

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE MATEMÁTICA
Curso de Bacharelado em Estatística

**ANÁLISE DA POPULAÇÃO ECONOMICAMENTE ATIVA NO BRASIL:
ESTUDO COMPARATIVO DOS SEXOS, NO PERÍODO DE 2005 A 2015**

Laíla Luana Campos

Uberlândia– MG

2016

Laíla Luana Campos

**ANÁLISE DA POPULAÇÃO ECONOMICAMENTE ATIVA NO BRASIL:
ESTUDO COMPARATIVO DOS SEXOS, NO PERÍODO DE 2005 A 2015**

Trabalho de Conclusão de Curso
apresentado ao Curso de Bacharelado em
Estatística da Universidade Federal de Uberlândia,
orientado pela Profª Drª Priscila Neves de Faria,
como requisito parcial para obtenção do título
técnico em Bacharelado em Estatística.

Uberlândia - MG

2016

**ANÁLISE DA POPULAÇÃO ECONOMICAMENTE ATIVA NO BRASIL:
ESTUDO COMPARATIVO DOS SEXOS NO PERÍODO DE 2005 A 2015**

Monografia aprovada para a obtenção do título de
Bacharel no Curso de Graduação em Estatística
da Universidade Federal de Uberlândia (MG)
pela banca examinadora formada por:

Uberlândia, 07 de julho de 2016



Prof. Dr. Lúcio Borges de Araújo, UFU/MG
Membro



Prof. Dr. Rogério de Melo Costa Pinto, UFU/MG
Membro



Profª Drª Priscila Neves Faria, UFU/MG
Orientador

Lista de abreviaturas e siglas

Análise de Variância (ANOVA)

Instituto Brasileiro de Geografia e Estatística (IBGE)

Instituto de Pesquisa Econômica Aplicada (IPEA)

População Economicamente Ativa (PEA)

População em Idade Ativa (PIA)

Pesquisa Mensal de Emprego (PME)

Pesquisa Nacional por Amostra de Domicílio (PNAD)

População Economicamente Inativa ou, População Não Economicamente Ativa (PNEA)

Programa das Nações Unidas para o Desenvolvimento (PNUD)

SUMÁRIO

RESUMO

ABSTRACT

1 INTRODUÇÃO	1
2 REFERENCIAL TEÓRICO	3
2.1. População Economicamente Ativa	3
2.1.1. As diferenças entre os sexos	5
2.1.2. Evolução da População Economicamente Ativa Feminina no Brasil	5
2.2. Experimentos Fatoriais	6
2.3. Testes de Comparações de Médias	7
2.3.1. Teste <i>t</i> -Student	7
2.3.2. ANOVA	8
2.3.3. Teste Tukey	9
2.4. Análise de Agrupamentos	9
2.4.1. Medidas de Similaridade e Dissimilaridade	11
2.4.2. Métodos Hierárquicos Aglomerativos	13
2.5. Número de clusters da partição final	17
3 METODOLOGIA	18
3.1. Testes de Comparação entre Médias	19
3.1.1. Teste <i>t</i>	19
3.1.2. ANOVA	20
3.1.3. Teste de Tukey	21
3.2. Análise de Agrupamentos	22
3.2.1. Medidas de Similaridade e Dissimilaridade	22
3.2.2. Métodos Hierárquicos e número de clusters da partição final	23
4 RESULTADOS	25
5 DISCUSSÕES	33
6 CONCLUSÃO	34
REFERÊNCIAS	35
ANEXO A: script utilizado no software R para análise dos dados	40

RESUMO

A População Economicamente Ativa (PEA) corresponde à faixa populacional responsável pela produção de riquezas a partir do trabalho e com remuneração especificamente voltada para o cumprimento de suas atividades. No presente estudo, foram utilizadas as técnicas de Delineamento Aleatorizado em Blocos em Esquema Fatorial, Análise de Variância e de Análise Multivariada de Agrupamentos Hierárquicos, aplicados a Distância Euclidiana e o Método de Ligação Média para a comparação entre os sexos masculino e feminino durante os anos de 2005 a 2015, em seis regiões metropolitanas, sendo elas: Belo Horizonte (MG), Rio de Janeiro (RJ), São Paulo (SP), Porto Alegre (RS), Salvador (BA) e Recife (PE). Dentre os resultados obtidos, observou-se que a diferença entre a quantidade de pessoas que compõem a PEA ainda é significativa entre os sexos.

Palavras-chaves: Agrupamentos, delineamento aleatorizado, regiões metropolitanas, mercado de trabalho.

ABSTRACT

The Economically Active Population (PEA) is the population group responsible for the production of wealth from the working and with the remuneration specifically geared to carry out its activities. In the present study, we used the Randomized Blocks in Factorial Scheme, Variance Analysis and the Multivariate Analysis of Hierarchical Cluster technique applied to Euclidean Distance and Average Connection Method for comparison between male and female during the years 2005 to 2015 in six metropolitan areas, which are: Belo Horizonte (MG), Rio de Janeiro (RJ), São Paulo (SP), Porto Alegre (RS), Salvador (Bahia) and Recife (PE). In one of the results, it was found that there is difference between the sexes in the labor force in metropolitan areas.

Keywords: Groupings, randomized blocks, metropolitan regions, the labour market.

1 INTRODUÇÃO

Todo país, seja subdesenvolvido ou desenvolvido, possui uma População Economicamente Ativa (PEA), que é uma parcela do contingente populacional e representa todas as pessoas que trabalham ou que estão procurando emprego. São essas pessoas que produzem para o país e que integram o sistema produtivo.

A População em Idade Ativa (PIA) é dividida em: PEA e População Não Economicamente Ativa ou mesmo inativa (PNEA). De acordo com Freitas (2016), no caso específico do Brasil, a PEA soma aproximadamente 79 milhões de pessoas ou 46,7%, índice muito baixo, uma vez que o restante da população, cerca de 53,3%, fica à mercê do sustento dos economicamente ativos. Em diversos países, o índice é superior, aproximadamente 75% atuam no setor produtivo. A utilidade das informações sobre a PEA está na obtenção de dados sobre o real nível de desemprego existente no país, ou seja, o grupo de pessoas que possui condições de trabalho, mas não consegue emprego. Além disso, é possível avaliar a respeito da menor presença da mulher no mercado de trabalho e a idade média de inclusão do jovem no meio trabalhista.

Nota-se que a PEA no Brasil vem sofrendo mudanças em relação aos sexos masculino e feminino. Nas últimas 6 décadas, apresentou um grande crescimento: a PEA total passou de 17,1 milhões de pessoas, em 1950, para quase 93,5 milhões de pessoas, em 2010 (aumento de 5,5 vezes). A PEA masculina passou de 14,6 milhões para 52,8 milhões (incremento de 3,6 vezes), enquanto a PEA feminina teve uma elevação extraordinária, passando de 2,5 milhões, em 1950, para 40,7 milhões, em 2010 (crescimento de 16,3 vezes). Parte desta expansão deve-se a mudanças metodológicas nos instrumentos de coleta do censo e das pesquisas domiciliares, mas a tendência de aumento de longo prazo é inegável e as mulheres são a principal força por detrás do crescimento da força de trabalho (ALVES, 2013).

