1 INTRODUÇÃO

As aposentadorias e os outros benefícios concecidos pela Previdência Social formam grande parte da renda dos municípios mineiros, influenciando sua economia e reduzindo desigualdades sociais Franca. Entretanto, o impacto dessa política pública não é uniforme, pois os municípios são bastante diversos entre si em relação à estrutura etária, à taxa de fecundidade, às atividades econômicas predominantes etc. O objetivo da presente pesquisa é construir uma tipologia dos municípios de acordo com a sua relação com a previdência social.

Em 3.546 dos 5.561 municípios brasileiros, em 2004, o pagamento de benefícios superou o Fundo de Participação dos Municípios (FPM), como era o caso das quatro melhores cidades em desenvolvimento humano no Brasil: São Caetano do Sul - SP, em que o pagamento de benefícios foi $27,5$ vezes maior do que o FPM; Águas de São Pedro - SP, $2,6$ vezes; Niterói - RJ, $38,4$; e Florianópolis - SC, $13,6$. Além disso, o recebimento desses beneficios evita o êxodo para grandes cidades, fixando as pessoas em seus municípios de origem Franca.

Ainda segundo França (2004), nos municípios de até 5 mil habitantes, os benefícios representavam $20,3\%$ da renda das famílias, sendo que a média brasileira era de $7,2\%$. Para a população rural, a previdência social é uma forma de redistribuição de renda, já que essas pessoas dificilmente contribuíram diretamente para a Previdência Social, fazendo com que este seja como um programa de renda mínima para os idosos no país. O acesso a esses benefícios melhora de forma significativa a qualidade de vida dos domicílios.

Atualmente há razoável disponibilidade de dados municipais no Brasil, seja através das pesquisas do Instituto Brasileiro de Geografia e Estatística (IBGE) (em especial, os Censos Demográficos e o Perfil dos Estados e dos Municípios Brasileiros) seja através dos registros administrativos de diversos órgãos públicos (Ministério da Saúde, Ministério da Educação, Ministério da Previdência Social etc.). 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 socioeconômicos municipais, uma vez que a realidade municipal é multidimensional e muitas dessas variáveis estão intimamente relacionadas.

Nesse sentido, o objetivo deste trabalho é propor uma classificação dos municípios minerios, com ênfase na mesorregião sul e sudoeste, em relação aos seus gastos com previdência social e a algumas variáveis demográficas. Especificamente, será utilizada a análise de componentes principais para reduzir a dimensionalidade dos dados.Em seguida, será realizada a análise de agrupamento. O que se espera é que os grupos de municípios obtidos apresentem grande homogeneidade interna e grande heterogeneidade extena em relação às variáveis analisadas. Serão utilizados diferentes métodos de agrupamento e os resultados obtidos serão comparados. Por último, diferentes critérios para definição do número de grupos na partição final da análise de agrupamento serão analisados.

2 Revisão de Literatura

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.

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.

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, que corresponde ao aumento, em termos relativos, da população idosa. 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 Gabriel. 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.

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 Gabriel. 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. 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.

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 2, 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

Figura 2: *
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 wong. 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% Gabriel. 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 wong.

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 wong.

O gráfico 3 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. Em 1960, a RDT avançou 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 3: Razão de dependência do Brasil, 1940 a 2050

Figura 4: *
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 prev.

Por último, o gráfico 5 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 wong.

Figura 5: Índice de envelhecimento do Brasil, 1950 a 2050

Figura 6: *
Fonte: elaboração própria a partir a partir de dados do Instituto Brasileiro de Geografia e Estatística)

2.2 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.

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 as técnicas de interdependência estão a análise fatorial e análise de agrupamento hair.

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:

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

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.

A obtenção da matriz de dados a partir de uma amostra aleatória, como expressa na forma da equação (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 Daniel.

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

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

em que $r_{ij} = \dfrac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$ Daniel.

