UNIVERSIDADE FEDERAL DE ALFENAS
UNIFAL-MG

LARISSA GONÇALVES SOUZA

O envelhecimento populacional nos municípios do Sul/Sudoeste de Minas Gerais: análise de agrupamento

ALFENAS - MG
2016

LARISSA GONÇALVES SOUZA

Classificação dos municípios brasileiros segundo gastos previdenciários utilizando componentes principais e análise de agrupamento

Dissertação parcial apresentada à Universidade Federal de Alfenas, como parte dos requisitos para obtenção do título de Mestre em Estatística Aplicada e Biometria.
Orientadora: Patrícia de Siqueira Ramos
Coorientador: Lincoln Frias

ALFENAS - MG
2016

RESUMO

Nas últimas décadas, o Brasil passou de um cenário em que experimentava elevadas taxas de mortalidade e fecundidade para outro com menores taxas, a mudança conhecida como transição demográfica. Essas transformações, que conduzem ao envelhecimento populacional, não ocorreram de maneira uniforme em todas as regiões do país, produzindo diferenciais demográficos. Nesse sentido, o objetivo deste trabalho é propor uma classificação dos municípios da mesorregião Sul/Sudoeste de Minas Gerais em relação ao processo de envelhecimento populacional. Em um primeiro momento, foi realizada a análise de agrupamento utilizando as variáveis originais. Em seguida, o agrupamento foi realizado com os escores dos componentes principais, com o intuito de comparar os resultados obtidos. Os dados utilizados no trabalho são provenientes do Censo Demográfico 2010 do IBGE, consultados por meio do Atlas do Desenvolvimento Humano no Brasil. As variáveis selecionadas 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. Para identificar os grupos de municípios com características semelhantes foram aplicados cinco métodos hierárquicos de agrupamento (ligação simples, ligação completa, distância média, centroide e *Ward*) e o método não hierárquico das k-médias. Por último, foram avaliados diferentes critérios para a definição do número de grupos na partição final da análise de agrupamento.

Palavras-chave: Envelhecimento. Análise de agrupamento. Variáveis originais. Componentes principais. Número de grupos.

SUMÁRIO

1 Introdção

Nas últimas décadas, o Brasil tem passado de forma gradual e progressiva a apresentar uma nova configuração de seu regime demográfico, caracterizado pelo envelhecimento de sua população (CAMARANO, 2014). De forma geral, todo o país, antes do início do processo, possuía um perfil demográfico representado por muitas mortes, 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 considerado uma preocupação. 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 (GURALNIK et al., 1995). 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.

Atualmente, há razoável disponibilidade de dados demográficos de municípios do Brasil, 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 propor uma classificação dos municípios da mesorregião Sul/Sudoeste de Minas Gerais em relação ao processo de envelhecimento populacional. Em um primeiro momento, foi realizada a análise de agrupamento utilizando as variáveis originais. Em seguida, o agrupamento foi realizado com os escores dos componentes principais, com o intuito de comparar os resultados obtidos. Para identificar os grupos de municípios com características semelhantes foram aplicados cinco métodos hierárquicos de agrupamento (ligação simples, ligação completa, distância média, centroide e *Ward*) e o método não hierárquico das k-médias. O que se espera é que os municípios escolhidos apresentem grande homogeneidade interna e grande heterogeneidade externa, em relação às variáveis analisadas. Por último, diferentes critérios para definição do número de grupos na partição final da análise de agrupamento são analisados. De forma geral, esse trabalho é um bom teste para as técnicas de agrupamento, por apresentar certa homogeneidade entre os grupos obtidos (coeficientes de variação baixos), sem outliers e muitas observações.

2 Revisão de Literatura

O objetivo desse capítulo é apresentar a transição demográfica brasileira e suas principais características. Em seguida, como consequência desse processo, examinar o envelhecimento populacional brasileiro. Por fim, são apresentados aspectos da análise multivariada, análise de componentes principais, análise de agrupamento e critérios para determinação do número de grupos.

2.1 A transição demográfica no Brasil

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 reduzir 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 apresentava 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 do gráfico 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.

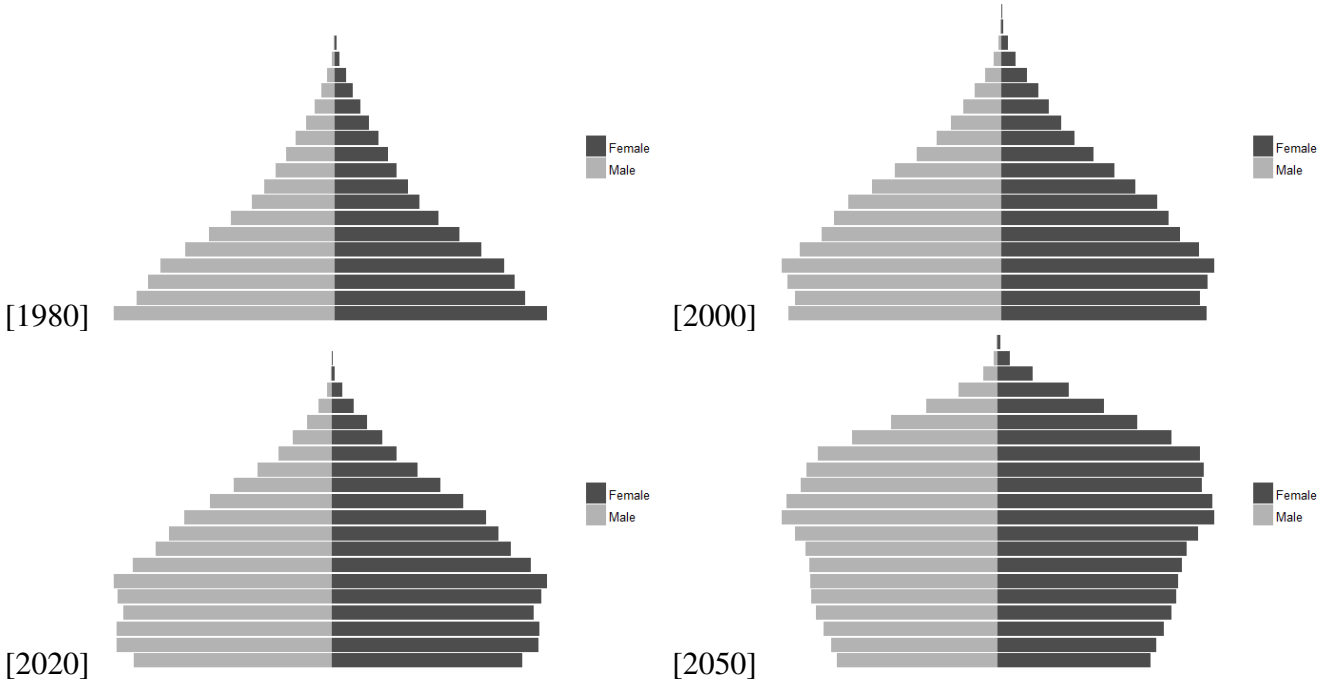


Figura 1 – Pirâmides etárias absolutas do Brasil, 1980-2050

Fonte: elaboração própria a partir de dados do *United States Census Bureau*, fonte 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. 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).

O gráfico 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.

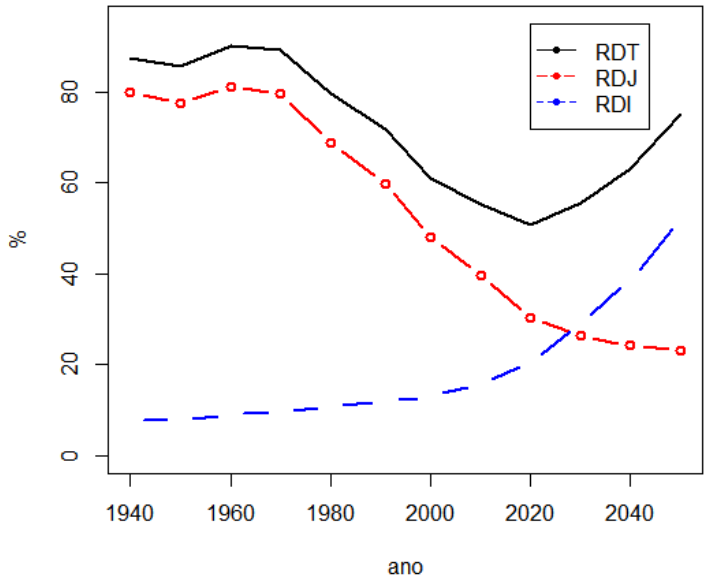


Figura 2 – Razão de dependência do Brasil, 1940 a 2050

Fonte: elaboração própria 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, o gráfico 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).

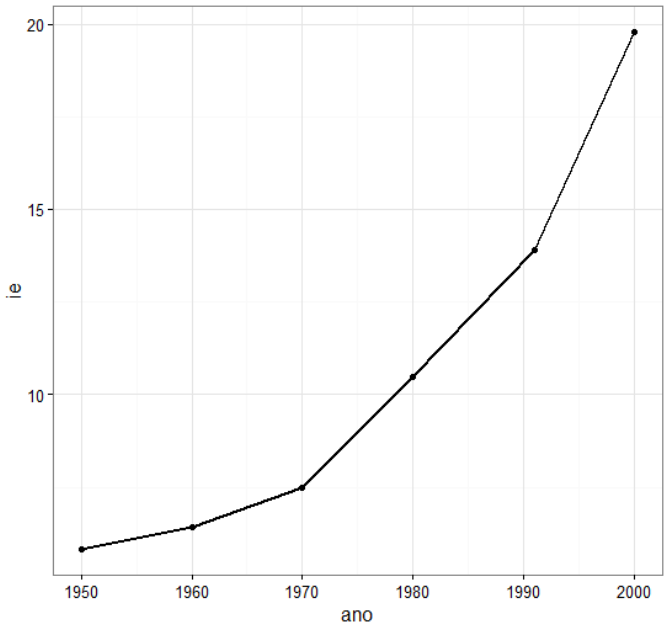


Figura 3 – Índice de envelhecimento do Brasil, 1950 a 2050

Fonte: elaboração própria a partir 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 populacinal. 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.

O Gráfico 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.

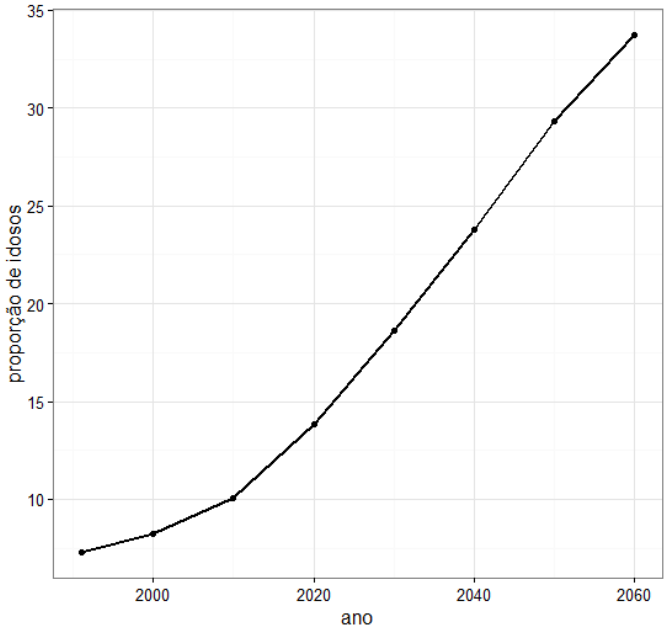


Figura 4 – Evolução da proporção de idosos com 60 anos ou mais na população brasileira - 1991 a 2060.

Fonte: elaboração própria a partir 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 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 Análise multivariada

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 equação (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:

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

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, $S_{ij} = S_{ji}$.

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

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

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 2 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}, \tag{2.2}$$

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 uni-

variada. 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).

A presente proposta empregará técnicas exploratórias. A análise de componentes principais será usada como forma de reduzir a dimensionalidade dos dados, simplificando a sua estrutura de covariâncias antes de aplicar a análise de agrupamento (*clusters*) que ajudará a identificar os grupos de municípios com perfis similares quanto à presença da previdência social. As duas técnicas são apresentadas a seguir.

2.4 Análise de componentes principais

Muitas variáveis distintas frequentemente são consideradas para realização de uma análise, o que dificulta não só a visualização da associação entre elas, mas também resulta em elevados níveis de correlação e multicolinearidade. Nesse contexto, a análise de componentes principais (ACP) é uma técnica multivariada utilizada para transformar os dados multivariados de forma que poucas dimensões expliquem a maior parte das informações contidas no conjunto de dados original (LATTIN; CARROLL; GREEN, 2011).

O objetivo do método é explicar a estrutura de covariâncias das p variáveis, $\boldsymbol{X}^T = \begin{bmatrix} X_1 & X_2 & \dots & X_p \end{bmatrix}$, por meio da construção de combinações lineares das variáveis originais. Os componentes principais são as p combinações lineares obtidas, $\boldsymbol{Y}^T = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_p \end{bmatrix}$, e são não correlacionados entre si. Entretanto, como a intenção é reduzir o número de variáveis, a informação contida nelas é substituída pela informação contida em k componentes principais, em que $k < p$. Os k componentes são ordenados de forma que os primeiros deles já conta-

bilizem a maior parte da variação presente em todas as variáveis originais (MINGOTI, 2005; EVERITT; HOTHORN, 2011).

A análise de componentes principais é uma técnica principalmente exploratória. Há métodos inferenciais para se testar hipóteses sobre componentes principais populacionais a partir de uma amostra aleatória de observações, mas eles são menos frequentes na literatura especializada (EVERITT; HOTHORN, 2011).