É uma tarefa difícil fazer projeções de longo prazo dos estudos relacionados à PEA no Brasil, esse não é o tipo de tarefa que se pode resolver, exclusivamente, através do uso de pacotes estatísticos a fim de ajustar curvas; é necessário que se concebam diferentes cenários associados ao conhecimento da área (WAJNMAN & RTOS-NETO, 1994). O mais importante é que a pesquisa trabalhe com dados atuais e, principalmente, aplicando no estudo da PEA uma metodologia distinta das convencionais. Não foram encontrados muitos trabalhos publicados com aplicações de estatística multivariada na PEA até o momento, sendo que, o único, utiliza análise de cluster e compara a PEA com

o PIB nos municípios gaúchos (KOETZ et al., 2004). Os demais artigos relacionados à PEA utilizam Estatística Descritiva (HOFFMANN & MENDONÇA, 2003), Regressão (MENEZES FILHO & SORZAFÁVE, 2000) e decomposição do Índice de Gini para avaliar a contribuição do rendimento das mulheres na determinação da desigualdade do rendimento domiciliar per capita no Brasil (HOFFMANN & LEONE, 2004).

Alguns estudos atuais abordando a análise da PEA em relação ao gênero trataram de fazer comparações. Lameiras (2013) concluiu que ao longo de poucos meses no ano de 2013 a PEA feminina mostrou uma taxa de crescimento em aceleração, configurando uma trajetória oposta à observada pela PEA masculina. Leone & Baltar (2008) analisaram entre 2004 e 2006 as tendências do mercado de trabalho destacando o crescimento da PEA feminina que, em seus resultados, teve um aumento maior que a masculina neste período. No entanto, mostram os autores que a segregação das mulheres no mercado de trabalho continua forte e que as diferenças de rendimento entre homens e mulheres são maiores para níveis de instrução mais elevados. Mais recentemente, Lavinhas & Cordilha (2014), estudando assimetrias no mercado de trabalho, estabeleceram um modelo relacionando diferenciais de renda, de horas trabalhadas e de permanência no emprego, com diversos fatores econômicos. Entre os resultados obtidos, mostrou-se que o mercado de trabalho é mais receptivo para mulheres quando os salários são mais baixos, mas que, por outro lado, os diferenciais de salários entre homens e mulheres são reduzidos quando há maior estabilidade no emprego.

Devido ao grande aumento da participação feminina no mercado de trabalho, é importante a comparação com o sexo masculino. A Análise de Agrupamentos, metodologia da Estatística Multivariada, além de avaliar semelhanças entre indivíduos em relação às variáveis em estudo, também tem sido frequentemente utilizada avaliando a similaridade entre regiões em relação a um período (anos, meses, etc). Em climatologia, isto foi verificado por Chierice Junior & Landim (2014), que classificaram os anos em grupos discretos, com características semelhantes, usando o Método Aglomerativo Hierárquico, a medida de Distância Euclidiana e o Método de Ligação de Ward; Uvo (2003) estudou a precipitação ao longo do norte da Europa, a qual foi representada por totais de precipitação mensais de 125 estações meteorológicas distribuídas por Noruega, Suécia, Finlândia, Dinamarca, Rússia ocidental e os Estados Bálticos, sendo que os dados abrangeram o período de 1967 a 1996; Keller- Filho et al. (2005) utilizou o agrupamento das observações pluviais segundo um critério de similaridade do comportamento probabilístico da precipitação pluvial, em sua evolução

ao longo de um período anual e Comunello et al. (2013), utilizou a análise de agrupamentos para avaliar se regiões climáticas eram homogêneas através de dados anuais, inclusive aqueles oriundos de anos bissextos, padronizados em 36 períodos, cada qual considerado como uma variável classificatória. Em economia, Catela (2013) verificou que a estratificação dos municípios por nível de renda per capita, assim como por salário médio recebido pelos trabalhadores, é necessária para se analisar a dinâmica transacional das cidades entre tais estratos e também para determinar os condicionantes dessa transição relativa a um período (neste caso, em municípios fluminenses, de 1998 a 2008).

A análise de grupos ou de cluster é uma técnica exploratória de análise multivariada que permite agrupar indivíduos ou variáveis em grupos homogêneos ou compactos relativamente a uma ou mais características comuns. Dessa forma, cada observação pertencente a um determinado cluster é similar a todas as outras pertencentes a esse cluster, e é diferente das observações pertencentes a outros clusters (MANTOVANI, 2006). Desta forma, existe a possibilidade que as variáveis em estudo sejam substituídas por anos e que, portanto, assim como é possível agrupar as variáveis, estes também possam ser agrupados.

Diante do exposto, o propósito do presente estudo é contribuir para as pesquisas envolvendo a comparação e a evolução da participação de cada sexo no mercado de trabalho entre os anos de 2005 e 2015, abrangendo seis das mais importantes regiões metropolitanas do Brasil: Belo Horizonte (MG), Rio de Janeiro (RJ), São Paulo (SP), Porto Alegre (RS), Salvador (BA) e Recife (PE).

2 REFERENCIAL TEÓRICO

2.1. População Economicamente Ativa

A PEA é um conceito elaborado para designar a população que está inserida no mercado de trabalho ou que, de certa forma, está procurando se inserir nele para exercer algum tipo de atividade remunerada (PENA, 2016)¹. Corresponde à faixa populacional

¹PENA, R. F. A. "**População Economicamente Ativa – PEA**"; *Brasil Escola*. Disponível em <<http://brasilecola.uol.com.br/geografia/populacao-economicamente-ativa-pea.htm>>. Acesso em 19 de março de 2016.

responsável pela produção de riquezas a partir do trabalho, com remuneração especificamente voltada para o cumprimento de suas atividades (PENA, 2016)². São essas pessoas que produzem para o país e que integram o sistema produtivo (FREITAS, 2016). A PEA envolve o que o Instituto Brasileiro de Geografia e Estatística (IBGE) classifica como população ocupada e população desocupada, é a população empregada ou que possui condições de trabalhar e que realiza algum esforço para isso (PENA, 2016)³. Em contrapartida à PEA, tem-se a PNEA, correspondendo a crianças, idosos e aqueles aposentados por algum tipo de invalidez ou limitação, bem como a faixa da população que não deseja procurar emprego, mesmo dispondo de todas as condições para tal (PENA, 2016)².

Não há um método consolidado em todo o mundo para definir aqueles que fazem parte da PEA, existindo uma discussão, no plano internacional, sobre a faixa de idade que a comporia. Nos países subdesenvolvidos, o índice engloba a população que possui de 10 a 60 anos, e já nos países desenvolvidos, geralmente, considera-se apenas aquela que possui mais de 15 anos de idade (PENA, 2016) ¹. No Brasil e também em muitos outros países, a idade mínima é de 15 anos (PENA, 2016) ³.

A PEA brasileira é um índice muito baixo, uma vez que o restante da população fica à mercê do sustento dos economicamente ativos, apesar de não considerar aqueles que não trabalham com contrato formal ou carteira assinada (PENA, 2016) ³. Em diversos países, o índice é superior, aproximadamente 75% atuam no setor produtivo. Quanto à participação de cada sexo no mercado de trabalho, os homens representam 58% e as mulheres 42% daqueles que desenvolvem atividades em distintos setores da economia (FREITAS, 2016). Destes, 66,4% dos habitantes do sexo masculino do Brasil compõem a PEA, e 49% dos habitantes do sexo feminino, conforme aponta o Instituto de Pesquisa Econômica Aplicada (IPEA) (PENA, 2016) ². Devido ao grande aumento da participação feminina no mercado de trabalho, é importante a comparação como tem

²PENA, R. F. A. **“População Economicamente Ativa – PEA”**. Disponível em: <<http://alunosonline.uol.com.br/geografia/populacao-economicamente-ativa-pea.html>> Acesso em 19 de março de 2016.