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

Segundo Mingoti, a análise multivariada se divide em dois grupos principais: técnicas exploratórias e técnicas de inferência estatística, como também ocorre na análise univariada. O primeiro possui um grande apelo prático por não dependerem do conhecimento da forma matemática da distribuição de probabilidade que gerou os dados amostrais e permitem a detecção de padrões. Exemplos de técnicas desse tipo são 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.

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.

2.3 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 reescrever os dados multivariados de forma que poucas dimensões expliquem a maior parte das informações contidas no conjunto de dados originais James.

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 contabilizem a maior parte da variação presente em todas as variáveis originais Mingoti,Everitt.

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.

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 Daniel. Isso ocorre porque a técnica possibilita que cada componente não esteja correlacionado com todos os outros, retirando a multicolinearidde em uma análise de dependência James.

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. 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. A obtenção de autovalores λ e autovetores $\boldsymbol{e}$ de uma matriz quadrada $p \times p$ são tais que $\boldsymbol{A}\boldsymbol{e} = \lambda\boldsymbol{e}$. Para maiores detalhes, consultar, por exemplo, onlineDaniel.

Ainda de acordo com onlineEveritt, 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 autovetor. 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.

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, sempre o primeiro componente principal tem a maior proporção de explicação da variância total de $\boldsymbol{X}$ Mingoti.

De acordo com onlineEveritt, os primeiros k componentes, em que $k < p$, explicam uma proporção da variância total,

$$\frac{\sum_{i=1}^{k} Var(Y_i)}{\text{Variância total de } X} = \frac{\sum_{i=1}^{k} \lambda_i}{\text{traço}(\boldsymbol{S})} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j}. \tag{3}$$

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.

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 (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.

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.

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 $\boldsymbol{X}$ James. Mais detalhes sobre critérios podem ser vistos também em onlinekhatree2000.

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. Os escores dos primeiros dois componentes principais podem ser plotados em um diagrama de dispersão para identificar agrupamentos ou outros tipos de padrão existente nos dados.

Para calcular os escores dos componentes de cada observação i, se os componentes foram obtidos

a partir da matriz de covariâncias amostral $\boldsymbol{S}$, 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.

2.4 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. Portanto, dada uma amostra de tamanho n, com cada objeto medido segundo p variáveis, a análise de conglomerados classifica os objetos em grupos com elevado grau de homogeneidade interna e heterogeneidade externa.

Diferente da análise de componentes principais que visa reduzir a dimensionalidade da matriz de dados, a análise de conglomerados busca reduzir o número de objetos, ou seja, o número de linhas em uma matriz de observações por variáveis. De acordo com onlineGordon, a classificação de dados em grupos pode ser realizada com o objetivo de simplifica-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.

Como forma de ilustração, considere um conjunto de dados fictícios em que há $n = 23$ observações (pessoas) e registros sobre suas idades e rendas mensais, ou seja, há $p = 2$ variáveis, ambas medidas na escala contínua. O objetivo é agrupar as pessoas em relação às duas variáveis. O diagrama de dispersão obtido se encontra na Figura 1 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.

Há dois objetivos possíveis de um agrupamento: agrupar as n observações em um número desconhecido de grupos 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 Daniel.

De acordo com onlineBartholomew qualquer processo de agrupamento tem como base duas etapas:

1. Obter a distância 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 correspondem às distâncias, ao passo que as de similaridades complementam as distâncias, assim quanto maior a medida de similaridade entre dois objetos menor será a de dissimilaridade e mais próximos eles serão Daniel.

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 $4 \times$

Figura 7: Diagrama de dispersão de dados fictícios de idade e renda de várias pessoas.