É preciso adotar um critério para reter apenas parte dos componentes, de maneira que grande parte da variância total seja explicada pelo conjunto pequeno de novas variáveis. Se o valor de k for pequeno e a quantidade de variação explicada pelos k componentes for grande, haverá uma simplificação da estrutura de covariâncias das variáveis originais. Essa técnica pode, então, ser utilizada como uma etapa intermediária para auxiliar em outras técnicas, como em problemas de multicolinearidade em regressão linear, por exemplo (FERREIRA, 2011). Isso ocorre porque a técnica possibilita que cada componente não esteja correlacionado com todos os outros, retirando a multicolinearidde em uma análise de dependência (LATTIN; CARROLL; GREEN, 2011).

A suposição de normalidade das p variáveis não é imprescindível para a aplicação da técnica, mas, se ocorrer, os componentes principais obtidos são, além de não correlacionados, independentes e normais. Os componentes podem ser obtidos a partir da matriz de covariâncias ou a partir da matriz de correlações das variáveis originais. Essa é uma questão discutida por alguns autores. Em geral, recomenda-se obter os componentes a partir da matriz de covariâncias amostral quando as variáveis estão na mesma escala e a partir da matriz de correlações amostral nos outros casos, que é o que ocorre mais frequentemente em situações práticas (EVERITT; HOTHORN, 2011). Já outros autores, como Khatree e Naik (2000) questionam essa escolha e argumentam que é preciso levar outras questões em conta.

O primeiro componente principal Y_1 é a combinação linear

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p,$$

cuja variância amostral é a maior dentre todas as outras combinações lineares. É importante usar uma restrição nos valores desses coeficientes, geralmente $\boldsymbol{a}_1^T\boldsymbol{a}_1 = 1$, ou seja, a soma dos quadrados desses valores deve ser igual a 1. Isso deve ser feito porque a variância de Y_1 poderia crescer de forma ilimitada apenas aumentando os coeficientes $\boldsymbol{a}_1^T = [a_{11} \quad a_{12} \quad \ldots \quad a_{1p}]$. A variância amostral de Y_1 é dada por $\boldsymbol{a}_1^T\boldsymbol{S}\boldsymbol{a}_1$, sendo $\boldsymbol{S}$ a matriz de covariâncias amostral

das X variáveis e $\boldsymbol{a}_1$ é o autovetor da matriz $\boldsymbol{S}$ associado ao maior autovetor λ dessa matriz (EVERITT; HOTHORN, 2011). A obtenção de autovalores λ e autovetores $\boldsymbol{e}$ de uma matriz quadrada $p \times p$ é tal que $\boldsymbol{Ae} = \lambda \boldsymbol{e}$. Para maiores detalhes, consultar, por exemplo, (FERREIRA, 2011).

Ainda de acordo com Everitt e Hothorn (2011), o segundo componente principal, Y_2 é definido como a combinação linear

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p,$$

ou seja, $Y_2 = \boldsymbol{a}_2^T \boldsymbol{X}$, em que $\boldsymbol{a}_2^T = [a_{21} \quad a_{22} \quad \ldots \quad a_{2p}]$ e $\boldsymbol{X}^T = [X_1 \quad X_2 \quad \ldots \quad X_p]$, que possui a maior variância sujeito às condições

$$\boldsymbol{a}_2^T \boldsymbol{a}_2 = 1,$$
$$\boldsymbol{a}_2^T \boldsymbol{a}_1 = 0,$$

em que a segunda condição garante que Y_1 e Y_2 são não correlacionados. De forma similar, todos os outros componentes serão obtidos.

O vetor de coeficientes que define o i-ésimo componente principal, $\boldsymbol{a}_i$ é o autovetor de $\boldsymbol{S}$ associado com o seu i-ésimo maior autovalor. A variância do i-ésimo componente principal é dada por λ_i, sendo os $\lambda_1, \lambda_2, \ldots, \lambda_p$ os autovalores de $\boldsymbol{S}$ sujeitos à restrição $\boldsymbol{a}_i^T \boldsymbol{a}_i = 1$ (EVERITT; HOTHORN, 2011).

A proporção da variância total de $\boldsymbol{X}$ explicada pelo i-ésimo componente principal é definida por

$$\frac{Var(Y_i)}{\text{Variância total de } X} = \frac{\lambda_i}{\text{traço}(\boldsymbol{S})} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}.$$

Além disso, as variâncias total e generalizada de $\boldsymbol{X}$ podem ser descritas pelas variâncias total e generalizada de $\boldsymbol{Y}$:

$$\text{traço}(\boldsymbol{S}) = \sum_{j=1}^{p} \lambda_j = S_1^2 + S_2^2 + \cdots + S_p^2 \quad \text{e} \quad |\boldsymbol{S}| = \prod_{j=1}^{p} \lambda_j.$$

Dessa forma, os vetores $\boldsymbol{X}$ e $\boldsymbol{Y}$ são equivalentes em relação a essas duas medidas de

variação. Além disso, o primeiro componente principal sempre tem a maior proporção de explicação da variância total de $\boldsymbol{X}$ (MINGOTI, 2005).

De acordo com Everitt Hothorn (2011), os primeiros k componentes, em que $k < p$, explicam uma proporção da variância total,

$$\frac{\sum_{i=1}^{k} Var(Y_i)}{\text{Variância total de } X} = \frac{\sum_{i=1}^{k} \lambda_i}{\text{traço}(\boldsymbol{S})} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j}. \tag{2.3}$$

Os componentes principais podem ser obtidos a partir da matriz de covariâncias amostrais $\boldsymbol{S}$ ou a partir da matriz de correlações amostrais $\boldsymbol{R}$. Extrair os componentes da matriz de covariâncias deve ser preferido quando as variáveis originais estão na mesma escala, o que é raro ocorrer. Extrair os componentes como os autovetores de $\boldsymbol{R}$ é equivalente a calcular os componentes das variáveis originais para depois padronizar cada um para ter variância igual a 1 (EVERITT; HOTHORN, 2011).

Um passo importante da aplicação da técnica de ACP é a escolha de quantos componentes serão retidos. Um critério muito utilizado é avaliar a representatividade dos k primeiros componentes, de acordo com a equação (2.3). Define-se qual o valor de porcentagem da variação é pretendido (mínimo de 70%, por exemplo) e escolhem-se quantos componentes forem necessários para atingir essa representatividade. Porém, é necessário ter cautela com a escolha do número k, pois a utilidade prática dos componentes principais diminui com o aumento desse valor (MINGOTI, 2005).

Um método gráfico que pode auxiliar na escolha do valor de k é o *scree plot*, em que é representado o valor k no eixo x e a porcentagem da variação explicada no eixo y. Assim, busca-se o ponto em que não há grande variação no eixo y, indicando que a inclusão de mais componentes não auxiliará muito na interpretação (EVERITT; HOTHORN, 2011).

A Regra de Kaiser também pode ser utilizada com o objetivo de responder a questão de quantos componentes devem ser retidos na aplicação da técnica de ACP. O método consiste em reter os componentes principais com autovalores maiores que 1. A regra foi proposta com base na ideia de que qualquer componente principal deveria explicar pelo menos tantas variações quanto qualquer uma das variáveis originais $\boldsymbol{X}$ (LATTIN; CARROLL; GREEN, 2011). Mais detalhes sobre critérios podem ser vistos também em (KHATREE; NAIK, 2000).

Os valores numéricos dos componentes, denominados escores, podem ser calculados para cada elemento amostral e, em seguida, esses valores podem ser analisados utilizando outras técnicas como análise de variância e análise de regressão (MINGOTI, 2005). Os escores dos primeiros dois componentes principais podem ser plotados em um diagrama de dispersão para identificar agrupamentos ou outros tipos de padrão existentes nos dados.

Para calcular os escores dos componentes de cada observação i, se os componentes foram obtidos a partir da matriz de covariâncias amostrais $\boldsymbol{S}$, deve-se obter

$$Y_{i1} = \boldsymbol{a}_1^T \boldsymbol{X}_i, \qquad Y_{i2} = \boldsymbol{a}_2^T \boldsymbol{X}_i, \qquad \ldots, \qquad Y_{ik} = \boldsymbol{a}_k^T \boldsymbol{X}_i,$$

em que k é o número de componentes retidos e $\boldsymbol{X}_i$ é o vetor de variáveis $p \times 1$ para a observação i.

2.5 Análise de agrupamento

A técnica de 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.

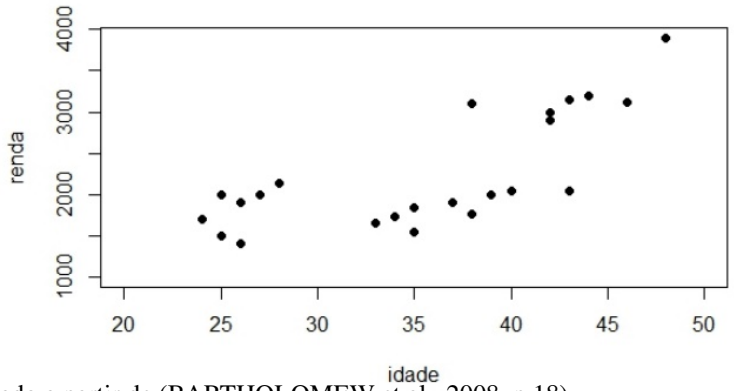
De acordo com Gordon (1999), a classificação de dados em grupos pode ser realizada com o objetivo de simplificá-los e realizar predições. A partir do método, é possível detectar o relacionamento e estrutura do conjunto de dados. Em muitas aplicações, os pesquisadores podem estar interessados na descrição 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 no gráfico 5 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 5 – Diagrama de dispersão de dados fictícios de idade e renda de várias pessoas.



Fonte: 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 *clusters* 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 cor-

respondem 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 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

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

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).

Para realizar o procedimento de agrupamento é necessário que a medida de similaridade ou dissimilaridade seja definida *a priori*. Na literatura há muitos tipos de distâncias que podem ser calculadas entre pares de observações, como a distância euclidiana, distância de Mahalanobis, distância euclidiana média e métrica *p de Minkowski*. Essas medidas são de dissimilaridade, isso significa que 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).

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

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

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 aplicado componentes principais para 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 média são uma generalização da distância euclidiana. Dessa forma, seja a distância generalizada entre dois elementos $X_{i.}$ e $X_{j.}$ definida por:

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

A seleção dessa matriz define a distância utilizada. Se $\boldsymbol{A} = \boldsymbol{S^{-1}}$, isto é, a matriz de covariância da população 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). Quando $\boldsymbol{A} = \boldsymbol{I}$ temos a distância euclidiana. E, por último, quando $\boldsymbol{A} = \boldsymbol{D^{-1}}$ temos a distância padronizada.

Segundo Ferreira (2011), os métodos de agrupamento 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. No método hierárquico divisivo, 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.

2.5.1 Técnicas hierárquicas aglomerativas

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). Esses métodos estão divididos em: ligação simples (vizinho mais próximo), ligação completa (vizinho mais distante), média das distâncias, 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:

Passo 1. Por se tratar de métodos hierárquicos aglomerativos, inicialmente os n objetos são alocados em n grupos de tamanho 1, representados por $G_1, G_2,...,G_n$. Então a distância entre dois grupos, para construção da matriz de proximidades, é obtida através da distância entre dois objetos:

$$d_{G_iG_j} = d_{ij}$$

Passo 2. Selecionar na matriz de distâncias os dois grupos G_i e G_j que possuem menor distância;

Passo 3. Unir os dois grupos G_i e G_j em um novo denominado de G_{n+s}, sendo $s = 1$, um índice do processo iterativo;

Passo 4. Definir a distância entre o novo agrupamento G_{n+s} e todos os agrupamentos G_k, de acordo com a técnica hierárquica aglomerativa escolhida;

Passo 5. Reiniciar o processo a partir do passo 2 até que se chegue a um agrupamento final.

Portanto, esse é o processo iterativo geral, agora serão apresentados e definidos os métodos hierárquicos aglomerativos. A diferença de um método para outro se encontra na definição das distâncias entre grupos. Como trata-se de técnicas hierárquicas aglomerativas, para cada método, considere inicialmente n grupos de tamanho um.

Ligação simples

Também denominado de ligação por vizinho mais próximo, nesse método a similaridade entre dois agrupamentos é definida como o mínimo entre as distâncias de dois objetos, de tal

forma que a distância entre o novo agrupamento G_{n+s} e todos os agrupamentos G_k é definida por:

$$d_{G_{n+s}G_k} = min\{d_{G_iG_k}, d_{G_jG_k}\}$$

Assim, inicialmente o algoritmo coloca no primeiro grupo os dois objetos mais parecidos entre si, isto é, aqueles de menor distância. Em seguida, a distância é avaliada novamente e um novo grupo é formado por dois objetos, ou um novo objeto é adicionado ao primeiro grupo formado e, assim sucessivamente. O processo é finalizado quando há um único grupo de tamanho n.

Uma vantagem da aplicação dessa técnica é o pouco esforço computacional exigido pelo algoritmo. Contudo, a definição de similaridade une um objeto a um agrupamento de acordo com a menor distância avaliada par a par, isso pode resultar em dois objetos próximos em um agrupamento e o mesmo objeto relativamente longe de todos os outros inseridos no mesmo grupo. Dessa forma, a ligação por vizinho mais próximo tende a formar agrupamentos longos e encadeados, com formatos não convexos (LATTIN; CARROLL; GREEN, 2011). Como alternativa foram propostas outras técnicas, que produzem soluções distintas e tentam eliminar essa tendência do método de ligação simples.

Ligação completa

A distância entre dois agrupamentos é definida como o máximo entre as distâncias calculadas para os pares de grupos, de tal forma que:

$$d_{G_{n+s}G_k} = max\{d_{G_iG_k}, d_{G_jG_k}\}$$

Dessa forma, por esse método, dois objetos são considerados similares de acordo com o menor valor de máximo (MINGOTI, 2005). Em virtude dessa definição de similaridade, a técnica é também conhecida como método do vizinho mais distante. Essa escolha garante que o novo objeto alocado a um grupo esteja próximo não somente de um elemento, mas de todos os outros. Dessa forma, é razoável dizer que os grupos resultantes da ligação completa geralmente são convexos e tendem a ser de aproximadamente mesmo diâmetro. Contudo, o método pode ser sensível à *outliers* (LATTIN; CARROLL; GREEN, 2011; MINGOTI, 2005).