³PENA, R. F. A. **“O Conceito de População Economicamente Ativa (PEA)”**. Disponível em <<http://mundoeducacao.bol.uol.com.br/geografia/o-conceito-populacao-economicamente-ativa-pea.htm>>. Acesso em 19 de março de 2016.

se dado essa ocorrência se comparar com o sexo masculino, abrangendo a pesquisa para os vinte e sete estados da federação.

2.1.1. As diferenças da PEA entre os sexos

Ao final do século XX, a força de trabalho feminina cresceu expressivamente em todo o mundo, sendo que o Brasil foi um dos países em que se pode observar uma das maiores taxas de crescimento: de 179,4 % entre 1970 e 1990, de acordo com a FEE (2003). A esse crescimento corresponde o ingresso cada vez maior de mulheres nas mais variadas atividades econômicas, sem que se veja uma alteração significativa nos padrões de desigualdade vigentes na valorização do trabalho feminino em comparação ao trabalho do homem (KOETZ et al., 2004). Segundo Hirata (1998), ainda hoje cabem às mulheres as ocupações de mais baixo *status*, com menores possibilidades de ascender profissionalmente, bem como remunerações menores, além de serem demitidas com frequências maiores e permanecerem mais tempo desempregadas que os homens.

Em 2015, o *Gender Gap Index*, do Fórum Econômico Mundial, numa das medidas utilizadas para se inferirem as diferenças entre gêneros nos diferentes países, o Brasil ocupa a 85ª posição em um ranking de 145 países, com a nota 0,686. Nesse ranking, a nota zero representa a completa igualdade entre gêneros e um, a completa desigualdade. Dessa forma, o Brasil se encontra na parcela de países que trata de forma bastante diferenciada homens e mulheres.

2.1.2. Evolução da População Economicamente Ativa Feminina no Brasil

Contemplar a evolução histórica da trajetória pessoal e profissional feminina é importante, quando se pensa na igualdade de direitos na sociedade (SCHLICKMANN & PIZARRO, 2012). Segundo Aquilini & Costa (2003), a entrada da mulher no mercado de trabalho brasileiro foi tardia, assim como em outros países. No Brasil, a Constituição Federal autorizou a instituição da cidadania e dos direitos humanos para as mulheres brasileiras só em 1988 (ALTTIMAN & COSTA, 2009). De acordo com Espíndola (2008), a década de 90 foi importante para o fortalecimento das mulheres no mercado de trabalho e, segundo Bruschini (2007), nas últimas décadas do século XX, o Brasil passou por importantes mudanças sociodemográficas. As transformações nos padrões de comportamento e no papel da mulher na sociedade foram fortalecidas pelos

movimentos feministas, fazendo com que mais mulheres atuassem na vida pública, além de que o aumento da escolaridade e a facilidade para ingressar nas universidades possibilitaram o acesso das mulheres a novas oportunidades. Tudo isto facilitou a entrada da mulher no mercado de trabalho, mas, no entanto, com salários ainda menores que o dos homens, ocupando cargos inferiores e com menos responsabilidades (SCHLICKMANN & PIZARRO, 2012).

Desde o início da década de 2000, o crescimento da participação feminina no mercado de trabalho chama atenção pela alavancagem, sendo, em diversos aspectos, até mesmo maior que a taxa masculina, o que indica uma diminuição gradual da diferença de gênero no que tange à ocupação. No início do século XXI, as mulheres ocupam postos no topo das grandes empresas, comandam tropas e governam países (SIMÕES & HASHIMOTO, 2012), porém, segundo o Programa das Nações Unidas para o Desenvolvimento (PNUD, 2009), em relação à renda, no Brasil, em 2008, elas receberam pouco mais da metade do salário dos homens, apesar de terem maior escolarização.

2.2. Experimentos Fatoriais

Experimentos fatoriais são realizados quando se deseja analisar dois ou mais tipos de tratamentos no mesmo experimento, sendo os tratamentos a combinação dos fatores. Vários fatores são testados ao mesmo tempo sobre uma variável resposta, como por exemplo, a temperatura, a concentração, as máquinas, os extratos, os operadores, etc.

Estes experimentos são úteis no estudo da interação entre os fatores. O número de fatores a ser estudado deve ser pequeno, sob pena de não ser possível entender o resultado da análise. Até três fatores, pode-se representar através de um gráfico tridimensional, mas, a partir de 4 fatores se torna difícil a interpretação da interação.

As diferentes categorias de um mesmo tratamento são chamadas níveis, e o número destes não complica o experimento. O estudo de experimentos fatoriais é útil quando nele se pode estudar as interações, portanto deve-se ter repetições. Os tratamentos são aplicados às unidades experimentais, enquanto que os blocos são apenas uma estratificação dessas unidades. Mas, às vezes pode-se encarar possíveis blocos como fatores.

O tipo mais simples é o Fatorial 2×2 , isto é, quando dois fatores são comparados, cada um com dois níveis de experimentação. Assim, um fatorial 3×3 significa que 2 tratamentos estão sendo experimentados com 3 níveis cada um - o que pode ser visto como 9 comparações. Já no fatorial $2 \times 2 \times 2$ temos 3 fatores em comparação, cada um com dois níveis. Se o número de níveis é igual para todos os fatores, os experimentos fatoriais também são referidos como 2^2 , 3^2 , 2^3 , etc. O expoente indica o número de fatores e a base indica o número de níveis. Também podem ser feitos experimentos 3×4 , (2 fatores, um com três níveis e outro com quatro níveis), 2×3 , e assim por diante.

Esses experimentos podem ser planejados em blocos, isto quando as unidades experimentais não forem homogêneas. Nesse caso, primeiro se agrupam as parcelas similares, que vão formar os blocos; depois os tratamentos são sorteados dentro dos blocos.

2.3. Testes de Comparações de Médias

Quando há a necessidade de se testar a significância das diferenças entre duas ou mais médias amostrais, ou seja, testar a hipótese nula de que as médias amostrais são todas iguais entre si, pode-se utilizar alguns Testes de Hipóteses. Existem vários Testes com esta finalidade, entre eles estão o Teste Z, o Teste t de Student e a Análise de Variância.

2.3.1. Teste t-Student

A aplicação do Teste t é indicada quando o tamanho amostral é igual ou inferior a 30 elementos. Para amostras com tamanho superior a 30, recomenda-se o teste Z. O uso do Teste t pressupõe que a característica em análise é normalmente distribuída com variância populacional desconhecida. Pode ser utilizado para comparar duas médias, sejam elas vindas de uma amostra e de uma população, ou de duas amostras pareadas ou mesmo de duas amostras independentes. No Teste t de Student para duas amostras independentes, testa-se uma variável difere entre dois grupos independentes de sujeitos.

O Teste t pode servir para testar contrastes envolvendo duas ou mais médias, porém este teste exige que as comparações a serem realizadas sejam escolhidas *a priori*, ou seja, antes de serem examinados os dados; podem-se testar no máximo, tantos contrastes quantos são os graus de liberdade para tratamentos, e estes contrastes devem

ser ortogonais. A ortogonalidade entre os contrastes indica independência linear na comparação estabelecida por um contraste com a comparação estabelecida pelos outros contrastes. Entre I médias de um fator, podem ser obtidos $I - 1$ contrastes ortogonais.

O nível de significância α é válido para um único contraste, e não para uma série deles; e somente se o contraste for estabelecido *a priori* e não sugerido pelos dados, pois, pode ficar caracterizado uma estatística de ordem ao querer comparar a maior com a menor média, o que acarretaria certa dependência entre as médias.