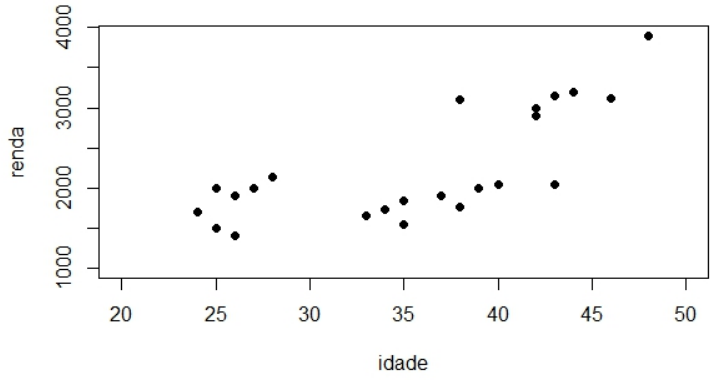

Figura 8: *
Fonte: modificado a partir de (BARTHOLOMEW et al., 2008, p.18).

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.

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.

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 $X_{i1}, X_{i2}, \ldots, X_{ip}$, e j, com os valores $X_{j1}, X_{j2}, \ldots, X_{jp}$. 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 James. De acordo com onlineDaniel, 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_p e X_k definida por:

$$D_{ij} = [(X_{ip} - X_{ij})'B(X_{ip} - X_{ij})]^{1/2}$$

em que B_{pxp} é uma matriz de ponderação, positiva definida. A seleção dessa matriz define a distância utilizada. Se B_{pxp} é a matriz de covariância (S_{pxp}^{-1}) 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. 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 Daniel. Além disso, a métrica produz agrupamentos compactos e convexos James e elimina o efeito o efeito de domínio na classificação das variáveis de maior variabilidade Daniel. A distância Euclidiana média, por sua vez, é definida quando a matriz B_{pxp} é igual a diag($\frac{1}{p}$) Mingoti.

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 cluster. Portanto, a cada passo se perde um grupo que é unido ao mais similar. 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.

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. Esses métodos estão divididos em: Método de Ligação Simples, Método de Ligação Completa, Método da Média das Distâncias, Método do Centróide e Método de Ward Mingoti. Com exceção do Método de Ward, as demais técnicas seguem um processo iterativo geral, denominado método de grupo de pares James, 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.

Método de 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 *cluster* formado e, assim sucessivamente. O processo é finalizado quando há um único *cluster* 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 *cluster*. Dessa forma, a ligação por vizinho mais próximo tende a formar agrupamentos longos e encadeados, com formatos não convexos James. Como alternativa foram propostas outras técnicas, que produzem soluções distintas e tentam eliminar essa tendência do método de ligação simples.

Método de 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:

$$d_{C_{n+s}C_k} = max\{d_{C_iC_k}, d_{C_jC_k}\}$$

Dessa forma, por esse método dois objetos são considerados similares de acordo com o menor valor de máximo Mingoti. 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 cluster esteja próximo não somente de um elemento, mas de todos os outros. Dessa forma, é razoável dizer que os conglomerados resultantes da ligação completa geralmente são convexos e tendem a ser de aproximadamente mesmo diâmetro. Contudo, o método pode ser sensível a discrepância nos dados James, Mingoti.

Método de Ligação Média

Nesse método a distância entre dois agrupamentos é definida como a média das distâncias entre todos os pares de objetos Mingoti. Assim, a distância média entre o conglomerado C_k e o novo conglomerado C_{n+s} pode ser representada por:

$$d_{C_{n+s}C_k} = \frac{n_i d_{C_iC_k} + n_j d_{C_iC_k}}{n_i + n_j}$$

em que $n_i + n_j$ é o número de objetos no conglomerado C_{n+s}. O método de ligação média tende a formar conglomerados com melhores partições do que os de ligação simples e completa James. Além disso, os conglomerados resultantes possuem aproximadamente a mesma variância interna Everitt.

Método centróide

A distância entre dois agrupamentos é definida a partir da distância entre dois centróides, calculados com base na média entre os objetos de cada grupo Mingoti. De acordo com James, 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:

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

Método de Ward

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

Passo 1. Inicialmente os n objetos são alocados em n *clusters* 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 SS_i de um conglomerado G_i é definida por:

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

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* conglomerado 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:

$$SSR = \sum_{i=1}^{g_k} SS_i$$

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:

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

Assim como no método do Centróide, 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 maior 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, os agrupamentos obtidos geralmente possuem mesmo número de observações Mingoti.

