 1970, o país experimentou uma redução acelerada da mortalidade, que conduziu ao aumento da esperança de vida e a um rápido crescimento populacional, principalmente nas décadas de 1950 e 1960. Como os níveis de fecundidade permaneceram elevados, enquanto a taxa de mortalidade decrescia, a taxa de crescimento da população brasileira se elevou significativamente nessa fase (CAMARANO, 2014).

Ainda de acordo com a autora, a partir de 1970, o Brasil experimentou a segunda fase da transição, caracterizada pela redução dos níveis de fecundidade. O processo ocorreu principalmente, devido à inserção da mulher no mercado de trabalho, mudanças econômicas e o

planejamento familiar. O resultado foi um crescimento vegetativo em níveis menores em relação à fase anterior e o início do processo de envelhecimento da população. Os primeiros países a experimentarem o processo de transição demográfica, localizados no oeste da Europa, demoraram mais de um século para reduzirem suas taxas de mortalidade e fecundidade e isso ocorreu devido à reduzida velocidade de queda dessas taxas (BORGES; CAMPOS; SILVA, 2015). Quando comparados aos países desenvolvidos é possível observar que os dois movimentos, tanto de redução da mortalidade quanto da fecundidade, ocorreram em um espaço de tempo muito curto no Brasil e em muitos dos outros países em desenvolvimento (CAMARANO, 2014).

A taxa de fecundidade total passou de 6,2 filhos/mulher, em 1950, para 1,7, em 2012, atingindo níveis inferiores do que o que garantiria a reposição da população que é de 2,1 filhos/mulher. Outra grande mudança ocorreu com a esperança de vida ao nascer, que era 45,4 anos em 1950, e hoje é 75,2 anos, graças à contínua queda dos níveis de mortalidade. Entretanto, essas transformações não ocorreram de forma uniforme em todas as regiões do país, produzindo diferenciais demográficos que resultam nas Unidades da Federação encontrarem-se em diferentes fases do processo (BORGES; CAMPOS; SILVA, 2015). Em 1970, as regiões Norte e Nordeste ainda apresentavam valores altos de mortalidade infantil e de número médio de filhos por mulher, enquanto as regiões Sudeste, Sul e Centro-Oeste já apresentavam queda nesses índices. Apesar da diminuição da taxa de mortalidade infantil ter tido diferentes ritmos nas cinco regiões, em todas houve uma queda de 70%, entre 1980 e 2010 (VASCONCELOS; GOMES, 2012). Além da redução dos níveis de mortalidade, houve uma mudança nos níveis de fecundidade. Em 2000, apenas a região Norte apresentava número médio superior a 3,0 filhos/mulher. Já em 2010, todas as outras regiões apresentavam níveis de fecundidade menores do que o nível de reposição, de 2,1 filhos por mulher (VASCONCELOS; GOMES, 2012).

A contínua redução dos níveis de fecundidade provoca modificações na estrutura etária da população, conduzindo ao processo de transição da estrutura etária. A queda da componente fecundidade altera a proporção de jovens e idosos de uma população. A partir da Figura 1, que ilustra as pirâmides etárias absolutas da população brasileira de 1980 a 2050, é possível ver a redistribuição dos grupos etários ao longo do anos.



(a) 1980 (b) 2000

(c) 2020 (d) 2050

Figura 1 – Pirâmides etárias absolutas do Brasil, 1980-2050.
Fonte: Da autora a partir de dados do *United States Census Bureau,* disponível em: *www.census.gov/population/international/data/idb*

Com a acelerada transformação na estrutura etária do país, a pirâmide etária de 1980, típica de uma população extremamente jovem, caracterizada por uma base larga e um topo estreito (muitas crianças e jovens e poucos idosos), está sendo gradualmente substituída por uma mais estreita na base e larga no topo, típica de uma população em processo de envelhecimento (CARVALHO; WONG, 2008). Atualmente, o segmento populacional que mais cresce no país é o de idosos, estima-se que entre 2012 e 2022, a taxa de crescimento desse segmento ultrapassem 4% (BORGES; CAMPOS; SILVA, 2015). Esse aspecto é visto como um desafio, pois o crescimento rápido de um segmento populacional não produtivo e o menor crescimento do segmento produtivo podem desequilibrar a divisão de recursos na sociedade, gerando sérios problemas econômicos e previdenciários.

As alterações nas relações intergeracionais podem ser analisadas também através da razão de dependência total (RDT) e do índice de envelhecimento (IE). A RDT corresponde à relação entre a população considerada inativa (crianças e jovens de 0 a 14 anos e idosos acima de 65 anos) e a população potencialmente ativa (adultos de 15 a 64 anos). O indicador mede,

em termos relativos, a parcela da população potencialmente inativa que deve ser sustentada pela potencialmente ativa. Quanto maior seu valor, maior o grau de dependência econômica da população. A RDT pode ainda ser decomposta em razão de dependência dos jovens (RDJ) e razão de dependência dos idosos (RDI). O índice de envelhecimento, por sua vez, mede o número de pessoas idosas de 65 ou mais anos de idade, para cada 100 crianças e jovens de 0 a 14 anos de idade. Assim quanto maior seu valor, mais envelhecida a população (CARVALHO; WONG, 2008).

Em virtude da transição da estrutura etária, no Brasil é esperado que a RDI (relação entre os idosos acima de 65 anos e dos adultos de 15 a 64 anos) alcance níveis mais elevados nos próximos anos, assim como haja uma considerável redução na RDJ (relação entre crianças e jovens de 0 a 14 anos e adultos de 15 a 64 anos), até sua estabilização (CARVALHO; WONG, 2008).

A Figura 2 mostra a série de razões de dependência do Brasil, no período de 1940 a 2050. Nas décadas de 1950 e 1960 houve um aumento da razão de dependência total relacionado, principalmente, com o aumento da razão de dependência dos jovens. Esse processo ocorreu devido à queda da mortalidade, que atingiu em um primeiro momento, prioritariamente os grupos etários das crianças (CAMARANO, 2014). Em 1960, a RDT alcançou 90 indivíduos inativos para cada 100 pessoas em idade ativa. Contudo, a partir de 1970, o indicador começa a reduzir continuamente até 2020. Os dados sugerem que esse processo está ocorrendo devido à queda da fecundidade, que conduz à redução dos níveis de natalidade e, consequentemente, diminui a parcela jovem da população, o que implica na redução da RDJ. De 1940 a 2020, a RDJ passará de 79 para 30 crianças e jovens para cada 100 pessoas potencialmente ativas.

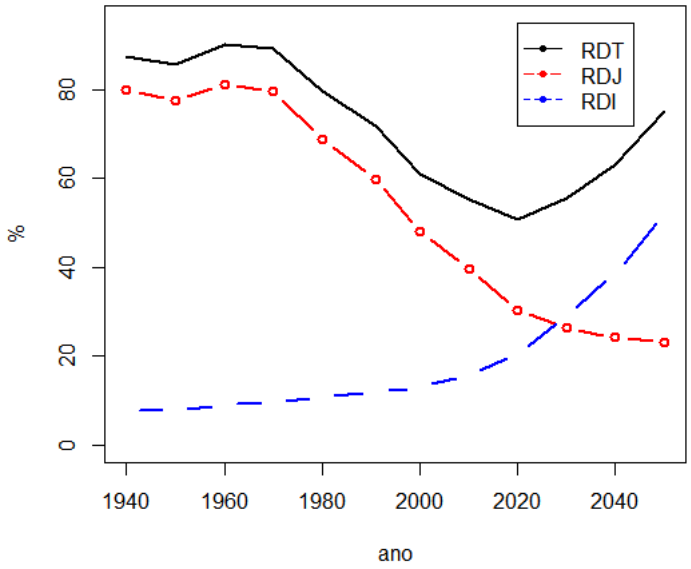


Figura 2 – Razão de dependência do Brasil, 1940 a 2050.
Fonte: Da autora a partir de dados do Instituto Brasileiro de Geografia e Estatística.

A RDI, por sua vez, nesse mesmo período se elevará de 9 para 20 idosos para cada 100 pessoas potencialmente ativas, alcançando 52, em 2050. A queda contínua da RDJ combinada com o aumento da RDI resultará no aumento da RDT, a partir de 2030. O processo implicará em menos trabalhadores ativos para cada inativo. Essa situação é preocupante devido ao pacto intergeracional do modelo previdenciário brasileiro, em que a geração de trabalhadores ativos custeia os benefícios pagos aos inativos (BRASIL, 2009).

Por último, a Figura 3 ilustra o índice de envelhecimento do Brasil no período de 1950 a 2050. Em 1950, havia 5 idosos de 65 ou mais anos de idade, para cada 100 indivíduos de 0 a 14 anos. Em 2050, as projeções indicam que o valor do indicador será aproximadamente 34 vezes maior, alcançando 172 idosos. A evolução do indicador aponta para o rápido processo de envelhecimento populacional, o que reforça a preocupação com os desafios relacionados à saúde e à assistência social gerados pelo processo de transição demográfica (CARVALHO; WONG, 2008).

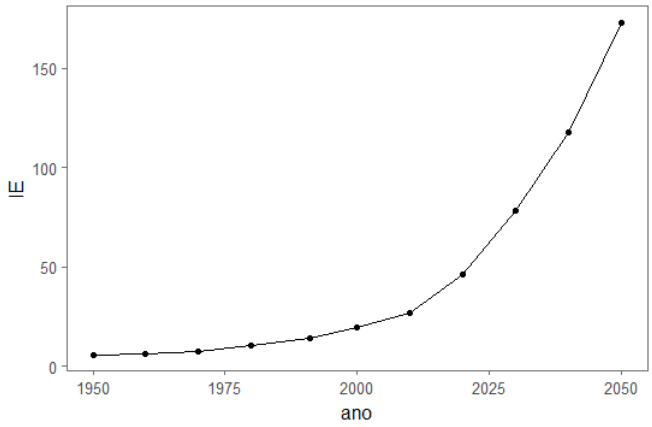


Figura 3 – Índice de envelhecimento do Brasil, 1950 a 2000.
Fonte: Da autora a partir de dados do Instituto Brasileiro de Geografia e Estatística.

2.2 ENVELHECIMENTO POPULACIONAL

O envelhecimento populacional corresponde ao aumento, em termos relativos, da população idosa. Portanto, o fenômeno está relacionado à mudança na estrutura etária da população. De fato, contradizendo o senso comum, o início do processo acontece com a queda sustentada dos níveis de fecundidade e não de mortalidade (CARVALHO; WONG, 2008). De acordo com Carvalho e Garcia (2003), a queda da mortalidade até o momento tem produzido um efeito de rejuvenescimento populacional. Isso ocorre porque incialmente a redução atingiu prioritariamente os mais jovens. Além disso, houve um aumento do número de mulheres sobreviventes até o final do período reprodutivo, o que conduziu a um aumento do número de nascimentos. O resultado foi uma mudança na estrutura etária no sentido de seu rejuvenescimento, com o aumento da proporção de jovens. Ainda segundo os autores, a redução da mortalidade contribui para o processo de envelhecimento apenas quando esta se concentrar nos grupos etários dos idosos.

Nas últimas décadas no Brasil, já tem sido observado um crescimento populacional mais elevado dos idosos em relação aos demais segmentos da população (CAMARANO, 2014). Ainda de acordo com a autora, essa alteração na estrutura etária produz uma série de preocupações para a sociedade. Com a contínua redução dos níveis de fecundidade tem sido observada uma redução da população em idade produtiva. Ao mesmo tempo, a redução da mortalidade contribuirá para que os idosos sobrevivam por um período de tempo maior. Esse cenário gera uma série de consequências para a sociedade, o Estado e as famílias.

A Figura 4 mostra a evolução da proporção de idosos com 60 anos ou mais no total da população do Brasil, no perído de 1991 a 2060. A proporção da população em idade avançada aumenta a cada ano, como consequência do processo de transição demográfica. Em 1991, 7,3% da população total era formada por homens e mulheres com 60 anos ou mais de idade, em 2060, esse valor aumentará para 33,7%. Esse comportamento associado à redução da proporção da população em idade ativa gera uma série de desafios à sociedade.

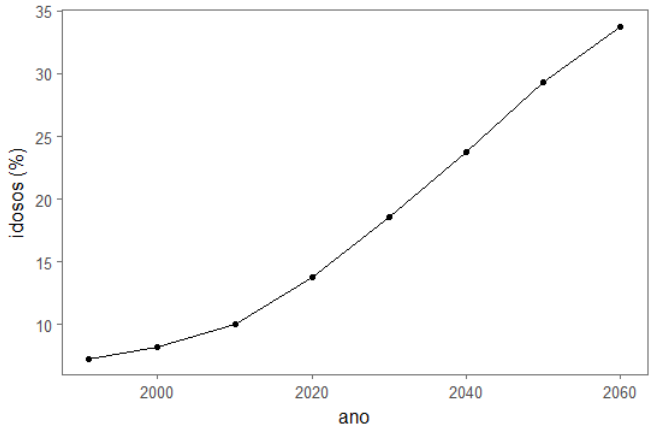


Figura 4 – Evolução da proporção de idosos com 60 anos ou mais na população brasileira, 1991 a 2060.

Fonte: Da autora a partir de dados do Instituto Brasileiro de Geografia e Estatística.

O futuro das aposentadorias é umas das principais preocupações geradas pelo envelhecimento populacional no Brasil. Isso acontece porque o principal regime previdenciário do país é o Regime Geral de Previdência Social (RGPS) (REIS; SILVEIRA; BRAGA, 2013), que possui caráter contributivo e é estruturado sob o modelo de repartição simples. A estrutura desse modelo de previdência garante que as contribuições dos trabalhadores ativos financiem os benefícios pagos aos inativos.

Além disso, a principal fonte de financiamento do sistema são as contribuições incidentes sobre a remuneração do trabalhadores, portanto a composição demográfica da população tem um impacto direto na sustentatibilidade financeira e atuarial do regime. Dessa forma, o desafio gerado pelo novo regime demográfico encontra-se na redução da relação entre trabalhadores ativos e inativos. Isso ocorre devido ao aumento da proporção de idosos em relação aos jovens. Portanto, haverá um número crescente de inativos sustentados por um número cada vez menor de ativos. Ao mesmo tempo, com o aumento da esperança de vida em todas as idades, não só a redução das contribuições será vista como um desafio, mas também o fato de que o beneficiário permanecerá recebendo o benefício por mais tempo (BRASIL, 2009).

Outra preocupação é o impacto do processo de envelhecimento populacional nos gastos com saúde. Os custos dos serviços de saúde geralmente são maiores para os idosos e isso pode ser explicado pelas maiores taxas de internação e o alto custo do tratamento de doenças crônicas (MARINHO; CARDOSO; ALMEIDA, 2014). Além disso, o número de idosos com doenças crônicas não letais tem crescido continuamente, o que sugere que eles necessitarão de cuidados com a saúde por um longo período. Outro desafio encontra-se no fato de que frequentemente os idosos doentes apresentam debilitações que os impedem de desenvolver atividades da vida diária, o que conduz a maior demanda por cuidadores de idosos. No entanto, essa necessidade acontece em um cenário em que o número de idosos aumenta e a oferta de possíveis cuidadores diminui, devido à queda da fecundidade.

Os pontos apresentados são apenas algumas das consequências trazidas com o envelhecimento, que impactam diretamente nas transferências de recursos do Estado para sociedade. A compreensão geral do processo pode auxiliar nessa tarefa. Portanto, questões referentes aos desafios gerados pelo processo têm sido frequentemente discutidos na literatura.

2.3 CARACTERIZAÇÃO DA MESORREGIÃO SUL/SUDOESTE DE MINAS GERAIS

O Brasil é formado por 137 mesorregiões, sendo 12 delas localizadas em Minas Gerais. A Figura 5 mostra o mapa das mesorregiões de Minas Gerais. Dentre elas, a SSM é a segunda em número de municípios no Brasil (possui 146, enquanto a Noroeste Rio-Grandense possui 216) e primeira em Minas Gerais. Os 146 municípios dessa mesorregião estão divididos em 10 microrregiões. Em termos de população, a mesorregião ocupa a 16ª posição (2.438.611 em 2010) no *ranking* brasileiro e 2ª posição em Minas Gerais, onde a metropolitana de Belo Horizonte é a mais populosa (6.236.117 no mesmo ano). Como o esperado, as mesorregiões mais populosas do país são aquelas onde estão localizadas as capitais dos estados. Com exceção dessas, apenas as mesorregiões de Campinas e do Centro Sul Baiano (onde estão Vitória da Conquista e Jequié) são mais populosas do que a SSM.

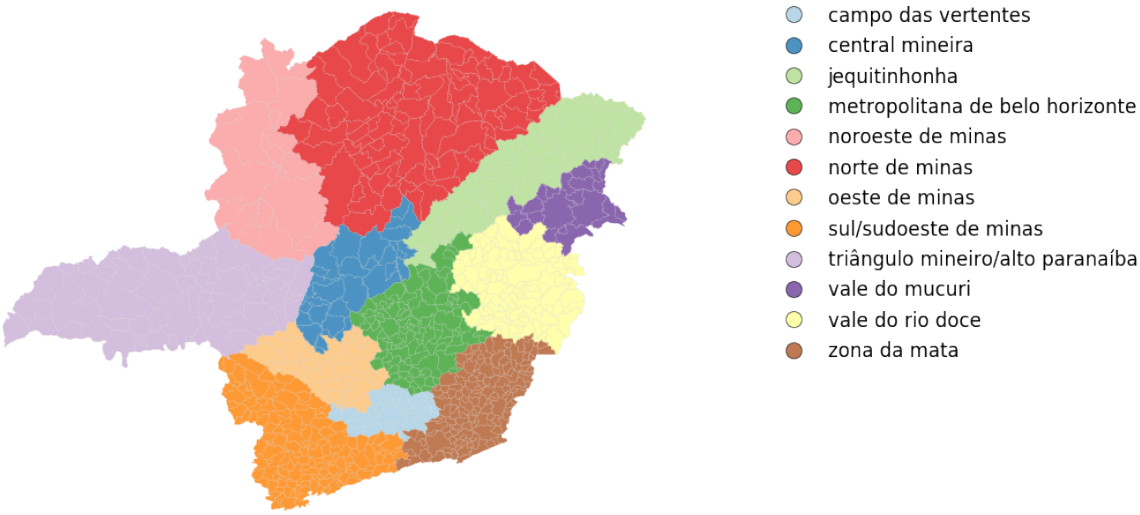


Figura 5 – Mapa das mesorregiões de Minas Gerais.
Fonte: Da autora.

Mesmo sendo tão populosa e dividida em tantos municípios, a mesorregião SSM é bastante homogênea em relação ao tamanho das populações. Uma maneira de verificar essa propriedade é através do coeficiente de variação (CV) de todas as mesorregiões do país. A média nacional é 1,68%, enquanto o CV da SSM é 1,41%. Isso mostra que a variação no tamanho da população entre as cidades da SSM está abaixo da média do Brasil. O município com a menor população da mesorregião é Consolação (1.727), enquanto aquele com maior número de habitantes é Poços de Caldas (152.435). Além disso, a mediana mostra que pelo menos metade dos municípios da SSM possuem menos que 9.515 habitantes. A Figura 6 apresenta o mapa da mesorregião SSM e seus municípios.

Figura 6 – Mapa da mesorregião Sul/Sudoeste de Minas Gerais.
Fonte: Da autora.

Em relação à situação econômica, dados mais recentes de 2013 mostram que a mesorregião SSM ocupou a 23ª posição no Brasil em termos de Produto Interno Bruto (PIB), com um valor de R$ 53,80 bilhões. Em relação aos municípios, aqueles que registram os maiores valores de PIB da mesorregião são Poços de Caldas, Pouso Alegre e Extrema. No entanto, pela ótica do PIB *per capita*, Extrema passa a ocupar a primeira posição, seguida por Itaú de Minas e Itamonte.

Comparando a distribuição dos setores do PIB da mesorregião SSM com todas as mesorregiões do Brasil é possível observar que o setor de serviços responde por quase metade do PIB da SSM (45%), valor acima do terceiro quartil (43%), considerando todas as mesorregiões do Brasil. Em seguida, aparece o setor de indústria, responsável por 20% do PIB, acima da mediana nacional (18%). O setor que responde pela terceira maior parcela do PIB é o da administração pública (15%), que está abaixo da mediana nacional (18%). Por último, o PIB da agropecuária representa 7% do PIB na mesorregião, valor abaixo da mediana nacional (9%). Dessa forma, o valor adicionado bruto da mesorregião SSM, representado pela soma dos percentuais dos

setores, resulta em 87% do PIB. O valor dos impostos responde pelos 13% restantes.