2.3.2. ANOVA

A Análise de Variância (*Analysis of Variance* - ANOVA) é uma metodologia utilizada para comparar três ou mais tratamentos. Devore (2006) definiu ANOVA como sendo um teste que contempla um conjunto de situações experimentais e procedimentos estatísticos para a análise de respostas quantitativas de unidades experimentais. Segundo Freund & Simon (2000), a Análise da Variância expressa uma medida da variação total em um conjunto de dados, como uma soma de termos, cada um dos quais é atribuído a uma fonte ou causa específica. Existem algumas considerações importantes a serem feitas antes de aplicar a ANOVA: as amostras analisadas devem ser escolhidas de maneira aleatória, os resíduos devem ser independentes, seguir uma distribuição normal e possuir variâncias homogêneas.

O Teste F da ANOVA é um procedimento robusto sendo que o que importa é a normalidade das médias amostrais; assim, a ANOVA torna-se mais segura à medida que os tamanhos das amostras aumentam, devido ao efeito do Teorema Central do Limite. A ANOVA considera que a variabilidade das observações, medida pelo desvio padrão, seja a mesma em todas as populações. Não é fácil verificar a suposição de as populações terem desvios-padrão iguais. Testes estatísticos de igualdade dos desvios-padrão são tão sensíveis à ausência de normalidade que, na prática, têm pouco valor. A solução é contar com a robustez da ANOVA. A ANOVA não é muito sensível a violações da suposição, particularmente quando todas as amostras têm tamanhos iguais ou semelhantes e nenhuma das amostras é muito pequena. Ao planejar um estudo, é importante tomar amostras do mesmo tamanho de todos os grupos que pretende comparar. Os desvios-padrão amostrais estimam os desvios-padrão da população, logo, deve-se verificar antes de fazer a ANOVA de que os desvios-padrão amostrais são semelhantes entre si. Espera-se que haja certa variação entre eles devido ao acaso.

De acordo com Downing & Clark (2000), o teste ANOVA é aplicado para testar se vários grupos de observações provêm de distribuições com a mesma média; na análise de variância calcula-se uma estatística F; se a hipótese nula é verdadeira e as médias de todos os grupos são realmente iguais, então a estatística F terá distribuição F e; se a estatística F calculada é superior ao valor crítico, então se rejeita a hipótese nula, de que as médias dos grupos sejam as mesmas. A razão (ou distribuição) F trata de um modelo de distribuição contínua útil para inferências estatísticas e é dada pelo quociente entre duas variáveis aleatórias independentes com distribuições qui-quadrado (FONSECA & MARTINS, 1996). A probabilidade da cauda (valor-p), isto é, a probabilidade de obter-se um valor, no mínimo, igual ao valor F observado quando a hipótese nula é verdadeira (FREUND & SIMON, 2000).

2.3.3. Teste de Tukey

Tukey (1953) desenvolveu um processo onde o experimento é usado como unidade no estabelecimento do nível de significância, ou seja, para um conjunto de médias verdadeiras iguais, um teste ao nível de 5% é aquele em que 5% dos experimentos darão uma ou mais diferenças significativas falsas e 95% darão diferenças não significativas, em média. É a taxa de erro tipo I por experimento.

O Teste de Tukey pode ser utilizado para comparar a totalidade dos contrastes entre duas médias, ou seja, para os $I(I-1)/2$ contrastes do tipo $C = m_i - m_u$; para $1 \leq i \leq u \leq I$, em que I é o número de níveis do fator em estudo. Este teste baseia-se na diferença mínima significativa (d.m.s.) representada por Δ .

O Teste de Tukey é válido para a totalidade dos contrastes de duas médias. Ele exige, em princípio, balanceamento. Mas, no caso dos tratamentos apresentarem números de repetições diferentes, o resultado obtido por este teste é apenas uma aproximação. É exato para testar a maior diferença, nos demais casos é conservador.

2.4. Análise de Agrupamentos

A Análise de Agrupamentos, também denominada *Cluster Analysis*, possui uma variedade de nomes que decorre em parte do uso de métodos de agrupamentos nas mais diversas áreas, como psicologia, biologia, sociologia, economia, engenharia, administração, entre outras. Bevilacqua (2013) utilizou metodologia da Estatística

Multivariada para estudar segmentações em marketing, e Reis et al. (2012) aplicou a Análise de Agrupamentos a fim de identificar e avaliar impactos socioeconômicos.

O método de análise de agrupamentos é uma técnica multivariada cuja finalidade primária é classificar objetos (indivíduos, empresas, produtos, etc), numa estrutura natural (Hair et al., 2005; Mingoti, 2005; Lattin et al., 2011; Johnson & Wichern, 1992; Cruz & Regazzi, 1994). Assim, cada objeto é muito semelhante aos outros no agrupamento em relação a algum critério de seleção predeterminado.

Quando não existe nenhum tipo de informação pré-existente, não há uma classificação “ótima” para um conjunto de dados, os métodos de *clustering* servem como uma ferramenta para exploração de dados que, se usados de forma adequada, podem particionar os dados sem o efeito de quaisquer preconceitos e noções arbitrárias (LINDEN, 2009). Como afirmam Johnson & Wichern (1992), o método corresponde também a uma importante técnica exploratória que busca identificar uma estrutura de agrupamentos com o intuito de avaliar a dimensionalidade dos dados, identificar *outliers* e fornecer interessantes hipóteses acerca de associações. De acordo com Hair et al. (1995), citado por Souza (2000), o critério essencial é a maximização das diferenças entre os grupos em relação à variação dentro dos grupos.

Monta-se uma matriz de dados das características de interesse, e, para verificar se a unidade A é mais parecida com B do que com C são construídas estimativas de distância. A escolha do método para o agrupamento exige o conhecimento de suas propriedades e dos objetivos da pesquisa, sendo que, entre as técnicas utilizadas, as mais comuns são as hierárquicas aglomerativas, nas quais classificam-se as unidades com sucessivas fusões em grupos, reduzindo a um único grupo ao final. As unidades são alocadas em uma matriz de distância, e iniciam-se as fusões com base nas proximidades. Os resultados são resumidos em um dendrograma, que corresponde a um gráfico bidimensional combinando a ocorrência da fusão com a estimativa de distância das unidades agrupadas.

Diante das diferentes técnicas existentes, determinando diversas medidas de distância e critérios de agrupamento, podem ser empregadas e comparadas diversas técnicas, devendo ser escolhida a que melhor se adeque aos dados. As distâncias podem ser obtidas pelo quadrado das distâncias euclidianas, potencializando as diferenças, ou como foi realizado por Hoffmann (1992), atribuindo-se ponderações aos escores fatoriais com base na contribuição destes para explicar a variância.

2.4.1. Medidas de Similaridade e Dissimilaridade

Quando se usam métodos de agrupamentos hierárquicos aglomerativos, é preciso agrupar os dados com base na proximidade mútua ou na similaridade. Segundo Mingoti (2005), é necessário considerar medidas que descrevam a similaridade entre elementos amostrais de acordo com as características que neles foram medidas. Se, para cada elemento amostral, têm-se informações de p variáveis armazenadas em um vetor, a comparação de diferentes elementos amostrais pode ser feita através de medidas matemáticas (métricas), que possibilitam a comparação de vetores, de observações dos elementos amostrais e agrupam aqueles de menor distância. Quando as variáveis em estudo possuem propriedades métricas e/ou quantitativas, um modo óbvio de se refletir a proximidade de dois objetos é com uma medida de distância. Ainda de acordo com Mingoti (2005), em outras situações, a pesquisa envolve a análise de variáveis qualitativas, e, nestes casos, há duas formas de tratamento: ou transformam-nas em variáveis quantitativas e usam-se medidas de similaridades, ou trabalha-se com coeficientes de similaridades e, neste caso, comparam-se os elementos de acordo com a presença ou ausência de certas características, que podem ser representadas por variáveis binárias do tipo 0 ou 1.