De acordo com onlineEverittet, os agrupamentos obtidos 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 9 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 *clusters* 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 9: Terminologia utilizada na descrição de dendrogramas

Figura 10: *
Fonte: elaboração própria a partir de Everitt.

Portanto, no caso representado pela figura acima, os objetos denominados de A e C foram os primeiros a serem unidos em um único *cluster*, 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 *cluster* 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.

Outra técnica de agrupamento que pode ser utilizada é *K-Means*. No entanto, diferente daquelas já apresentadas, essa técnica é do tipo não hierárquica ou de partição James. As principais características desses métodos são: aplicação do processo à matriz de dados X e número de grupos k definido *a priori* Daniel. 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 foram 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.

Método das K-Médias

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

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 onlineMingoti, o processo iterativo do método pode ser descrito como:

Passo 1. Inicialmente são escolhidos k centróides, 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 centróide inicial, então o objeto é unido ao grupo de menor distância;

Passo 3. Os valores dos centróides 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), diferentes números de grupos podem produzir diferentes partições finais Daniel. Diversas sugestões para decisão das sementes são apresentadas na literatura, onlineMingoti 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 *cluster*, utilizado no método de Ward, é próximo ao critério do quadrado da soma de erros de partição do método $k-means$ James. 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 centróide 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.

Não há um consenso sobre o melhor método para escolha do número de grupos inicial ou de seus centróides, contudo é aconselhável que o processo seja realizado com diferentes escolhas para busca da melhor solução de agrupamento Daniel, James.

2.4.1 Número de grupos

A etapa final do processo de agrupamento hierárquico é definir a partição do conjunto de dados. Essa não é uma tarefa simples e existem vários métodos propostos para definir o número k de agrupamentos ou em qual passo o algoritmo de agrupamento deve ser interrompido. Apesar de não haver um consenso, alguns critérios podem ser utilizados para auxiliar na decisão final Mingoti. Diferentes técnicas podem ser aplicadas tanto nos resultados dos agrupamentos resultantes de métodos hierárquicos como regra de parada do algoritmo, como para obter o número de grupos *a priori* para aplicação de procedimentos não hierárquicos milligan1985.

Ainda de acordo com onlinemilligan1985, 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 *clusters* menor do que o apropriado. Apesar dos dois tipos de erros serem indesejáveis, o segundo produz consequências consideradas mais sérias, pois informação é perdida na união de diferentes grupos.

Quando métodos hierárquicos de agrupamento são utilizados, um dendrograma é obtido e deve-se decidir em qual altura o corte deve ser realizado, o que vai gerar um determinado número de grupos. A questão é decidir o ponto de corte. Uma maneira informal de tomar essa decisão é avaliar as mudanças na altura do dendrograma nos diferentes passos e escolher a maior mudança observada. Porém, mesmo com um número de observações não muito grande (como 15 ou 20), não é simples decidir onde está essa maior mudança Everitt.

Outros critérios informais utilizam um gráfico da medida *versus* o número de grupos em que se busca identificar grandes mudanças no gráfico para um determinado número k e um ponto de parada é sugerido. Porém, esses critérios são subjetivos. Várias técnicas mais formais têm sido propostas e alguns trabalhos avaliaram suas propriedades, tais como onlinemilligan1985 e onlinedimitriadou2002.

onlinemilligan1985 apresentaram um estudo de simulação de Monte Carlo para comparar 30 critérios de determinação de número de grupos na literatura de agrupamento. A ideia foi testar as regras de decisão de parada que buscam eliminar a subjetividade presente nos métodos gráficos. Com o intuito de obter diferentes partições finais, um conjunto de dados fictícios foram analisados por 4 métodos de agrupamentos hierárquicos diferentes. O trabalho identificou como as duas melhores técnicas as propostas por onlinecalinski1974, também conhecida como pseudo F, e onlineduda1973, denominada pseudo T^2. No entanto, o autor ressalta 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. onlineEveritt apresentam um resumo desses e de outros critérios que podem auxiliar nessa decisão.

