MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Alfenas - UNIFAL-MG

Projeto de Pesquisa

Classificação dos municípios brasileiros em relação aos gastos com previdência utilizando análise de agrupamento

Autor: Larissa Gonçalves Souza

Orientadora: Patrícia de Siqueira Ramos

Co-orientador: Reinaldo Antônio Gomes Marques

Alfenas - MG

2015

Sumário

1 Introdução

Comumente, ao se analisar um conjunto de dados, o pesquisador precisa decidir se, em média, os tratamentos aplicados produziram resultados iguais ou qual tratamento produz melhores resultados. Uma análise de variância visa fundamentalmente verificar se existe uma diferença significativa entre as médias dos tratamentos, mas não indica quais as médias se diferem. Para investigar onde se encontram as diferenças existem os testes de comparações múltiplas.

Há muitos testes de comparações múltiplas e eles diferem quanto ao controle de erro tipo I e poder. Para escolher o melhor teste, deve ser levada em conta as qualidades estatísticas dos procedimentos (tipo de erro que é controlado e a forma como esse erro é controlado). Os testes de comparações múltiplas de médias podem ser conservadores ou liberais. Alguns testes, como t de Student e Duncan, possuem elevadas taxas de erro tipo I por experimento, sendo liberais, enquanto outros, como Scheffé e Tukey, possuem taxas de erro tipo I por experimento inferiores ao nível nominal de significância e são considerados conservadores.

O que se espera é um teste que controle o erro tipo I na maior parte das situações, mas apresente altas taxas de poder. Métodos de reamostragem *bootstrap* têm sido utilizados em alguns estudos sobre teste de comparações múltiplas das médias para melhorar o seu desempenho, Ramos e Ferreira (2009) utilizaram *bootstrap* para um dos procedimentos de comparações múltiplas de Caliński & Corsten e seu desempenho foi considerado superior ao do teste original.

A idéia básica de *bootstrap*, na ausência de qualquer conhecimento sobre a população, é realizar reamostragem com reposição de tamanho n da amostra original. A distribuição *bootstrap* de algum estimador de interesse é utilizada no lugar da distribuição teórica deste mesmo estimador, em função da dificuldade de desenvolvê-la ou do desconhecimento da distribuição da população de onde foi obtida a amostra aleatória.

Em estudos de desempenho de testes estatísticos, devido à dificuldade de se obter analiticamente informações sobre as taxas de erro tipo I e poder, a simulação Monte Carlo é uma alternativa viável para comparar os testes de comparações múltiplas.

O teste SNK é um teste que controla as taxas de erro tipo I por experimento sob H_0 mas se torna liberal sob H_0 parcial. Além disso, seu poder é superior aos testes de Tukey, t protegido de Bonferroni e Scheffé. Portanto, é um teste com boas qualidades mas que

podem ser melhoradas com o uso do *bootstrap*.

Assim, o objetivo desse projeto é propor uma versão *bootstrap* do teste de comparações múltiplas SNK, comparar suas taxas de erro tipo I com o teste SNK original e com outros testes com boas propriedades, tais como Scott-Knott e os testes de Caliński & Corsten.

O objetivo deste trabalho é propor uma classificação dos municípios brasileiros (sul de MG?) em relação aos seus gastos com previdência social e a algumas variáveis demográficas. O que se espera é que os grupos de municípios obtidos apresentem grande homogeneidade interna e grande heteogeneidade externa em relação às variáveis analisadas. Serão utilizados diferentes métodos de análise de agrupamento e os resultados obtidos serão comparados.

2 Revisão de Literatura

2.1 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, p.21).

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):

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \dots & X_{pp} \end{bmatrix}, \tag{1}$$

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; HOTHORN, 2011,p.2).

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 (FERREIRA, 2008, p.28).

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

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

Outras estatísticas descritivas, como a matriz de somas de quadrados e produtos e a matriz de correlações, podem ser consideradas, dependendo do objetivo da pesquisa (FERREIRA, 2008, p.28-29).

Segundo Mingoti (2005, p.21), 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 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.

2.2 Análise de componentes principais

O objetivo da técnica de análise de componentes principais (ACP) é explicar a estrutura de covariâncias das p variáveis, $\boldsymbol{X}^T = [X_1 \quad X_2 \quad \dots \quad X_p]$, por meio da construção de combinações lineares das variáveis originais. Os componentes principais são as p combinações lineares obtidas, $\boldsymbol{Y}^T = [Y_1 \quad Y_2 \quad \dots \quad Y_p]$, 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, 2005, p.59; EVERITT; HOTHORN, 2011, p.61).

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, p.63).

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, 2008, p.395-396).

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, p.66). Já outros autores, como Kathree; 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, p.64). 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, Ferreira (2008).

Ainda de acordo com os autores, 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; HOTHORN, 2011, p.64).

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, 2005, p.62).