Se as variáveis são quantitativas, para que se possa proceder ao agrupamento dos elementos, é necessário que se decida a priori a medida de similaridade ou dissimilaridade que será utilizada, pois existem várias diferentes, e cada uma delas produz um determinado tipo de agrupamento. De acordo com Mingoti (2005), as medidas de distâncias são medidas de dissimilaridade e, logo, quanto menor os seus valores, mais similares serão os elementos que estão sendo comparados. Segundo a mesma autora, considerando um conjunto de dados de n elementos amostrais, tendo-se medido p -variáveis aleatórias em cada um deles, para cada elemento amostral j , tem-se o vetor de medidas X_j , definido por:

$$X_j = [X_{1j}, X_{2j}, \dots, X_{pj}]', \text{ sendo } j = 1, 2, \dots, n.$$

Estabelecida a medida de parença, é construída a matriz de distâncias, que é uma matriz quadrada com o valor das medidas (similaridade ou dissimilaridade) entre todos os elementos. Uma ilustração da matriz é apresentada na Figura 1. É importante

destacar que qualquer medida de distância usada para variáveis quantitativas pode ser transformada num coeficiente de similaridade.

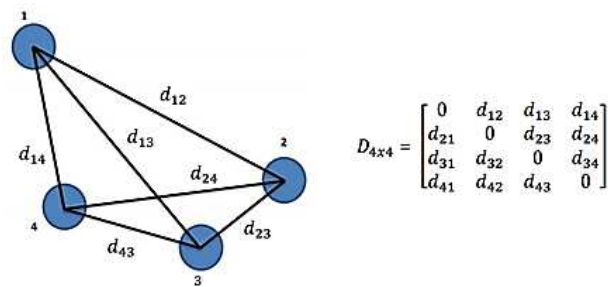


Figura 1: Matriz de distâncias entre 4 elementos.

Fonte: MINGOTI, S.A., 2005.

A distância euclidiana não preserva a ordem das distâncias com a mudança de escala, e, como as variáveis X são frequentemente medidas em unidades diferentes, é comum proceder-se à padronização das variáveis antes de se obter o valor das distâncias. Assim, atribui-se peso igual a cada variável padronizada, de maneira que elas sejam igualmente importantes, após a padronização, na determinação da proximidade relativa dos objetos no espaço. A transformação de dados mais difundida é a padronização das variáveis, de acordo com a fórmula $Z_j = \frac{X_i}{\hat{\sigma}_j}$, em que $\hat{\sigma}_j$ é o desvio padrão associado à j -ésima característica.

As medidas definidas pela métrica euclidiana (ou qualquer uma das métricas de Minkowhki) não levam em considerações quaisquer padrões de covariância que existam nos dados. A métrica proposta por Mahalanobis é ajustada pela covariância, e capta se os pontos A e B estão no mesmo contorno de isodensidade. Uma vantagem do D^2 de Mahalanobis é que ele efetua uma rotação e reescala os eixos da coordenada de tal modo que a distribuição multivariada resultante possua matriz de covariância igual à matriz identidade I . Isto é útil quando o objetivo da análise de agrupamentos for o de gerar agrupamentos compactos e convexos.

Um coeficiente de correlação nem sempre é uma medida apropriada de similaridade, ele é uma medida de covariância que é também um tipo de proximidade, mas não necessariamente de similaridade. Quando as variáveis são quantitativas, uma medida de relacionamento natural é o coeficiente de correlação de Pearson, o qual expressa a similaridade de duas variáveis relativas à relação linear. Quanto maior for o

valor do coeficiente de correlação em valor absoluto, maior será a proximidade entre as variáveis, isto é, mais linearmente relacionadas elas estarão.

Segundo Bussab et al. (1990), a Correlação Cofenética equivale à correlação de Pearson entre a matriz de similaridade original e aquela obtida após a construção do dendrograma, sendo que, quanto mais próxima de 1, menor será a distorção provocada pelo agrupamento dos indivíduos. Landim (2000) afirma que, como o dendrograma é uma simplificação em duas dimensões de uma relação n -dimensional, é inevitável que algumas distorções quanto à similaridade apareçam. A medida de tal distorção pode ser obtida pelo Coeficiente de Correlação Cofenético entre os valores da matriz inicial de similaridade e aqueles derivados do dendrograma.

Sprent (1993), citado por Mingoti (2005), diz que Coeficientes de Associação Não Paramétricos como o de Spearman e o de Kendall também podem ser utilizados para o agrupamento das variáveis quantitativas. Existem algumas medidas de similaridade específicas para variáveis categóricas, como o coeficiente qui-quadrado, o de contingência de Pearson e o de concordância de Kappa.

2.4.2. Métodos Hierárquicos Aglomerativos

A análise de cluster é útil na classificação de elementos e em situações nas quais se deseja reduzir a dimensionalidade da amostra ou o número de variáveis a serem avaliadas. De acordo com Mingoti (2005), uma análise mais elaborada de cada conglomerado é mais informativa do que uma análise do conjunto de n observações como um todo, mas nada impede que outras técnicas da estatística multivariada sejam utilizadas na sumarização da informação de cada conglomerado obtido na partição. Na análise de cluster é possível utilizar todas as n observações amostrais, mesmo se incompletas, ao contrário das outras técnicas de estatística multivariada. Os elementos amostrais podem ser comparados de modo ponderado, nas variáveis para as quais existem informações completas. Os métodos de agrupamentos são sensíveis à presença de valores discrepantes (*outliers*), portanto, é fundamental que uma análise exploratória dos dados seja feita antes para verificar a existência de observações desse tipo.

As técnicas de agrupamentos são classificadas em hierárquicas (aglomerativas ou divisivas) e não hierárquicas. Os Métodos Hierárquicos são, na maioria das vezes, utilizados em análise exploratória dos dados com o intuito de identificar os possíveis

agrupamentos e o valor provável do número deles. Já para o uso dos Métodos Não Hierárquicos, é necessário que o valor do número de grupos seja pré-especificado.

De acordo com Mingoti (2005), as técnicas hierárquicas aglomerativas partem do princípio de que no início do processo de agrupamento cada elemento do conjunto de dados observado é considerado como um conglomerado isolado, assim, tem-se n conglomerados. Em cada estágio do algoritmo, os elementos mais similares vão sendo combinados e passam a constituir um único agrupamento. Cada novo *cluster* formado é um agrupamento de conglomerados formados nos estágios anteriores, até que todos os elementos considerados pertençam a um único grupo, diminuindo o número de conglomerados a cada estágio do processo, o que é observado na Figura 2. Se dois elementos aparecem juntos num mesmo cluster em algum estágio do processo de agrupamento, eles permanecerão juntos em todos os estágios subsequentes. Os Métodos Hierárquicos Aglomerativos de Análise de *Cluster* também podem ser úteis na seleção de variáveis mais importantes para caracterizar um determinado tipo de situação.