Distância média

Nesse método, a distância entre dois agrupamentos é definida como a média das distâncias entre todos os pares de objetos (MINGOTI, 2005). Assim, a distância média entre o grupo G_k e o novo grupo G_{n+s} pode ser representada por:

$$d_{G_{n+s}G_k} = \frac{n_i d_{G_iG_k} + n_j d_{G_iG_k}}{n_i + n_j},$$

em que $n_i + n_j$ é o número de objetos no grupo G_{n+s}. O método de ligação média tende a formar grupos com melhores partições do que os de ligação simples e completa (LATTIN; CARROLL; GREEN, 2011). Além disso, os grupos resultantes possuem aproximadamente a mesma variância interna (EVERITT; HOTHORN, 2011).

Centroide

A distância entre dois agrupamentos é definida a partir da distância entre dois centroides, calculados com base na média entre os objetos de cada grupo (MINGOTI, 2005). De acordo com (LATTIN; CARROLL; GREEN, 2011), considerando um agrupamento G_k e um novo agrupamento formado pela união dos grupos G_i e G_j, denominado de G_{n+s}, então a distância ao quadrado entre eles é definida como:

$$d^2(G_k,G_{n+s}) = \frac{n_{G_i} d_{G_k,G_i^2} + n_{G_J} d_{G_k,G_J^2}}{n_{G_i} + n_{G_j}} - \frac{n_{G_i} n_{G_j} d_{G_i,G_j^2}}{(n_{G_i} + n_{G_j})^2}$$

Ward

O método de Ward, diferente dos anteriores, não segue o processo iterativo geral, pois não busca a menor distância entre dois grupos (passo 2), mas a menor soma de quadrados mínimos dentro do grupo, ou seja, a menor variância interna. Os grupos resultantes geralmente possuem o mesmo número de objetos, são convexos e compactos (LATTIN; CARROLL; GREEN, 2011). Segundo (MINGOTI, 2005), o processo iterativo dessa técnica segue os seguintes passos:

Passo 1. Inicialmente os n objetos são alocados em n grupos de tamanho 1, representados por $G_1, G_2,...,G_n$.

Passo 2. Em 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:

$$SQ_i = \sum_{j=1}^{n_i} (\boldsymbol{X_{ij}} - \boldsymbol{\bar{X}_i})'(\boldsymbol{X_{ij}} - \boldsymbol{\bar{X}_i})$$

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

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

$$SQT = \sum_{i=1}^{g_k} SQ_i,$$

em que, g_k é o número de grupos no passo k.

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

$$d(G_p,G_i) = \left[\frac{n_p ni}{n_p + n_i}\right] (\boldsymbol{\bar{X}_p} - \boldsymbol{\bar{X}_i})'(\boldsymbol{\bar{X}_p} - \boldsymbol{\bar{X}_i})$$

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 grupo que está sendo comparado. Dessa forma, o fator $[\frac{n_p ni}{n_p+n_i}]$ pondera a distância de dois grupos de tamanhos diferentes. Quanto maiores os valores de n_i e n_p, maior será o fator de ponderação e, portanto, maior a distância entre os vetores de médias comparados. Para aplicação do método de Ward é necessário que as p-variáveis sejam quantitativas para que seja possível o cálculo dos vetores de médias. Além disso, geralmente os agrupamentos obtidos possuem o mesmo número de observações (MINGOTI, 2005).

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 6 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.

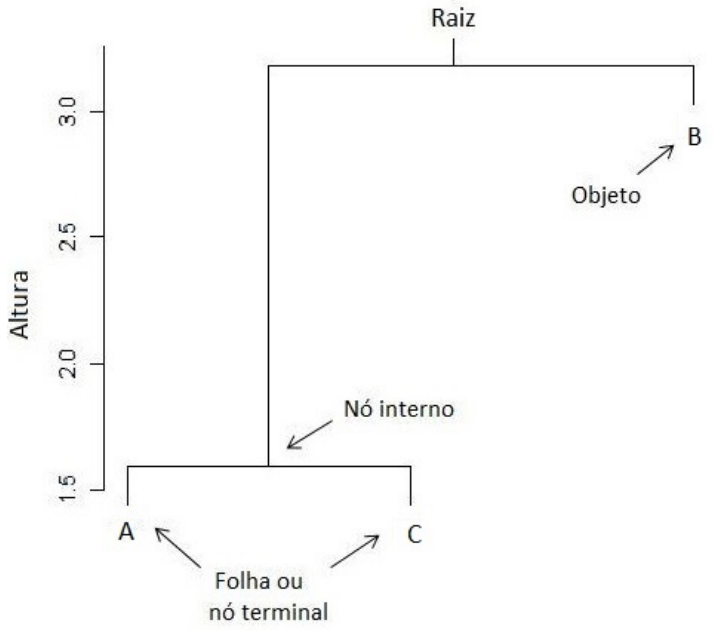


Figura 6 – Terminologia utilizada na descrição de dendrogramas

Fonte: elaboração própria a partir de (EVERITT; HOTHORN, 2011)

.

Portanto, no caso representado pela figura 6, 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. 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. Portanto, é necessário decidir uma regra de parada do processo, para obtenção de k grupos.

2.5.2 Técnicas não hierárquicas: k-médias

Outra técnica de agrupamento que pode ser utilizada é k-médias (k-means). No entanto, diferente daquelas já apresentadas, essa técnica é do tipo não hierárquica ou de partição (LATTIN; CARROLL; GREEN, 2011). As principais características desses métodos são: aplicação do processo à matriz de dados $\boldsymbol{X}$ e número de grupos k definido *a priori* (FERREIRA, 2011). A implementação dos métodos não hierárquicos é realizada a partir de algoritmos computacionais do tipo iterativo, por isso são vistos como mais adequados que os métodos hierárquicos na análise de um conjunto de dados com um grande número de observações. Como o próprio nome diz, esses métodos não seguem a propriedade da hierarquia, isso significa que mesmo que 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 dendrogamas para representação dos agrupamentos formados passo a passo (MINGOTI, 2005).

A técnica procura uma partição das n observações em *k* agrupamentos ($G_1, G_2, \cdots$, 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

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

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 muito grande 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 como:

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

Passo 2. Uma medida de distância é aplicada para comparar cada objeto a cada centroide inicial, então o objeto é unido ao grupo de menor distância;

Passo 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;

Passo 4. As etapas 2 e 3 são repetidas 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 (etapa 1) (FERREIRA, 2011). Diversas sugestões para decisão 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 no método de Ward, 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 é aconselhável que o processo seja realizado com diferentes escolhas para busca da melhor solução de agrupamento (FERREIRA, 2011; LATTIN; CARROLL; GREEN, 2011).

2.5.3 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 agrupamentos hierárquicos aglomerativos e a inicial nos agrupamentos não hierárquicos. Isso acontece porque 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, 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 diferentes. 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 7 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 8, 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.

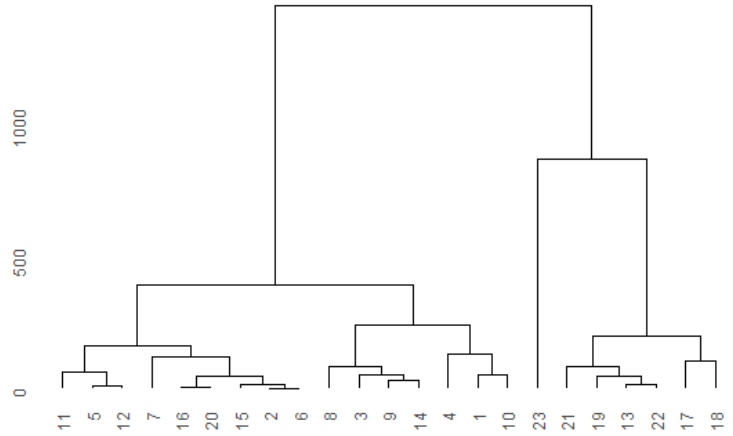


Figura 7 – Ilustração de corte no dendograma com 23 observações

Fonte: elaboração própria

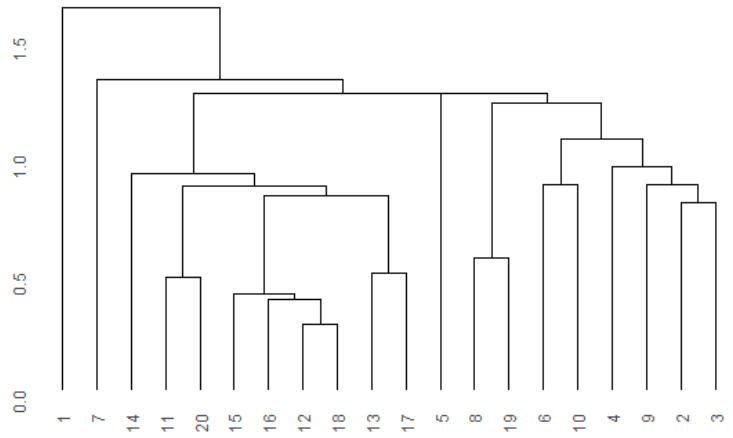


Figura 8 – Ilustração de dendograma com 20 observações

Fonte: elaboração própria

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. Porém, esses critérios 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).

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 foi analisado por 4 métodos de agrupamentos hierárquicos diferentes, sendo eles vizinho mais próximo, vizinho mais distante, distância média e método de *Ward*.

O trabalho citado identificou como as cinco técnicas como 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.

Ainda de acordo com os autores, 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.

global: o ex avaliar alguma medida ao longo de todo o conjunto de dados e optimizá-lo em função do número de grupos

local: este último considerar pares individuais de clusters e de teste se eles devem ser reunidas

Gordon (1999) chama atenção para dois tipos de métodos, denominados de local e glo-

bal. O primeiro considera os pares individuais de grupos e testa se eles realmente devem ser unidos. O último avalia todo o conjunto de dados. Portanto, não oferecem indicação se os dados ralmente deveriam ser agrupados. Uma das estatísticas que possui essa característica, por exemplo, é a de Calinski e Harabasz (1974).

Segundo Tibshirani, Walther e Hastie (2001), dentre os métodos heurísticos para solucionar o problema do número de grupos, inclui-se o *scree plot*. A proposta do método é que se plote 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. Nesse sentido, segundo os autores, 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 fazem simulação de 5 cenários para avaliação de 6 critérios de escolha do número de grupos. As estatística 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*. *(Tenho dúvida na diferença entre elas)

O trabalho de Fang e Wang (2012) proprõe um método *bootstrap* e o compara com validação cruzada (WANG, 2010), estatística *silhouette* (ROUSSEEUW; KAUFMAN, 1990), *gap* (TIBSHIRANI; WALTHER; HASTIE, 2001) e uma abordagem denominada *jump*, proposta por Sugar e James (2011). Os resultados mostram que o método proposto por eles obteve bom desempenho na maior parte das situações. Outro resultado interessante foi que a estatística *gap* foi 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.

Fujita et al. (2014) propõem um método não paramétrico baseado na estatística *silhouette* de Rousseeuw e Kaufman (1990). Os autores o comparam com 7 critérios, sendo eles *Bayesian Information Criterion* (BIC) (CELEUX; GOVAERT, 1992), Calinski e Harabasz (1974), Krzanowski e Lai (1988), *silhouette* de Rousseeuw e Kaufman (1990), estatística *gap* (TIBSHIRANI; WALTHER; HASTIE, 2001), *prediction strength* (TIBSHIRANI; WALTHER, 2005) e o método *jump* (SUGAR; JAMES 2003). Foram considerados 7 cenários de simulação bem variados. Os autores concluíram que o método proposto é mais adequado que os outros quando os dados são correlacionados, não normais e há muitas variáveis ou com grupo dominante.

Diante dessa diversidade de abordagens, é crucial utilizar não apenas um método para definir o número de grupos, mas avaliar os resultados obtidos com diferentes critérios (EVERITT; HOTHORN, 2011). Além disso, as especificidades do problema analisado devem ser levadas em conta para que se decida qual critério fornece grupos cuja interpretação seja mais útil (CARVALHO; MATA; RESENDE, 2008). As estatísticas a serem usadas nesse trabalho são apresentadas na seção 3.2.

3 Dados e metodologia

O objetivo principal deste capítulo é descrever as variáveis do estudo, a fonte de dados utilizada e os procedimentos aplicados para a análise dos dados.

3.1 Base de dados e variáveis do estudo

As bases de dados utilizadas neste trabalho são proveniente 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 9 variáveis demográficas dos 146 municípios da mesorregião Sul/Sudoeste de Minas Gerais, apresentados no Anexo A, que se relacionam ao processo de envelhecimento populacional. Essa mesorregião foi escolhida por ser os três campi da Universidade Federal de Alfenas (Unifal-MG) se encontram. Na Tabela 1 são apresentadas essas variáveis, suas definições de acordo com o Atlas do Desenvolvimento Humano no Brasil e as siglas que serão utilizadas.

Tabela 1 – Descrição das variáveis demográficas

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	mortalidade infantil	número de crianças que morrem antes de completar um ano de vida em cada 1000 crianças nascidas vivas.
mort5	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: elaboração própria a partir de definições do Atlas Brasil.

3.2 Metodologia

Nesse trabalho foi utilizada a técnica multivariada de análise de agrupamento para identificar os grupos de municípios com características semelhantes. As análises foram feitas usando a linguagem R (TEAM et al., 2013) por meio do programa *RStudio* (STUDIO, 2012). Primeiro foram utilizados os dados originais e, posteriormente, os escores dos componentes principais, para fins de comparação.