De acordo com Everitt; Hothorn (2011, p.65), os primeiros k componentes, em que

$k < p$, explicam uma proporção da variância total,

$$\frac{\sum_{i=1}^{k} Var(Y_i)}{\text{Variância total de } X} = \frac{\sum_{i=1}^{k} \lambda_i}{\text{traço}(\boldsymbol{S})} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j}. \tag{2}$$

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, p.68).

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). 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, p. 89).

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, p.72). Mais detalhes sobre critérios podem ser vistos também em (KATHREE; 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, p.60). 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.

2.3 Análise de agrupamento (*Cluster Analysis*)

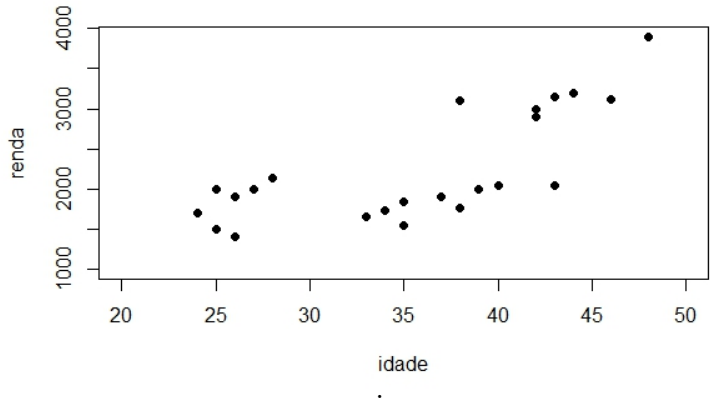
A técnica de análise de agrupamento (AA), também conhecida como análise de conglomerados, classificação ou *cluster analysis*, objetiva agrupar elementos da amostra de tal forma que (a) as observações de um mesmo grupo sejam similares entre si em relação às variáveis medidas e que (b) as observações de grupos diferentes sejam heterogêneas em relação a essas mesmas variáveis (MINGOTI, 2005, p.155).

"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, p.165).

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 (FERREIRA, 2008, p.341).

Figura 1: Diagrama de dispersão de dados fictícios de idade e renda de várias pessoas.



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

Segundo Ferreira (2008, p.342), 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). 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 o método 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, p.166).

Agrupamentos obtidos a partir de métodos hierárquicos, aglomerativos ou divisivos, podem ser representados por um diagrama bidimensional muito utilizado, o dendrograma. Esse gráfico ilustra as fusões ou divisões realizadas a cada passo do processo. Maiores detalhes podem ser obtidos em Everitt et al. (2011, p.72).

Ao se tentar realizar um agrupamento é muito importante saber quão próximas ou distantes estão as observações. Muitos métodos de agrupamento iniciam com uma matriz de distâncias $n \times n$ que refletem uma medida de similaridade ou dissimilaridade entre os

n elementos da amostra. Dois elementos são considerados próximos quando sua distância é pequena ou sua similaridade é grande (EVERITT et al., 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., 2011, p.19).

Há muitos tipos de distâncias que podem ser calculadas entre pares de observações, mas 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}.

Uma técnica de agrupamento que pode ser utilizada é *k-means*, que 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 minimize 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 $\bar{X}_j^{(l)} = \dfrac{1}{n_i} \sum_{i \in G_l} X_{ij}$ é a média dos indivíduos no grupo G_l em relação à variável j (EVERITT; HOTHORN, 2011, p.175).

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.

2.3.1 Número de grupos

A etapa final do processo de agrupamento 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, 2005).

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; HOTHORN, 2011, p.170).

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 sugeridas e alguns trabalhos avaliaram suas propriedades, tais como Milligan; Cooper (1985) e Dimitriadou et al. (2002).

O trabalho de Milligan; Cooper (1985) identificou como as duas melhores técnicas as propostas por Calinski and Harabasz (1974), também conhecida como pseudo F, e Duda and Hart (1973), denominada pseudo T^2. Everitt et al. (2011) mostra um resumo desses critérios e comenta sobre outros que podem auxiliar na decisão.

Outro método que também utiliza a matriz de dissimilaridade é o *silhouette plot* proposto por Kaufman; Rousseeuw (1990). For each object i they define an index , which compares object i?s separation from its cluster against the heterogeneity of the cluster

(for the exact definition of this index on the basis of the dissimilarities, see Kaufman and Rousseeuw, 1990). When has a value close to 1, the heterogeneity of object i?s cluster is much smaller than its separation and object i is taken as ?well classified?. Similarly, when is close to ?1 the opposite relationship applies and object i is taken to be ?misclassified?. When the index is near zero it is not clear whether the object should have been assigned to its current cluster or a neighbouring cluster. In the silhouette plot the are displayed as horizontal bars, ranked in decreasing order for each cluster (an examplewill be shown in the next section, see Figure 5.5). The silhouette plot is a means of assessing the quality of a cluster solution, enabling the investigator to identify ?poorly? classified objects and so distinguishing clear-cut clusters from weak ones. Silhouette plots for cluster solutions obtained from different choices for the number of groups can be compared, and the number of groups chosen so that the quality of the cluster solution is maximized. In this respect the average silhouette width ? the average of the over the entire data set ? can be maximized to provide a more formal criterion for selecting the number of groups. Kaufman and Rousseeuw (1990) also give some guidance as to the desirable size of the silhouette width; they consider a reasonable classification to be characterized by a silhouette width above 0.5 and point out that a small silhouette width, say an average width below 0.2, should be interpreted as a lack of substantial cluster structure.