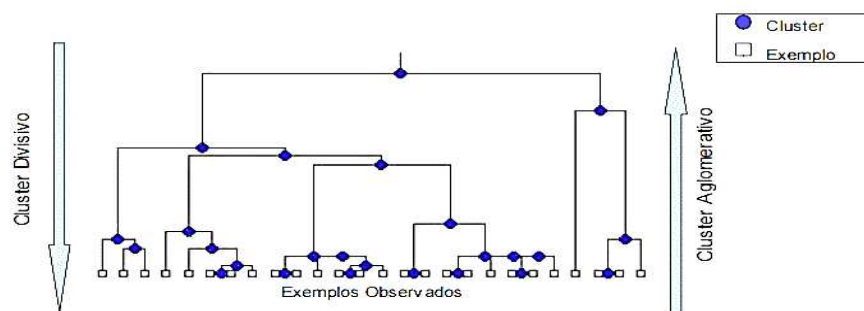


Figura 2: Modelo de um dendrograma, seus componentes e diferença entre Clusters Hierárquicos Aglomerativo e Divisivo.

Fonte: FREITAS, 2006.

Ainda segundo Mingoti (2005) e também Lattin et al. (2011), devido à propriedade de hierarquia, é possível construir um gráfico chamado Dendrograma, o qual é um diagrama bidimensional em forma de árvore e representa a história de agrupamento, por ser gerado pela sequência iterativa descrita anteriormente. O eixo vertical indica o nível de similaridade ou dissimilaridade e no eixo horizontal são marcados os elementos amostrais numa ordem conveniente relacionada à história de agrupamento. As linhas verticais, partindo dos elementos amostrais agrupados, têm altura correspondente à distância do agrupamento ou ao nível de similaridade.

Dentre os vários Métodos de Agrupamentos Hierárquicos Aglomerativos, o Método da Ligação Simples ou Método do Vizinho Mais Próximo (*Single Linkage*) utiliza a distância mínima entre dois objetos de um conjunto n de grupos distintos como sendo a distância entre os grupos. O próximo grupo é representado pela menor distância entre o primeiro grupo determinado e o objeto mais próximo a este. Os passos seguem até o encadeamento de todos os objetos em um único agrupamento, este com diferentes arranjos de objetos em um dado nível da escala de distâncias. Segundo Lattin et al. (2011), uma desvantagem da Ligação Simples é sua tendência a produzir agrupamentos longos e escadeados, com formatos não convexos, devido ao fato de adicionar um objeto a um agrupamento desde que esteja próximo a qualquer um dos outros objetos do agrupamento, mesmo que esteja relativamente longe de todos os outros. Se os verdadeiros agrupamentos subjacentes forem não convexos, essa propriedade não será necessariamente ruim; no entanto, na maioria dos casos, as modas que ocorrem naturalmente tendem a ser convexas e compactadas.

Já o Método da Ligação Completa ou Método do Vizinho Mais Distante (*Complete Linkage*) é exatamente oposto ao da Ligação Simples, em que no primeiro passo considera-se a distância entre dois grupos como sendo entre os objetos de maior distância, estes definindo grupos polarizados. Com a redução das distâncias entre os grupos e objetos, estes passam a formar agrupamentos com menor distância, encadeando-se, o que assegura que cada objeto adicionado ao agrupamento esteja próximo de todos os objetos no agrupamento, e não somente de um. De acordo com Lattin et al. (2011), uma desvantagem do Método da Ligação Completa é que ela pode ser altamente sensível a discrepâncias nos dados.

O Método da Ligação Média (*Average Linkage*), também conhecida como UPGMA, trata de uma variação dos procedimentos descritos anteriormente, sendo que, neste, segundo Mingoti (2005), a distância entre dois grupos é representada pela média da distância entre todos os pares de objetos pertencentes a cada grupo. De acordo com Lattin et al. (2011), neste método chega-se mais perto de um ajuste de árvore que satisfaça o critério de minimização dos Mínimos Quadrados, e o dendrograma é estabelecido pelos indivíduos com maior similaridade. Se os dados forem não métricos, a média pode ser substituída pela mediana, e, neste caso, o método é chamado de Ligação Mediana. Na Figura 3, tem-se um esquema do que ocorre:

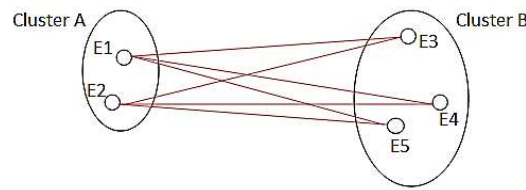


Figura 3: Gráfico das distâncias médias entre dois clusters.

Fonte: METZ & MONARD (2006).

em que

$$d_{AB} = \frac{(dE_1, E_3 + dE_1, E_4 + dE_1, E_5 + dE_2, E_3 + dE_2, E_4 + dE_2, E_5)}{6}$$

O Método do Centróide utiliza a distância entre os vetores de médias (chamados de centroides) dos grupos que estão sendo comparados. De acordo com Mingoti (2005), em cada etapa do agrupamento é necessário voltar aos dados originais para o cálculo da matriz de distâncias, sendo agrupados os conglomerados que apresentam o menor valor de distância. É robusto para discrepâncias, mas pode ser superado pela Ligação Média.

O Método de Ward, também chamado de Método de Variância Mínima, é baseado na redução da informação resultante, dada a inclusão de um conjunto de objetos em um grupo. Segundo Lattin et al. (2011), esta redução de informação é determinada pela Soma Total do Quadrado do Erro de cada objeto, em função da média do grupo a que este, supostamente, pertença. Esta regra de inclusão envolve todos os pares possíveis, sendo definidos como pertencente a um dado grupo o objeto que contribua o mínimo com o aumento da Soma de Quadrado do Erro, isto é, a variância mínima dentro do grupo. Tende a produzir agrupamentos de tamanhos iguais, convexos e compactos.

Existe uma semelhança entre o Método de Ward e o Método do Centróide no que se refere ao uso dos vetores de médias amostrais como representante da informação global dos conglomerados em cada passo do processo de agrupamento. No entanto, a distância usada no Método de Ward leva em consideração a diferença dos tamanhos dos conglomerados que estão sendo comparados, enquanto que o Método dos Centróides não possui qualquer fator de ponderação em relação a isso (MINGOTI, 2005).

Vale ressaltar que os Métodos de Ligação Simples, Completa e da Média podem ser utilizados tanto para variáveis quantitativas, quanto qualitativas, ao contrário dos Métodos do Centróide e de Ward, que são apropriados apenas para variáveis quantitativas, já que têm como base a comparação de vetores de médias. Espera-se que

haja certa consistência entre as soluções obtidas por métodos diferentes (MINGOTI, 2005).

2.5. Número de clusters da partição final

A escolha do número final de grupos em que o conjunto de dados deve ser repartido pode ser subjetiva, mas existem alguns métodos que auxiliam na determinação deste número. De acordo com Mingoti (2005), o propósito é encontrar o número de clusters que esteja associado à partição natural dos elementos que estão sendo comparados e agrupados, ou seja, determinar o ponto de corte dos agrupamentos.

Um agrupamento gerado é único e o melhor para um dado k , porém, nem todo agrupamento gerado é útil para efeitos de classificação. Um determinado k pode refletir um número de classes inadequado para a divisão de uma determinada população de observações, sendo necessário encontrar o valor de k que melhor reflita a divisão natural da população de dados observada (WANGEINHEM, 2016). O critério mais simples utilizado para decidir qual o número de grupos a adotar é o corte do dendrograma pela análise subjetiva dos diferentes níveis do mesmo, o que torna esse procedimento naturalmente enviesado pelas necessidades e opiniões dos analistas e pesquisadores (MARTINS et al., 2004).