Segundo Hair et al. (2009), em análise multivariada há suposições estatísticas que devem ser avaliadas. Algumas técnicas são mais afetadas do que outras na violação de determinadas suposições. Mas basicamente, as principais suposições estatísticas exigidas em diversas técnicas multivariadas são normalidade, linearidade e homocedasticidade. Contudo, na análise de agrupamento elas não possuem tanta importância. Nesse caso, especial atenção deve ser dada à representadidade da amostra e multicolinearidade entre as variáveis estudadas.

Ainda de acordo com os autores, a amostra deve ser representativa para que os resultados encontrados possam ser generalizados para a população. A multicolinearidade, por sua vez, tem um impacto negativo porque ela atua como um mecanismo de ponderação implícita. Portanto, devem ser avaliadas as correlações entre as variáveis.

Nesse sentido, inicialmente foi realizada uma análise exploratória dos dados que auxiliou nas aplicação das técnicas multivariadas posteriores. Primeiro, foi obtida a matriz de correlações entre as variáveis, duas a duas, de forma a identificar os pares de variáveis mais associadas entre si. Tal matriz tem a forma apresentada na equação (2.2). Em seguida, foi verificada a suposição de normalidade dos dados. Segundo FERREIRA (2011), na literatura há diversos testes para verificar a normalidade multivariada dos dados. Dentre eles, foram escolhidos os procedimentos de Mardia e o teste Shapiro-Wilk para o caso multivariado.

Após essa etapa de formulação do problema, escolha das variáveis e análise exploratória dos dados, foi iniciado o processo de agrupamento. Os próximos passos consistiram na seleção de uma ou mais medidas de similaridade ou dissimilaridade, tratamento dos dados, escolha dos métodos de agrupamento, definição do número de grupos e, por último, avaliação e interpretação dos resultados.

A medida de similaridade ou dissimilaridade tem como objetivo medir o quão próximos estão duas observações ou grupos. Em um primeiro momento, foi utilizada a distância de Mahalanobis, que representa uma medida padronizada da distância euclidiana, indicada quando

as variáveis estão altamente correlacionadas. Segundo Hair et al. (2009), essa medida ajusta as correlações entre as variáveis e pondera igualmente cada variável do conjunto de dados original. Ainda de acordo com o autor, a distância de Mahalanobis, além de padronizar os dados em termos dos desvios padrão, soma a variância-covariância dentro dos grupos, executando o processo de ajuste de correlações entre variáveis.

Em seguida, os agrupamentos foram realizados utilizando também a distância euclidiana. Como há diferentes distâncias disponíveis na literatura e cada uma produz um resultado de agrupamento diferente, foi decidido aplicar as duas distâncias mais utilizadas, para fins de comparação. No agrupamento hierárquico com as variáveis originais é recomendado que, caso elas estejam em escalas diferentes, sejam padronizadas antes de calcular a distância euclidiana. Isso acontece porque variáveis em escalas diferentes apresentam maior dispersão e produzem um impacto maior sobre o valor da medida de distância. Portanto, nesse caso é aconselhável que se padronize os dados antes que essas medidas sejam calculadas. Como a medida de Mahalanobis já faz o processo de padronização, em um primeiro momento não foi necessário executá-la. Portanto, primeiramente foi verificado se a matriz de dados com m colunas (variáveis) e n linhas (observações) precisava ser tratada. Como as variáveis demográficas estão expressas em diferentes unidades/medidas foi necessária a padronização. Foi utilizada a função genérica *scale* da linguagem R, que centraliza e padroniza pela subtração da média e divisão pelo desvio padrão para cada variável. Dessa maneira, cada variável é convertida em escores padrão com uma média 0 e um desvio padrão 1. Os dados padronizados também foram utilizados na aplicação do método não hierárquico das k-médias.

O próximo passo foi a escolha dos métodos de agrupamento. Foram adotados os cincos métodos hierárquicos apresentados na seção 2.5, sendo eles: ligação simples, ligação completa, distância média, centroide e Ward. Em seguida, também foi utilizado o método não hierárquico das k-médias. Em qualquer uma das abordagens, seja hierárquica aglomerativa ou não hierárquica, é preciso definir o número ótimo de grupos. No primeiro caso, a regra de corte encerra o processo de agrupamento com a solução considerada representativa da estrutura de dados analisada. No segundo caso, o critério requer que o número de grupos seja definido antes de sua aplicação, para o cálculo dos centroides iniciais. Assim, os métodos para definição do número de grupos mostrados na seção 2.5.3 foram utilizados e foi verificado se, e em quais condições, houve convergência entre eles no caso dos dados analisados. O pacote *NbClust* foi usado para determinar o número ótimo de grupos. De forma geral, ele determina o número adequado de

grupos por meio de até 30 métodos diferentes, sendo possível escolher o número de grupos, os métodos de agrupamento e também diferentes medidas de distâncias.

Como não existe um método considerado ideal, entre os diversos disponíveis, são aplicados diferentes critérios e analisado para qual número de grupos a maioria converge. Foi utilizado como base o trabalho de Milligan e Cooper (1985) e Tibshirani e Walther (2005). Basicamente foram selecionados os 9 primeiros métodos considerados os melhores do artigo no trabalho de Milligan e Cooper (1985), sendo eles: pseudo F (CALINSKI; HARABASZ, 1974), Je(2)/Je(1) (DUDA; HART et al., 1973), C-*Index* (HUBERT; LEVIN, 1976), *gamma* (BAKER; HUBERT, 1975), Beale (BEALE, 1969), estatística CCC *(Cubic Clustering Criterion)*, ponto bisserial, *Gplus* (ROHLF, 1974), Davies and Bouldin (DAVIES; BOULDIN, 1979) e, por último, a estatística *gap* (TIBSHIRANI; WALTHER, 2005). Na aplicação do método das k-médias, também foi usada uma combinação de um método hierárquico aglomerativo e não hierárquico. Em outras palavras, a técnica hierárquica aglomerativa de Ward foi aplicada e o número ótimo de grupos obtido foi utilizado para calcular as sementes.

Milligan e Cooper (1985) e Charrad et al. (2014) apresentam uma síntese desses e outros critérios. A seguir são apresentados os métodos selecionados para a escolha do número de grupos que serão usados nesse trabalho.

1. Pseudo F

A estatística proposta por Calinski e Harabasz (1974) é definida por:

$$\text{Pseudo}F = \frac{SQE/(g-1)}{SQD/(n-g)}, \tag{3.1}$$

em que SQE é a soma de quadrados total entre os g grupos da partição, definida como $SQE = \sum_{i=1}^{g} n_i(\bar{X}_{i.} - \bar{X})^T(\bar{X}_{i.} - \bar{X})$. SQD é a soma de quadrados total dentro dos grupos da partição, dada por $SQD = \sum_{i=1}^{g}\sum_{j=1} n_i(X_{ij} - \bar{X}_{i.})^T(X_{ij} - \bar{X}_{i.})$ e n é o número de observações.

A proposta do critério se baseia na ideia de que em cada etapa do algoritmo um teste F de análise de variância esteja sendo realizado, para fins de comparação dos vetores de médias dos grupos que estão sendo obtidos na etapa (MINGOTI, 2005). A Pseudo F não aumenta monotonicamente, e sim alcança um valor de máximo para determinado número de grupos. Dessa forma, acima de um valor específico de k pode ocorrer decréscimo no valor da estatística, o que indica que aumentar o número de grupos a partir de determinado k não contribui para reduzir a heterogeneidade interna do grupo (LATTIN; CARROLL; GREEN, 2011). Logo, até um deter-

minado valor, quanto maior o valor da pseudo F, melhor é a partição no sentido de aumentar a homogeneidade dentro do grupo e heterogeneidade entre grupos. Isso acontece porque o valor máximo da estatística pseudo F está relacionado com o menor valor-p do teste e, portanto, a hipótese nula de igualdade dos vetores de médias populacionais com maior significância é rejeitada (MINGOTI, 2005; LATTIN; CARROLL; GREEN, 2011). De forma geral, quanto maior o valor de F, mais significativo é o teste. Como consequência, o processo resulta em uma partição final com maior heterogeneidade entre elementos de grupos diferentes. É como se em cada etapa do algoritmo de agrupamento fosse aplicado um teste para comparação dos vetores de médias dos dois grupos que se uniram.

2. Je(2)/Je(1)

Proposta por Duda, Hart et al. (1973), Je(2) representa a soma dos quadrados dos erros dentro do grupo, quando os dados são divididos em dois grupos, e Je(1) dá os erros ao quadrado quando apenas um grupo é formado. O número ótimo de grupo é tal que

$$\text{Je(2)/Je(1)} \geq 1 - \frac{2}{\pi p} - z\sqrt{\frac{2(1-\frac{8}{\pi^2 p})}{n_m p}} = \text{valor\_crítico}, \tag{3.2}$$

em que z é um escore da normal padrão, p o número de variáveis e n_m o tamanho da amostra dentro do grupo m.

3. C*Index*

O critério foi revisado por Hubert e Levin (1976) e é calculado da seguinte maneira:

$$C\textit{Index} = [d_w - min(d_w)]/[max(d_w) - min(d_w)], \tag{3.3}$$

em que d_w é a soma das distâncias dentro do grupo. O menor valor do índice é usado para indicar o número ideal de grupos.

4. *Gamma*

Esse critério é uma adaptação da estatística *gamma* de *Goodman and Kriskal* para uso em análise de agrupamento (BAKER; HUBERT, 1975). As comparações são feitas considerando a dissimilaridade dentro dos grupos e entre os grupos. A estatística é dada por:

$$gamma = \frac{s(+) - s(-)}{s(+) + s(-)}, \tag{3.4}$$

em que $s(+)$ representa o número de comparações concordantes, isto é, quando a dissimilaridade dentro do grupo é estritamente menor que a dissimilaridade entre grupos. Ao passso que,

$s(-)$ é o número de comparações discordantes, aquelas que a dissimilaridade dentro do grupo é estritamente maior do que entre grupos. O número ideal de grupos está associado ao valor máximo do índice *gamma*.

5. Beale

Proposta por Beale (1969), o critério propõe o uso de uma razão F para testar a hipótese de um g_1 *versus* g_2 grupos nos dados, tal que $g_2 > g_1$. A estatística é calculada a partir da equação a seguir

$$Beale = F = \frac{\left(\frac{v_{kl}}{W_k + W_l}\right)}{\left(\frac{n_m - 1}{n_m - 2}\right)} 2\left(\frac{2}{p} - 1\right), \tag{3.5}$$

em que $v_{kl} = W_m - Wk - W_l$, sendo W uma matriz de dispersão dentro do grupo m, k e l. A estatística assume que os grupos c_k e c_l se uniram para formar o grupo c_m. O número ótimo de grupos é obtido pela comparação de F com uma distribuição $F_p(n_m - 2)p$. A hipótese nula de um único grupo é rejeitada para altos valores de significância de F.

6. Estatística CCC (*Cubic Clustering Criterion*)

O coeficiente CCC foi proposto por Sarle (1983). A estatística é dada pelo produto de dois termos, conforme equação a seguir

$$CCC = ln[\frac{1 - E(R^2)}{1 - R^2}]\frac{\sqrt{\frac{np*}{2}}}{(0{,}001 + E(R^2))^(1{,}2)}, \tag{3.6}$$

em que R^2 é a proporção da variância explicada pelos grupos e seu valor esperado é determinado sob a suposição de que os dados foram amostrados a partir de uma distribuição uniforme p-dimensional. O valor máximo do índice é usado para indicar o número ideal de grupos da partição final. Ainda, segundo Johnson (1998), esse número estaria associado a valores da estatística maiores do que 3.

7. Ponto bisserial

Nesse critério, uma correlação ponto bisserial é calculada entre as entradas da matriz de distâncias original e uma matriz correspondente que consiste em entradas 0 ou 1, que indicam se duas observações estão ou não no mesmo grupo. O valor máximo é utilizado para sugerir o número ótimo de grupos do conjunto de dados. A estatística é definida conforme equação a

seguir:

$$\text{ponto bisserial} = \frac{[\bar{S_b} - \bar{S_w}][N_w N_b / N_t^2]^{1/2}}{S_d}, \tag{3.7}$$

em que $\bar{S_w} = S_w/N_w$, $\bar{S_b} = S_b/N_b$ e S_d é o desvio padrão de todas as distâncias. Além disso,

- N_w é o número total de par de observações dentro de um mesmo grupo, dado por $N_w = \sum_{k=1}^{q} \frac{n_k(n_k-1)}{2}$.
- N_b é o total de pares de observações pertencentes a grupos diferentes $N_b = N_t - N_w$.
- N_t é o total de pares de observações no conjunto de dados, dado por $N_t = \frac{n(n-1)}{2}$
- S_w é a soma das distâncias dentro do grupo $S_w = \sum_{k=1}^{q} \sum_{i,j \epsilon C_k} d(x_i,x_j), i < j.$
- S_b é a soma das distâncias entre grupos $S_b = \sum_{k=1}^{q-1} \sum_{l=k+1}^{1} \sum_{i \epsilon C_k j \epsilon C_l} d(x_i,x_j).$

8. *Gplus*

Esse critério foi proposto por Rohlf (1974) e é dado por

$$gplus = \frac{2s(-)}{N_t(N_t-1)}, \tag{3.8}$$

em que $s(-)$ é o número de comparações discordantes, ou seja, o número de vezes em que duas observações que estavam no mesmo grupo tinha uma distância maior do que duas observações não agrupadas no conjunto de dados. Os valores mínimos de índice são usados para determinar o número grupos na partição final.

9. Índice DB

O índice proposto por Davies e Bouldin (1979) é uma função da relação da soma de dispersão dentro de cada grupo e da separação entre grupos. A proposta é um quadro geral de medidas de separação de grupos. Esse índice geral é definido como a média dos índices calculados a partir de cada grupo individual. Assim, um índice de um grupo individual é tomado como o valor máximo da comparação emparelhada envolvendo esse grupo e outras partições da solução. Cada comparação par a par é calculada como a razão entre uma medida de dispersão dentro do grupo e a distância entre os centroides de dois grupos como medida de separação. O valor mínimo da estatística é usado para indicar o número de grupos apropriado.