A estatística Pseudo F proposta por onlinecalinski1974 é definido por:

$$PseudoF = \frac{tr[B/(K-1)]}{tr[W/(n-K)]}$$

em que, B é a matriz da soma de quadrados entre agrupamentos, W representa a matriz da soma de quadrados dentro dos grupos, K corresponde ao número de grupos e n é o número de objetos. A estatística pode ser utilizada tanto para determinar o número de grupos no caso dos métodos hierárquicos como nos não-hierárquicos Daniel.

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 *clusters* que estão sendo obtidos na etapa Mingoti. O 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 conglomerado James. Logo, até um determinado valor, quanto maior a 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 relacionada com a menor probabilidade de significância do teste e, portanto, a hipótese nula de igualdade dos vetores de médias populacionais com maior significância é rejeitada Mingoti,James. 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.

O critério Pseudo T^2 proposto por onlineduda1973 é representado por:

$$PseudoT^2 = \frac{B_{ip}}{[\sum_{j\epsilon G_i} \parallel X_{ij} - \bar{X}_i \parallel^2] + [\sum_{j\epsilon G_p} \parallel X_{pj} - \bar{X}_p \parallel^2](n_i + n_p - 2)^{-1}}$$

onde

$$\parallel X_{kj} - \bar{X_k} \parallel = [(X_{kj} - \bar{X_k})'(X_{kj} - \bar{X_k})^{\frac{1}{2}}, k = i, p$$

A estatística é definida quando dois grupos são alocados formando um novo agrupamento, por exemplo se G_k é a união de G_p e G_i. Segundo onlineMingoti, a cada etapa do algoritmo o valor da estatística é calculada, assim como no caso da Pseudo F e, então é plotado um gráfico do tipo etapa *versus* valor da Pseudo T^2. Com base nesse critério, o número de grupos apropriado é aquele correspondente ao valor de máximo da estatística. Esse valor é aquele que rejeita a igualdade entre vetores de médias, o que sugere que a união dos dois grupos referentes a esse passo não deveria ser realizada. Portanto, geralmente o pesquisador escolhe ou o valor k de grupos correspondente a esse passo ou o anterior.

Um método que utiliza a matriz de dissimilaridade é o *silhouette plot* proposto por onlinekaufman2009. Neste método, para cada observação i, é definido um índice $s(i)$ entre -1 e 1. Quando este valor é próximo de 1 indica que a observação foi bem classificada no grupo, se próximo de -1 indica o contrário. Quando o valor assumido é próximo de 0 não está claro se a observação deveria estar no seu grupo ou em outro. O gráfico mostra os valores de $s(i)$ em forma de barras horizontais, ordenadas de forma decrescente para cada agrupamento. Comparar *silhouette plots* para soluções obtidas com diferentes números de grupos pode auxiliar na escolha deste número, levando a melhores agrupamentos.

A estatística GAP foi criada por onlineTibshirani2001 com o mesmo propósito. Segundo o autor, dentre os métodos heurísticos para solucionar o problema do número de grupos, inclui-se o *Elbow Method*. 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, a estatística proposta formaliza a ideia de procurar pelo "*elbow*"no gráfico do número de grupos $versus$ algum critério de otimização, ou seja, procurar por uma grande mudança na inclinação do gráfico a partir da qual não há grandes ganhos no critério.

Para compreensão do método, suponha um conjunto de dados $\{x_{ij}\}, i = 1, 2, 3, ..., n, j = 1, 2, 3, ..., p$, composto por n observações independentes e p variáveis, alocado em k grupos $G_1, G_2, ..., G_k$, em que G_s indica as observações no grupo s, e n_s representa o número de objetos no grupo s. Então a soma das distâncias de todas as observações do agrupamento s é dada por:

$$W_k = \sum_{s=1}^{k} \frac{1}{2n_r} D_r$$

em que, $D_r = \sum_{i,j \epsilon G_s} d_{ij}$ e d_{ij} é a distância Euclidiana entre as observações i e j. Então, a medida de erro W_k é a soma das médias das distâncias entre os objetos dos k agrupamentos ponderadas pelo fator $1/2$.