A estatística GAP foi criada com o mesmo propósito ((Tibshirani et al., 2001)).

É crucial não utilizar 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, 2011).

Além disso, quando estudos socioeconômicos estão em questão, além dos métodos computacionais, a informação do pesquisador da área deve ser levada em conta para que se tenha uma interpretação útil dos grupos formados (CARVALHO et al., 2007).

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, três foram escolhidas e todas as que serão utilizadas estão na Tabela 1.

Tabela 1: Variáveis escolhidas a partir do Ministério da Previdência Social (MPS) e do Atlas Brasil.

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

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 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.

3.2 Análise de componentes principais (ACP)

A ACP será utilizada como primeiro passo do estudo, com o objetivo de reduzir a dimensionalidade dos dados, conforme explicitado na seção 2.2. Pretende-se reduzir o conjunto das XXXXX 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.

3.3 Análise de agrupamento (AA)

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 2.3, a AA oferece várias opções para a escolha da medida de distância, do método de agrupamento e do número de grupos.

Todas as rotinas necessárias para a análise dos dados serão realizadas utilizando o programa *R* em sua versão 3.2.0 (R CORE TEAM, 2015).

4 Resultados Esperados

O testes SNK original é um teste exato sob H_0 completa mas se torna liberal sob H_0 parcial. Espera-se que essa característica seja melhorada em sua versão *bootstrap* proposta neste trabalho.

Espera-se que o teste de comparações múltiplas SNK *bootstrap* controle as taxas de erro tipo I sob H_0 completa, ou seja, que elas se mantenham em torno do nível de significância α estabelecido e tenha valores altos de poder sob H_1. Sob H_0 parcial espera-se que as taxas de erro tipo se mantenham controladas dentro dos grupos de médias iguais e que o poder seja alto entre os grupos de médias diferentes. Se isso ocorrer haverá uma melhora do teste SNK original.

5 Cronograma

O projeto será concluído em 24 meses, no período de maio de 2013 a abril de 2015. 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.

7 Referências

Calinski, R. B. and Harabasz, J. (1974) A dendrite method for cluster analysis. Communica- tions in Statistics, 3, 1?27.

Dimitriadou, E., Dolnicar, S. and Weingessel, A. (2002) An examination of indexes for determining the number of clusters in binary data sets. Psychometrika, 67, 137?159.

Duda, R. O. and Hart, P. E. (1973) Pattern Classification and Scene Analysis. John Wiley & Sons, Inc., New York.

Milligan, G. W. and Cooper, M. C. (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50, 159?179.

BARTHOLOMEW, D. J.; STEELE, F; MOUSTAKI, I.; GALBRAITH, J. I. **Analysis of Multivariate Social Science Data**, 2008.

CAMARANO, A. A. Introdução. In: CAMARANO, A. A. **Novo regime demográfico**: uma nova relação entre população e desenvolvimento. Rio de Janeiro: Ipea, 2014, p.15-40.

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

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster Analysis**, UK: John Wiley & Sons, 5.ed., 2011.

FERREIRA, D. F. **Estatística Multivariada**. Lavras: UFLA, 2008.

KHATREE, R.; NAIK, D. N. **Multivariate Data Reduction and Discrimination with SAS Software**. SAS Institute Inc., 2000.

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

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria, 2015. Disponível em: <http://www.R-project.org/>. Acesso em: 01 agosto 2015.

REIS, P.; SILVEIRA, S.; BRAGA, M. 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.

SANTOS, J. P. C.; SILVA, K. M. G. C.; PEREIRA, S. B. M. Tipologia dos municípios bai-

anos com base em análise multivariada. **Textos para discussão - Superintendência de Estudos Econômicos e Sociais da Bahia**, n.2, 2011.

SOARES, S. S. D. Apresentação. In: CAMARANO, A. A. **Novo regime demográfico**: uma nova relação entre população e desenvolvimento. Rio de Janeiro: Ipea, 2014, p.1-2.

VASCONCELOS, A. M. N.; GOMES, M. M. F. Transição demográfica: a experiência brasileira. **Epidemiologia e Serviços de Saúde**, Brasília, v. 21, n. 4, p. 539-548, out./dez. 2012.

Patrícia de Siqueira Ramos

Alfenas,