10. Estatística *gap*

O critério proposto por Tibshirani, Walther e Hastie (2001) é dado por:

$$gap(q) = \frac{1}{B}\sum_{b} log(W*_{qb}) \breve{} log(W_q),$$

em que B é um conjunto de dados gerados de uma distribuição uniforme e W_{qb} é uma matriz de dispersão dos dados dentro de q grupos.

Assim a escolha do número de grupos via a Estatística *gap* é dada por:

$$\hat{k}\text{= menor valor de } k \text{ tal que } gap(k) \geq gap(k+1) - s_{k+1},$$

em que $s_q = sd_q\sqrt{(}1+1/B)$ e $Sd_q = [(1/B)\sum_b\{log(W*_{qb}) - \bar{l}\}^2]^{\frac{1}{2}}$ com $\bar{l} = (1/B)\sum_b log(W*_{qb})$.

Portanto, o número de grupos é escolhido de tal forma que o valor k seja o menor que siga a condição da desigualdade acima.

Após obter os agrupamentos com os dados originais, a mesma análise foi feita com os escores dos componentes principais. Primeiramente, a ACP foi utilizada para reduzir a dimensionalidade dos dados, conforme explicitado na seção 2.4. 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. Dessa forma, um pequeno número dessas novas variáveis explicam boa parte da variação presente nos dados originais.

O número de componentes retidos foi definido através da avaliação da representatividade dos k primeiros componentes conjuntamente com o método gráfico *scree plot*. Se apenas dois componentes, Y_1 e Y_2, explicarem boa parte da variação presente nos dados, é possível obter um gráfico bidimensional com os valores das observações, os escores, de Y_1 e Y_2. Além disso, é interessante que os componentes principais tenham uma interpretação prática. Para isso, após a definição do número de componentes utilizados, as correlações entre cada variável original e cada componente foram calculadas, os chamados *loadings*. Os valores dessas correlações, bem como seus sinais, indicam como cada componente pode ser interpretado.

Ao final, os resultados dos agrupamentos obtidos com os dados originais e os escores dos componentes principais foram comparados entre si, bem como os resultados obtidos dos diferentes métodos de agrupamentos. Além disso, como suporte para avaliação dos grupos obtidos foi calculada a mediana das variáveis de cada grupo e o coeficiente de variação (CV) dentro do grupo. Essas medidas podem ser usadas para identificar as variáveis que separam os

grupos. Nesse sentido, a mediana auxilia na identificação das variáveis que mais contribuem para a divisão das observações entre os grupos. O CV, por sua vez, 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.

4 Resultados parciais

4.1 Análise descritiva das variáveis

O primeiro passo consistiu em obter um resumo estatístico das variáveis originais, apresentado na Tabela 2. 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, por exemplo, esperava viver em média $78{,}15$ anos. Esses valores sinalizam um regime de baixa mortalidade e alta esperança de vida na mesorregião estudada.

Tabela 2 – Resumo estatístico das variáveis originais

	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: elaboração própria.

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 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, 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. Isso pode ser visto na Tabela 3, em que os municípios de Passos e Itajubá, que apresentam os menores níveis de mortalidade, são responsáveis pelas maiores esperanças de vida. 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 encontrada na mesorregião estudada é $0{,}96$, ao passo que a menor é $0{,}92$. Em relação a probabilidade de sobrevivência até 60 anos de idade, o maior valor registrado é $0{,}87$ e o menor é $0{,}80$.

Em relação à taxa de envelhecimento, 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 $10{,}32\%$. A taxa de participação dos idosos (65 anos ou mais de idade) em relação a população total do município. Uma taxa elevada reflete, principalmente, a redução dos níveis de fecundidade e o aumento da esperança de vida dos idosos.

A Tabela 3 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.

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 tam-

Tabela 3 – 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: elaboração própria.

bém estão associados às menores taxas de mortalidade infantil (10,35 e 10,50 óbitos por mil nascidos) e até aos 5 anos de idade (12,11 e 12,12 óbitos por mil nascidos vivos), seguidos pela cidade de Poços de Caldas (11,27 e 13,18 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 encontram-se em Itajubá (0,96 e 0,88), São Loureço (0,96 e 0,87) e Passos (0,95 e 0,86), respectivamente.

As três menores taxas de fecundidade total 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 a população total do município (t_env) foram Córrego do Bom Jesus (14,85%), Senador José Bento (13,65%) e Pratápolis (12,97%).

A Tabela 4, 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.

Como pode ser observado na Tabela 4, 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

Tabela 4 – 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	São Tomé das Letras	5,65
Divisa Nova	52,86	Bandeira do Sul Ibitiúra de Minas Toledo	92,40	79,71	Carmo da Cachoeira	7,44
Serrania	52,64	Natércia	92,49	79,92	São Bento Abade	5,90

Fonte: elaboração própria.

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 $0{,}80$). Os municípios Bandeira do Sul, Ibitiúra de Minas e Toledo foram responsáveis pela segunda menor esperança de vida ($73{,}14$ 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 ($0{,}80$). 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) e 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 ($0{,}80$).

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 São Tomé das Letras ($5{,}65\%$), Carmo da Cachoeira ($7{,}44\%$) 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 investigação da

normalidade multivariada, apenas com o intuito de descrever os dados. O primeiro resultado obtido foi o do teste de Mardia que rejeitou a hipótese de normalidade multivariada com um valor-p menor que $0{,}001$. O mesmo resultado foi obtido pelo teste de Shapiro Wilk multivariado, que também rejeitou a hipótese de normalidade dos dados com um valor-p menor que $0{,}001$. Em seguida, foi analisada a multicolinearidade entre as variáveis. A Figura 9 apresenta as correlações entre cada um dos pares de variáveis demográficas selecionadas. 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.

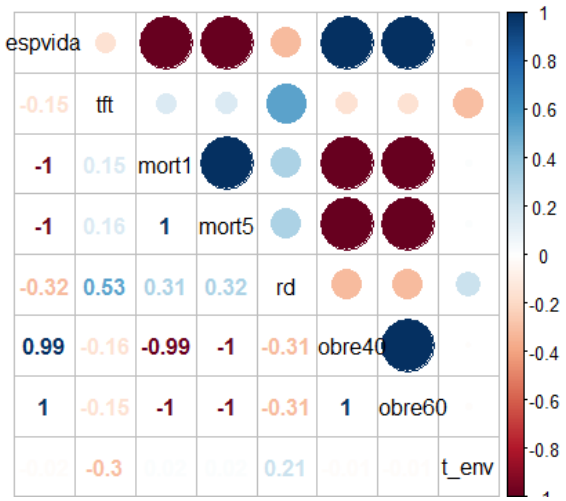


Figura 9 – Correlações entre as variáveis

Fonte: elaboração própria

A variável esperança de vida está altamente 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 perfeita positiva entre si, porque 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 correlação perfeita positiva entre si. Por fim, as variáveis tft e rd estão positivamente correlacionadas, isso pode estar associado ao comportamento 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.

Dada a correlação perfeita entre as variáveis sobre40 e sobre60, bem como entre mort1 e mort5, foi necessário fazer a escolha de quais variáveis permaneceriam na análise. Desse forma, a variável mort1 foi escolhida por ser mais representativa que mort5. Assim como a variável sobre60 é considerada mais informativa que sobre40.

4.2 Agrupamentos com as variáveis originais

Inicialmente, os municípios da mesorregião Sul/Sudoeste de Minas Gerais foram agrupados de acordo com as 6 variáveis demográficas utilizando a distância de Mahalanobis e os métodos hierárquicos aglomerativos: ligação simples, ligação completa, distância média, centroide e *Ward*. Foram obtidos os dendrogramas e os resultados da escolha do número final de grupos para cada caso, de acordo com os 10 critérios selecionados na seção 3.2.

A primeira técnica de análise de agrupamento utilizada foi a de ligação simples. As uniões entre observações ou grupos realizadas a cada passo do processo estão ilustradas na Figura 10, que mostra o dendrograma do agrupamento por esse método. Como mencionado na seção 2.5.3, uma maneira informal de seleção do número final de grupos é o corte no dendrograma. Portanto, avaliando o ponto de maior mudança, é possível observar que o corte produziria dois agrupamentos finais. O primeiro composto pelos municípios Itajubá e São Lourenço e o outro pelos demais 144 municípios. Esse resultado supõe que, em relação às variáveis consideradas na análise, os municípios Itajubá e São Lourenço são mais parecidos entre si e diferentes em relação a todos os outros da mesorregião estudada.

Para eliminar a subjetividade presente no corte do dendrograma foram considerados também os critérios de avaliação de número de grupos, já mencionados anteriormente. A Tabela 5 mostra os resultados desses métodos para a técnica ligação simples. Dentre os 10 métodos avaliados, seis sugerem dois grupos como o ideal. Portanto, o resultado da maioria dos critérios coincide com o sugerido pelo corte no dendrograma. Dentre os demais métodos, apenas o índice DB resultou em 10 grupos, enquanto a estatística pseudo F, *gamma* e ponto bisserial sugerem quatro grupos como o ideal. Com essa divisão, São Lourenço e Itajubá continuariam formando um agrupamento, no entanto, Varginha passaria a constituir um único grupo, assim como os municípios de Passos, Poços de Caldas e Pouso Alegre se uniriam, e as outras 140 cidades permaneceriam no outro grupo.

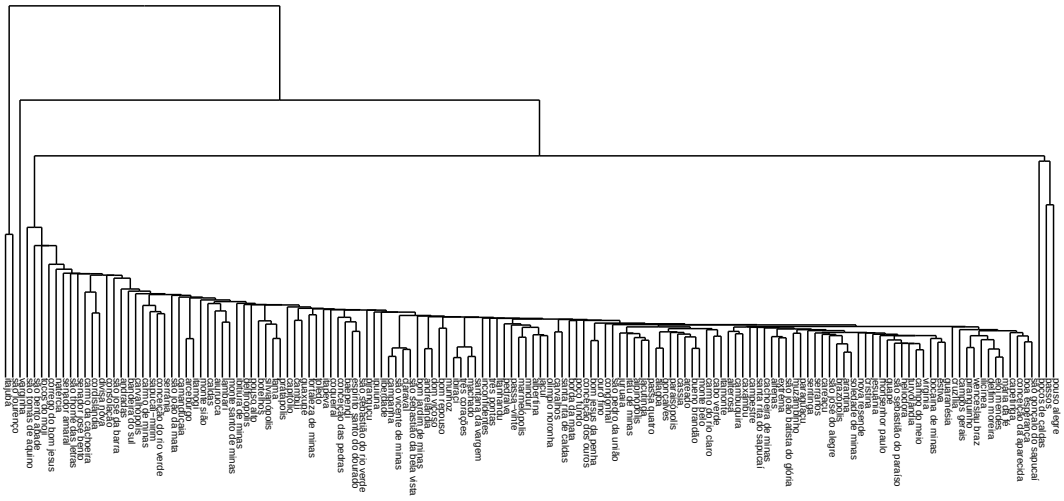


Figura 10 – Dendrograma pelo método ligação simples e distância de Mahalanobis

Fonte: elaboração própria

Tabela 5 – Número ideal de grupos para o método ligação simples e distância de Mahalanobis segundo a maioria dos critérios usados

Critério	Número de grupos	Critério	Número de grupos
pseudo F	4	CCC	2
Je(2)/Je(1)	2	ponto bisserial	4
índice C	2	*Gplus*	2
gamma	4	índice DB	10
Beale	2	*gap*	2
Número ideal de grupos		2	

Fonte: elaboração própria.

O próximo método usado para agrupar as observações foi o de ligação completa. A Figura 11 apresenta o dendrograma do agrupamento. Ao avaliar as mudanças na altura do gráfico nos diferentes passos é possível perceber que o ponto com maior mudança no nível de fusão sugere um corte que produz dois agrupamentos. O primeiro é formado por três municípios, sendo eles Pouso Alegre, Passos e Poços de Caldas e o segundo contém os 143 municípios restantes. Os resultados de seis critérios que avaliam o número ótimo de grupos, apresentados na Tabela 6, também sugerem 2 grupos como o ideal. Apenas a estatística pseudo F resultou em 12 grupos e os demais métodos em 4 grupos, que são constituídos pelos mesmos municípios do método anterior de ligação simples.

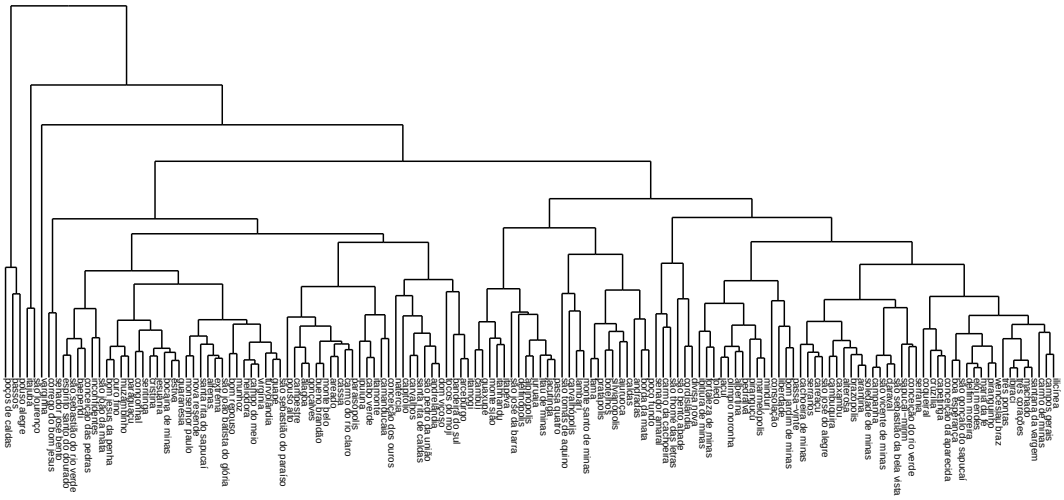


Figura 11 – Dendrograma pelo método ligação completa e distância de Mahalanobis

Fonte: elaboração própria

Tabela 6 – Número ideal de grupos para o método ligação completa e distância de Mahalanobis segundo a maioria dos critérios usados

Critério	Número de grupos	Critério	Número de grupos
pseudo F	12	CCC	2
Je(2)/Je(1)	2	ponto bisserial	4
índice C	2	*Gplus*	4
gamma	4	índice DB	2
Beale	2	*gap*	2
Número ideal de grupos		2	

Fonte: elaboração própria.