Em relação ao comportamento das variáveis demográficas na messoregião estudada, a Tabela 1 apresenta a média dessas variáveis de todas as mesorregiões do Brasil e da mesorregião SSM, em 2010. A esperança de vida ao nascer da SSM (75,46 anos) é 2,36 anos maior que a média de todas as messoregiões brasileiras (73,09 anos). A mesorregião Distrito Federal registrou a maior esperança de vida ao nascer do Brasil (77,35 anos), ao passo que a Oeste Maranhense a menor (69,03 anos). A taxa de fecundidade total da SSM foi de 1,95 filhos por mulher, abaixo do nível de reposição e da média do país (2,19 filhos por mulher). A mesorregião Norte do Amapá ocupou a primeira posição dentre aquelas com maiores níveis de fecundidade (4,29 filhos por mulher) e a Nordeste Rio-Grandese a menor (1,61 filhos por mulher).

Tabela 1 – Média das variáveis demográficas para o Brasil e para a mesorregião Sul/Sudoeste de Minas Gerais, 2010.

variável	Brasil	SSM
espvida	73,09	75,46
tft	2,19	1,95
mort1	19,25	14,68
mort5	21,53	17,09
sobre40	93,78	93,85
sobre60	82,75	83,15
t_env	8,40	9,46

Fonte: Da autora.

Em relação à mortalidade infantil, na SSM o indicador foi de 14,69 óbitos de menores de um ano de idade, por mil nascidos vivos. Esse valor foi abaixo da média nacional (19,25). Nesse ano, o Vale do Itajaí foi responsável pela menor mortalidade infantil (11,56) e o Oeste Maranhense pela maior (32,87). O mesmo aconteceu com a mortalidade até os 5 anos de idade, a taxa de 17,09 óbitos de menores de 5 anos de idade, por mil nascidos vivos, da mesorregião SSM é menor que a do Brasil (21,53). A mesorregião brasileira Grande Florianópolis obteve a menor taxa de mortalidade até os 5 anos de idade (13,53) e a Oeste Maranhense a maior (35,81). A probabilidade de sobrevivência até os 40 anos, por sua vez, é um pouco maior na SSM (93,85%) que no Brasil (93,78%). A mesorregião Norte Fluminense foi responsável pela menor probabilidade (91,30%) e Norte de Roraima a maior (96,01%). O mesmo aconteceu com a probabilidade de sobrevivência até 60 anos de idade, que também foi maior para a SSM (83,15%) do que para o Brasil (82,75%). A mesorregião que apresentou o menor valor dessas variável foi a Vale do Mucuri (78,62%) e maior foi a Metropolitana de Recife (87,06%). Por fim, a taxa de envelhecimento também foi maior na SSM (9,46%) que no Brasil (8,40%). A me-

sorregião Norte do Amapá registrou a menor taxa (3,23%) e a Centro Ocidental Rio-Grandense a maior (12,08%). O comportamento dessas variáveis evidencia que, em média, os indicadores da SSM possuem melhores desempenhos que os do Brasil.

2.4 ESTATÍSTICA MULTIVARIADA E ANÁLISE DE AGRUPAMENTO

Os dados levantados em uma pesquisa são considerados multivariados quando os valores referentes a cada unidade amostral ou observação se referem a diversas variáveis aleatórias ao mesmo tempo, levando cada observação a ser multidimensional. Na maioria das pesquisas, os dados são multivariados mas, muitas vezes, o pesquisador opta por analisar cada variável separadamente. Porém, em geral, as variáveis são correlacionadas entre si e, quanto maior o número de variáveis, mais complexa se torna a análise univariada. Ao se utilizar a análise multivariada, as variáveis são analisadas ao mesmo tempo, fornecendo uma avaliação muito mais ampla do conjunto de dados, encontrando-se padrões e levando-se em conta a correlação entre as variáveis (MINGOTI, 2005).

Nesse sentido, a análise multivariada corresponde ao conjunto de técnicas que analisam duas ou mais variáveis correlacionadas entre si simultaneamente, permitindo que se discrimine a influência ou relevância de cada uma delas. Os métodos multivariados são divididos como métodos de dependência e interdependência. Caso no estudo haja variáveis dependentes e independentes é aconselhável que se use uma das técnicas de dependência, tais como regressão múltipla, análise discriminante ou regressão logística. Por sua vez, se não existir uma discriminação preliminar de quais variáveis são dependentes e independentes, as técnicas de interdependência devem ser aplicadas. Dentre elas estão a análise fatorial e análise de agrupamento (HAIR et al., 2009).

A representação de dados multivariados se dá como em planilhas eletrônicas. Se há uma amostra aleatória de tamanho n e, para cada unidade amostral ou observação, os valores de p variáveis foram observados, cria-se uma matriz de dados $\boldsymbol{X}$ com dimensão n (linhas) por p colunas:

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}, \tag{2.1}$$

em que cada unidade amostral é representada por uma linha da matriz de dados $\boldsymbol{X}$, sendo um vetor com p elementos (variáveis), e cada variável é representada por uma coluna de $\boldsymbol{X}$, sendo um vetor com n elementos, as observações (EVERITT; HOTHORN, 2011).

A obtenção da matriz de dados a partir de uma amostra aleatória, como expressa na forma da definição (2.1), pode não ser muito informativa, principalmente se o tamanho amostral n for grande e houver um número excessivo de variáveis p. Torna-se interessante utilizar medidas resumo dos dados amostrais, da mesma forma que é feito no caso univariado, calculando-se a média, mediana, desvio padrão etc., de forma a sintetizar os dados da amostra obtida (FERREIRA, 2011).

Uma medida de tendência central muito utilizada é a média amostral que, no caso multivariado, torna-se o vetor de médias amostral de dimensão $p \times 1$, em que cada elemento é a média de cada variável:

$$\bar{\boldsymbol{X}} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix}. \tag{2.2}$$

Para medir a dispersão dos dados, no lugar da variância amostral, utiliza-se a matriz de covariâncias amostral $\boldsymbol{S}$ de dimensão $p \times p$. Sua diagonal principal é composta pelas variâncias das p variáveis e os elementos fora da diagonal são as covariâncias entre as variáveis. Essa

matriz é simétrica, ou seja, $S_{ij} = S_{ji}$.

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{bmatrix}. \tag{2.3}$$

A correlação é também uma medida de covariação entre duas variáveis, porém em uma escala padronizada, ou seja, seus valores variam entre -1 e +1. Valores próximos de +1 indicam que as variáveis estão fortemente correlacionadas de forma positiva, grandes valores de uma estão associados a grandes valores da outra. Já valores próximos de -1 indicam que as variáveis estão fortemente correlacionadas de forma negativa, indicando que grandes valores de uma estão associados a pequenos valores da outra. A matriz de correlações amostral é dada por

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}, \tag{2.4}$$

em que $r_{ij} = \dfrac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$ (FERREIRA, 2011).

Outras estatísticas descritivas, como a matriz de somas de quadrados e produtos, podem ser consideradas, dependendo do objetivo da pesquisa (FERREIRA, 2011).

Segundo Mingoti (2005), a análise multivariada se divide em dois grupos principais: técnicas exploratórias e técnicas de inferência estatística, como também ocorre na análise univariada. O primeiro possui um grande apelo prático por suas técnicas não dependerem do conhecimento da forma matemática da distribuição de probabilidade que gerou os dados amostrais e permitem a detecção de padrões. Exemplos desse tipo incluem análise de componentes principais, análise fatorial exploratória, análise de agrupamento (*clusters*), entre outras. O foco do segundo grupo de técnicas é a estimação de parâmetros, testes de hipóteses, análise de regressão multivariada etc., cujo objetivo é utilizar a amostra para realizar inferências sobre a população de onde essa amostra foi extraída.

As técnicas exploratórias são muitas vezes denominadas técnicas de sintetização por se concentrarem em condensar uma grande massa de dados em uma forma mais simples. Assim,

há uma redução significativa do volume de dados envolvido na análise ou uma redução da dimensionalidade (BARTHOLOMEW et al., 2008).

Dentre as técnicas exploratórias, a análise de agrupamento (AA), também conhecida como análise de conglomerados, classificação ou *cluster analysis* corresponde a um método que busca uma partição dos elementos de uma amostra em grupos de tal forma que (a) as observações de um mesmo grupo sejam similares entre si em relação às variáveis medidas e que (b) as observações de grupos diferentes sejam heterogêneas em relação a essas mesmas variáveis (MINGOTI, 2005). Portanto, dada uma amostra de tamanho n, com cada objeto medido segundo p variáveis, a análise de agrupamento classifica os objetos em grupos com elevado grau de homogeneidade interna e heterogeneidade externa.

De acordo com Gordon (1999), a classificação de dados em grupos pode ser realizada com o objetivo de simplificá-los e realizar predições. A partir do método, é possível detectar o relacionamento e estrutura do conjunto de dados. Em muitas aplicações, os pesquisadores podem estar interessados na descrição de um conjunto de dados maior e a atribuição de novos objetos, bem como fazer previsão e descobrir hipóteses para explicar a estrutura dos dados.

"Análise de agrupamento"é uma expressão genérica que compreende vários métodos numéricos que pretendem descobrir grupos de observações homogêneas. O que esses métodos tentam é formalizar o que os seres humanos conseguem fazer bem em duas ou três dimensões, por exemplo, por meio de diagramas de dispersão. Os grupos são identificados pela avaliação das distâncias entre os pontos (EVERITT; HOTHORN, 2011).

Como forma de ilustração, considere um conjunto de dados fictícios em que há $n = 23$ observações (pessoas) e registros sobre suas idades e rendas mensais, ou seja, há $p = 2$ variáveis, ambas medidas na escala contínua. O objetivo é agrupar as pessoas em relação às duas variáveis. O diagrama de dispersão obtido se encontra na Figura 7 e é possível identificar três agrupamentos, apenas pela análise visual. Se houvesse mais uma variável e, consequentemente mais uma dimensão, ainda seria possível imaginar um gráfico com os pontos. Porém, com mais de três dimensões, torna-se mais difícil a visualização dos dados (BARTHOLOMEW et al., 2008).

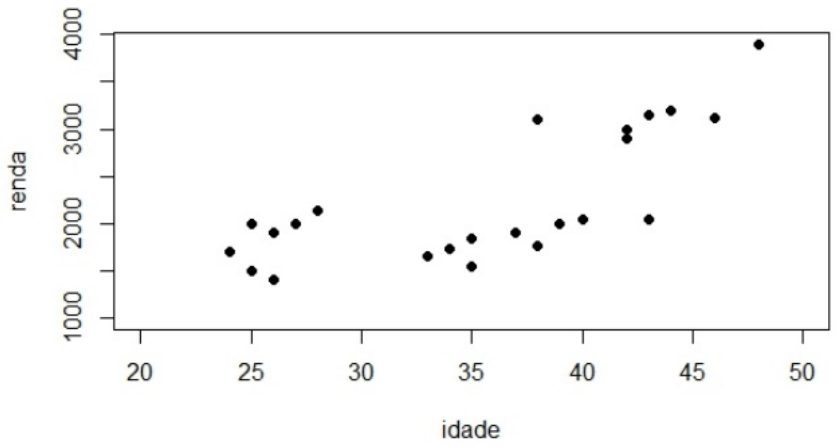


Figura 7 – Diagrama de dispersão de dados fictícios de idade e renda de várias pessoas.
Fonte: Da autora modificado a partir de (BARTHOLOMEW et al., 2008, p.18).

Há dois objetivos possíveis de um agrupamento: agrupar as n observações em um número de grupos desconhecidos ou classificar as observações em um conjunto predefinido de grupos. A análise de agrupamento deve ser utilizada no primeiro caso e análise discriminante no segundo. Na análise de agrupamento, geralmente, o número de grupos não é conhecido a princípio e encontrar o melhor agrupamento não é uma tarefa simples (FERREIRA, 2011).

De acordo com (BARTHOLOMEW et al., 2008), qualquer processo de agrupamento tem como base duas etapas:

1. Obter as distâncias de todos os pares de objetos para construção da matriz de proximidades;

2. Desenvolver um algoritmo para formação de grupos com base nessas distâncias.

As distâncias da etapa 1 são determinadas com base em medidas de similaridade ou dissimilaridade, que indicam a proximidade dos objetos. As medidas de dissimilaridade correspondem às distâncias, ao passo que as de similaridades complementam as distâncias, assim, quanto maior a medida de similaridade entre dois objetos menor será a de dissimilaridade e mais próximos eles serão (FERREIRA, 2011). A seguir são apresentadas algumas orientações sobre medidas de distâncias que podem ser utilizadas.

2.4.1 Distâncias

Para realizar o procedimento de agrupamento escolhido é necessário que a medida de similaridade ou dissimilaridade seja definida *a priori*. Alguns tipos comuns de distâncias que

podem ser calculadas entre os pares de observações são a distância euclidiana, distância euclidiana padronizada e distância de Mahalanobis, entre outras. Essas medidas são de dissimilaridade, ou seja, quanto menor seus valores, mais próximos ou similares são os objetos comparados. A escolha da métrica interfere diretamente no resultado final do agrupamento (MINGOTI, 2005).

A distância entre as observações i e j aparece na i-ésima linha e j-ésima coluna da matriz de distâncias. Por exemplo, se há $n = 4$ elementos na amostra, a matriz de distâncias terá dimensão 4×4 e será da forma

$$\begin{bmatrix} - & d_{12} & d_{13} & d_{14} \\ d_{21} & - & d_{23} & d_{24} \\ d_{31} & d_{32} & - & d_{34} \\ d_{41} & d_{42} & d_{43} & - \end{bmatrix}, \tag{2.5}$$

em que d_{ij} é a distância entre os elementos i e j. Geralmente, essa matriz é simétrica, ou seja, $d_{12} = d_{21}$, $d_{13} = d_{31}$, e assim por diante (BARTHOLOMEW et al., 2008).

Dentre as distâncias citadas, um tipo muito simples e comum é a distância euclidiana, calculada por

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}, \tag{2.6}$$

em que d_{ij} é a distância euclidiana entre os elementos i, com os valores X_{i1}, X_{i2}, ..., X_{ip}, e j, com os valores X_{j1}, X_{j2}, ..., X_{jp}. Na aplicação da distância euclidiana há dois caminhos diferentes. Como as variáveis geralmente são medidas em unidades distintas, é necessário que os dados sejam padronizados. Dessa forma, a cada variável padronizada é atribuído o mesmo peso. No entanto, caso seja aplicada a técnica de componentes principais para a redução da dimensionalidade dos dados, o peso difere de acordo com o componente. Nessa situação, é atribuído ao primeiro componente um peso maior na determinação da similaridade entre os objetos (LATTIN; CARROLL; GREEN, 2011). De acordo com FERREIRA (2011), o uso dessa métrica faz com que variáveis com maior variabilidade dominem a classificação e ordenação dos objetos, portanto é mais indicada para grupos de variáveis com escalas similares.

As distâncias de Mahalanobis e euclidiana padronizada são uma generalização da distância euclidiana. Dessa forma, seja a distância generalizada entre dois elementos $X_{i.}$ e $X_{j.}$

definida por:

$$d_{ij} = (\boldsymbol{X_{i.}} - \boldsymbol{X_{j.}})^T \boldsymbol{A} (\boldsymbol{X_{i.}} - \boldsymbol{X_{j.}}). \tag{2.7}$$

A seleção dessa matriz $\boldsymbol{A}$ define a distância utilizada. Quando $\boldsymbol{A} = \boldsymbol{I}$ temos a distância euclidiana e quando $\boldsymbol{A} = \boldsymbol{D^{-1}}$ temos a distância euclidiana padronizada. Se $\boldsymbol{A} = \boldsymbol{S^{-1}}$, isto é, a matriz de covariâncias da matriz de dados, obtém-se a distância de Mahalanobis. Nesse caso, são consideradas as diferenças de variâncias e relações lineares entre as variáveis, a partir das covariâncias (MINGOTI, 2005). A definição dessa métrica propõe a ideia de que objetos situados na mesma direção das correlações entre as variáveis são mais similares entre si do que aqueles situados na direção oposta (FERREIRA, 2011). Além disso, a métrica produz agrupamentos compactos e convexos (LATTIN; CARROLL; GREEN, 2011) e elimina o efeito o efeito de domínio na classificação das variáveis de maior variabilidade (FERREIRA, 2011).

Há outras medidas de similaridade e dissimilaridade propostas na literatura, tais como as distâncias: euclidiana média, quarteirão (*city-block*) ou Manhattan, de Chebychev, angular, Canberra, entre outras. Embora a distância euclidiana seja uma das mais utilizadas, a de Mahalanobis é a mais indicada na maioria das situações aplicadas por levar em conta a colinearidade existente entre as variáveis usadas para realizar o agrupamento (MOOI; SARSTEDT, 2011). Além disso, outras distâncias conhecidas em situações práticas, como as de Kolmogorov, de Hellinger, de Rao, entre outras, são funções da distância de Mahalanobis sob pressuposições de normalidade e homocedasticidade, e sob outras condições (MCLACHLAN, 1999).

Após a definição da medida de distância utilizada, é necessário escolher um método de agrupamento. Segundo FERREIRA (2011), esses métodos são divididos em hierárquicos (aglomerativos ou divisivos) e não hierárquicos. No primeiro tipo, há *n* grupos no início, cada um com uma observação, e no fim há um único grupo com todas as observações. A cada passo, cada observação ou grupo é unido a outra observação ou grupo (método hierárquico aglomerativo). A união ocorre com base no critério de similaridade, os objetos mais próximos entre si são alocados para um mesmo grupo, até que todos estejam em um único grupo. Portanto, a cada passo se perde um grupo, que é unido ao outro mais similar a ele. Nos métodos hierárquicos divisivos, há um único grupo com as n observações no início e, ao final, há *n* grupos. Nos métodos que não são hierárquicos é preciso definir o número *k* de grupos inicialmente para, em seguida, atribuir as n observações aos *k* grupos da melhor maneira possível. Sempre é preciso

usar uma alocação arbitrária no início do processo e, iterativamente, buscar a alocação ótima.

2.4.2 Técnicas hierárquicas aglomerativas

De acordo com Everitt et al. (2011), os agrupamentos resultantes a partir de métodos hierárquicos, aglomerativos ou divisivos, podem ser representados por um diagrama bidimensional muito utilizado, o dendrograma, também denominado de diagrama de árvore. Esse gráfico ilustra as fusões ou divisões realizadas a cada passo do processo. A Figura 8 mostra algumas das terminologias utilizadas para descrever os dendrogramas.

Ainda segundo os autores, o arranjo de nós e caules representam a topologia da árvore. O diagrama descreve o processo pelo qual foi obtida a hierarquia, assim há várias sub-árvores oriundas da raiz da árvore. O nó interno representa partições particulares, ou seja, os agrupamentos formados a partir dos nós terminais, que representam os objetos. A altura do nó interno corresponde ao ponto em que os objetos ou grupos foram unidos, ou seja, a proximidade entre eles. Dessa forma, a ordem de união dos grupos segue o princípio de ordem crescente da altura do nó.