À medida que se avança no algoritmo de agrupamento, ou seja, passa-se do estágio k para o estágio $k + 1$, a similaridade entre os conglomerados que estão sendo combinados nos respectivos passos vai decrescendo e, conseqüentemente, a distância entre eles vai aumentando. Desse modo, se for feito um gráfico do passo de cada estágio do processo, pode-se visualizar se há pontos de salto relativamente grandes em relação aos demais valores de distância. Estes pontos indicam o momento ideal de parada do algoritmo, isto é, o número de conglomerados final e a composição final dos grupos (SILVA & FISCH, 2013).

De acordo com Felix (2004), citado por Mingoti (2005), outro critério que pode ser utilizado é semelhante ao anterior, porém, em vez de se observar o comportamento da distância em cada estágio do agrupamento, observa-se o nível de similaridade. Neste caso, procura-se detectar pontos nos quais há um decréscimo acentuado na similaridade dos conglomerados unidos, pontos estes que indicam que o algoritmo de agrupamento deveria ser interrompido. O número de conglomerados final da partição seria aquele relacionado ao estágio em que o algoritmo foi interrompido.

Segundo Milligan & Cooper (1985), ao aplicar um índice para determinar o número ideal de grupos, podem ocorrer dois tipos de erros de decisão: o primeiro tipo de erro ocorre quando o índice indica a seleção de g grupos, sendo que na verdade há menos que g grupos no conjunto de dados. Já o segundo tipo de erro ocorre quando o índice indica menos grupos no conjunto de dados do que o real. Mesmo a severidade dos dois tipos de erros podendo mudar de acordo com o contexto do problema, tem-se que a ocorrência do segundo tipo de erro é considerada mais grave na maioria das análises, pois haverá perda de informação por fundir grupos distintos. Assim, os autores destacaram o fato de existirem poucas apresentações de métodos de comparação de desempenho de índices para determinar o número correto de grupos. Desse modo, Milligan & Cooper (1985) testaram um total de 30 medidas de validação de agrupamento visando determinar o número ideal de grupos para cada uma das medidas utilizadas no processo de agrupamento hierárquico, fazendo uso de dados artificiais com número de grupos conhecido (no intervalo de 2 a 5 grupos). Dentre os métodos avaliados, um dos que apresentaram melhor desempenho foi o índice de Calinski & Harabasz (CALINSKI & HARABASZ, 1974).

3 METODOLOGIA

As informações sobre a População Economicamente Ativa (PEA) são coletadas pelas pesquisas domiciliares produzidas pelo IBGE: são pesquisas censitárias, os censos demográficos (decenais) a partir de 1960, e pesquisas domiciliares, Pesquisa Nacional por Amostra de Domicílio (PNAD) (anual), desde seu início, em 1967, e Pesquisa Mensal de Emprego (PME), iniciada em 1980 e tornada sistemática de 1982 em diante.

A coleta desses dados padece de problemas que vão desde a operacionalização do conceito, principalmente nas economias onde as relações de produção não são totalmente independentes das atividades domésticas, até a dificuldade de enumeração, principalmente em relação a determinados segmentos da população como crianças, jovens e idosos.

De acordo com a Fundação IBGE (1990), as principais dificuldades de operacionalização do conceito são de três ordens: 1) definição do que seja uma atividade econômica, sendo reconhecida como tal a atividade que gera algum rendimento em espécie ou in natura, seja diretamente ao indivíduo que a exerce seja indiretamente através de sua unidade doméstica; 2) período em que a atividade

econômica deva ser exercida: pode-se considerar tanto o período de uma semana, o que facilita as respostas e evita erros "de memória", quanto o período de um ano que pode refletir melhor as condições de trabalho; 3) a terceira dificuldade diz respeito aos limites de idade que, em cada sociedade podem ser diferentes e, mesmo nas pesquisas do IBGE, até início da década de 2000, os limites inferiores de idade divergiam: na PNAD e censos demográficos, 10 anos, enquanto na PME, 15 anos. Os indicadores da PEA (sexo, idade, qualificação, distribuição pelos setores econômicos e pelas atividades rurais e urbanas) em seu conjunto são um importante indício do grau de desenvolvimento de uma economia.

Todas as informações supracitadas foram obtidas diretamente da página de Dados Abertos do Governo Federal (<http://dados.gov.br/tag/IBGE>) que são dados livremente disponíveis para todos utilizarem e redistribuírem como desejarem, sem restrição de licenças, patentes ou mecanismos de controle. Os dados encontram-se divididos por sexo, e são referentes ao número de pessoas compondo a PEA em seis regiões metropolitana, sendo elas São Paulo, Rio de Janeiro, Recife, Salvador, Porto Alegre e Belo Horizonte, entre os anos de 2005 e 2015.

No presente estudo, foram analisadas as médias da PEA dos meses, no período de 2005 a 2015, separadas por cada região metropolitana em estudo e subdividas por sexo. Foram realizadas as análises exploratória dos dados por meio de gráficos e de estatísticas descritivas, como as médias e os desvios padrões, prosseguindo com testes de hipóteses para comparação das médias, finalizando com Análise de Agrupamentos, quando necessário, seguida pela comparação das médias entre e dentre os grupos obtidos. Foram utilizados os softwares Minitab, para a ANOVA e o Teste de Tukey e o R, na Análise de *Clusters*, sendo que os comandos aplicados neste último seguem descritos no Anexo A.

3.1. Testes de Comparação entre Médias

3.1.1. Teste t

No presente estudo, foi aplicado o Teste t para comparar as médias da PEA entre os sexos dos agrupamentos formados. Este teste é utilizado quando o número de elementos em estudo é inferior a 30 e o desvio padrão populacional é desconhecido. Para duas amostras independentes, o valor de t calculado é dado por:

$$t_{calculado} = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

em que

$\bar{X}$ e $\bar{Y}$ são as médias amostrais de cada grupo,

μ_X e μ_Y são as médias populacionais de cada grupo,

s_X e s_Y são os desvios padrão para cada grupo,

n_X e n_Y são o número de elementos de cada grupo.

Os graus de liberdade (g.l.) são calculados da seguinte forma: $g.l. = n_X + n_Y - 2$.

Compara-se o valor t calculado com o valor t tabelado, rejeitando ou não a hipótese de igualdade entre as médias de acordo de onde esses valores se encontrarão na curva da Distribuição t .

3.1.2. ANOVA

Para a comparação de três ou mais médias, utiliza-se do Teste de Comparação entre Médias denominado de ANOVA. A Análise de Variância envolvendo o estudo de todos os níveis de um fator em todos os níveis do outro fator foi aplicada em Esquema Fatorial, no Delineamento Aleatorizado em Blocos. Esses dados serão representados pelo seguinte modelo matemático:

$$Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \beta_k + \varepsilon_{ijk}$$

em que, neste trabalho,

os fatores A e B são sexo e regiões metropolitanas, respectivamente, os meses são os blocos;

Y_{ijk} é a variável resposta, sendo, aqui, correspondente ao valor da PEA, ou seja, ao valor observado do i -ésimo fator A (sexo), no j -ésimo fator B (região metropolitana) e no k -ésimo bloco (mês), com $i = 1, 2$; $j = 1, 2, \dots, 6$ e $k = 1, 2, \dots, 12$;

μ é uma constante comum às respostas (normalmente a média geral);

a_i é o efeito do i -ésimo fator A, ou seja, o efeito do i -ésimo fator sexo;

b_j é o efeito do j -ésimo fator B, ou seja, o efeito do j -ésimo fator região metropolitana;

$(ab)_{ij}$ é o efeito da interação entre os fatores A e B, portanto, é o efeito da interação entre os fatores sexo e região metropolitana;

β_k é o efeito do k -ésimo bloco, ou seja, é o efeito do mês;

ε_{ijk} é erro experimental associado ao i -ésimo fator A (sexo), no j -ésimo fator B (região metropolitana) e no k -ésimo bloco (mês) assumido ser independente e $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Para se verificar as pressuposições para aplicação da ANOVA, existem vários testes que podem ser utilizados, entre eles, destaca-se o Teste de Bartlett, para a verificação da homogeneidade da variância, aplicado neste trabalho, como não houve violação da normalidade dos dados.