A proposta é padronizar o gráfico do $log(W_k)$ com o número de grupos K, em que W_k é minimizado para k grupos, comparando com a esperança da distribuição sob a hipótese nula dos dados. A estimativa do número ótimo de agrupamentos é o valor k em que a estatística GAP definida a seguir, é maior:

$$Gap_n(k) = E *_n \{log(W_k)\} - log(W_k)$$

em que $E*_n$ corresponde a esperança sobre o tamanho da amostra n para a distribuição de referência.

O processo de obtenção da estatística Gap é composto por 4 etapas:

Etapa 1. Agrupar os dados em agrupamentos $k = 1, 2, ..., K$ e calcular a medida de erro W_k;

Etapa 2. Obter B populações de referência e calcular $W*_{kb}, b = 1, 2, ..., B, k = 1, 2, ..., K$. Os autores mostraram que para cada valor k de interesse pode-se obter através de simulação de Monte Carlo as B populações de referência que seguem distribuição uniforme. Então a estatística Gap é estimada a partir da seguinte maneira:

$$Gap(k) = \frac{1}{B}\sum_b log(W*_{kb})?log(W_k)$$

Etapa 3. Seja $\bar{l} = (1/B)\sum_b log(W*_{kb})$, então o desvio padrão é dado por

$$Sd_k = [(1/B)\sum_b \{log(W*_{kb}) - \bar{l}\}^2]^{\frac{1}{2}}$$

e $s_k = sd_k\sqrt{(}1 + 1/B)$. A escolha do número de grupos via a Estatística Gap é dada por:

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

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

Diante dessa diversidade, é crucial utilizar não apenas um método para definir o número de grupos, mas avaliar os resultados obtidos com diferentes critérios. Além disso, alguns deles fazem suposições sobre a estrutura dos grupos e terão bom desempenho apenas se as suposições forem atendidas Everitt.

Além disso, quando estudos socioeconômicos estão em questão, adicionalmente aos métodos estatísticos, as especificidades do problema analisado devem ser levadas em conta para que se decida qual critério fornece grupos cuja interpretação seja mais útil carvalho2007.

3 Metodologia

3.1 Bases de dados e variáveis do estudo

As bases de dados utilizadas neste trabalho são provenientes do Ministério da Previdência Social (disponível em www.previdencia.gov.br) e do Atlas do Desenvolvimento Humano no Brasil 2013 (disponível em www.atlasbrasil.org.br), que utiliza os censos demográficos realizados pelo IBGE em 1991, 2000 e 2010 para calcular cerca de 230 variáveis para os 5.565 municípios brasileiros. Os dados estão tabulados em formato de planilhas `.xls`, o que facilita seu tratamento. A partir das variáveis demográficas presentes no Atlas, inicialmente três foram escolhidas - outras poderão ser acrescentadas de acordo com o que for sugerido pela literatura a ser analisada. A Tabela 1 apresenta a lista provisória das variáveis .

A escolha das variáveis do Atlas se deu devido à suposição que os valores do benefícios dos municípios devem ter relação com o número de habitantes, a idade de seus habitantes e a porcentagem de trabalho formalizado. Outras variáveis poderiam ser escolhidas para este fim, porém, posteriormente, mais algumas podem ser incluídas no estudo.