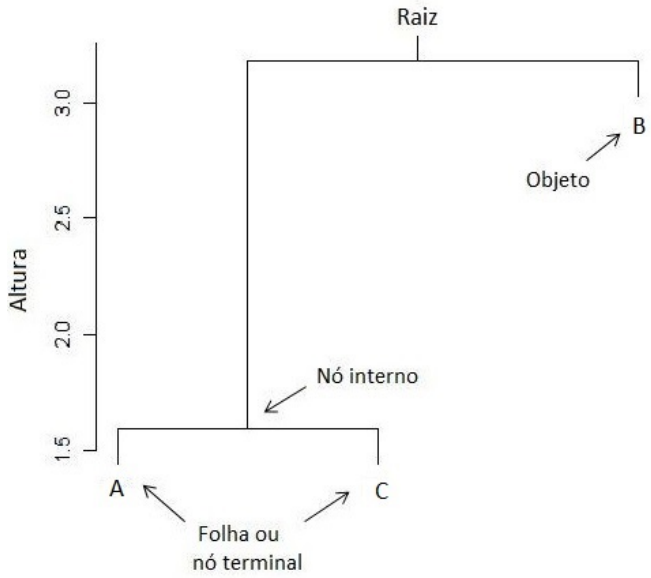


Figura 8 – Terminologia utilizada na descrição de dendrogramas.
Fonte: Da autora a partir de (EVERITT; HOTHORN, 2011)

Portanto, no caso representado pela Figura 8, os objetos denominados de A e C foram os primeiros a serem unidos em um único grupo, com nível de fusão de aproximadamente

1,7 (altura do nó). Esse valor corresponde à distância entre os elementos A e C nas variáveis medidas. Após essa fusão, a amostra formada por 3 elementos foi dividida em 2 grupos, o primeiro de tamanho 2 contendo os elementos A e C e, o segundo de tamanho 1, formado pelo elemento B. No próximo passo o elemento 3 é reunido ao primeiro grupo formado, com nível de fusão de aproximadamente 3,2, obtendo um único cluster de tamanho 3. Nesse exemplo foram considerados apenas 3 elementos para ilustrar a terminologia utilizada na descrição de dendrogramas, no entanto em uma análise real de muitos objetos, diversos grupos são obtidos.

Ao se utilizar um procedimento hierárquico aglomerativo, depois que uma fusão é realizada, ela não será mais desfeita. Assim, quando o método coloca dois elementos em um mesmo grupo, eles não mais aparecerão em grupos diferentes. Para o pesquisador encontrar a melhor solução com o número de agrupamentos ótimo, ele deverá adotar algum critério de divisão (EVERITT; HOTHORN, 2011).

O pesquisador tem a difícil tarefa de decidir em qual altura o corte no dendrograma deve ser realizado para escolha do número final de grupos. Isso ocorre porque o objetivo dos processos de agrupamentos hierárquicos é agrupar os *n* grupos de tamanho 1 em um único grupo com todas as observações. Contudo, o interesse do pesquisador é agrupar as observações em vários grupos e, para isso, é necessário decidir uma regra de parada do processo, para obtenção de k grupos. Esse tópico será discutido na seção 2.4.4.

Os principais métodos hierárquicos aglomerativos são: ligação simples (vizinho mais próximo), ligação completa (vizinho mais distante), ligação média, centroide e método de Ward (MINGOTI, 2005).

Com exceção de Ward, as demais técnicas seguem um processo iterativo geral, denominado método de grupo de pares (LATTIN; CARROLL; GREEN, 2011), que será descrito a seguir:

Em um método hierárquico aglomerativo, inicialmente, as n observações são alocadas em n grupos de tamanho 1 e, após todos os passos de fusão dos grupos, é formado um único grupo contendo os n elementos. O algoritmo básico para todos os métodos é semelhante (EVERITT; HOTHORN, 2011):

(Início) Grupos $G_1, G_2, \ldots, G_n$ - cada um contendo uma única observação.
(1) Unir os grupos G_i e G_j mais próximos entre si e diminuir o número de grupos de 1.
(2) Se o número de grupos é igual a 1, parar; senão, retornar ao passo (1).

Porém, antes do processo iniciar, a matriz de distâncias entre os objetos precisa ser obtida (conforme apresentado na seção 2.4.1) e deverá ser recalculada a cada novo passo, de forma a contabilizar todas as distâncias entre grupos. Com base nessa matriz, o algoritmo de agrupamento irá definir quais os grupos mais próximos entre si, ou seja, aqueles que apresentam menores distâncias.

Após selecionar a medida de dissimilaridade é preciso decidir qual método de agrupamento aplicar. Há vários métodos hierárquicos aglomerativos e eles diferem na forma com que definem a distância entre um grupo recém-formado a uma observação ou a outros grupos já existentes (MOOI; SARSTEDT, 2011). Os procedimentos aglomerativos incluem:

- Vizinho mais próximo (ligação simples): a distância entre dois grupos é a menor distância entre dois elementos dentro dos dois grupos.

- Vizinho mais distante (ligação completa): oposto ao vizinho mais próximo, define a distância entre dois grupos como sendo a maior distância entre quaisquer dois elementos dos dois grupos.

- Ligação média (distância média): a distância entre dois grupos é definida como a distância média entre todos os pares de elementos dos dois grupos.

- Centroide: o centro geométrico (centroide) de cada grupo é calculado primeiro. A distância entre os dois grupos é definida como a distância entre os dois centroides.

- Ward: método distinto dos demais, pois não combina os dois objetos mais semelhantes sucessivamente. Em vez disso, os objetos cuja fusão resulte na menor variância dentro do grupo são combinados.

O método de Ward, também denominado de método de variância mínima, foi proposto por Jr (1963). Diferentemente dos outros métodos hierárquicos aglomerativos, não segue o algoritmo básico de agrupamento apresentado. A razão dessa diferença é que ele não busca a menor distância entre dois grupos, mas sim a menor soma de quadrados dentro do grupo, ou seja, a menor variância interna. Num primeiro momento, ele permite a redução dos n conjuntos iniciais a $n-1$ conjuntos mutuamente exclusivos considerando a fusão dos dois grupos, dentre todos os $n(n-1)/2$ pares possíveis, que resulte na menor soma de quadrados. Esse processo é repetido até haver um único grupo.

Segundo (MINGOTI, 2005), o processo iterativo dessa técnica segue os seguintes passos:

(1) Inicialmente, as n observações são alocadas em n grupos de tamanho 1, representados por $G_1, G_2, \ldots, G_n$.
(2) A cada passo do processo de agrupamento, a soma dos quadrados dentro de cada grupo é calculada como a soma do quadrado da distância euclidiana de cada elemento do grupo em relação ao vetor de médias do grupo. Assim, a soma de quadrados SQ_i de um grupo G_i é definida por:

$$SQ_i = \sum_{j=1}^{n_i} (\boldsymbol{X_{ij}} - \bar{\boldsymbol{X}}_{\boldsymbol{i}})^T (\boldsymbol{X_{ij}} - \bar{\boldsymbol{X}}_{\boldsymbol{i}}), \tag{2.8}$$

em que, n_i é o número de elementos no grupo G_i quando se está no passo m do processo, $\boldsymbol{X_{ij}}$ é o vetor de observações do j-ésimo elemento amostral que pertence ao i-ésimo grupo e $\bar{\boldsymbol{X}}_{\boldsymbol{i}}$ é o vetor de médias do grupo.

No passo m, a soma de quadrados total dentro dos grupos é dada por:

$$SQT = \sum_{i=1}^{g_m} SQ_i, \tag{2.9}$$

em que, g_m é o número de grupos no passo m.

A distância entre dois grupos G_r e G_s é definida como a soma de quadrados entre eles, dada por:

$$d(G_r,G_s) = \left[\frac{n_r n_s}{n_r + n_s}\right] (\bar{\boldsymbol{X}}_{\boldsymbol{r}} - \bar{\boldsymbol{X}}_{\boldsymbol{s}})^T (\bar{\boldsymbol{X}}_{\boldsymbol{r}} - \bar{\boldsymbol{X}}_{\boldsymbol{s}}), \tag{2.10}$$

que é a soma de quadrados entre os grupos G_r e G_s. A cada passo de aplicação do processo iterativo, os dois grupos que minimizarem o valor de (2.10) são unidos.

Ainda de acordo com (MINGOTI, 2005), os métodos de Ward e da máxima verossimilhança são relacionados quando a distribuição dos dados é normal multivariada. Porém, para a aplicação do método de Ward não é necessário que os dados sejam normais multivariados, basta

que as variáveis sejam quantitativas, já que ele se baseia no cálculo de médias.

Assim como no método do centroide, a distância entre dois grupos é definida considerando os vetores de médias amostrais, no entanto, o método de Ward considera a diferença entre o número de elementos em cada um dos grupos que estão sendo comparados. Dessa forma, o fator $\left[\frac{n_r n_s}{n_r+n_s}\right]$ pondera a distância de dois grupos r e s de tamanhos diferentes. Quanto maiores os valores de n_r e n_s, maior será o fator de ponderação e, portanto, maior a distância entre os vetores de médias comparados (MINGOTI, 2005).

Em situações práticas, a tarefa de escolher um dentre os métodos hierárquicos aglomerativos pode se tornar difícil, já que os agrupamentos obtidos para o mesmo conjunto de dados podem ser bastante diferentes dependendo do método utilizado (MOOI; SARSTEDT, 2011). Para auxiliar nessa tarefa, Everitt et al. (2011) dividem estudos empíricos sobre os métodos de agrupamento em dois tipos: estudos de simulação (em que se sabe a estrutura dos dados e uma avaliação dos métodos é feita em relação à recuperação dessa estrutura) e estudos reais (em que o critério é a interpretabilidade dos grupos obtidos). O que deve ficar claro é que não há um método que deve ser recomendado e que é melhor do que todos os outros, mas algumas observações gerais podem ser feitas e que são descritas a seguir.

O método do vizinho mais próximo tende a ser menos custoso computacionalmente, mas menos satisfatório do que outros métodos por causa do efeito de encadeamento ou *chaining* (novos grupos formados tendem a se unir a uma nova observação simples e não a um grupo já existente) (FERREIRA, 2011). Já o vizinho mais distante, por ser baseado em distâncias máximas, é muito afetado por *outliers* e os grupos obtidos tendem a ser compactos (MOOI; SARSTEDT, 2011; LATTIN; CARROLL; GREEN, 2011; MINGOTI, 2005). Os métodos da ligação média e centroide tendem a formar grupos com melhores partições do que os da ligação simples e completa. Além disso, os grupos resultantes possuem aproximadamente a mesma variância interna (LATTIN; CARROLL; GREEN, 2011; EVERITT; HOTHORN, 2011; MOOI; SARSTEDT, 2011).

Everitt et al. (2011) enumeram alguns trabalhos em que o método de Ward retornou melhores resultados do que os outros, mas tende a impor uma estrutura esférica aos dados que pode não existir. No mesmo sentido, de acordo com FERREIRA (2011), muitos trabalhos apontam os métodos de Ward e da ligação média como tendo os melhores desempenhos de forma geral, mas seu desempenho depende dos dados. Blashfield (1976) comparou os métodos para conjuntos de dados simulados e concluiu que o método de Ward foi o mais preciso e é

o mais indicado para dados quantitativos contínuos. Geralmente, os grupos resultantes pelo método de Ward possuem o mesmo número de objetos, são convexos e compactos (LATTIN; CARROLL; GREEN, 2011). Assim, é uma boa escolha aplicar esse método se é esperado que haja grupos de iguais tamanhos ou se esses grupos obtidos favoreçam a interpretabilidade (MOOI; SARSTEDT, 2011).

Além das técnicas hierárquicas, há outros métodos de agrupamento que particionam as observações em um número específico de grupos utilizando como critério a minimização ou maximização de algum critério. Um desses métodos de otimização mais populares é o das k-médias, a ser descrito em seguida (EVERITT et al., 2011).

2.4.3 Técnicas não hierárquicas: k-médias

Como o próprio nome diz, os métodos hierárquicos não seguem a propriedade da hierarquia, isso significa que mesmo se dois objetos forem unidos em algum passo do processo, pode ser que eles não permaneçam no mesmo grupo na partição final. E, portanto, isso implica que não é possível construir dendrogramas para a representação dos agrupamentos formados passo a passo (MINGOTI, 2005). Há métodos não hierárquicos baseados em estimação de densidades, misturas de distribuição e partição. Os procedimentos de partição são os mais utilizados e um deles, o das k-médias, é o mais popular (FERREIRA, 2011).

O procedimento de agrupamento das k-médias (k-*means*) tem como principais características: aplicação do processo à matriz de dados $\boldsymbol{X}$ e número de grupos k definido *a priori* (FERREIRA, 2011). A técnica procura uma partição das n observações em *k* agrupamentos $(G_1, G_2, \ldots, G_k)$, em que G_i denota o conjunto de observações que está no *i*-ésimo grupo e k é dado por algum critério numérico de minimização. A implementação mais usada do método tenta encontrar a partição dos n elementos em k grupos que minimizem a soma de quadrados dentro dos grupos (SQDG) em relação a todas as variáveis. Esse critério pode ser escrito como

$$SQDG = \sum_{j=1}^{p} \sum_{l=1}^{k} \sum_{i \in G_l} (X_{ij} - \bar{X}_j^{(l)})^2, \tag{2.11}$$

em que $\bar{X}_j^{(l)} = \dfrac{1}{n_i} \displaystyle\sum_{i \in G_l} X_{ij}$ é a média dos indivíduos no grupo G_l em relação à variável j

(EVERITT; HOTHORN, 2011).

Ainda de acordo com os autores, apesar de o problema parecer relativamente simples, ele não é tão direto. A tarefa de selecionar a partição com a menor soma de quadrados dentro dos grupos se torna complexa porque o número de partições possíveis se torna elevado mesmo com um tamanho amostral não tão grande. Por exemplo, para n = 100 e k = 5, o número de partições é da ordem de 10^{68}. Esse problema levou ao desenvolvimento de algoritmos que não garantem encontrar a solução ótima, mas que levam a soluções satisfatórias.

Segundo Mingoti (2005), o processo iterativo do método pode ser descrito pelos seguintes passos:

(1) Inicialmente são escolhidos k centroides, denominados de "sementes", calculados com base no número de grupos escolhido *a priori*;

(2) Uma medida de distância é aplicada para comparar cada objeto a cada centroide inicial, sendo que o objeto é unido ao grupo de menor distância;

(3) Os valores dos centroides são recalculados considerando cada grupo formado, então o passo 2 é repetido com os novos vetores de médias calculados para os novos grupos;

(4) Os passos 2 e 3 são repetidos até que não haja mais realocação dos objetos entre os grupos.

O agrupamento final obtido através do método das k-médias depende diretamente da escolha das sementes (passo 1) (FERREIRA, 2011). Diversas sugestões para a definição das sementes são apresentadas na literatura, Mingoti (2005) apresenta algumas propostas, sendo elas: aplicação de técnicas hierárquicas aglomerativas, escolha aleatória ou via observação dos valores discrepantes do conjunto de observações.

Segundo a autora, as sementes iniciais podem ser escolhidas com base no número de grupos obtidos após a aplicação de uma técnica hierárquica aglomerativa. Nesse caso, o vetor de médias de cada grupo é calculado e utilizado como semente para o uso do método das k-médias. O método de Ward é frequentemente utilizado para selecionar os centroides iniciais porque o critério de fusão de grupos com base na menor soma de quadrados dentro do grupo, utilizado nesse método, é próximo ao critério do quadrado da soma de erros de partição do método k-médias (LATTIN; CARROLL; GREEN, 2011). A segunda sugestão se baseia na escolha aleatória a partir de um procedimento de amostragem aleatória simples repetido m vezes, produzindo para cada grupo o centroide das m sementes selecionadas. Outra regra de decisão se baseia na seleção de k elementos discrepantes, em relação às p-variáveis no conjunto

de dados, como sementes de um agrupamento inicial (MINGOTI, 2005).

Não há um consenso sobre o melhor método para escolha do número de grupos inicial ou de seus centroides, contudo, alguns autores aconselham que o processo seja realizado com diferentes escolhas para buscar a melhor solução de agrupamento (FERREIRA, 2011; LATTIN; CARROLL; GREEN, 2011). Pena, Lozano e Larranaga (1999) apresentam um estudo comparativo de diferentes métodos de inicialização para as k-médias e a escolha aleatória aparece como um dos melhores dentre os comparados, por tornar o procedimento das k-médias mais efetivo e independente do agrupamento inicial.

De acordo com Mooi e Sarstedt (2011), o método não hierárquico das k-médias é superior aos métodos hierárquicos por ser menos afetado por *outliers* e por variáveis não relevantes para o agrupamento. E, por ser um método mais eficiente computacionalmente, pode ser aplicado a grandes conjuntos de dados. O bom desempenho do método das k-médias em relação aos métodos hierárquicos também foi confirmado por LIMA (2001 apud MINGOTI, 2005), que comparou métodos hierarquicos e não hirárquicos, via simulação de Monte Carlo. Os resultados mostraram que o método *Fuzzy* apresentou o melhor desempenho, seguido pela técnica das k-médias e Ward. No entanto, uma desvantagem apontada por Mooi e Sarstedt (2011) no uso do método das k-médias é a necessidade de definir previamente o número de grupos e esse ponto será discutido na seção 2.4.4.

2.4.4 Número de grupos

Nas aplicações de métodos de agrupamento, o pesquisador precisa, em algum momento, decidir o número apropriado de grupos, independente do método utilizado. Essa é a etapa final nos métodos de agrupamento hierárquicos aglomerativos e a inicial nos agrupamentos não hierárquicos. As técnicas hierárquicas aglomerativas iniciam o procedimento com k observações separadas em k grupos. A cada passo, o algoritmo reúne duas observações ou grupos e ao final, um único grupo com as k observações é obtido. Portanto, é preciso que uma regra de corte seja estabelecida, para que o número ideal de grupos seja escolhido. No uso de métodos não hierárquicos, como o das k-médias, a escolha do número de grupos acontece antes da aplicação do método porque, por definição, essas técnicas exigem que o número de grupos seja escolhido *a priori* (MINGOTI, 2005).

De acordo com Milligan e Cooper (1985), a escolha do número apropriado de grupos está sujeita a dois tipos de erros. O primeiro acontece quando a regra de parada seleciona um número k de grupos maior do que o adequado. O segundo tipo ocorre quando a regra de decisão conduz a escolha de um número de grupos menor do que o apropriado. Apesar dos dois tipos de erros serem indesejáveis, o segundo produz consequências consideradas mais sérias, pois informação é perdida. De forma geral, essa não é considerada uma tarefa simples. A seguir, serão apresentadas diferentes abordagens propostas na literatura (MINGOTI, 2005).

Um método gráfico utilizado para a escolha do número adequado de agrupamentos é o corte no dendrograma. A questão, entretanto, é decidir onde o corte deve ser feito. Segundo Everitt e Hothorn (2011), uma maneira informal de tomar essa decisão é avaliar as mudanças na altura do gráfico nos diferentes passos e escolher a maior observada. A ideia é que a maior mudança representa uma maior diferença no nível de fusão, o que sugere que o grupo pode se tornar menos homogêneo internamente com essa união.

A Figura 9 ilustra um exemplo onde 23 observações foram agrupadas pelo método hierárquico distância média. O ponto de maior mudança é facilmente identificado, o que produz uma divisão final com 3 agrupamentos. Mas nem sempre é fácil visualizar onde o corte deve ser realizado. Na Figura 10, que ilustra o dendrograma de 20 outras observações agrupadas utilizando o método hierárquico ligação simples, apesar do número de observações não ser muito grande, não é simples decidir onde o corte deve ser feito.

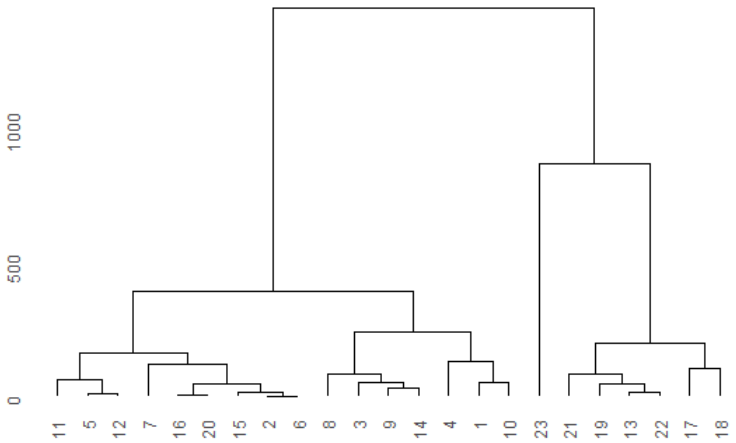


Figura 9 – Ilustração de corte no dendrograma com 23 observações.
Fonte: Da autora.