Em seguida, os municípios foram agrupados pelo método da distância média e a Figura 12 apresenta o dendrograma dessa análise. Diferente das técnicas anteriores, a maior modificação no nível de fusão sugere não dois grupos como o ideal, mas a divisão das observações em quatro agrupamentos. Um grupo formado por Itajubá e São Lourenço, um contendo apenas a cidade de Varginha, outro formado por Pouso Alegre, Passos e Poços de Caldas e, por último, um grupo com as demais 140 observações. Dentre os métodos de escolha do número de grupos, apenas *gamma* e ponto bisserial sugerem quatro grupos como o ideal. Mais uma vez, a maioria dos métodos indicam dois grupos como ideal. E, assim como no método de ligação simples, Itajubá e São Lourenço seriam mais parecidas entre si e diferentes de todos os outros municípios da mesorregião Sul/Sudoeste de Minas Gerais.

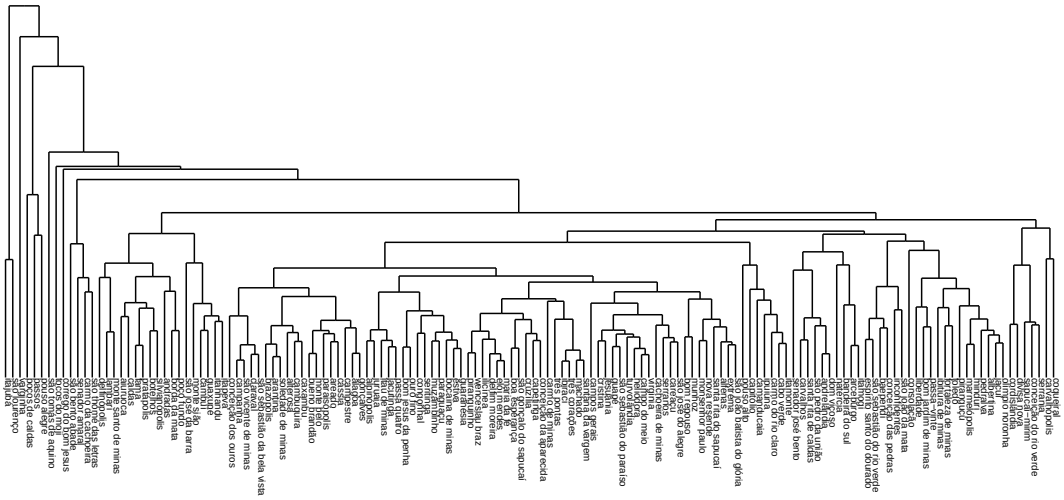


Figura 12 – Dendrograma pelo método distância média e distância de Mahalanobis

Fonte: elaboração própria

Tabela 7 – Número ideal de grupos para o método distância média e distância de Mahalanobis segundo a maioria dos critérios usados

Critério	Número de grupos	Critério	Número de grupos
pseudo F	14	CCC	2
Je(2)/Je(1)	2	ponto bisserial	4
índice C	2	*Gplus*	2
gamma	4	índice DB	7
Beale	2	*gap*	2
Número ideal de grupos		2	

Fonte: elaboração própria.

O próximo método hierárquico aglomerativo utilizado para agrupar as observações foi o do centroide e a Figura 13 apresenta seu o dendrograma. A altura do corte no dendrograma, considerando maior mudança no nível de união das observações ou grupos, gera três agrupamentos como uma possível solução final. Nesse caso, Varginha se diferenciaria em relação às variáveis medidas em tal medida que, novamente, estaria sozinha em um grupo. O outro grupo seria formado por Itajubá e São Lourenço e um grande grupo pelos demais municípios. Contudo, o resultado da análise dos critérios de número de grupos, apresentado na Tabela 8 converge para dois grupos finais, que seriam os mesmos obtidos com o método de ligação simples e distância média (um grupo com Itajubá e São Lourenço e o outro com os demais municípios). Os critérios pseudo F e índice DB sugerem 6 e 14 agrupamentos na partição final, respectivamente, e os métodos *gamma*, ponto bisserial e *Gplus* resultam em quatro como o número ideal.

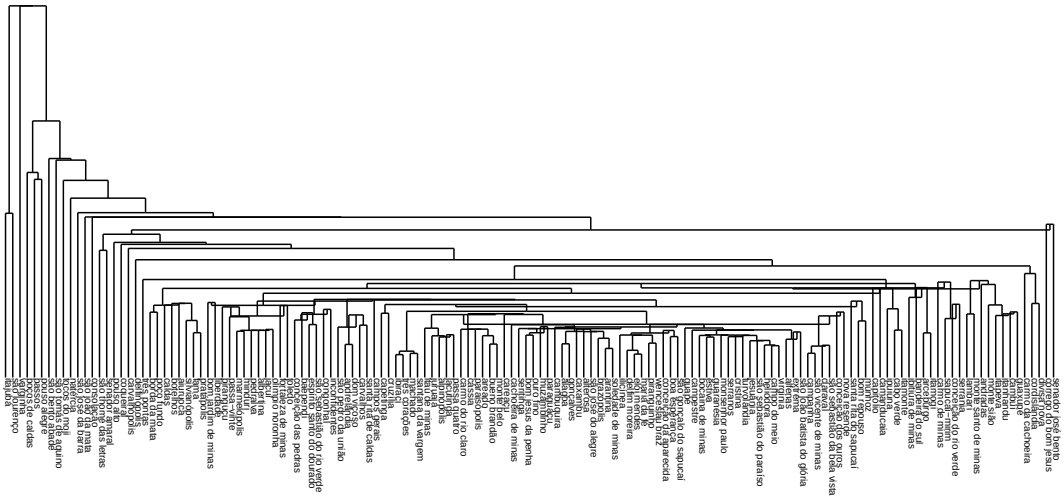


Figura 13 – Dendrograma pelo método centroide e distância de Mahalanobis

Fonte: elaboração própria

Tabela 8 – Número ideal de grupos para o método centroide e distância de Mahalanobis segundo a maioria dos critérios usados

Critério	Número de grupos	Critério	Número de grupos
pseudo F	6	CCC	2
Je(2)/Je(1)	2	ponto bisserial	4
índice C	2	*Gplus*	4
gamma	4	índice DB	14
Beale	2	*gap*	2
Número ideal de grupos		2	

Fonte: elaboração própria.

Por último, os dados foram agrupados pelo método de *Ward*. A Figura 14 apresenta o dendrograma do agrupamento. O corte no ponto que representa a maior diferença no nível de fusão, isto é, o que indica que o grupo pode se tornar menos homogêneo internamente com essa união (EVERITT et al., 2011), resulta na divisão dos municípios em cinco grupos. Nesse cenário, as observações estariam alocadas em grupos com 2, 3 26, 28 e 87 municípios. Essa divisão também sugere que Itajubá e São Lourenço formariam um único grupo, assim como Pouso Alegre, Passos e Poços de Caldas. No entanto, a maioria dos métodos de escolha de número de grupos converge para dois grupos como o ideal para a partição final, como pode ser observado na Tabela 9. Segundo esse resultado, 26 municípios estariam alocados em um grupo e 120 em outro grupo.

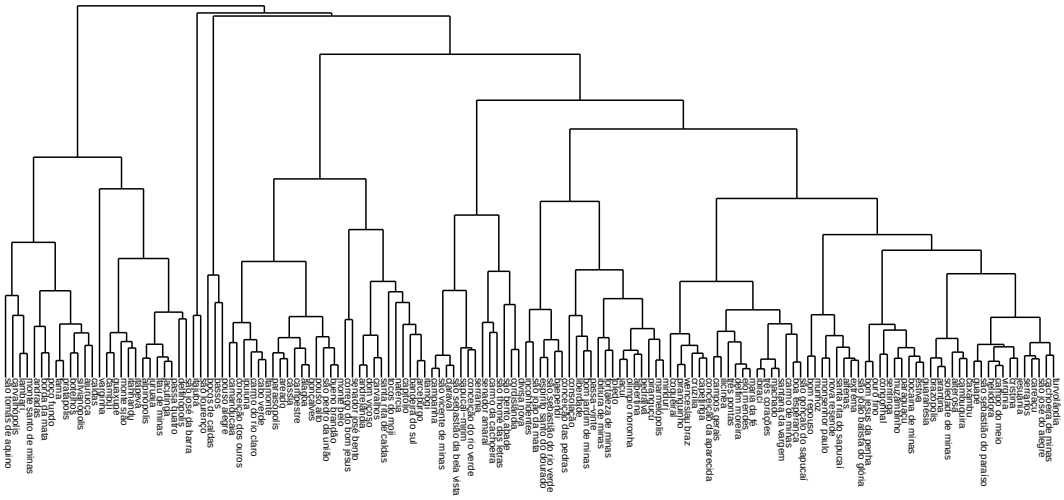


Figura 14 – Dendrograma pelo método *Ward* e distância de Mahalanobis

Fonte: elaboração própria

Tabela 9 – Número ideal de grupos para o método *Ward* e distância de Mahalanobis segundo a maioria dos critérios usados

Critério	Número de grupos	Critério	Número de grupos
pseudo F	2	CCC	2
Je(2)/Je(1)	2	ponto bisserial	4
índice C	2	*Gplus*	15
gamma	15	índice DB	2
Beale	2	*gap*	2
Número ideal de grupos		2	

A Tabela 10 mostra uma síntese do número de municípios em cada um dos dois grupos, de acordo com o método hierárquico usado. As técnicas ligação simples, distância média e centroide dividem as observações de forma que o primeiro grupo contém 144 municípios e o segundo grupo apenas dois, sendo eles Itajubá e São Lourenço. Portanto, os resultados desses métodos convergem para essa divisão das observações. O método ligação completa, por sua vez, divide os municípios em um grupo com 143 municípios e outro com três, sendo eles Passos, Poços de Caldas e Pouso Alegre. Por último, o método de *Ward* separa as observações em grupos com 26 e 120 municípios.

Tabela 10 – Número de observações nos grupos usando métodos hierárquicos aglomerativos e distância de Mahalanobis

Método	Grupo 1	Grupo 2
ligação simples	144	2
ligação completa	143	3
distância média	144	2
centroide	144	2
Ward	26	120

Fonte: elaboração própria

Como suporte para a avaliação desses resultados foi calculada a mediana das variáveis de cada grupo, bem como o coeficiente de variação (CV) dentro do grupo. A mediana, apresentada na Tabela 11, contribui para comparar os grupos em relação a cada variável, com o intuito de identificar quais variáveis poderiam estar contribuindo mais para a divisão entre os grupos. na contribui para avaliar a heterogeneidade entre grupos.

Como os agrupamentos pelos métodos ligação simples, distância média e centroide resultaram nos mesmos grupos de municípios, os resultados das medianas das variáveis e CVs são os mesmos. Os três métodos dividiram as observações em 2 grupos, um contendo Itajubá e São Lourenço e outro com os demais municípios. De forma geral, analisando a mediana, é razoável afirmar que a variável que mais contribuiu para a heterogeneidade entre os grupos foi a sobre60. Esse resultado pode estar associado ao fato que os municípios Itajubá e São Lourenço possuem as maiores probabilidades de sobrevivência até os 60 anos da mesorregião estudada, o que pode ter contribuído para esse resultado. Por outro lado, a que menos contribuiu foi a variável t_env.

Tabela 11 – Medianas das variáveis dos grupos obtidos por métodos hierárquicos aglomerativos e distância de Mahalanobis

Método	Grupo	espvida	tft	mort1	rd	sobre60	t_env
ligação simples	1	75,53	1,96	14,50	44,88	83,28	9,39
	2	77,67	1,67	11,00	42,92	87,06	9,50
ligação completa	1	75,53	1,96	14,50	44,92	83,28	9,47
	2	77,33	1,69	11,27	40,55	85,62	8,41
distância média	1	75,53	1,96	14,50	44,88	83,28	9,39
	2	77,67	1,67	11,00	42,92	87,06	9,50
centroide	1	75,53	1,96	14,50	44,88	83,28	9,39
	2	77,67	1,67	11,00	42,98	87,06	9,50
Ward	1	77,23	1,95	12,16	43,56	85,69	9,82
	2	75,24	1,96	14,95	44,99	82,86	9,38

Fonte: elaboração própria

Em relação ao método centroide, a variável com maior diferença no valor da mediana, ao se compararem os dois grupos, foi a rd, tendo sido a que mais contribuiu para a heterogeneidade entre os grupos. A variável que menos contribuiu foi a tft. Os resultados do método de *Ward* mostram que a variável sobre60 é a maior responsável pela heterogeneidade entre os grupos, enquanto a variável tft é a que menos colabora no sentido de aumentar a heterogeneidade entre os grupos.

Os coeficientes de variação de cada variável dos grupos são apresentados na Tabela 12. Os valores do CV auxiliam na avaliação da homogeneidade interna do grupo. Quanto menor seu valor, mais homogêneo é considerado o agrupamento em relação àquela variável. De forma geral, os resultados mostram que para todas as variáveis os valores registrados são baixos, o que indica que os grupos são homogêneos. Comparando todos os grupos obtidos pelos métodos de agrupamento, o mais homogêneo é o grupo 2 de ligação completa, composto por três municípios. O grupo 1 resultante do método de *Ward*, composto por 26 municípios, também é muito homogêneo principalmente em relação às variáveis espvida, mort1 e sobre60.