Como as variáveis de previdência estão expressas em diferentes unidades/medidas, torna-se necessária uma padronização. Será utilizado o valor por habitante, ou seja, as variáveis serão divididas pelo

Tabela 1: Lista provisória das variáveis

Sigla	Descrição	Fonte
QUANT	quantidade de benefícios em dezembro	MPS
ARREC	valor arrecadado	MPS
VAB	valor anual dos benefícios	MPS
VBD	valor dos benefícios em dezembro	MPS
POP	população residente total no município	Atlas
T_ENV	razão entre a população de 65 anos ou mais de idade e a população total multiplicado por 100	Atlas
P_FORMAL	razão entre o número de pessoas de 18 anos ou mais formalmente ocupadas e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100	Atlas

número de habitantes do município (POP). Outro tipo de transformação possível seria fazer com que todas variassem no intervalo de 0 a 1 ou de 0 a 100.

O primeiro passo da análise será a obtenção da matriz de correlações entre as variáveis, duas a duas, de forma a identificar os pares de variáveis mais associadas entre si. Tal matriz terá a forma apresentada na equação (2). Essa etapa compreende uma análise exploratória dos dados que auxiliará nas aplicação das técnicas multivariadas posteriores.

A ACP será utilizada para reduzir a dimensionalidade dos dados, conforme explicitado na seção **??**. Pretende-se reduzir o conjunto das variáveis originais correlacionadas entre si a um novo conjunto de variáveis, os componentes principais, não correlacionadas. O que se espera é que um pequeno número dessas novas variáveis expliquem boa parte da variação presente nos dados originais. Se apenas dois componentes, Y_1 e Y_2, já explicarem boa parte da variação presente nos dados, será 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 de quantos serão utilizados, as correlações entre cada variável original e cada componente serão calculadas, os chamados *loadings*. Os valores dessas correlações, bem como seus sinais, indicarão como cada componente poderá ser interpretado.

Após a aplicação da técnica de componentes principais será utilizada a análise de agrupamento para identificar os grupos de municípios com características semelhantes. Como foi mostrado na seção **??**, a AA oferece várias opções para a escolha da medida de distância, do método de agrupamento e do número de grupos.

Como medida de similaridade será utilizada a distância euclidiana, a mais aplicada em análise de agrupamento. Um dos métodos de agrupamento será o método hierárquico aglomerativo de Ward, o mais indicado quando as variáveis medidas estão na escala contínua.

Além do método aglomerativo também será aplicado o não hierárquico das k-médias, também muito popular. Este método requer que o número de grupos seja definido antes de sua aplicação. Assim, os métodos para definição do número de grupos mostrados na seção 2.4.1 serão utilizados e verificar-se-á se, e em quais condições, há convergência entre eles no caso dos dados analisados. Ao final, os resultados dos agrupamentos obtidos com os métodos Ward e k-médias serão comparados entre si.

Todas as rotinas necessárias para a análise dos dados serão realizadas utilizando o programa *R* em sua versão 3.2.0 r2015. R.

4 Resultados Esperados

O que se espera é que os municípios brasileiros sejam agrupados de forma satisfatória de acordo com as variáveis demográficas e previdenciárias escolhidas. Pretende-se identificar padrões nos dados e que semelhanças e diferenças entre os municípios sejam detectadas. O resultado pode ajudar na análise dos impactos e dos problemas da previdência social, contribuindo para novas pesquisas e para a melhoria dessa política pública.

5 Cronograma

O projeto será concluído em 24 meses, no período de março de 2015 a fevereiro de 2017. As atividades mensais a serem desenvolvidas compõem-se das seguintes etapas:

Atividades	Meses																							
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	X	X	X	X																				
2					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3										X	X	X	X	X	X	X								
4													X	X	X	X	X	X						
5																		X						
6															X	X	X	X	X	X	X	X	X	X
7																								X

1: Definição do Tema;

2: Revisão de literatura;

3: Implementação do trabalho;

4: Obtenção dos resultados;

5: Qualificação;

6: Redação da dissertação;

7: Finalização do trabalho/defesa da dissertação.

6 Disciplinas necessárias

As disciplinas necessárias para realizar este projeto são: Álgebra Linear Aplicada; Probabilidade; Inferência Estatística; Inglês Instrumental em Estatística Aplicada e Biometria; Estatística Computacional; Análise Multivariada.