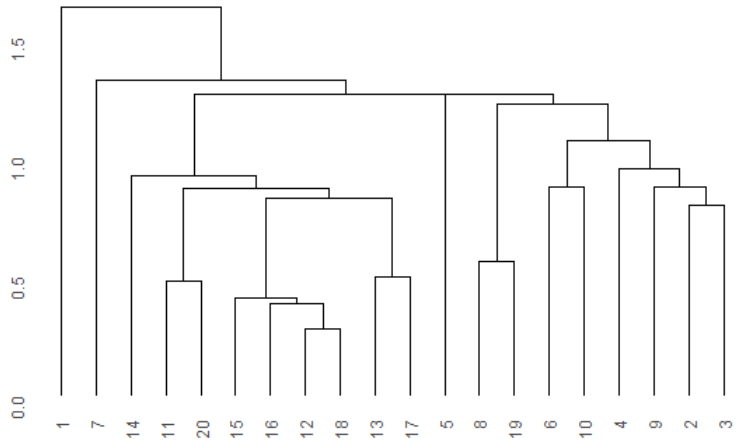


Figura 10 – Ilustração de corte no dendrograma com 20 observações.
Fonte: Da autora.

Outros critérios informais utilizam um gráfico da medida *versus* o número de grupos. Nesses métodos, o objetivo é identificar grandes mudanças no gráfico para um determinado número k. Segundo Tibshirani, Walther e Hastie (2001), dentre os métodos heurísticos está o *scree plot*. A proposta do método é que se faça um gráfico do número de grupos k da solução de agrupamento *versus* uma medida de erro W_k correspondente a cada passo. A medida W_k decresce monotonicamente conforme k aumenta, contudo, em um certo ponto, W_k cai abruptamente formando uma quebra do tipo "cotovelo", que indica o número de grupos que deve ser escolhido.

Como os critérios gráficos são subjetivos, várias técnicas mais formais têm sido propostas e alguns trabalhos avaliaram suas propriedades (MILLIGAN; COOPER, 1985; GORDON, 1998; DIMITRIADOU; DOLNICAR; WEINGESSEL, 2002; TIBSHIRANI; WALTHER; HASTIE, 2001; SUGAR; JAMES, 2011; FUJITA et al., 2014). Nesse sentido, segundo Tibshirani, Walther e Hastie (2001), a estatística *gap* foi proposta com o objetivo de formalizar a ideia de procurar pelo "cotovelo"no gráfico do número de grupos $versus$ algum critério de otimização. Os autores fizeram simulação de cinco cenários para a avaliação de seis critérios de escolha do número de grupos. As estatísticas comparadas foram as propostas por Calinski e Harabasz (1974), Krzanowski e Lai (1988), Hartigan (1975), Rousseeuw e Kaufman (1990) e duas variações da estatística *gap*. O trabalho de Fang e Wang (2012), além de propor um método, também o compara com outros, inclusive a estatística *gap*, que se saiu melhor quando os dados seguiam a normal multivariada e com o uso da distância euclidiana. O mesmo ocorre com outras distribuições simétricas, contudo, ela falha quando os dados seguem distribuições assimétricas.

Os autores Milligan e Cooper (1985) apresentaram um estudo de simulação de Monte Carlo para comparar 30 critérios de determinação do número de grupos. Os autores não incluíram nenhum método gráfico na análise, pois o objetivo foi testar as técnicas que buscam eliminar a subjetividade presente nesses métodos. Além disso, usaram dois critérios externos: índice de Jaccard e estatística Rand ajustada. Basicamente esses critérios usam informações externas ao processo de agrupamento para validação dos grupos obtidos. Nesse trabalho, a informação externa era a real estrutura dos grupos. Os autores utilizaram conjuntos de dados artificiais que continham 2, 3, 4 ou 5 grupos não sobrepostos, contendo 50 observações cada e bem separados. Com o intuito de obter diferentes partições finais, esses conjuntos de dados fictícios foram analisados por quatro métodos de agrupamento hierárquicos diferentes, sendo eles vizinho mais próximo, vizinho mais distante, distância média e método de Ward.

O trabalho citado identificou como as cinco técnicas com melhores desempenhos: em primeiro lugar a estatística pseudo F (CALINSKI; HARABASZ, 1974), em seguida o critério Je(2)/Je(1) (DUDA; HART et al., 1973), C-*Index* (HUBERT; LEVIN, 1976), Gamma (BAKER; HUBERT, 1975) e Beale (BEALE, 1969).

Esses também foram os 5 métodos identificados por Gordon (1998) como os melhores. A estatística pseudo F é a que aparece mais frequentemente na literatura como tendo o melhor desempenho na maioria das situações (MOOI; SARSTEDT, 2011; MINGOTI, 2005; EVERITT et al., 2011; LATTIN; CARROLL; GREEN, 2011) apresentando um bom custo-benefício, por levar em conta simplicidade e adequacidade.

Ainda de acordo com Gordon (1998), a estatística Traço de W apresentou desempenho ruim, apesar de ser uma das mais populares. A estatística $|T|/|W|$, proposta por Friedman e Rubin (1967), não acertou em nenhuma das 432 tentativas. No entanto, os autores ressaltaram que os resultados estão sujeitos a serem dependentes da estrutura de dados, ou seja, pode ser que a ordenação dos melhores testes seja modificada caso sejam testados com uma estrutura de dados diferentes. Os dados foram gerados com o uso da normal multivariada e isso pode ter contribuído para que alguns métodos não tenham um bom desempenho. Ou seja, em outras situações, o resultado poderia ser diferente.

Alguns autores sugerem uma combinação de métodos hierárquicos e das k-médias. Nessa abordagem, primeiro uma técnica hierárquica é aplicada para identificar o número de grupos *a priori* e, em seguida, o método das k-médias para classificar as observações (HAIR et al., 2009; MINGOTI, 2005). Nesse caso, as técnica de Ward ou ligação média são frequente-

mente utilizadas.

Diante dessa diversidade de abordagens e por não haver um critério adequado em todas as situações, é razoável levar em conta considerações práticas. Em alguns casos, pode haver alguma ideia *a priori* ou uma teoria que sugira uma estrutura nos dados. Entretanto, o mais importante é que os resultados sejam interpretáveis e tenham significado prático e útil (EVERITT et al., 2011; MOOI; SARSTEDT, 2011; CARVALHO; MATA; RESENDE, 2008).

3 DADOS E METODOLOGIA

3.1 DADOS

As bases de dados utilizadas neste trabalho são provenientes do Censo Demográfico 2010 realizado pelo IBGE, consultado a partir do Atlas do Desenvolvimento Humano no Brasil 2013 (disponível em www.atlasbrasil.org.br). Os dados estão tabulados em formato de planilhas, o que facilita seu tratamento. Para o desenvolvimento do trabalho foram escolhidas 8 variáveis demográficas, apresentadas na Tabela 2, que se relacionam ao processo de envelhecimento populacional. Os dados coletados são referentes aos 146 municípios da mesorregião Sul/Sudoeste de Minas Gerais (SSM), apresentados na Tabela 8 do Apêndice A. Essa mesorregião foi escolhida por ser onde os três *campi* da Universidade Federal de Alfenas (Unifal-MG) se encontram e com base na caracterização da mesorregião.

Tabela 2 – Siglas e descrições das variáveis demográficas do Atlas do Desenvolvimento Humano no Brasil.

sigla	variável	descrição
espvida	esperança de vida ao nascer	número médio de anos que um indivíduo espera viver a partir do nascimento, se permanecerem constantes ao longo da vida o nível e o padrão de mortalidade por idade prevalecentes no ano do Censo.
tft	taxa de fecundidade total	número médio de filhos que uma mulher deverá ter ao terminar o período reprodutivo, que compreende o grupo etário de 15 a 49 anos de idade.
mort1	taxa de mortalidade infantil	probabilidade de morrer antes de completar 1 ano de idade, por 1000 crianças nascidas vivas.
mort5	taxa de mortalidade até os 5 anos de idade	probabilidade de morrer entre o nascimento e a idade exata de 5 anos, por 1000 crianças nascidas vivas.
rd	razão de dependência	medida pela razão entre o número de pessoas com 14 anos ou menos e de 65 anos ou mais de idade (população considerada inativa) e o número de pessoas com idade de 15 a 64 anos (população potencialmente ativa) multiplicado por 100. O indicador mede, em termos relativos, a parcela da população potencialmente inativa (dependente) que deve ser sustentada pela potencialmente ativa.
sobre40	probabilidade de sobrevivência até 40 anos	probabilidade de uma criança recém-nascida viver até os 40 anos, se permanecerem constantes ao longo da vida o nível e o padrão de mortalidade por idade do ano do Censo.
sobre60	probabilidade de sobrevivência até 60 anos	probabilidade de uma criança recém-nascida viver até os 60 anos, se permanecerem constantes ao longo da vida o nível e o padrão de mortalidade por idade do ano do Censo.
t_env	taxa de envelhecimento	razão entre a população de 65 anos ou mais de idade e a população total multiplicado por 100.

Fonte: Da autora a partir de definições do Atlas Brasil.

As variáveis utilizadas na análise não são fornecidas diretamente nos Censos Demográficos, o Atlas do Desenvolvimento Humano no Brasil obtém essas variáveis a partir de técnicas indiretas propostas por Brass et al. (1968). Os indicadores de longevidade e mortalidade (espvida, mort1, mort5, sobre40 e sobre60) foram calculados usando dados da tábua de mortalidade infanto-juvenil construída para cada município. A taxa de fecundidade total foi obtida a partir da técnica de Brass de estimação. E, por último, as variáveis rd e t_env foram obtidas diretamente através de relações que consideram o tamanho da população de determinados grupos etários, fornecido pelo Censo Demográfico.

De forma geral, para a aplicação da técnica de mortalidade infanto-juvenil são necessários dados referentes ao total de filhos nascidos vivos e total de filhos na data do censo, por faixa etária das mulheres. Assim, por meio da técnica de Brass de estimação são obtidas as estimativas das probabilidades de morte e de sobrevivência, que serão utilizadas para a construção da tábua de mortalidade (BRASIL, 2016).

Inicialmente, adota-se uma coorte hipotética inicial de 100.000 nascimentos e aplica-se o conjunto estimado de probabilidades de sobrevivência (p_2 , p_3 e p_5) para obtenção do número de sobreviventes às exatas idades 2, 3 e 5 anos (l_2, l_3 e l_5). A partir dessas informações e utilizando uma tabela de referência ou padrão por transformação logito é possível estimar as demais funções da tábua de mortalidade (CARVALHO; SAWYER; RODRIGUES, 1994), sendo elas:

l_x = número de sobreviventes à exata idade x.

${}_nL_x$ = tempo a ser vivido, pelos sobreviventes da coorte, entre as idades x e (x+n).

${}_nd_x$ = número de óbitos, entre as idades x e (x+n), dos sobreviventes da coorte à exata idade x.

T_x = tempo a ser vivido a partir da exata idade x, pelos sobreviventes da coorte, até que a mesma se extinga.

e_x = número médio de anos a serem vividos, a partir da exata idade x, por um indivíduo representativo da coorte hipotética, caso ela experimente, ao longo de sua vida, a função de mortalidade vigente naquele local e determinado ano. A função é dada pela divisão T_x/l_x.

Considerando a metodologia proposta por Brass et al. (1968) e a construção da tábua de mortalidade, os indicadores coletados para o desenvolvimento desse trabalho foram calculados pelo Atlas do Desenvolvimento Humano no Brasil, da seguinte maneira:

1) Esperança de vida ao nascer

$$\text{espvida} = T_0/100.000 \tag{3.1}$$

2) Taxa de mortalidade infantil

$$\text{mort1} = \frac{d_0}{l_0} = \frac{l_0 - l_1}{l_0} \tag{3.2}$$

3) Taxa de mortalidade até os 5 anos de idade

$$\text{mort5} = \frac{l_0 - l_5}{l_0} \tag{3.3}$$

4) Probabilidade de sobrevivência até os 40 anos

$$\text{sobre40} = \frac{l_{40}}{l_0} \tag{3.4}$$

5) Probabilidade de sobrevivência até os 60 anos

$$\text{sobre60} = \frac{l_{60}}{l_0} \tag{3.5}$$

A técnica Brass de estimação também foi usada para estimar a taxa de fecundidade total dos municípios com base nos dados fornecidos pelo Censo Demográfico 2010. Com as informações sobre nascidos vivos durante os 12 meses anteriores à data do Censo (fecundidade corrente) e total de nascidos vivos (fecundidade retrospectiva ou parturição) foram estimadas as taxas específicas de fecundidade de cada município. A partir dessas taxas de mulheres em idade reprodutiva (15 a 49 anos) é obtida a tft, que corresponde ao número médio de filhos que uma mulher teria ao final do período reprodutivo se passasse pelas experiências de fecundidade observadas naquele período. As variáveis rd e t_env foram encontradas através das razões mencionadas na Tabela 2.

3.1.1 Relação das variáveis demográficas com o envelhecimento populacional

As variáveis selecionadas têm seu comportamento afetado em alguma medida com o processo de transição demográfica, que leva ao envelhecimento populacional. O aumento da longevidade, representado pelos ganhos no indicador esperança de vida, ocorre como consequência da redução dos níveis de mortalidade (CARVALHO; GARCIA, 2003). Portanto, é razoável dizer que quanto menor o nível de mortalidade de um município, maior será sua esperança de vida ao nascer. Além disso, esse aumento da longevidade se deve em grande parte à redução da mortalidade infantil (CAMARANO, 2014). Isso acontece porque em um primeiro momento a queda da mortalidade se concentrou nas idades mais jovens (CARVALHO; GARCIA, 2003). Portanto, à medida que se caminha no processo de transição demográfica, menores são os níveis de mortalidade infantil e de mortalidade até os 5 anos de idade. Ao mesmo tempo, a diminuição da mortalidade que de forma geral tem atingido todos os grupos etários, leva ao aumento tanto da probabilidade de sobrevivência até os 40 anos de idade, como até os 60 anos.

Outra componente que reduz de forma sustentada no processo de transição demográfica é a fecundidade, esse comportamento conduz ao início do processo de envelhecimento populacional (CARVALHO; WONG, 2008). Portanto, quanto menor a taxa de fecundidade total, mais envelhecido se torna o município. As variáveis razão de dependência e taxa de envelhecimento também estão associadas à fecundidade. Um aumento ou queda na razão de dependência pode ser resultado tanto de uma mudança na razão de dependência dos jovens, como na razão de dependência dos idosos.

Segundo Paiva e Wajnman (2005), no processo de transição demográfica é possível identificar 3 estágios diferentes relacionados ao comportamento da variável razão de dependência, devido às mudanças da estrutura etária da população. Na primeira fase da transição demográfica, a queda da mortalidade infantil gera um aumento na proporção de jovens na população e, consequentemente, contribui para o aumento da razão de dependência, via aumento da razão de dependência dos jovens. Em um segundo estágio, ocorre uma queda da razão de dependência, impulsionada pela redução da fecundidade, que provoca a redução da proporção de jovens na população. Portanto, nesses dois primeiros momentos, as modificações na razão de dependência são impulsionadas, em grande parte, pela razão de dependência dos jovens. Em todo esse período, a razão de dependência de idosos é pequena e exerce pouco peso sobre a razão de dependência. Contudo, com a contínua queda do nível de fecundidade e aumento da proporção da

população em idade avançada, a razão de dependência dos idosos irá aumentar, ultrapassando a de jovens e conduzindo ao aumento da razão de dependência. Esse terceiro estágio vai ocorrer quando a população estiver muito envelhecida. Em relação à taxa de envelhecimento, nesse contexto ela está representando a proporção da população de 65 anos ou mais na população. Ou seja, quanto maior seu valor maior a proporção de idosos naquele município.

Na literatura há diversos trabalhos que estudam o processo de transição demográfica e o consequente envelhecimento populacional. De forma geral, essas e outras variáveis são utilizadas nesses estudos, dentre elas: idade mediana, índice de envelhecimento, taxa bruta de natalidade, taxa bruta de mortalidade e taxa de crescimento anual (VASCONCELOS; GOMES, 2012; LUTZ; SANDERSON; SCHERBOV, 2008; CARVALHO; WONG, 2008). Neste trabalho, para garantir a qualidade e comparabilidade dos dados, foram utilizadas apenas as variáveis demográficas disponíveis no Atlas Brasil.

3.2 METODOLOGIA

Nesse trabalho foi utilizada a técnica multivariada de análise de agrupamento para identificar os grupos de municípios da mesorregião SSM com características semelhantes. As análises foram feitas usando a linguagem R (R Core Team, 2016) por meio do programa *RStudio* (RStudio Team, 2016). De acordo com Hair et al. (2009), existem suposições como a normalidade, linearidade e homecedasticidade que devem ser avaliadas na análise multivariada. No entanto, algumas técnicas são mais afetadas do que outras caso ocorra violação de alguma delas. No caso da análise de agrupamento, essas suposições não possuem tanta importância, pois exercem pouco peso sobre a análise. Os autores ressaltam ainda que, no uso de técnicas de agrupamento, especial atenção deve ser dada à multicolinearidade entre as variáveis estudadas.

Inicialmente foi realizada uma análise exploratória dos dados que auxiliou na aplicação das técnicas de agrupamento. Primeiro foi obtido um resumo estatístico das variáveis estudadas, em seguida foram apresentados os municípios com os três melhores e piores desempenhos nessas variáveis. A fim de identificar os pares de variáveis mais associadas entre si, foi calculada a matriz de correlações entre as variáveis, duas a duas. Tal matriz tem a forma apresentada na definição (2.4).

Após a etapa de formulação do problema, escolha das variáveis e análise exploratória

dos dados, foi iniciado o processo de agrupamento. Nesse sentido, o primeiro passo consistiu em selecionar qual seria o método de agrupamento para classificar os municípios. Foi escolhido o método não hierárquico das k-médias que demonstra um desempenho superior aos métodos hierárquicos (MOOI; SARSTEDT, 2011), como visto na seção 2.4.3. Em linhas gerais, com o uso dessa técnica, cada município foi alocado no grupo cujo vetor de médias amostral (centroide) era o mais semelhante ao vetor de valores observados para o respectivo município (MINGOTI, 2005).

O método das k-médias exige que seja definido *a priori* tanto o número de grupos desejado quanto os k centroides iniciais. Alguns critérios para escolha do número de grupos foram apresentados na seção 2.4.4, dentre eles o uso de uma técnica hierárquica aglomerativa. Nessa abordagem primeiro uma técnica hierárquica é aplicada para identificar o número de grupos *a priori* e, em seguida, são definidas as sementes iniciais e o método das k-médias é aplicado para classificar as observações. Nesse trabalho, foi escolhido o método de Ward para especificar o número de grupos pretendido. Além dele retornar melhores desempenhos que os outros métodos (EVERITT; HOTHORN, 2011), o método é considerado mais preciso e indicado para dados quantitativos contínuos (BLASHFIELD, 1976). As sementes iniciais foram selecionadas de forma aleatória dentro do conjunto de dados. Pena, Lozano e Larranaga (1999) apresentam um estudo comparativo de diferentes métodos de inicialização para as k-médias e a escolha aleatória aparece como um dos melhores dentre os comparados, por tornar o procedimento das k-médias mais efetivo e independente do agrupamento inicial. A documentação da função `kmeans` do `R` também sugere a escolha aleatória das sementes iniciais e que esse número seja 25. Portanto, primeiro foi aplicado o método de Ward para definir o número de grupos e, em seguida, o método das k-médias para a classificação dos municípios em grupos.

Para aplicação do método de Ward foi usada a distância de Mahalanobis para medir o quão próximos estão dois grupos ou observações. Essa medida é indicada quando as variáveis estão altamente correlacionadas, pois ela ajusta as correlações entre as variáveis e pondera igualmente cada variável do conjunto de dados original (HAIR et al., 2009). Além disso, assim como na abordagem não hierárquica, nesse caso também é necessário definir o número de grupos da partição final. Contudo, nos métodos hierárquicos aglomerativos o número é definido no final do processo de agrupamento, decidindo-se o ponto de corte do dendrograma. Diante disso, foram utilizados os critérios da maior diferença no nível de fusão e da interpretabilidade. O primeiro critério refere-se à maior altura no dendrograma, que indica que o grupo pode se

tornar menos homogêneo internamente com essa união (EVERITT et al., 2011). O segundo critério diz respeito à melhor solução cuja interpretação corresponda aos objetivos da análise.