Esses resultados são interessante pois, o que se espera é que, quanto maior o número de municípios no grupo, mais heterogêneo ele se torna. E isso não foi observado nos agrupamentos obtidos. Como mencionado, o grupo 2 do método de ligação completa, que contém 3 municípios foi mais homogêneo que outros grupos menores. O mesmo ocorreu com o grupo 1 do método de *Ward*, em relação às variáveis espvida, mort1 e sobre60.

Tabela 12 – Coeficiente de variação (CV) das variáveis dos grupos obtidos pelos métodos hierárquicos aglomerativos e distância de Mahalanobis

Método	Grupo	espvida	tft	mort1	rd	sobre60	t_env
ligação simples	1	1,70	13,12	13,27	6,57	2,24	15,96
	2	0,71	7,62	6,43	4,56	0,84	12,13
ligação completa	1	1,69	13,13	13,14	6,46	2,27	15,85
	2	0,62	6,20	5,78	3,20	0,76	10,70
distância média	1	1,70	13,12	13,27	6,57	2,24	15,96
	2	0,71	7,62	6,43	4,56	0,84	12,13
centroide	1	1,70	13,12	13,27	6,57	2,24	15,96
	2	0,71	7,62	6,74	4,56	0,84	12,13
Ward	1	0,57	11,08	4,87	7,46	0,72	17,98
	2	1,51	13,61	11,79	6,34	2,03	15,33

Fonte: elaboração própria

Cada método produziu um resultado diferente, o único consenso entre eles foi em relação ao número de grupos finais. Além disso, os grupos obtidos estão muito desbalanceados. Os

resultados apresentam um grande grupo composto pela maior parte dos municípios e um menor composto por 2, 3 ou 26 municípios. Essa situação sugere que a maioria dos municípios são muito parecidos entre si em relação às variáveis medidas e há um pequeno grupo distinto. No entanto, essa solução trouxe uma preocupação com a possibilidade do resultado não representar adequadamente as diferenças e semelhanças entre os municípios da mesorregião estudada. Portanto, com o intuito de tentar confirmar essa representação dos dados optou-se por usar uma medida de distância diferente para comparar os resultados obtidos. Foi escolhida a aplicação da distância euclidiana com dados padronizados.

A seguir serão apresentados os resultados obtidos para os métodos ligação simples, ligação completa, distância média, centroide e *Ward*, usando a distância euclidiana com os dados padronizados.

A Figura 15 mostra o dendrograma dos agrupamentos obtidos pelo método ligação simples. Ao se buscar no gráfico o ponto de maior mudança na altura do dendrograma é possível identificar dois pontos. Em um deles, o corte produz dois agrupamentos, um formado somente pelo município São Tomás de Aquino e o outro pelos demais municípios. Caso o corte fosse feito na outra altura, resultariam três agrupamentos, no entanto, um deles continuaria apenas com São Tomas de Aquino e um segundo formado por Córrego do Bom Jesus e, um terceiro grupo composto pelos outros 144 municípios da mesorregião estudada. A Tabela 13 apresenta os resultados dos critérios para escolha do número de grupos, considerando o método ligação simples. Como pode ser visto, a maioria dos métodos sugere dois grupos na partição final. Nesse cenário, as observações são agrupadas de tal forma que o primeiro grupo contenha 145 municípios e o outro grupo é formado por apenas um município (São Tomás de Aquino). Portanto, os resultados sugerem que há apenas um município que se diferencia de todos os outros 145 municípios em relação às variáveis medidas, de tal maneira que ele passa a constituir um único grupo.

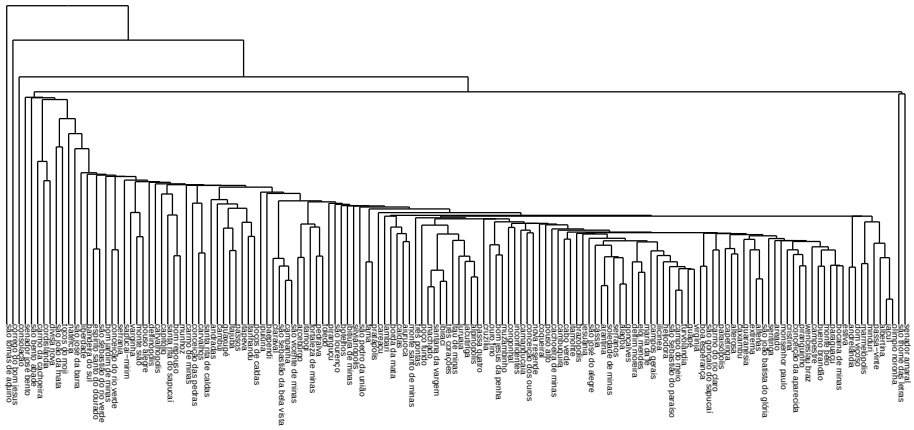


Figura 15 – Dendrograma pelo método ligação simples e distância euclidiana

Fonte: elaboração própria

Tabela 13 – Número de grupos pelo método ligação simples e distâncias euclidiana segundo a maioria dos critérios usados

Critério	Número de grupos	Critério	Número de grupos
pseudo F	10	CCC	2
Je(2)/Je(1)	2	ponto bisserial	13
índice C	7	*Gplus*	2
gamma	10	índice DB	14
Beale	2	*gap*	2
número ideal de grupos		2	

Fonte: elaboração própria.

O segundo método de agrupamento usado foi o de ligação completa. A Figura 16 apresenta o dendrograma desse agrupamento. Analisando a figura, é razoável dizer que o corte poderia ser feito na altura que produz três grupos finais. Contudo, os resultados dos critérios para escolha do número ótimo de grupos indicam dois grupos com o ideal. Dentre os 10 métodos selecionados, seis indicam esse valor, como pode ser visto na Tabela 14. Dessa forma, em um grupo estariam 42 municípios e, no outro, 104 municípios.

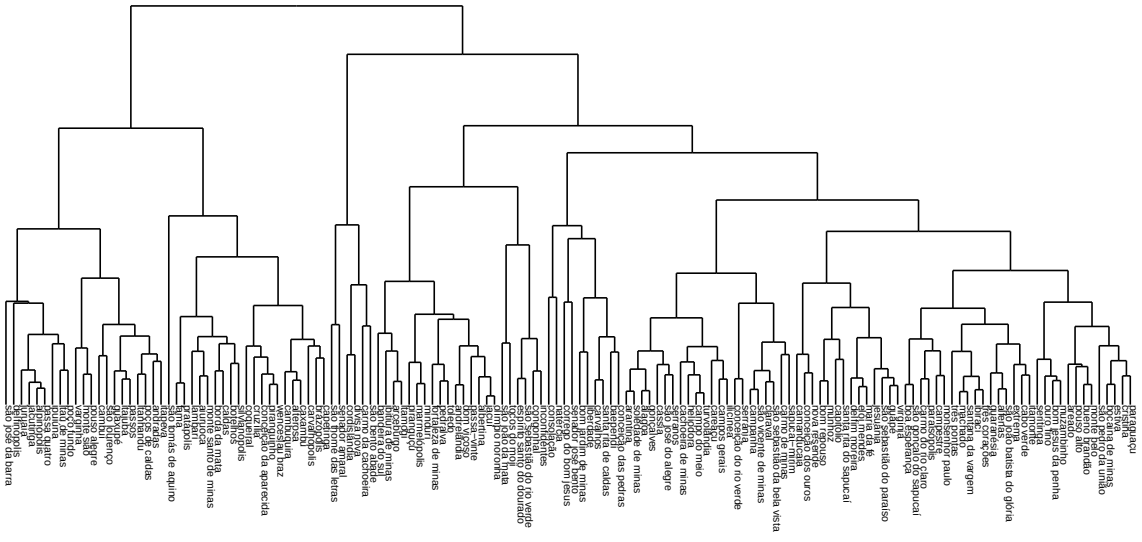


Figura 16 – Dendrograma pelo método ligação completa e distância euclidiana

Fonte: elaboração própria

Tabela 14 – Número de grupos pelo método ligação completa e distância euclidiana segundo a maioria dos critérios usados

Critério	Número de grupos	Critério	Número de grupos
pseudo F	2	CCC	2
Je(2)/Je(1)	2	ponto bisserial	7
índice C	2	*Gplus*	15
gamma	15	índice DB	15
Beale	2	*gap*	2
número ideal de grupos		2	

Fonte: elaboração própria.

Os próximos agrupamentos foram obtidos aplicando a técnica hierárquica aglomerativa distância média. A Figura 17 ilustra o dendrograma resultante da análise. Os pontos onde poderia ser realizado o corte no dendrograma sugerem quatro ou seis grupos para divisão das observações. Caso quatro grupos finais fossem escolhidos, os municípios se dividiriam em grupos de tal forma que, em um deles estariam 137 muncípios, no segundo, seis municípios (Carmo da Cachoeira, Cordislândia, Divisa Nova, São Bento Abade, São Thomé das Letras e Senador Amaral), no terceiro, dois municípios (Consolação e Córrego do Bom Jesus) e, por último, um grupo formado por um município (São Tomás de Aquino). Por outro lado, se os dados fossem divididos em seis grupos, os agrupamentos seriam constituídos da seguinte forma: 124 municípios, 13 municípios (Andradas, Cambuí, Guaxupé, Itajubá, Itanhandu, Itapeva, Monte Sião, Passos, Poços de Caldas, Pouso Alegre, São José da Barra, São Lourenço e Varginha), seis municípios (Carmo da Cachoeira, Cordislândia, Divisa nova, São Bento Abade, São Thomé das

Letras e Senador Amaral), um município (Consolação), um município (Córrego do Bom Jesus) e um município (São Tomás de Aquino). No entanto, os critérios selecionados para a escolha ótima do número de grupos sugerem dois como número de grupos ideal, como pode ser visto na Tabela 15. A divisão produz um primeiro agrupamento composto por 140 municípios e outro formado por seis municípios, sendo eles Carmo da Cachoeira, Cordislândia, Divisa nova, São Bento Abade, São Thomé das Letras e Senador Amaral.

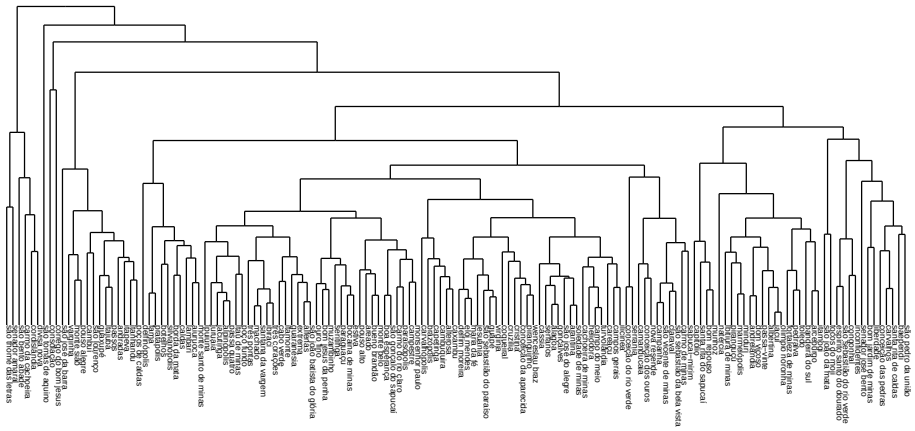


Figura 17 – Dendrograma pelo método distância média e distância euclidiana

Fonte: elaboração própria

Tabela 15 – Número de grupos pelo método distância média e distância euclidiana segundo a maioria dos critérios usados

Critério	Número de grupos	Critério	Número de grupos
pseudo F	14	CCC	
Je(2)/Je(1)	2	ponto bisserial	9
índice C	2	*Gplus*	2
gamma	15	índice DB	6
Beale	2	*gap*	2
número ideal de grupos		2	

Fonte: elaboração própria.

Os próximos agrupamentos foram obtidos usando o método do centroide, a Figura 18 apresenta o dendrograma obtido. O corte no ponto com maior mudança na altura do dendrograma divide as observações em dois grupos. O resultado sugere que São Tomás de Aquino se diferencia dos demais municípios a ponto de estar separado de todos os outros em um grupo. A Tabela 16 apresenta os resultados dos critérios de número de grupos. Dos 10 critérios avaliados, a maioria sugere dois grupos com o ideal. Apenas os métodos pseudo F e ponto bisserial indicam 13 agrupamentos, o índice C e *gamma* sugerem seis e oito grupos, respectivamente.

Portanto, considerando a divisão das observações em dois grupos, como já mencionado, o resultado seria um grupo com um município e outro formado por 145 municípios.

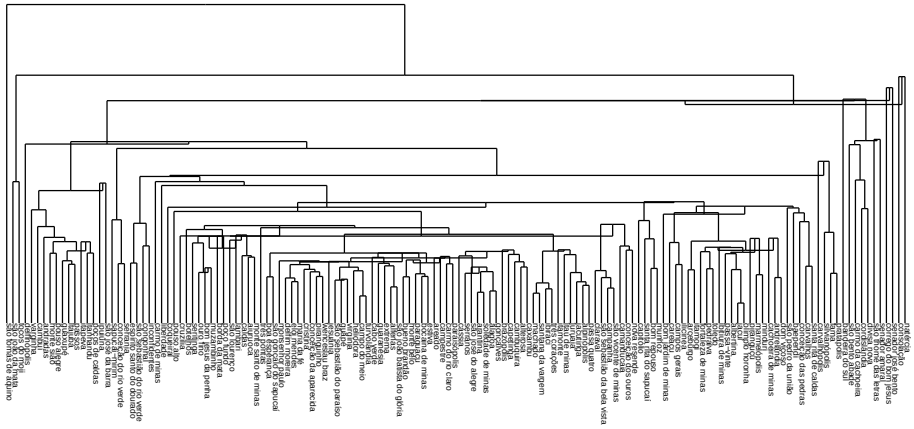


Figura 18 – Dendrograma pelo método centroide e distância euclidiana

Fonte: elaboração própria

Tabela 16 – Número de grupos pelo método centroide e distância euclidiana segundo a maioria dos critérios usados

Critério	Número de grupos	Critério	Número de grupos
pseudo F	13	CCC	2
Je(2)/Je(1)	2	ponto bisserial	13
índice C	6	*Gplus*	2
gamma	8	índice DB	2
Beale	2	*gap*	2
número ideal de grupos		2	

Fonte: elaboração própria.