A técnica de componentes principais (ACP) foi utilizada para reduzir a dimensionalidade dos dados. O método foi aplicado com a intenção de reduzir o conjunto das variáveis originais correlacionadas entre si a um novo conjunto de variáveis, os componentes principais, não correlacionadas. Os escores, valores numéricos dos componentes, dos dois primeiros componentes pricinpais foram plotados em um diagrama de dispersão para identificar a existência de grupos. Primeiro, esse procedimento foi aplicado antes da análise de agrupamento e, em seguida, com a partição proposta pelo método das k-médias.

Por fim, os grupos resultantes foram representados no mapa da mesorregião SSM com o intuito de verificar a proximidade geográfica dos municípios inseridos em cada grupo. Foram utilizados gráficos *boxplots* com o intuito de analisar a variabilidade dos dados estudados e comparar os grupos obtidos em relação às variáveis. Os *boxplots* permitem verificar o primeiro quartil, mediana, terceiro quartil, provavéis *outliers* e a variabilidade dos dados alocados dentro de cada grupo, de acordo com cada variável. O eixo horizontal do gráfico representa os grupos na ordem de 1 a 4 e o eixo vertical a variável analisada. Além disso, como suporte para avaliação dos grupos obtidos foram calculadas algumas medidas estatísticas como a mediana, o coeficiente de variação (CV), máximo e mínimo das variáveis de cada agrupamento. O CV é uma medida de dispersão relativa usada para avaliar a homogeneidade interna do agrupamento. Quanto menor seu valor, mais homogêneo é considerado o grupo em relação àquela variável. Por último, foram selecionadas as variáveis população (pop) e rendimento médio dos ocupados (renocup), com o intuito de caracterizar o comportamento dos grupos mais envelhecidos e menos envelhecidos em relação a população e renda. A Figura 11 mostra um fluxograma da metodologia utilizada nesse trabalho.

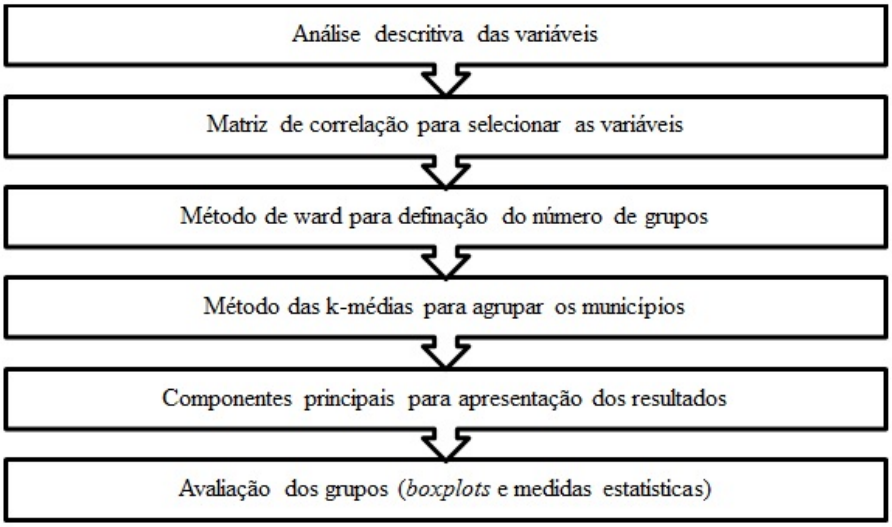


Figura 11 – Fluxograma da metodologia aplicada no trabalho.
Fonte: Da autora.

4 RESULTADOS E DISCUSSÃO

4.1 ANÁLISE DESCRITIVA DAS VARIÁVEIS

O primeiro passo consistiu em obter um resumo estatístico das variáveis demográficas, apresentado na Tabela 3. O resultado mostra que a menor esperança de vida ao nascer (espvida), no ano 2010, da mesorregião Sul/Sudoeste de Minas Gerais, foi 73,03 anos enquanto a maior 78,15 anos. Portanto, um indivíduo nascido, em 2010, no município com a maior esperança de vida ao nascer, por exemplo, esperava viver em média 78,15 anos. Esses valores sinalizam um regime de baixa mortalidade e alta esperança de vida ao nascer na mesorregião estudada.

Tabela 3 – Resumo estatístico das variáveis demográficas da mesorregião Sul Sudoeste de Minas Gerais, 2010.

	espvida	tft	mort1	mort5	rd	sobre40	sobre60	t_env
mínimo	73,03	1,33	10,35	12,11	37,68	92,33	79,54	5,46
1 quartil	74,44	1,79	13,40	15,63	43,26	93,23	81,67	8,49
mediana	75,56	1,95	14,45	16,86	44,85	93,92	83,32	9,39
média	75,46	1,95	14,69	17,09	45,24	93,85	83,15	9,45
3 quartil	76,28	2,08	16,18	18,86	47,27	94,35	84,36	10,32
máximo	78,15	2,70	18,50	21,55	53,20	95,99	87,58	14,85

Fonte: Da autora.

Em relação à taxa de fecundidade total (tft), a menor registrada foi de 1,33 filhos por mulher, ao passo que a maior foi de 2,70 filhos por mulher. Além disso, os dados mostram

que pelo menos 109 dos 146 municípios estudados já experimentam taxas de fecundidade total abaixo do nível de reposição (2,1 filhos por mulher), representado pelo terceiro quartil 2,08.

Uma das maneiras de medir a relação intergeracional dos municípios é por meio da razão de dependência total (rd). Quanto maior a razão, maior o peso da população considerada inativa (0 a 14 anos e 65 anos e mais de idade) sobre a população ativa (15 a 64 anos de idade). A mediana mostra que 50% dos valores estão abaixo de 44,85 e 50% acima. O menor valor do indicador encontrado registrou 37,68 dependentes para cada 100 pessoas potencialmente ativas e o maior atinge 53,20 dependentes. Como mencionado na seção 2.1, a razão de dependência total é formada pela soma entre a razão de dependência dos jovens e a razão de dependência dos idosos. Quando o Brasil era caracterizado por um regime de alta fecundidade e redução da mortalidade, a razão de dependência dos jovens tinha um peso muito maior que a dos idosos no cálculo da razão de dependência total. Atualmente, com a queda das taxas de fecundidade e o consequente envelhecimento da população, é razoável dizer que a razão de dependência dos idosos tem sido a principal responsável pelos aumentos observados na razão de dependência total.

Em relação ao indicador mortalidade infantil (mort1), o menor valor registrado é 12,11 que corresponde ao número de crianças que morreram antes de completar um ano de vida em cada 1000 crianças nascidas vivas. O maior valor registrado foi 18,50. A mortalidade até os 5 anos de idade (mort5), por sua vez, registrou o valor de 12,11 como o mínimo e 21,55 como o máximo. Os menores níveis de mortalidade podem ser associados a maior longevidade. Os municípios de Passos e Itajubá, por exemplo, apresentam os menores níveis de mortalidade e são responsáveis pelas maiores esperanças de vida ao nascer. Além disso, como a mortalidade infantil está contabilizada na mortalidade até os 5 anos de idade, os municípios associados aos maiores e menores valores desse indicador são os mesmos para os dois casos.

A maior probabilidade de sobrevivência até 40 anos de idade (sobre40) encontrada na mesorregião estudada é 0,96, ao passo que a menor é 0,92, aproximadamente. Em relação a probabilidade de sobrevivência até 60 anos de idade (sobre60), o maior valor registrado é 0,87 e o menor é 0,80, aproximadamente. Em relação à taxa de envelhecimento (t_env), a mediana do indicador foi de 9,39%, sendo o valor máximo 14,85% e o mínimo 5,46%. Além disso, 109 dos 146 municípios estudados experimentam taxa de envelhecimento menores que 10,32%. Uma taxa elevada reflete, principalmente, a redução dos níveis de fecundidade e o aumento da esperança de vida dos idosos.

A Tabela 4 apresenta os municípios que registram as três maiores esperanças de vida ao nascer, razão de dependência, sobrevida até os 40 e 60 anos de idade e taxas de envelhecimento; e também aqueles responsáveis pelas três menores taxas de fecundidade total, taxas de mortalidade infantil e até os 5 anos de idade.

Tabela 4 – Municípios com os melhores indicadores demográficos da mesorregião Sul/Sudoeste de Minas Gerais, 2010.

município	espvida	município	tft	município	mort1	mort5
Passos	78,15	São Sebastião do Rio Verde São João da Mata	1,33	Passos	10,35	12,11
Itajubá	78,06	Espírito Santo do Dourado	1,39	Itajubá	10,50	12,12
Guaxupé	77,81	Inconfidentes	1,41	Poços de Caldas	11,27	13,18
município	rd	município	sobre40	sobre60	município	t_env
Tocos do Moji	37,68	Itajubá	95,99	87,58	Córrego do Bom Jesus	14,85
São João da Mata	39,05	São Lourenço	95,62	86,63	Senador José Bento	13,65
Varginha	39,18	Passos	95,25	86,54	Pratápolis	12,97

Fonte: Da autora.

Os municípios responsáveis pelas maiores esperanças de vida ao nascer são Passos (78,15 anos), Itajubá (78,06 anos) e Guaxupé (77,81 anos). Os dois primeiros municípios também estão associados às menores taxas de mortalidade infantil (10,35 e 10,50 óbitos por mil nascidos, respectivamente) e até aos 5 anos de idade (12,11 e 12,12 óbitos por mil nascidos vivos, respectivamente), seguidos pela cidade de Poços de Caldas (11,27 (mort1) e 13,18 (mort5) óbitos por mil nascidos vivos). Como o indicador taxa de mortalidade até 1 ano de idade é levado em consideração na conta da taxa de mortalidade até os 5 anos de idade, é razoável que os municípios responsáveis pelos menores valores nesses indicadores sejam os mesmos. A mesma situação ocorre com os indicadores sobrevida até os 40 anos e até os 60 anos de idade. As maiores probabilidade de sobrevivência tanto até 40 anos, quanto até 60 anos de idade encontram-se em Itajubá (0,96 e 0,88), São Lourenço (0,96 e 0,87) e Passos (0,95 e 0,86), respectivamente.

As três menores taxas de fecundidade total da SSM foram registradas por São Sebastião do Rio Verde e São João da Mata, (1,33 filhos por mulher), Espírito Santo do Dourado (1,39 filhos por mulher) e Inconfidentes (1,41 filhos por mulher). Em todos esses municípios o indicador já está muito abaixo do nível de reposição (2,1 filhos por mulher) e abaixo da taxa registra para o estado de Minas Gerais no mesmo ano, 1,8 filhos por mulher. Em relação à razão de dependência total, os municípios associados ao menor peso da população considerada inativa sobre a população potencialmente ativa foram Tocos do Moji (37,68%), São João da Mata (39,05%) e Varginha (39,18%). Por último, os municípios com as menores taxas de participação dos idosos (65 anos ou mais de idade) em relação à população total do município (t_env) foram Córrego do Bom Jesus (14,85%), Senador José Bento (13,65%) e Pratápolis (12,97%).

A Tabela 5, por sua vez, apresenta os municípios que registram as três menores esperanças de vida ao nascer, razão de dependência, sobrevida até os 40 e 60 anos de idade, taxas de envelhecimento e os municípios associados as três maiores taxas de fecundidade total e mortalidade infantil e até os 5 anos de idade.

Tabela 5 – Municípios com os piores indicadores demográficos da mesorregião Sul/Sudoeste de Minas Gerais, 2010.

município	espvida	município	tft	município	mort1	mort5
Carmo da Cachoeira Divisa Nova São Bento Abade São Tomé das Letras	73,03	São Bento Abade	2,70	Carmo da Cachoeira Divisa Nova São Bento Abade São Tomé das Letras	18,50	21,55
Bandeira do Sul Ibitiúra de Minas Toledo	73,14	Senador Amaral	2,55	Bandeira do Sul Ibitiúra de Minas Toledo	18,4	21,34
Natércia	73,28	Carmo da Cachoeira	2,53	Natércia Fortaleza de Minas	18,1	21.06 21,00
município	rd	município	sobre40	sobre60	município	t_env
São Tomás de Aquino	53,20	Carmo da Cachoeira Divisa Nova São Bento Abade São Tomé das Letras	92,33	79,54	Senador Amaral	5,46
Divisa Nova	52,86	Bandeira do Sul Ibitiúra de Minas Toledo	92,40	79,71	São Tomé das Letras	5,65
Serrania	52,64	Natércia	92,49	79,92	São Bento Abade	5,90

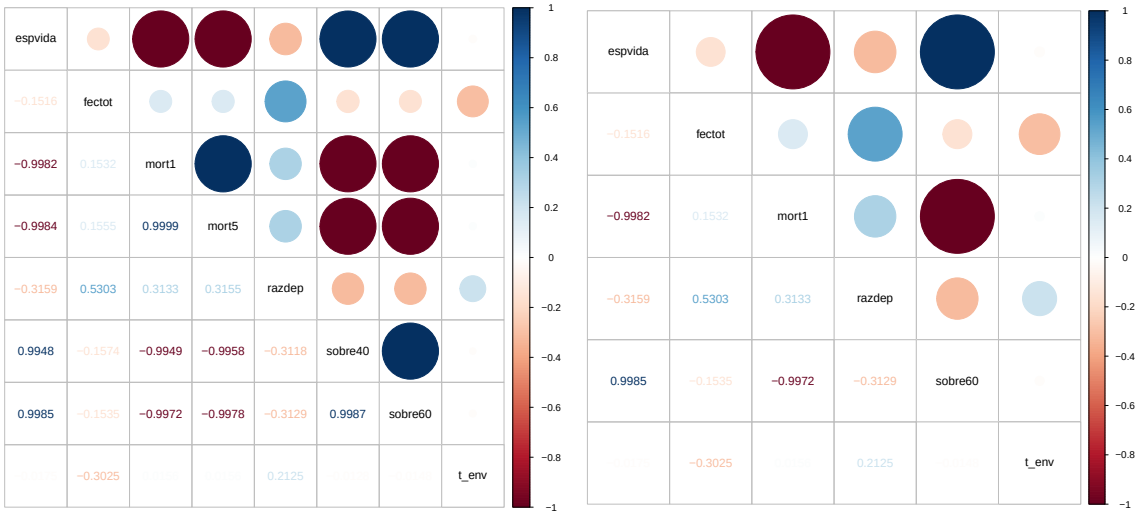
Fonte: Da autora.

Como pode ser observado na Tabela 5, os municípios Carmo da Cachoeira, Divisa Nova, São Bento Abade e São Tomé das Letras registraram a menor esperança de vida ao nascer (73,03 anos); a maior taxa de mortalidade infantil e até os os 5 anos de idade (18,50 óbitos por mil nascidos vivos e 21,55 óbitos por mil nascidos vivos, respectivamente); e menores probabilidade de sobrevivência até 40 e 60 anos (0,92 e 0,80), respectivamente. Os municípios Bandeira do Sul, Ibitiúra de Minas e Toledo foram responsáveis pela segunda menor esperança de vida ao nascer (73,14 anos), segunda maior taxa de mortalidade infantil (18,40 óbitos por mil nascidos vivos) e taxa de mortalidade até os 5 anos de idade (21,34 óbitos por mil nascidos vivos). Esses municípios também são os que apresentam a segunda posição entre as menores probabilidade de sobrevida até os 40 anos de idades (0,92) e até os 60 anos de idade (0,80), aproximadamente. Por último, o município Natércia ocupava a terceira posição entre as menores esperanças de vida ao nascer (73,28 anos), maiores taxa de mortalidade infantil (18,1 óbitos por mil nascidos vivos), taxa de mortalidade até os 5 anos de idade (21,06 óbitos por mil nascidos vivos) e menores probabilidades de sobrevida até os 40 anos (0,92) e até os 60 anos (0,80), aproximadamente.

Em relação à taxa de fecundidade total, as três maiores registradas estão em níveis muito acima do de reposição, São Bento Abade (2,70 filhos por mulher), Senador Amaral (2,55 filhos por mulher) e Carmo da Cachoeira (2,53 filhos por mulher). Os municípios com as maiores

razões de dependência foram São Tomás de Aquino (53,20%), Divisa Nova (52,86%) e Serrania (52,64%). As cidades com as menores taxas de participação dos idosos sobre a população total (taxa de envelhecimento) foram Senador Amaral (5,46%), São Tomé das Letras (5,65%) e São Bento Abade (5,90%). De forma geral, é razoável supor que os municípios com os melhores indicadores estão em um estágio mais avançado da transição demográfica, registrando menores níveis de mortalidade e fecundidade.

O próximo passo da análise exploratória do conjunto de dados foi a análise da correlação entre as variáveis. A Figura 12 apresenta as correlações entre cada um dos pares de variáveis demográficas selecionadas, antes e após a retirada de variáveis. Quanto maior o diâmetro do círculo, mais próximo de 1 é o valor do coeficiente de correlação e, portanto, maior o grau de associação entre as variáveis. Em relação às cores, a associação positiva é indicada pela cor azul e a negativa pela cor vermelha.



(a) Antes. (b) Depois.

Figura 12 – Correlações entre as variáveis (antes e após a retirada de variáveis).
Fonte: Da autora.

A variável espvida é correlacionada negativamente com as variáveis mort1 e mort5 e positivamente com as variáveis sobre40 e sobre60. Quanto menor a mortalidade, maior será a esperança de vida ao nascer e quanto maior a probabilidade de sobrevida, maior a esperança de vida ao nascer. As variáveis mort1 e mort5 apresentam correlação de 0,9999, ou seja, quanto maior a mortalidade infantil, maior será a mortalidade até os 5 anos de idades. Do mesmo modo, essas variáveis são correlacionadas negativamente com as probabilidades de sobrevida até aos 40 e 60 anos. As variáveis sobre40 e sobre60 também apresentam alta correlação (0,9987). Por fim, as variáveis tft e rd estão positivamente correlacionadas. Isso pode estar associado ao fato

de que a menor tft, em um primeiro momento, reduz a razão de dependência dos jovens, o que conduz à redução da razão de dependência total. As demais variáveis apresentam correlação quase nula.

De forma geral, as variáveis da análise são altamente correlacionadas. As maiores correlações encontradas foram entre as variáveis sobre40 e sobre60, bem como entre mort1 e mort5. Nesse sentido, foi necessário fazer a escolha de quais variáveis permaneceriam na análise. Como mort1 e mort5 representam informações sobre nível de mortalidade, foi decidido retirar mort5 da análise porque a mortalidade infantil é considerada mais representativa nesse contexto. Do mesmo modo, sobre60 permaneceu no estudo por ser considerada mais informativa que sobre40. As demais variáveis, apesar de apresentarem altas correlações, continuaram na análise porque elas representam informações interessantes, que poderiam ser úteis para os agrupamentos.

4.2 AGRUPAMENTOS

Inicialmente, com o intuito de visualizar a existência de possíveis agrupamentos nos dados, foram ilustrados os dois primeiros componentes principais nos eixos horizontal e vertical da Figura 13. Dessa forma, cada município foi representado pelo seu valor numérico dos componentes, os chamados escores. Os dois primeiros componentes principais explicaram 76% da variância total. Além disso, é possível observar que muitas observações (municípios) ficaram posicionadas muito próximas e outras mais distantes entre si. No entanto, a disposição dos municípios no gráfico não deixa clara a divisão das observações entre grupos.

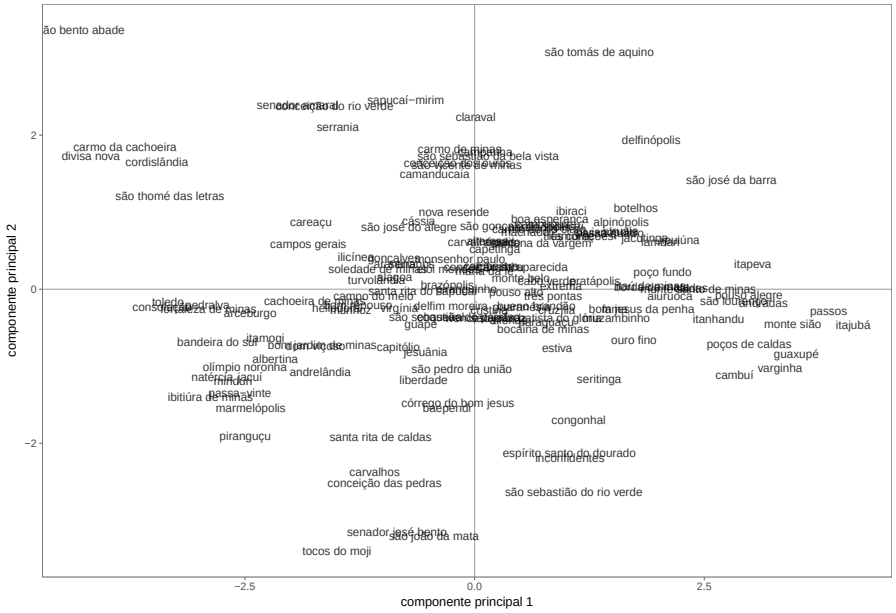


Figura 13 – Dispersão dos municípios em função dos escores dos componentes principais.
Fonte: Da autora.

Em seguida, foi obtido o dendrograma da análise de agrupamento pelo método de Ward e distância de Mahalanobis, apresentado na Figura 14. O corte foi realizado na altura que classifica as observações em quatro agrupamentos. Também foram examinadas as partições com cinco e sete grupos, mas considerando a interpretabilidade dos resultados, não foram encontradas evidências de que seriam mais adequadas. Portanto, o número predefinido de grupos usado para aplicação do método das k-médias foi quatro, encontrado a partir do método hierárquico de Ward.

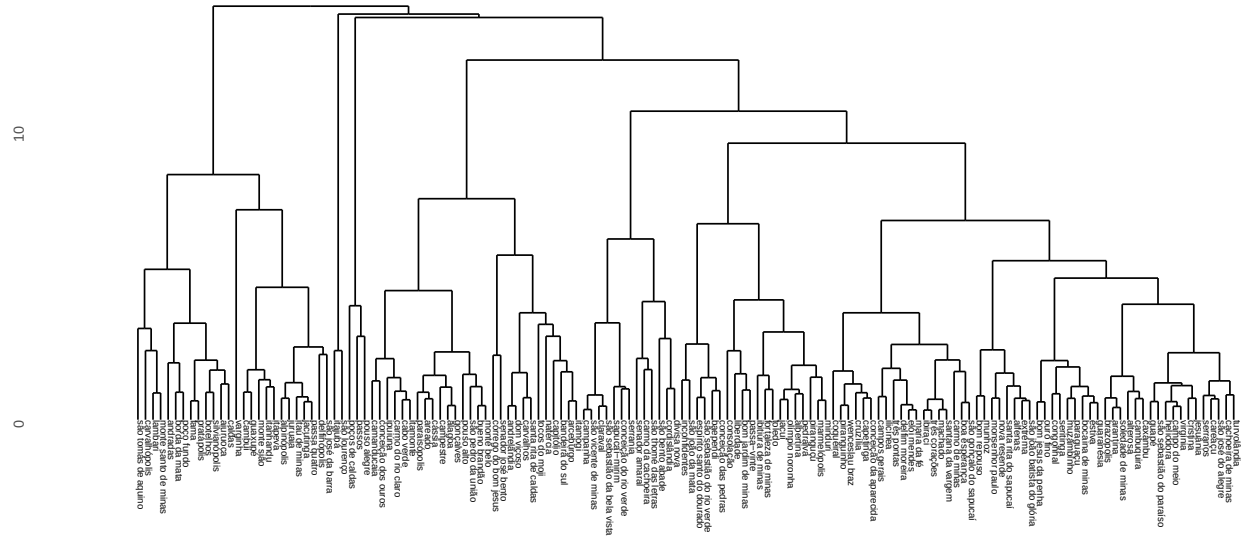


Figura 14 – Dendrograma pelo método de Ward e distância de Mahalanobis.
Fonte: Da autora.

Com o intuito de propor uma visualização espacial da partição dos dados, a Figura 15 mostra a dispersão dos municípios em função dos escores dos componentes principais dos quatro grupos obtidos pelo método das k-médias. Por esse método, os municípios foram divididos quanto ao processo de envelhecimento populacional, da seguinte forma: o grupo 1 (G1) é o menor agrupamento, composto por 17 municípios; o grupo 2 (G2) contém 46 municípios; no grupo 3 (G3) estão inseridos 36 municípios e, por último, o grupo 4 (G4) é formado por 47 municípios. As Tabelas 9, 10, 11 e 12 do Apêndice B apresentam os nomes dos municípios classificados dentro de cada grupo.

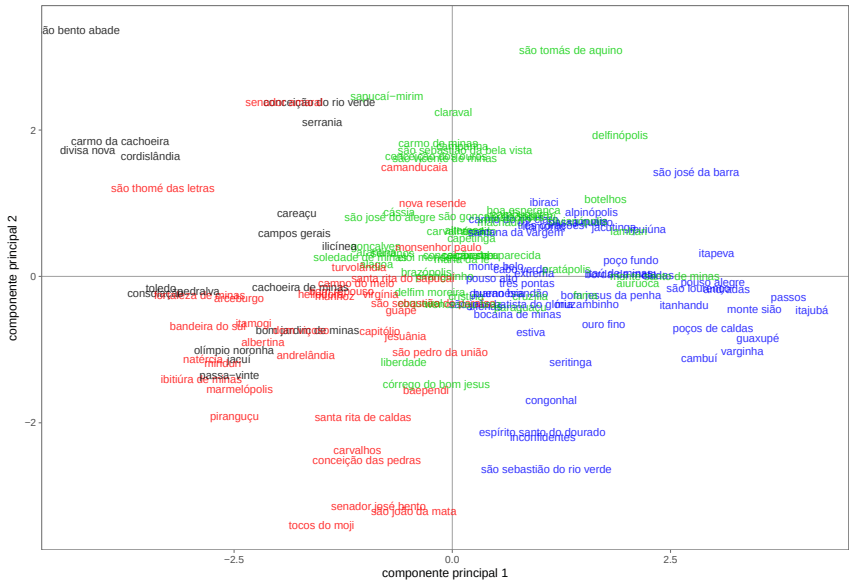


Figura 15 – Dispersão dos municípios em função dos escores dos componentes principais dos quatro grupos obtidos pelo método das k-médias.

Fonte: Da autora.

A Figura 16 apresenta os quatro grupos no mapa da mesorregião SSM. De forma geral, com exceção do grupo 1, é possível perceber uma certa tendência de municípios vizinhos pertencerem a um mesmo grupo. Contudo, não houve concentração dos municípios de cada grupo em uma única região do mapa. Por isso, ao longo de toda a extensão territorial da mesorregião são encontrados municípios em diferentes estágios do processo de envelhecimento populacional.

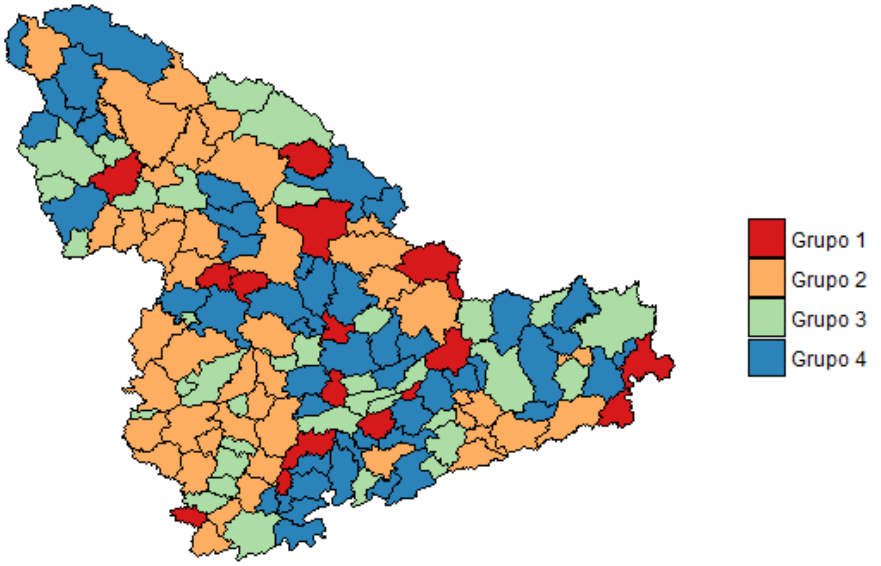


Figura 16 – Mapa dos municípios da mesorregião Sul/Sudoeste de Minas Gerais classificados em quatro grupos pelo método das k-médias.
Fonte: Da autora.

Para auxiliar na análise dos resultados foram utilizados gráficos *boxplots* apresentados na Figura 17 e algumas medidas estatísticas mostradas na Tabela 6. A partir desses resultados foi possível encontrar o perfil demográfico de cada agrupamento resultante. Foram considerados grupos mais envelhecidos aqueles que apresentaram os maiores valores das variáveis espvida e sobre60 e os menores de tft, mort1 e rd. Embora exista um consenso de que quanto maior o valor da t_env, mais envelhecido é o município, ela não foi decisiva na separação dos grupos.

Dois dos quatro grupos se destacaram por possuírem um perfil bem definido. No primeiro grupo estão os municípios menos envelhecidos, ou seja, aqueles que apresentam os piores desempenhos nos indicadores demográficos estudados. No segundo grupo estão os municípios mais avançados no processo de envelhecimento populacional. O terceiro e quarto grupos são caracterizados por apresentarem valores intermediários nas variáveis estudadas. O terceiro grupo apresenta um comportamento mais próximo do primeiro na maior parte das variáveis, enquanto o quarto grupo é mais próximo do grupo 2. Com isso, os grupos 2, 4, 3 e 1 representam a ordem dos mais envelhecidos para os menos envelhecidos.

A partir da comparação dos resultados das medidas estatísticas e *boxplots* é possível observar que o grupo 1 registra a menor mediana de espvida (73,56 anos). Além disso, o maior valor do indicador dentro desse grupo foi de 75,01 anos (Conceição do Rio Verde), que ainda está muito abaixo dos outros agrupamentos. A menor espvida do G1 foi de 73,03 anos, que foi registrada por Divisa Nova, Carmo da Cachoeira e São Bento Abade. Por outro lado, o grupo 2

foi responsável pelos maiores valores de mediana (76,60 anos), mínimo (75,45 anos) e máximo (78,15 anos). Os municípios de Alfenas e Passos representam os valores de mínimo e máximo do G2, respectivamente. Portanto, um indivíduo de qualquer município do segundo grupo esperava viver, em média, mais anos do que aqueles pertencentes aos demais. Por último, ainda considerando a variável espvida, o G2 apresentou a maior variabilidade nos dados, representada pela diferença entre o terceiro e primeiro quartil. O indicador espvida representa a longevidade do município, nesse sentido quanto menor seu nível de mortalidade, maior será espvida.

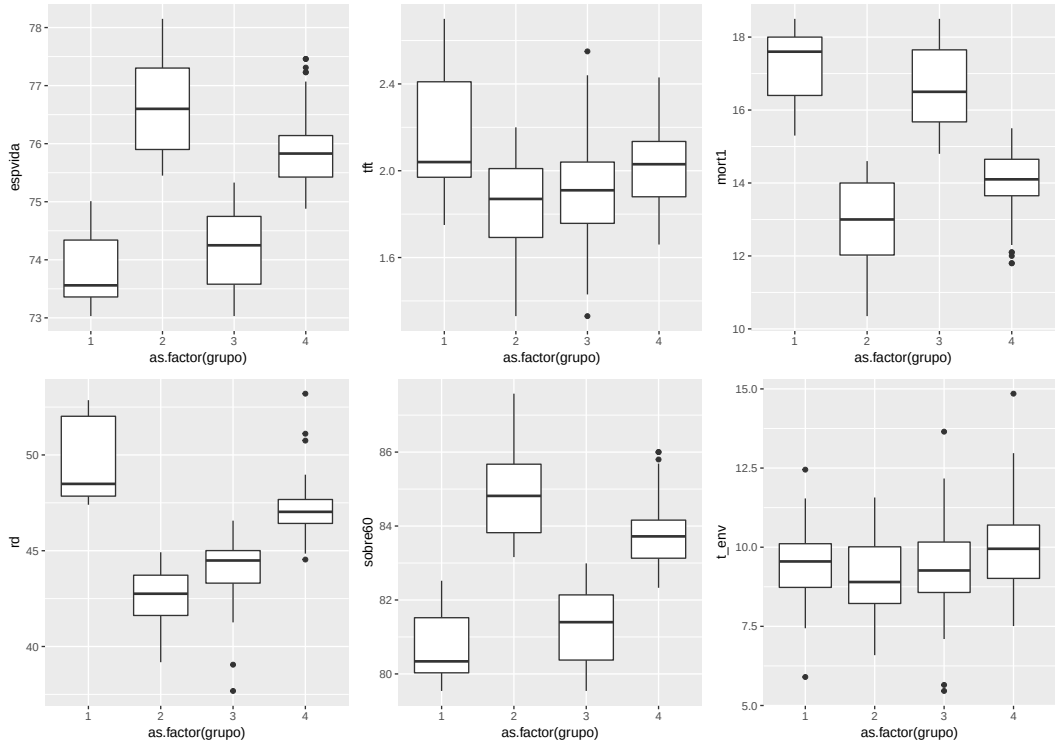


Figura 17 – *Boxplots* dos grupos da SSM de acordo com as variáveis: esperança de vida ao nascer (espvida), taxa de fecundidade total (tft), mortalidade infantil (mort1), razão de dependência (rd), probabilidade de sobrevivência até 60 anos (sobre60) e taxa de envelhecimento (t_env)

Fonte: Da autora.

Tabela 6 – Resumo estatístico das variáveis demográficas dos grupos obtidos pelo método das k-médias.

medida	grupo	espvida	tft	mort1	rd	sobre60	t_env
Mediana	1	73,56	2,04	17,60	48,49	80,34	9,55
	2	76,60	1,87	13,00	42,75	84,81	8,90
	3	74,25	1,91	16,50	44,49	81,40	9,26
	4	75,83	2,03	14,10	47,03	83,72	9,95
CV	1	0,93	13,84	6,53	4,31	1,28	16,23
	2	0,99	12,04	8,84	3,57	1,31	13,72
	3	0,94	14,38	6,84	4,25	1,29	17,73
	4	0,94	9,89	7,32	3,44	1,22	14,86
Mínimo	1	73,03	1,75	15,30	47,40	79,54	5,90
	2	75,45	1,33	10,35	39,18	83,16	6,59
	3	73,03	1,33	14,80	37,68	79,54	5,46
	4	74,88	1,66	11,80	44,54	82,33	7,51
Máximo	1	75,01	2,70	18,50	52,86	82,52	12,45
	2	78,15	2,20	14,60	44,92	87,58	11,57
	3	75,33	2,55	18,50	46,57	82,99	13,65
	4	77,46	2,43	15,50	53,20	86,00	14,85

Fonte: Da autora.

No caso da variável tft, o primeiro grupo apresentou mediana, mínimo e máximo superiores aos outros agrupamentos, o que contribuiu para que ele fosse classificado como o menos envelhecido. Pelo menos 50% dos municípios experimentaram tft maior que 2,04 filhos por mulher. Além disso, é necessário mencionar que 6 dos 17 municípios do grupo tiveram tft maior que o nível de reposição (2,10 filhos por mulher). O município de Jacuí registrou o menor valor do indicador (1,75 filhos por mulher). Em contrapartida, São Bento Abade foi responsável pela maior tft (2,70 filhos por mulher). Os resultados revelam ainda que o G1 foi o que apresentou maior variabilidade nos dados, considerando essa variável. Sob a perspectiva do grupo 2, o comportamento do nível de fecundidade é muito diferente. O menor nível foi de São Sebastião do Rio Verde (1,33 filhos por mulher) e o maior de Carmo do Rio Claro (2,20 filhos por mulher). Pelo menos metade dos municípios desse grupo possuem uma tft maior que 1,87 filhos por mulher. Além disso, apenas seis dos seus 46 municípios estão acima do nível de reposição e com valores muito próximos dele. Os indicadores desses seis municípios estão compreendidos entre 2,12 e 2,20 filhos por mulher. Vale ressaltar que quanto menor a tft do município, mais envelhecido ele se torna.

O primeiro grupo também apresentou o pior desempenho em relação à variável mort1. Esse comportamento é confirmado pelos altos valores registrados pela mediana (17,60 óbitos por mil nascidos vivos), mínimo (15,30 óbitos por mil nascidos vivos) e máximo (18,50 óbitos

por mil nascidos vivos). Como mencionado anteriormente, a longevidade de uma população está associada aos seus níveis de mortalidade. Isso fica evidente ao se observar que o município com a menor mortalidade infantil é o mesmo que apresentou a maior espvida (Conceição do Rio Verde). Da mesma forma, os municípios Divisa Nova, Carmo da Cachoeira e São Bento Abade são responsáveis pelos maiores níveis de mortalidade infantil e menores espvida. Por outro lado, o segundo grupo registrou mediana, mínimo e máximo inferiores aos demais. Pelo menos metade dos municípios apresentam mortalidade infantil inferior à 13 óbitos por mil nascidos vivos. O município de Passos experimentou a menor mortalidade infantil (10,35 óbitos por mil nascidos vivos) e Alfenas a maior (14,60 óbitos por mil nascidos vivos).

No que diz respeito ao indicador rd, no primeiro grupo foram observados os maiores valores de mediana, (48,49%), mínimo (47,40%) e máximo (52,86%). Os municípios de Cachoeira de Minas e Divisa Nova experimentaram o menor e maior valores do indicador, respectivamente. Deve-se lembrar que, à medida que a população avança no processo de envelhecimento, primeiro é esperado uma redução da rd. Logo, no ano estudado, os valores altos da variável não indicam um bom desempenho. Isso reafirma a desvantagem do G1 em relação aos outros. Por outro lado, o segundo grupo teve mediana, mínimo e máximo inferiores aos demais. Pelo menos 50% dos municípios desse grupo apresentaram rd menor que 42,75%, limitado por um valor mínimo de 39,18% (Varginha) e máximo de 44,92% (Passos).

Em relação à variável sobre60, dentre os municípios alocados no primeiro grupo, estão os que registraram o menor valor da mesorregião SSM. Os resultados apontam que Divisa Nova, São Bento Abade e Carmo da Cachoeira representam o mínimo de 79,54%. Esse agrupamento também registra valores de mediana (80,34%) e máximo (82,52%) inferiores aos outros três agrupamentos. Do mesmo modo, o G2 mais uma vez se destacou por um comportamento melhor que os demais. Pelo menos 50% dos municípios desse grupo experimentaram sobre60 maiores que 84,81%. O menor valor foi experimentando por Alfenas (83,16%) e o maior por Itajubá (87,58%).

E, finalmente, a análise da variável t_env mostrou que o primeiro grupo apresentou mediana de 9,55%. Essa foi a única variável desse agrupamento com pontos discrepantes. Os municípios de Consolação (12,45%) e São Bento Abade (5,90%) foram esses *outliers*. No segundo grupo, a mediana foi de 8,90%. O município de São José da Barra foi o responsável pelo menor valor do indicador (6,59%) e Poços de Caldas o maior (11,57%). O terceiro e quarto agrupamento serão analisados a seguir, contudo já é possível adiantar que as estatísticas dessa

variável pouco se diferenciaram entre os grupos. Em razão disso, mesmo usada como critério para estudar o envelhecimento populacional, essa variável pouco contribuiu para entender a classificação dos municípios entre os grupos.

No que diz respeito ao terceiro e quarto grupos, observou-se que os municípios classificados neles possuem um comportamento intermediário nos indicadores considerados. Em relação às variáveis espvida, mort1 e sobre60, o G4 teve um comportamento semelhante ao G2. Ao passo que o G3 registrou valores mais aproximados do G1, ou seja, caracterizando um grupo menos envelhecido. Em pelo menos 50% dos municípios do G3, os indivíduos esperavam viver em média mais do que 74,25 anos, enquanto no G4 esse valor foi de 75,83 anos. Além disso, esses valores foram limitados por um máximo de 75,33 anos (São Sebastião do Paraíso) no terceiro grupo e 77,46 anos no quarto (Monte Santo de Minas e São Tomás de Aquino). O valor mínimo foi de 73,03 anos no terceiro agrupamento (São Thomé das Letras) e 74,88 anos no quarto (Soledade de Minas). O G4 além de ser caracterizado pela menor variação nos dados, é o único que apresenta pontos discrepantes, sendo eles São Tomás de Aquino e Monte Santo de Minas (77,46 anos), Delfinópolis (77,31 anos), Lambari e Aiuruoca (77,23 anos).