Por último, as observações foram agrupadas pelo método de *Ward*. A Figura 19 mostra o dendrograma da análise. Fica evidente a separação das observações em dois grandes grupos, com aproximadamente o mesmo número de municípios. Os resultados dos critérios de escolha do número de grupos, apresentados na Tabela 17, confirmam o sugerido pelo corte do dendrograma. Com essa divisão, 77 municípios ficam inseridos em um grupo e os outros 69 municípios no outro grupo.

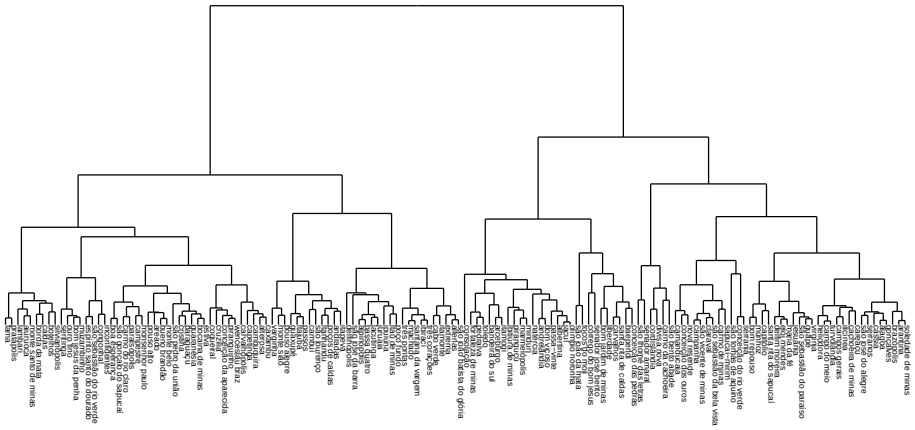


Figura 19 – Dendrograma pelo método de *Ward* e distância euclidiana

Fonte: elaboração própria

Tabela 17 – Número de grupos pelo método de *Ward* e distância euclidiana segundo a maioria dos critérios usados

Critério	Número de grupos	Critério	Número de grupos
pseudo F	2	CCC	2
Je(2)/Je(1)	2	ponto bisserial	3
índice C	10	*Gplus*	15
gamma	15	índice DB	15
Beale	2	*gap*	2
número ideal de grupos		2	

Fonte: elaboração própria.

A Tabela 18 apresenta a divisão do número de municípios em cada um dos dois grupos, de acordo com o método de agrupamento usando distância euclidiana. As técnicas ligação simples e centroide dividem as observações de forma que o primeiro grupo contém 145 municípios e o segundo grupo um município (São Tomás de Aquino). Portanto, os resultados desses métodos convergem para essa divisão das observações. O método da ligação completa, por sua vez, divide os municípios em um grupo mais balanceado, em relação aos outros métodos, assim como o método de *Ward*. Por último, a divisão sugerida pelo método da distância média mostrou uma tendência de separar em um único grupo os municípios que estão entre aqueles com piores desempenhos nos indicadores esperança de vida ao nascer, mortalidade infantil e probabilidade de sobrevida até os 60 anos.

Tabela 18 – Número de observações nos grupos usando métodos hierárquicos aglomerativos e distância euclidiana

Método	Grupo 1	Grupo 2
ligação simples	145	1
ligação completa	42	104
distância média	140	6
centroide	145	1
Ward	74	72

Fonte: elaboração própria

A Tabela 19 apresenta as medianas das variáveis dos grupos obtidos por métodos hierárquicos aglomerativos e distância euclidiana. Como os métodos ligação simples e centroide dividiram as observações em 2 grupos, em que o grupo 1 contém 145 municípios e o grupo 2 apenas um município (São Tomás de Aquino), as variáveis de cada grupo possuem as mesmas medianas e, portanto, a mesma interpretação. Nesses casos, a variável que mais contribui para a heterogeneidade entre os grupos foi a rd e a que menos contribuiu foi a tft. Os valores das medianas das variáveis dos grupos usando a distância média estão bem diferentes. Isso sugere que, por esse método, praticamente todas as variáveis estão contribuindo consideravelmente para a divisão entre o grupos, com exceção da tft que registra uma diferença pequena entre os grupos. Em relação aos agrupamentos resultantes do método ligação completa, a variável que exerceu um peso maior para os grupos se tornarem heterogêneos foi a mort1, a variável tft, por sua vez, foi a que menos se alterou entre os grupos. Por último, as medianas dos dois grupos resultantes do método de *Ward* estão próximas em todas as variáveis medidas. A maior diferença é encontrada na variável mort1, o que sugere que essa variável é a que mais contribui para a divisão entre os grupos.

Tabela 19 – Medianas das variáveis dos grupos obtidos por métodos hierárquicos aglomerativos e distância euclidiana

Método	Grupo	espvida	tft	mort1	rd	sobre60	t_env
ligação simples	1	75,53	1,95	14,50	44,85	83,28	9,38
	2	77,46	2,43	11,80	53,20	86,00	10,70
ligação completa	1	76,96	1,92	12,45	44,45	85,32	9,76
	2	75,03	1,97	15,25	44,88	82,55	9,36
distância média	1	75,69	1,94	14,30	44,83	83,52	9,51
	2	73,03	2,49	18,50	49,66	79,54	6,67
centroide	1	75,53	1,95	14,50	44,85	83,28	9,38
	2	77,46	2,43	11,80	53,20	86,00	10,70
Ward	1	76,22	1,91	13,50	43,79	84,28	9,44
	2	74,41	1,98	16,20	46,36	81,64	9,36

Fonte: elaboração própria.

A Tabela 20 apresenta o coeficiente de variação das variáveis dos grupos obtidos pelos métodos hierárquicos aglomerativos e distância euclidiana. Os resultados mostram que os valores registrados para todas as variáveis dos grupos resultantes dos métodos ligação simples e centroide são baixos, o que indica certa homogeneidade interna. Com exceção da rd, as variáveis do primeiro grupo, obtido pelo método ligação completa, registram menor dispersão que o segundo grupo. Em relação ao método distância média, de forma geral, o segundo grupo formado é considerado mais homogêneo que o primeiro, em relação a quase todas variáveis demográficas, as únicas exceções ocorrem com as variáveis rd e t_env. Por fim, os agrupamentos referentes à técnica de *Ward* também se mostram bastante homogêneos. Em todas as variáveis medidas, o primeiro grupo, composto por 77 municípios registra coeficientes de variação abaixo daqueles do segundo grupo, portanto, o grupo 1 apresenta homogeneidade interna maior que o grupo 2.

Tabela 20 – Coeficiente de variação (CV) das variáveis dos grupos obtidos pelos métodos hierárquicos aglomerativos e distância euclidiana

Método	Grupo	espvida	tft	mort1	rd	sobre60	t_env
ligação simples	1	1,72	13,08	13,49	6,44	2,28	15,91
ligação completa	1	0,99	10,68	8,95	7,42	1,29	15,69
	2	1,36	13,91	10,28	6,20	1,83	15,91
distância média	1	1,65	12,10	13,05	6,31	2,18	14,79
	2	0,67	4,72	4,45	7,07	0,94	25,25
centroide	1	1,72	13,08	13,49	6,44	2,28	15,91
Ward	1	0,99	10,78	8,48	5,67	1,29	14,46
	2	1,23	14,39	8,99	6,46	1,67	17,32

Fonte: elaboração própria.

De forma geral, utilizando diferentes medidas de distância, os resultados encontrados foram completamente diferentes. Com o uso da distância de Mahalanobis houve uma tendência em alocar em um grupo menor os municípios com tendência a registrar melhores desempenhos nos indicadores considerados. O contrário aconteceu com o uso da distância euclidiana, que apresentou tendência em separar em um grupo menor os municípios com piores desempenhos. Em relação ao município de São Tomás de Aquino, ele entre os 10 com maiores esperanças de vida, probabilidade de sobrevida até 60 anos e ocupa a 11ª posição das menores taxas de mortalidade infantil. Por outro lado, ele registra alta taxa de fecundidade total, o que conduz à elevada razão de dependência ocupando a primeira posição na mesorregião estudada. No entanto, os grupos continuaram muito desbalanceados na maioria dos métodos.

Na literatura não há um consenso sobre qual o melhor método de agrupamento, mas sim peculiaridades de cada técnica que precisam ser consideradas em sua aplicação. O método da ligação simples, por exemplo, geralmente não apresenta um bom desempenho e tende a produzir grupos longos e encadeados, com formatos não convexos. Isso acontece porque o método possui a desvantagem de unir um objeto a um grupo desde que ele esteja próximo de qualquer um dos objetos desse grupo, mesmo que esteja distante de todos os outros (LATTIN; CARROLL; GREEN, 2011). De acordo com Hair et al. (2009), o método da ligação completa tenta eliminar esse problema da ligação simples e produz grupos mais compactos, convexos e homogêneos, mas é considerado muito sensível a discrepâncias nos dados (LATTIN; CARROLL; GREEN, 2011). Em contrapartida, os agrupamentos produzidos pelo método da distância média são conhecidos por possuírem aproximadamente mesma variância interna e por formar melhores

partições que os métodos de ligação simples e completa (MINGOTI, 2005; HAIR et al., 2009). Em relação ao método centroide, uma das vantagens associadas é que o método é robusto para discrepâncias, mas, ainda assim, a distância média pode obter um melhor desempenho nesse sentido (LATTIN; CARROLL; GREEN, 2011). Por fim, o método de *Ward* é conhecido por tender a resultados com grupos formados aproximadamente pelo mesmo número de observações em cada um deles (LATTIN; CARROLL; GREEN, 2011). Além disso, é o mais indicado quando o pesquisador está trabalhando com dados contínuos (EVERITT et al., 2011). De forma geral, esses métodos podem resultar em diferentes agrupamentos. No entanto, é esperado que os resultados obtidos apresentem se não certa semelhança, pelo menos certa consistência (MINGOTI, 2005).

Os próximos passos são a aplicação do método não hierárquico k-médias e os agrupamentos utilizando os escores dos componentes principais.

5 Anexo A - Lista de municípios da mesorregião Sul/Sudoeste de Minas Gerais

Tabela 21 – Municípios da mesorregião Sul/Sudoeste de Minas Gerais

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: elaboração própria.

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.; STEELE, F.; GALBRAITH, J.; MOUSTAKI, I. **Analysis of multivariate social science data**. Boca Raton: CRC press, 2008.

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

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

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?** [S.l.]: Instituto de Pesquisa Econômica Aplicada (Ipea), 2014.

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 população brasileira: um enfoque demográfico. **Cad. saúde pública**, v. 19, n. 3, p. 725–733, 2003.

CARVALHO, J. A. M. d.; WONG, L. R. A transição da estrutura etária da população brasileira na primeira metade do século xxi. p. 597–605, 2008.

CELEUX, G.; GOVAERT, G. A classification em algorithm for clustering and two stochastic versions. **Computational statistics & Data analysis**, Elsevier, v. 14, n. 3, p. 315–332, 1992.

CHARRAD, M.; N., G.; BOITEAU, V.; A., N. Nbclust: An r package for determining the relevant number of clusters in a data set. **Journal of Statistical Software**, v. 61, n. 6, 2014.

DAVIES, D.; BOULDIN, D. Acluster separation measure: Ieee transactions on pattern analysis and machine intelligence. **PAMI-1**, v. 2, p. 224–227, 1979.

DIMITRIADOU, E.; DOLNICAR, S.; WEINGESSEL, A. An examination of indexes for de-

termining the number of clusters in binary data sets. **Psychometrika**, Springer, v. 67, n. 1, p. 137–159, 2002.

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**. [S.l.]: Springer Science & Business Media, 2011.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster analysis**. 5. ed. UK: John Wiley and 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.

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.; TAKAHASHI, D. Y.; PATRIOTA, A. G.; SATO, J. R. 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.

GURALNIK, J.; FRIED, L.; SIMONSICK, E.; KASPER, J.; LAFFERTY, M. National institute of aging, national institutes of health pub. 1995.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. [S.l.]: Bookman Editora, 2009.

HARTIGAN, J. A. **Clustering algorithms**. [S.l.: s.n.], 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.

JOHNSON, D. E. **Applied multivariate methods for data analysts**. [S.l.]: Duxbury Resource Center, 1998.

KHATREE, R.; NAIK, D. N. Multivariate data reduction and discrimination with sas software. **nc: SAS Institute Inc.**, 2000.

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**, American Economic Association, v. 17, n. 4, p. 167–190, 2003.

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

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.

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.

ROHLF, F. J. Methods of comparing classifications. **Annual Review of Ecology and Systematics**, JSTOR, p. 101–113, 1974.

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

SARLE, W. S. **Cubic clustering criterion**. [S.l.]: SAS Institute, 1983.

STUDIO, R. Rstudio: integrated development environment for r. **RStudio Inc, Boston, Massachusetts**, 2012.

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

TEAM, R. C. et al. R: A language and environment for statistical computing. Vienna, Austria, 2013.

TIBSHIRANI, R.; WALTHER, G. Cluster validation by prediction strength. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 14, n. 3, p. 511–528, 2005.

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. **Epidemiologia e Serviços de Saúde, Coordenação-Geral de Desenvolvimento da Epidemiologia em Serviços/Secretaria de Vigilância em Saúde/Ministério da Saúde**, v. 21,

n. 4, p. 539–548, 2012.

WANG, J. Consistent selection of the number of clusters via crossvalidation. **Biometrika**, Biometrika Trust, v. 97, n. 4, p. 893–904, 2010.