Quanto à variável mort1, os resultados indicam que o quarto grupo foi o que apresentou menor variabilidade. Além do mais, também foram identificados pontos discrepantes nesse agrupamento, sendo eles São Tomás de Aquino e Monte Santos de Minas (11,80 óbitos por mil nascidos vivos), Delfinópolis (12 óbitos por mil nascidos vivos), Lambari e Aiuruoca (12,10 óbitos por mil nascidos vivos). A mediana de 14,10 óbitos por mil nascidos vivos foi menor que do terceiro grupo (16,50 óbitos por mil nascidos vivos). Os resultados de mínimo e máximo reafirmam o melhor desempenho do G4 comparado ao G3. O limite inferior do quarto agrupamento (11,80 óbitos por mil nascidos vivos) é menor que do terceiro (14,80 óbitos por mil nascidos vivos). Eles são representados pelos municípios de Monte Santo de Minas e São Tomás de aquino (grupo 4) e São Sebastião do Paraíso (grupo 3). Do mesmo modo, o limite superior é menor para o quarto agrupamento (15,50 óbitos por mil nascidos vivos) do que para o terceiro (18,50 óbitos por mil nascidos vivos. Os municípios de Soledade de Minas e São Thomé das Letras são os responsáveis por esses indicadores, respectivamente.

No caso da variável sobre60, o quarto grupo foi o único com pontos discrepantes, sendo eles São Tomás de Aquino e Monte Santo de Minas (86%) e Delfinópolis (85,80%). Quando comparado aos grupos 1 e 3, é possível observar que o quarto agrupamento apresentou melhor desempenho na sobre60. O ponto de mínimo desse grupo foi de 82,33% (Liberdade) e o de

máximo foi de 86% (São Tomás de Aquino e Monte Santo de Minas). Pelo menos 50% dos municípios pertencentes a esse agrupamento apresentaram sobre60 maior que 83,72%. Por fim, o G3 registrou valores de mínimo, máximo e mediana inferiores aos grupos 2 e 4. Isso também contribuiu para esse agrupamento ser classificado como menos envelhecido, assim com o primeiro. A mediana foi de 81,40% com valores limitados por um mínimo de 79,54% (São Thomé das Letras) e máximo de 82,99% (Nova Resende).

Apesar das estatísticas da t_env serem muito parecidas entre os agrupamentos e, em função disso, contribuírem menos para a divisão dos grupos, o quarto grupo parece apresentar um comportamento melhor que o terceiro. A mediana desse agrupamento foi de 9,95%, ao passo que do terceiro foi de 9,26%. O menor valor do indicador registrado no G4 foi de 7,51% (Carmo de Minas) e o maior, um ponto discrepante no conjunto de observações, foi de 14,85% (Córrego do Bom Jesus). O município Senador Amaral foi responsável pelo valor mínimo no G3 (5,46%) e Senador José Bento pelo máximo (13,65%). Os dois municípios também são pontos discrepantes, assim como São Thomé das Letras (5,65%).

Em contrapartida, quando se analisam os agrupamentos em relação às variáveis tft e rd, verifica-se que o grupo 3 apresentou melhores resultados que o 4. Pelo menos 50% dos municípios do terceiro agrupamento experimentaram tft abaixo de 1,91 filhos por mulher, enquanto no quarto agrupamento esse valor alcançou 2,03 filhos por mulher. Os valores de mínimo e máximo do G3 são dois pontos discrepantes, sendo eles Senador Amaral (2,55 filhos por mulher) e São João da Mata (1,33 filhos por mulher). No G4, os municípios de Liberdade e São Tomás de Aquino representam o mínimo (1,66 filhos por mulher) e máximo (2,43 filhos por mulher), respectivamente. Por último, no que diz respeito à rd, os municípios classificados no G3 apresentam valores inferiores aos do G4. O mínimo do terceiro agrupamento foi de 37,68 (Tocos do Moji), enquanto do quarto é de 44,54 (Campestre). O município de Turvolândia registrou o valor máximo (46,57) no terceiro grupo, ao passo que São Tomás de Aquino (53,20) no quarto grupo.

Para avaliar a homogeneidade interna dos grupos foi usado o coeficiente de variação (CV) de cada variável. Quanto menor seu valor, mais homogêneo é considerado o agrupamento em relação àquela variável. De forma geral, os resultados mostram que todos os grupos apresentaram valores baixos em relação a todas as variáveis, o que aponta para homogeneidade interna. A variável espvida foi responsável pelos menores valores de CV, enquanto as variáveis t_env e tft pelos maiores. Por último, as variáveis população (pop) e rendimento médio dos ocupados

(renocup) foram selecionadas para auxiliar na caracterização dos grupos obtidos. A Figura 18 mostra os *boxplots* dos grupos de acordo com essas variáveis e a Tabela 7 apresenta a mediana, o CV, o mínimo e máximo dos grupos para cada uma dessas variáveis.

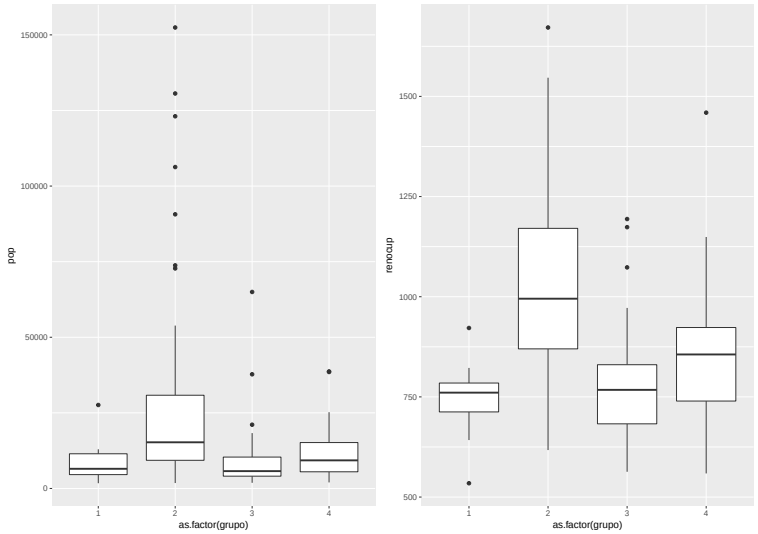


Figura 18 – *Boxplots* dos grupos da SSM de acordo com as variáveis população (pop) e rendimento médio dos ocupados (renocup).

Fonte: Da autora.

Tabela 7 – Resumo estatístico das variáveis população (pop) e rendimento médio dos ocupados (renocup) dos grupos obtidos pelo método das k-médias.

medida	grupo	pop	renocup
Mediana	1	6.501	760,59
	2	15.263	995,09
	3	5.729	767,55
	4	9.289	855,90
CV	1	74,82	11,72
	2	120,16	23,32
	3	120,22	18,97
	4	73,18	17,99
Mínimo	1	1.727	534,51
	2	1.789	617,29
	3	1.868	563,11
	4	1.995	558,92
Máximo	1	27.600	921,96
	2	152.435	1671,92
	3	64.980	1193,90
	4	38.688	1459,40

Fonte: Da autora.

A variável pop foi escolhida com o intuito de conhecer o tamanho da população dos municípios dentro de cada grupo. Os resultados mostram que pelo menos 50% dos municípios

do G1 (grupo menos envelhecido) possuem uma população menor que 6.501 habitantes, com um mínimo de 1.727 (Consolação) e máximo de 27.600 (Campos Gerais), que é um ponto discrepante no conjunto de dados. Portanto, o que se observa é um grupo composto por pequenos municípios. No grupo mais envelhecido (G2), por sua vez, estão inseridos grande parte dos municípios maiores como Poços de Caldas, Pouso Alegre, Varginha e Passos, que representam os únicos municípios com mais de 100.000 habitantes da mesorregião. No entanto, nesse grupo também estão inseridos municípios que estão entre os menores da SSM em termos de população, como por exemplo, São Sebastião do Rio Verde (2.107 habitantes), Bom Jesus da Penha (3.842 habitantes), Espírito Santo do Dourado (4.426 habitantes), entre outros. Pelo menos 50% dos municípios do segundo grupo apresentam população menor que 15.263 habitantes, com um mínimo de 1.789 (Seritinga) e máximo de 152.435 (Poços de Caldas). Além disso, esse é o grupo que possui maior variabilidade nos dados e também o maior número de *outliers*, como pode ser obervado pelo *boxplot*.

A migração interna, que corresponde aquela que ocorre dentro de um mesmo território, ou seja, entre regiões, estados e municípios do país, contribui pra explicar a classificação de pequenos municípios em um grupo mais avançado no processo de envelhecimento populacional. De forma geral, o que se observa é uma migração de jovens para os municípios maiores e mais desenvolvidos, em busca de trabalho e renda (CAMPOS; BARBIERI, 2013). Diante disso, a migração interna exerce papel importante na definição da estrutura etária da população. Isso ocorre porque a mudança dos jovens para os grandes centros contribui para intensificar o processo de envelhecimento populacional nos municípios menores (WONG; CARVALHO, 2006). A saída deles, além de aumentar a proporção de idosos na população, também contribui pra reduzir o nível de fecundidade da população.

Por outro lado, a estrutura etária das cidades maiores também se torna, em parte, reflexo da imigração dos jovens. Isso ajuda a explicar a razão de os municípios de Pouso Alegre, Varginha e Três Corações estarem entre aqueles que experimentam as menores proporções de pessoas com 65 anos (t_env) da mesorregião SSM. Os jovens migrantes que ingressam nesses municípios contribuem, em um primeiro momento, para aumentar o denominador da taxa (população total) e, portanto, reduzir a proporção de idosos. Essa análise sobre a variável t_env é interessante, pois ela mostra como uma visão geral sobre as variáveis analisadas separadamente pode comprometer a realidade dos municípios. Caso o envelhecimento da população fosse estudado usando apenas a proporção de idosos com mais de 65 anos de idade, Varginha que possui

t_env de 7,17% seria considerada menos envelhecida que Consolação (12,45%), por exemplo. Entretanto, considerando as demais variáveis demográficas na análise multivariada, Varginha foi classificada no grupo mais envelhecido e Consolação no menos envelhecido.

Campos e Barbieri (2013) também discutem o processo de migração dos idosos, concluindo que é possível observar duas direções nos fluxos migratórios desse grupo. A primeira refere-se a idosos que migram de municípios maiores para os menores, em busca de segurança e qualidade de vida. A segunda diz respeito aos idosos de cidades menores que migram para cidades maiores para acompanhar familiares e/ou na tentativa de obter melhor assistência à saude. Entretanto, a migração dos jovens é muito mais expressiva e exerce um peso muito maior para o envelhecimento da população do município de origem.

No que diz respeito aos demais grupos, o terceiro (G3) registrou mediana de 5.729 habitantes, mínimo de 1.868 (Senador José Bento) e máximo de 64.980 (São Sebastião do Paraíso), que representa um *outlier*. Esse foi o agrupamento com a menor variabilidade nos dados. O G4 apresentou mediana de 9.289 habitantes, Serranos é o menor município desse grupo, com 1.995 habitantes, enquanto Machado é o maior (38.688). Além disso, os grupos não parecem tão homogêneos em relação à variável pop, principalmente os grupos 2 e 3 que registraram um alto coeficiente de variação.

A variável rendimento médio dos ocupados, que representa em reais a média dos rendimentos de todos os trabalhos das pessoas ocupadas de 18 anos ou mais de idade, também foi analisada em cada grupo obtido. Berquó e Cavenaghi (2006) mostram a contínua redução do declínio da taxa de fecundidade total no Brasil e os diferenciais produzidos nesse processo, através das variáveis educação e renda. Ainda de acordo com as autoras, as mulheres com baixo rendimento domiciliar e escolaridade apresentavam maiores níveis de fecundidade do que aquelas em situação melhor. Essa associação negativa da renda e da escolaridade com o nível de fecundidade foi estudada considerando uma amostra de mulheres em idade reprodutiva (15 a 49 anos) e, especificamente, as variáveis rendimento domiciliar *per capita* e diferentes faixas de anos de estudo. Dada a dificuldade de obter e caracterizar os grupos em relação à escolaridade das mulheres, foi escolhida apenas uma variável relacionada à renda.

Em linhas gerais, foi observado que o grupo mais envelhecido (G2) apresentou melhor desempenho sobre a variável renocup e também maior variabilidade nos dados, como pode ser visto pelo *boxplot*. Em pelo menos 50% dos municípios desse agrupamento o rendimento médio dos ocupados é maior que R$ 995,09, Juruaia foi o município responsável pelo maior

rendimento médio dos ocupados (R$ 1.671,96) e Pedralva pelo menor (R$ 617,29). Por outro lado, a mediana do G1 (menos envelhecido) foi de R$ 760,59, o mínimo de R$ 534,51 (Consolação) e máximo de R$ 921,96. Os resultados revelam que nesse grupo são registradas as menores estatísticas relacionadas ao renocup. Os grupos 3 e 4 apresentaram comportamento intermediário, no entanto o G4 registrou melhor desempenho que o G3 em relação a essa variável. As medianas nesses grupos foram de R$ 767,55 (G3) e R$ 855,90 (G4). Os municípios de Senador José Bento e Alagoa foram responsáveis pelos menores valores da variável renocup (R$ 563,11 e R$ 558,92) nos grupos 3 e 4, respectivamente. Por outro lado, Camanducaia e Caxambu registraram os maiores valores (R$ 1.193,90 e R$ 1.459,40) nesses mesmos agrupamentos. Portanto, comparando os quatro grupos em relação à variável renda, foi observado que, de forma geral, os municípios dos grupos mais envelhecidos registram um rendimento médio dos ocupados maior que aqueles pertecentes aos grupos menos envelhecidos.

Para avaliar a homogeneidade interna dos grupos em relação às variáveis pop e renocup foram calculados seus coeficientes de variação. Os resultados mostram que as duas variáveis registraram valores um pouco mais elevados que aquelas utilizadas na análise multivariada. No entanto, o que mais chama atenção são os altos valores da variável pop, o que indica que os grupos são pouco homogêneos em relação ao tamanho da população, ou seja, os grupos são formados por municípios com tamanhos populacionais muito diferentes.

5 CONSIDERAÇÕES FINAIS

Um tema recorrente para a teoria da transição demográfica diz respeito ao seu momento de início, magnitudade e velocidade, que são diferentes para os diversos países e regiões do mundo. Nesse sentido, eventualmente todos os municípios do Brasil passariam pela transição demográfica e o consequente envelhecimento da população, contudo, não de maneira homogênea. Isso fica nítido com a classificação dos municípios da mesorregião Sul/Sudoeste de Minas Gerais em mais de um agrupamento, representando os diferentes estágios no processo de envelhecimento populacional. O método das k-médias propôs uma divisão dos municípios em quatro grupos: o primeiro (G1) constituído por municípios menos envelhecidos, o segundo grupo (G2) formado por municípios que se encontram em um estágio mais avançado do processo de envelhecimento da população e, por último, os grupos G3 e G4 que assumiram posições intermediárias entre os dois primeiros (G3 menos envelhecido que o G4). Aproximadamente 64% dos municípios foram classificados nos grupos considerados mais envelhecidos (G2 e G4), o que corresponde a 93 dos 146 municípios.

Esses resultados podem servir como subsídios para os formuladores de políticas públicas, pois diferentes dinâmicas populacionais demandam políticas públicas específicas para cada grupo de municípios, de acordo com sua estrutura etária. Um grupo de municípios menos envelhecido possivelmente demandará mais recursos direcionados às crianças e à população economicamente ativa. Portanto, é razoável supor que maior atenção seja dada à educação e mercado de trabalho, por exemplo. Por outro lado, os municípios mais envelhecidos precisam se dedicar mais à assistência para a população idosa tanto no âmbito da saúde, como no de infraestrutura, para que seja capaz de atender às necessidades de um envelhecimento ativo e saudável (WONG; CARVALHO, 2006). Diante disso é necessário que a transição da estrutura etária seja considerada para a alocação eficiente de recursos destinados à população (BRITO, 2007).

A análise simultânea das diferentes variáveis permitiu uma avaliação muito mais ampla do processo de envelhecimento populacional na mesorregião SSM. Os grupos foram considerados mais envelhecidos ou menos envelhecidos entre si com base na comparação do desempenho das variáveis em cada um deles. Contudo, o quão envelhecido cada grupo está frente ao processo de transição demográfica, não foi objeto de estudo desse trabalho. Além disso, esse trabalho possui limitações quanto às variáveis e ao método de agrupamento. Para garantir a qualidade

e comparabilidade dos dados foram usadas apenas variáveis demográficas disponíveis no Atlas do Desenvolvimento Humano (construído a partir dos censos demográficos do IBGE). Em relação aos métodos de agrupamento, como não há uma classificação evidente dos municípios em relação ao processo de envelhecimento populacional, foi escolhido o método das k-médias, por demonstrar um desempenho superior aos métodos hierárquicos (MOOI; SARSTEDT, 2011). No entanto, a investigação do problema a partir de outros métodos e/ou distâncias podem resultar em diferentes agrupamentos.

Nesse sentido, como agenda para trabalhos futuros a inclusão de diferentes variáveis como por exemplo idade mediana, o índice de envelhecimento e outros indicadores demográficos e socioeconômicos poderia fornecer mais informações e até mesmo organizar os municípios em grupos diferentes. Além disso, outras combinações de critérios para o uso do metodo das k-médias ou de diferentes técnicas de agrupamentos podem contribuir para o estudo do processo de envelhecimento populacional.

REFERÊNCIAS

BAKER, F. B.; HUBERT, L. J. Measuring the power of hierarchical cluster analysis. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 70, n. 349, p. 31–38, 1975.

BARTHOLOMEW, D. J. et al. **Analysis of multivariate social science data**. Boca Raton: CRC press, 2008.

BEALE, E. **Cluster analysis**. London: Scientific Control System, 1969.

BERQUÓ, E.; CAVENAGHI, S. Fecundidade em declínio: breve nota sobre a redução no número médio de filhos por mulher no brasil. **Novos Estudos - CEBRAP**, São Paulo, n. 74, p. 11–15, 2006.

BLASHFIELD, R. K. Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. **Psychological Bulletin**, v. 83, n. 3, p. 377–388, 1976.

BORGES, G. M.; CAMPOS, M. B. de; SILVA, L. G. de Castro e. Transição da estrutura etária no brasil: oportunidades e desafios para a sociedade nas próximas décadas. In: ERVATTI, L. R.; BORGES, G. M.; JARDIM, A. d. P. (Ed.). **Mudança demográfica no Brasil no início do século XXI subsídios para as projeções da população**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística IBGE, 2015, (Estudos e análises). p. 138–151.

BRASIL. Atlas do desenvolvimento humano no brasil 2013. 2016.

BRASIL. Ministério da Previdência Social. **Previdência Social: reflexões e desafios**. Brasília: MPS, 2009. v. 30, 232 p.

BRASS, W. et al. **Tire Demography of Tropical Africa**. [S.l.]: Princeton: Princeton University Press, 1968.

BRITO, F. A transição demográfica no brasil: as possibilidades e os desafios para a economia e a sociedade. **Texto para discussão**, CEDEPLAR/UFMG, Belo Horizonte, n. 318, p. 28, 2007.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics-theory and Methods**, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974.

CAMARANO, A. A. O. **Novo regime demográfico: uma nova relação entre população e desenvolvimento?** Rio de Janeiro: Instituto de Pesquisa Econômica Aplicada (Ipea), 2014.

CAMPOS, M. B. de; BARBIERI, A. F. Considerações teóricas sobre as migrações de idosos. **Revista Brasileira de Estudos da População**, São Paulo, v. 30, p. 69–84, 2013.

CARVALHO, A. X. Y.; MATA, D. D.; RESENDE, G. M. Clusterização dos municípios

brasileiros. In: **Dinâmica dos Municípios**. Brasília: Instituto de Pesquisa Econômica Aplicada (IPEA), 2008. p. 181–207.

CARVALHO, J. A. M. d.; GARCIA, R. A. O envelhecimento da populaçäo brasileira: um enfoque demográfico. **Cad. saúde pública**, Rio de Janeiro, v. 19, n. 3, p. 725–733, 2003.

CARVALHO, J. A. M. D.; SAWYER, D. O.; RODRIGUES, R. do N. **Introdução a alguns conceitos básicos e medidas em demografia**. 1994.

CARVALHO, J. A. M. d.; WONG, L. R. A transição da estrutura etária da população brasileira na primeira metade do século XXI. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 24, n. 3, p. 597–605, 2008.

DIMITRIADOU, E.; DOLNICAR, S.; WEINGESSEL, A. An examination of indexes for determining the number of clusters in binary data sets. **Psychometrika**, Springer, v. 67, n. 1, p. 137–159, 2002.

DOBRIANSKY, P. J.; SUZMAN, R. M.; HODES, R. J. Why population aging matters: A global perspective. **National Institute on Aging, National Institutes of Health, US Department of Health and Human Services, US Department of State**, 2007.

DUDA, R. O.; HART, P. E. et al. **Pattern classification and scene analysis**. [S.l.]: Wiley New York, 1973. v. 3.

EVERITT, B.; HOTHORN, T. **An introduction to applied multivariate analysis with R**. New York: Springer-Verlag, 2011.

EVERITT, B. S. et al. **Cluster analysis**. 5. ed. United Kingdom: John Wiley & Sons, 2011.

FANG, Y.; WANG, J. Selection of the number of clusters via the bootstrap method. **Computational Statistics & Data Analysis**, Elsevier, v. 56, n. 3, p. 468–477, 2012.

FERREIRA, D. F. **Estatística multivariada**. 2. ed. Lavras: Editora UFLA, 2011. 675 p.

FRIEDMAN, H. P.; RUBIN, J. On some invariant criteria for grouping data. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 62, n. 320, p. 1159–1178, 1967.

FUJITA, A. et al. A non-parametric statistical test to compare clusters with applications in functional magnetic resonance imaging data. **Statistics in medicine**, Wiley Online Library, v. 33, n. 28, p. 4949–4962, 2014.

GORDON, A. D. Cluster validation. In: **Data science, classification, and related methods**. [S.l.]: Springer, 1998. p. 22–39.

____. **Classification**. 2. ed. United States of America: Chapman and Hall, 1999.

HAIR, J. F. et al. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009.

HARTIGAN, J. A. **Clustering algorithms**. New York: John Wiley & Sons, 1975.

HUBERT, L. J.; LEVIN, J. R. A general statistical framework for assessing categorical clustering in free recall. **Psychological bulletin**, American Psychological Association, v. 83, n. 6, p. 1072, 1976.

JR, J. H. W. Hierarchical grouping to optimize an objective function. **Journal of the American statistical association**, Taylor & Francis, v. 58, n. 301, p. 236–244, 1963.

KRZANOWSKI, W. J.; LAI, Y. A criterion for determining the number of groups in a data set using sum-of-squares clustering. **Biometrics**, JSTOR, p. 23–34, 1988.

LATTIN, J.; CARROLL, J. D.; GREEN, P. E. **Análise de Dados Multivariados**. São Paulo: Cengage Learning, 2011.

LEE, R. The demographic transition: three centuries of fundamental change. **The journal of economic perspectives**, Berkeley, v. 17, n. 4, p. 167–190, 2003.

LIMA-COSTA, M. F.; VERAS, R. Saúde pública e envelhecimento. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 19, n. 3, p. 700–701, 2003.

LIMA, J. O. **Uma comparação do método fuzzy e redes neurais artificiais com os procedimentos de agrupamentos hierárquicos e não hierárquicos tradicionais**. 2001. Dissertação (Mestrado) — Universidade Federal de Minas Gerais - UFMG, Belo Horizonte.

LUTZ, W.; SANDERSON, W.; SCHERBOV, S. The coming acceleration of global population ageing. **Nature**, v. 451, n. 7179, p. 716–719, 2008.

MARIN, M. J. S.; PANES, V. C. B. Envelhecimento da população e as políticas públicas de saúde. **Revista do Instituto de Políticas Públicas de Marília**, Marília, v. 1, n. 1, p. 26–34, 2015.

MCLACHLAN, G. J. Mahalanobis distance. **Resonance**, Springer, v. 4, n. 6, p. 20–26, 1999.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, Springer, v. 50, n. 2, p. 159–179, 1985.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.

MOOI, E.; SARSTEDT, M. **A Concise Guide to Market Research**. Germany: Springer-Verlag Berlin Heidelberg, 2011.

PAIVA, P. d. T. A.; WAJNMAN, S. Das causas às consequências econômicas da transição demográfica no brasil. **Revista Brasileira de Estudos da População**, São Paulo, v. 22, n. 2, p. 13–15, 2005.

PENA, J. M.; LOZANO, J. A.; LARRANAGA, P. An empirical comparison of four

initialization methods for the k-means algorithm. **Pattern recognition letters**, San Sebastián, v. 20, n. 10, p. 1027–1040, 1999.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2016. Disponível em: <https://www.R-project.org/>.

REIS, P. R. d. C.; SILVEIRA, S. D. F. R.; BRAGA, M. J. Previdência social e desenvolvimento socioeconômico: impactos nos municípios de pequeno porte de minas gerais. **RAP: Revista Brasileira de Administração Pública**, v. 47, n. 3, 2013.

ROUSSEEUW, P. J.; KAUFMAN, L. **Finding Groups in Data**. [S.l.]: Wiley Online Library, 1990.

RStudio Team. **RStudio: Integrated Development Environment for R**. Boston, MA, 2016. Disponível em: <http://www.rstudio.com/>.

SUGAR, C. A.; JAMES, G. M. Finding the number of clusters in a dataset. **Journal of the American Statistical Association**, Taylor & Francis, 2011.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 63, n. 2, p. 411–423, 2001.

VASCONCELOS, A. M. N.; GOMES, M. M. F. Transição demográfica: a experiência brasileira. **Epidemiol. Serv. Saúde**, Brasília, v. 21, n. 4, p. 539–548, 2012.

WONG, L. L. R.; CARVALHO, J. A. O rápido processo de envelhecimento populacional do brasil: sérios desafios para as políticas públicas. **Revista Brasileira de Estudos da População**, São Paulo, v. 23, n. 1, p. 5–26, 2006.

ANEXOS

ANEXO A - Municípios da mesorregião Sul/Sudoeste de Minas Gerais.

Tabela 8 – Municípios da mesorregião Sul/Sudoeste de Minas Gerais.

Municípios			
Aiuruoca	Coqueiral	Itajubá	Pouso Alto
Alagoa	Cordislândia	Itamogi	Pratápolis
Albertina	Córrego do Bom Jesus	Itamonte	Santa Rita de Caldas
Alfenas	Cristina	Itanhandu	Santa Rita do Sapucaí
Alpinópolis	Cruzília	Itapeva	Santana da Vargem
Alterosa	Carmo do Rio Claro	Itaú de Minas	São Bento Abade
Andradas	Carvalhópolis	Jacuí	São Gonçalo do Sapucaí
Andrelândia	Carvalhos	Jacutinga	São João Batista do Glória
Arantina	Cássia	Jesuânia	São João da Mata
Arceburgo	Caxambu	Juruaia	São José da Barra
Areado	Claraval	Lambari	São José do Alegre
Baependi	Conceição da Aparecida	Liberdade	São Lourenço
Bandeira do Sul	Conceição das Pedras	Machado	São Pedro da União
Boa Esperança	Conceição do Rio Verde	Maria da Fé	São Sebastião da Bela Vista
Bocaina de Minas	Conceição dos Ouros	Marmelópolis	São Sebastião do Paraíso
Bom Jardim de Minas	Congonhal	Minduri	São Sebastião do Rio Verde
Bom Jesus da Penha	Consolação	Monsenhor Paulo	São Thomé das Letras
Bom Repouso	Delfim moreira	Monte Belo	São Tomás de Aquino
Borda da Mata	Delfinópolis	Monte Santo de Minas	São Vicente de Minas
Botelhos	Divisa Nova	Monte Sião	Sapucaí-Mirim
Brazópolis	Dom Viçoso	Munhoz	Senador Amaral
Bueno Brandão	Elói Mendes	Muzambinho	Senador José Bento
Cabo verde	Espírito Santo do Dourado	Natércia	Seritinga
Cachoeira de Minas	Estiva	Nova Resende	Serrania
Caldas	Extrema	Olímpio Noronha	Serranos
Camanducaia	Fama	Ouro Fino	Silvianópolis
Cambuí	Fortaleza de Minas	Paraguaçu	Soledade de Minas
Cambuquira	Gonçalves	Paraisópolis	Tocos do Moji
Campanha	Guapé	Passa Quatro	Toledo
Campestre	Guaranésia	Passa-Vinte	Três Corações
Campo do Meio	Guaxupé	Passos	Três Pontas
Campos Gerais	Heliodora	Pedralva	Turvolândia
Capetinga	Ibiraci	Piranguçu	Varginha
Capitólio	Ibitiúra de Minas	Piranguinho	Virgínia
Careaçu	Ilicínea	Poço Fundo	Wenceslau Braz
Carmo da Cachoeira	Inconfidentes	Poços de Caldas	
Carmo de Minas	Ipuiúna	Pouso Alegre	

Fonte: Da autora.

ANEXO B - Grupos obtidos pelo método das k-médias.

Tabela 9 – Municípios da mesorregião Sul/Sudoeste de Minas Gerais classificados no grupo 1 (G1) pelo método das k-médias.

municípios		
Bom Jardim de Minas	Consolação	Passa-Vinte
Cachoeira de Minas	Cordislândia	Pedralva
Campos Gerais	Divisa Nova	São Bento Abade
Careaçu	Ilicínea	Serrania
Carmo da Cachoeira	Jacuí	Toledo
Conceição do Rio Verde	Olímpio Noronha	

Fonte: Da autora.

Tabela 10 – Municípios da mesorregião Sul/Sudoeste de Minas Gerais classificados no grupo 2 (G2) pelo método das k-médias.

municípios		
Alfenas	Guaxupé	Passos
Alpinópolis	Ibiraci	Poço Fundo
Andradas	Inconfidentes	Poços de Caldas
Bocaina de Minas	Ipuiúna	Pouso Alegre
Bom Jesus da Penha	Itajubá	Pouso Alto
Borda da Mata	Itamonte	Santana da Vargem
Bueno Brandão	Itanhandu	São João Batista do Glória
Cabo Verde	Itapeva	São José da Barra
Caldas	Itaú de Minas	São Lourenço
Cambuí	Jacutinga	São Sebastião do Rio Verde
Carmo do Rio Claro	Juruaia	Seritinga
Congonhal	Monte Belo	Três Corações
Espírito Santo do Dourado	Monte Sião	Três Pontas
Estiva	Muzambinho	Varginha
Extrema	Ouro Fino	
Guaranésia	Passa Quatro	

Fonte: Da autora.

Tabela 11 – Municípios da mesorregião Sul/Sudoeste de Minas Gerais classificados no grupo 3 (G3) pelo método das k-médias.

municípios		
Albertina	Fortaleza de Minas	Piranguçu
Andrelândia	Guapé	Santa Rita de Caldas
Arceburgo	Heliodora	Santa Rita do Sapucaí
Baependi	Ibitiúra de Minas	São João da Mata
Bandeira do Sul	Itamogi	São Pedro da União
Bom Repouso	Jesuânia	São Sebastião do Paraíso
Camanducaia	Marmelópolis	São Thomé das Letras
Campo do Meio	Minduri	Senador Amaral
Capitólio	Monsenhor Paulo	Senador José Bento
Carvalhos	Munhoz	Tocos do Moji
Conceição das Pedras	Natércia	Turvolândia
Dom Viçoso	Nova Resende	Virgínia

Fonte: Da autora.

Tabela 12 – Municípios da mesorregião Sul/Sudoeste de Minas Gerais classificados no grupo 4 (G4) pelo método das k-médias.

municípios		
Aiuruoca	Claraval	Monte Santo de Minas
Alagoa	Conceição da Aparecida	Paraguaçu
Alterosa	Conceição dos Ouros	Paraisópolis
Arantina	Coqueiral	Piranguinho
Areado	Córrego do Bom Jesus	Pratápolis
Boa Esperança	Cristina	São Gonçalo do Sapucaí
Botelhos	Cruzília	São José do Alegre
Brazópolis	Delfim Moreira	São Sebastião da Bela Vista
Cambuquira	Delfinópolis	São Tomás de Aquino
Campanha	Elói Mendes	São Vicente de Minas
Campestre	Fama	Sapucaí-Mirim
Capetinga	Gonçalves	Serranos
Carmo de Minas	Lambari	Silvianópolis
Carvalhópolis	Liberdade	Soledade de Minas
Cássia	Machado	Wenceslau Braz
Caxambu	Maria da Fé	

Fonte: Da autora.

ANEXO C - Código utilizado para as análises.

```
# Dados do Sul/Sudoeste de MG sobre envelhecimento
# AA usando variáveis originais e distância de Mahalanobis
# métodos: Ward e k-médias

# fase preliminar ------------------------------------------------------
# carregar dados Sul/Sudoeste de Minas
load("sul_dem.rda")

# pacotes utilizados
library(corrplot)     # correlações
library(ecodist)      # Ward - distância de Mahalanobis
library(ggplot2)      # gráficos
library(ggthemes)     # gráficos
library(ggdendro)     # dendrogramas
library(dplyr)        # manipulação de dados
library(gridExtra)    # apresentação de gráficos combinados

# limpar dados
# tirar pop
sul <- sul[,-13]
# nomes dos municípios como rótulos
sul <- as.data.frame(sul)
row.names(sul) <- sul$nome.mun
# resumo estatístico
summary(sul[,-c(1:4)])
# CV
Xb <- apply(sul[,-c(1:4)],2,mean)
S <- apply(sul[,-c(1:4)],2,sd)
CV <- S/Xb*100
CV
# correlações
graphics.off()
sul.r <- cor(sul[,-c(1:4)])
corrplot.mixed(sul.r,tl.col = "black",
               number.digits = 4, number.font = 0.1,
               tl.cex = 1, is.corr = F)
# retirar as variáveis correlacionadas com outras (mort5 e sobre40)
# e que, de alguma forma, calculam a mesma coisa
sul <- sul[,-c(8,10)]
# correlações após retirar as variáveis
graphics.off()
sul.r <- cor(sul[,-c(1:4)])
corrplot.mixed(sul.r,tl.col = "black",
               number.digits = 4, number.font = 0.1,
               tl.cex = 1, is.corr = F)
```

```
# variáveis originais - Mahalanobis ------------------------------------
# matriz de distâncias - Mahalanobis
dis <- (distance(sul[,-c(1:4)],method="mahalanobis"))^0.5
# agrupamento hierárquico de Ward
wr <- hclust(dis, method = "ward.D2")
ggdendrogram(wr)
# dendrograma com opções - Ward
hc <- dis %>% hclust(method = "ward.D2")
ggdendrogram(hc)
ddata <- hc %>% as.dendrogram() %>% dendro_data()
ggdendrogram(hc) + geom_text(size= 2, aes(x = x, y = y, label = label,
                                          angle = -90, vjust=0, hjust = 0), data= label(ddata)) +
  scale_y_continuous(expand = c(0.6, 0)) +
  scale_x_continuous(expand = c(0.1, 0)) +
  theme(axis.text.x = element_blank())

# gráficos - dois primeiros CPs - Mahalanobis
X_cp <- princomp(sul[,-c(1:4)], cor=T)
summary(X_cp) # os dois primeiros contabilizam 76% da variância total
escores <- X_cp$scores[,1:2]
escores <- as.data.frame(escores)
graf_CP <- ggplot(data=escores,aes(x=escores[,1],y=escores[,2],label=rownames(escores)))
graf_CP +  geom_hline(yintercept=0,colour="gray65") +
  geom_vline(xintercept=0,colour="gray65") +
  geom_text(colour="black",alpha=0.8,size=4) +
  ggtitle(" ") + theme_few() + ylab("componente principal 2") +
  xlab("componente principal 1")

# k-médias - k = 4
set.seed(1)
k_4 <- kmeans(sul[,-c(1:4)], 4, nstart=25)
labk4 <- k_4$cluster
graf_CP +  geom_hline(yintercept=0,colour="gray65") +
  geom_vline(xintercept=0,colour="gray65") +
  geom_text(colour=labk4,alpha=0.8,size=4) +
  ggtitle(" ") + theme_few() + ylab("componente principal 2") +
  xlab("componente principal 1")

# avaliar grupos obtidos
table(labk4)
# municípios dentro de cada grupo
sapply(unique(labk4),function(g)row.names(sul[,-c(1:4)])[labk4 == g])
# avaliar medidas estatisticas - variáveis originais -
aggregate(sul[,-c(1:4)],list(labk4),mean)
aggregate(sul[,-c(1:4)],list(labk4),median)
aggregate(sul[,-c(1:4)],list(labk4),min)
aggregate(sul[,-c(1:4)],list(labk4),max)
# CV
Xb <- aggregate(sul[,-c(1:4)],list(labk4),mean)
```

```
S <- aggregate(sul[,-c(1:4)],list(labk4),sd)
CV <- S/Xb*100
CV

# boxplots
# k-médias, k = 4
# espvida
sul4_km <- sul
sul4_km$grupo <- labk4
a <- ggplot(data = sul4_km, aes(as.factor(grupo), espvida)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(min(sul$espvida),max(sul$espvida)))
# tft
b <- ggplot(data = sul4_km, aes(as.factor(grupo), tft)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(min(sul$tft),max(sul$tft)))
# mort1
c <- ggplot(data = sul4_km, aes(as.factor(grupo), mort1)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(min(sul$mort1),max(sul$mort1)))
# rd
d <- ggplot(data = sul4_km, aes(as.factor(grupo), rd)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(min(sul$rd),max(sul$rd)))
# sobre60
e <- ggplot(data = sul4_km, aes(as.factor(grupo), sobre60)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(min(sul$sobre60),max(sul$sobre60)))
# t_env
f <- ggplot(data = sul4_km, aes(as.factor(grupo), t_env)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(min(sul$t_env),max(sul$t_env)))
# combinando os gráficos
grid.arrange(a,b,c,d,e,f,ncol=3)

# outras variáveis não usadas no agrupamento:
# pop e renocup
load("sul_dem_n.Rda")
# resumo estatístico
aggregate(dados_n,list(labk4),median)
aggregate(dados_n,list(labk4),min)
aggregate(dados_n,list(labk4),max)
# CV
Xb <- aggregate(dados_n,list(labk4),mean)
S <- aggregate(dados_n,list(labk4),sd)
CV <- S/Xb*100
CV
# boxplots
# k-médias, k = 4
```

```
# pop
dados_n$grupo <- labk4
a <- ggplot(data = dados_n, aes(as.factor(grupo), pop)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(min(dados_n$pop),max(dados_n$pop)))
# renocup
b <- ggplot(data = dados_n, aes(as.factor(grupo), renocup)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(min(dados_n$renocup),max(dados_n$renocup)))
grid.arrange(a,b,ncol=2)

# mapas
library(maptools)
library(rgdal)
library(dplyr)
library(rgeos)

# criar data frame com nome do município e a qual grupo pertence
z <- as.data.frame(labk4)
x <- data.frame(nome.mun=row.names(z), grupo=as.factor(z$labk4))
x
# incluir as outras informações dos municípios
sul$nome.mun <- as.factor(sul$nome.mun)
x$nome.mun <- as.factor(x$nome.mun)
xx <- inner_join(sul, x, by="nome.mun")

# faça o download dos arquivos necessários para criar o mapa em:
# ftp://geoftp.ibge.gov.br/organizacao_do_territorio/malhas_territoriais/malhas_municipais/
municipio_2015/UFs/MG/mg_municipios.zip
# descompacte os arquivos em uma pasta e ajuste o endereço abaixo:
mgm <- readOGR(dsn="C:/Users/Larissa/Documents/Dissertacao_Larissa/mapa",
               layer="31MUE250GC_SIR")
mgm@data <- rename(mgm@data, codmun7=CD_GEOCMU)
mgm$codmun7 <- as.factor(mgm$codmun7)
mga <- right_join(mgm@data, xx, by="codmun7")
mgf <- fortify(mgm, region="codmun7")

mga$id <- mga$codmun7
mgf <- inner_join(mgf, mga, by = "id")
ggplot(mgf, aes(long, lat, group=group, fill=grupo)) +
  geom_polygon(colour='black') + coord_equal() + theme_void() +
  scale_fill_brewer(name="",palette = "Spectral", labels=c("Grupo 1", "Grupo 2", "Grupo 3", "Grupo 4"))
```