3.1.3. Teste de Tukey

O Teste de Tukey foi utilizado após evidência de que pelo menos uma das médias da PEA divergia entre os sexos e regiões metropolitanas através da ANOVA, com o intuito de verificar quais eram estas médias divergentes. O Teste de Tukey é realizado da seguinte forma:

Seja: $L = \bar{y}_i - \bar{y}'_i$

L será dita significativa a um nível α se

$$|L| > \Delta$$

onde Δ , a diferença significativa mais honesta, é o critério do Teste de Tukey.

$$\Delta = q_{\alpha, a, f} \frac{s(L)}{\sqrt{2}}, \text{ ou } \Delta = q_{\alpha, a, f} \sqrt{\frac{s^2}{n}}, \text{ para } n_i = n'_i = n$$

em que

$q_{\alpha, a, f}$ = amplitude total estudentizada para uso no teste Tukey

Esse valor de q dependente de α , a e f , é tabelado. Para um dado experimento (se ele for balanceado), tem-se apenas um valor de q e, portanto, apenas um critério Δ . Esta é uma vantagem sobre o teste Duncan, no que diz respeito à facilidade de aplicação do teste.

O critério Δ de Tukey (1953) pode também ser usado para se calcular um conjunto de intervalos de confiança para as comparações. Assim, a diferença entre as médias populacionais $\mathcal{L} = \mu_i - \mu'_i$ ou $\mathcal{L} = \alpha_i - \alpha'_i$, pode ser colocada no intervalo de confiança a $100(1 - \alpha)\%$:

$$L - \Delta \leq \mathcal{L} \leq L + \Delta$$

Quando os tratamentos têm números diferentes de repetições, uma alternativa à fórmula geral mostrada anteriormente é usar uma aproximação, sendo que esta

aproximação é boa se os números n_i e n'_i não forem muito diferentes. Nesse caso, se duas médias μ_i e μ'_i tem números de repetições n_i e n'_i , respectivamente, a fórmula pode ser substituída por

$$\Delta = q_{\alpha,a,f} \sqrt{\frac{s^2}{\min(n_i, n'_i)}}.$$

3.2. Análise de Agrupamentos

3.2.1. Medidas de Similaridade e Dissimilaridade

A Distância Euclidiana entre dois elementos X_l e X_k , $l \neq k$, é definida por meio da expressão:

$$d_{X_l, X_k} = \sqrt{\sum_{i=1}^p (X_{il} - X_{ik})^2}$$

ou seja, os dois elementos amostrais são comparados em cada variável pertencente ao vetor de observações.

O Coeficiente de Correlação de Pearson é dado por:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}.$$

A Correlação Cofenética mede o grau de ajuste entre a matriz de similaridade original (matriz S) e a matriz resultante da simplificação proporcionada pelo método de agrupamento (matriz C). No caso, C é aquela obtida após a construção do dendrograma. Tal correlação é calculada usando:

$$r_{cof} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (c_{ij} - \bar{c})(s_{ij} - \bar{s})}{\sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (c_{ij} - \bar{c})^2} \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (s_{ij} - \bar{s})^2}}$$

em que

c_{ij} : valor de similaridade entre os indivíduos i e j , obtidos a partir da matriz cofenética;

s_{ij} : valor de similaridade entre os indivíduos i e j , obtidos a partir da matriz de similaridade;

$$\bar{c} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n c_{ij};$$

$$\bar{s} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n s_{ij}$$

3.2.2. Métodos Hierárquicos e número de clusters da partição final

No Método de Ward, a distância entre os conglomerados C_l e C_i é dada por:

$$d_{(C_l, C_i)} = \left[\frac{n_l n_i}{n_l + n_i} \right] (\bar{X}_l - \bar{X}_i)' (\bar{X}_l - \bar{X}_i)$$

que é a soma de quadrados entre os clusters C_l e C_i . Em cada passo do algoritmo de agrupamento, os dois conglomerados que minimizam a distância dada acima são combinados.

Calinski & Harabasz (1974) sugerem para cada passo do agrupamento, o cálculo da estatística chamada Pseudo F iniciada pela definição de C_i e C_l que são os grupos unidos num determinado estágio, o nível de similaridade é definido por:

$$S_{il} = 1 - \frac{d_{ij}}{\max\{d_{jk}, j, k = 1, 2, \dots, n\}} \times 100$$

em que $\max\{d_{jk}, j, k = 1, 2, \dots, n\}$ é a maior distância entre os n elementos amostrais na matriz de distâncias $D_{n \times n}$ do primeiro estágio do processo de agrupamento.

Calinski & Harabasz (1974) definiram que para cada passo do agrupamento, o cálculo da estatística Pseudo F seja calculada por meio da seguinte equação:

$$F = \frac{SSB/(g^* - 1)}{SSR/(n - g^*)} = \left(\frac{n - g^*}{g^* - 1} \right) \left(\frac{R^2}{1 - R^2} \right)$$

em que

SSB é a soma de quadrados total entre os g^* grupos da partição, sendo que:

$$SSB = \sum_{i=1}^{g^*} n_i (\bar{X}_i - \bar{X})' (\bar{X}_i - \bar{X})$$

SSR é a soma de quadrados total dentro dos grupos da partição (Soma de Quadrados Residual):

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

R^2 é o coeficiente da partição, dado por:

$$R^2 = \frac{SSB}{SST_c}$$

SST_c é a soma de quadrados total corrigida para a média global em cada variável, expresso por:

$$SST_c = \sum_{i=1}^{g^*} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})' (X_{ij} - \bar{X})$$

g^* é o número de grupos relacionado com a partição do respectivo estágio de agrupamento; n é o número de elementos amostrais; $X'_{ij} = (X_{i1j} X_{i2j} \dots X_{ipj})$ é o vetor de medidas observadas para o j – ésimo elemento amostral do i – ésimo grupo; $\bar{X}'_i = (\bar{X}_{i1} \bar{X}_{i2} \dots \bar{X}_{ip})$ é o vetor de médias do i – ésimo grupo; e $\bar{X}' = (\bar{X}_{.1} \bar{X}_{.2} \dots \bar{X}_{.p})$ é o vetor de médias global, sem levar em conta qualquer posição, em que $\bar{X}_{.l} = \frac{1}{n} \sum_{i=1}^{g^*} \sum_{j=1}^{n_i} X_{ilj}$, $l = 1, 2, \dots, p$.

Mingoti (2005) afirma que pode ser mostrado que a estatística Pseudo F tem distribuição F com $p(g^* - 1)$ e $p(n - g^*)$ graus de liberdade quando os n elementos amostrais constituem uma amostra aleatória de uma distribuição normal p -variada e quando os elementos amostrais são alocados aleatoriamente em grupos. Porém, sabe-se que na prática a alocação aleatória não acontece, uma vez que a partição dos elementos amostrais é feita através de métodos de agrupamento com critérios de similaridade previamente definidos por métricas matemáticas. Apesar desse fato, tem-se que a ideia por trás deste critério é como se em cada passo do algoritmo de agrupamento estivesse sendo feito um Teste F de Análise de Variância para a comparação dos vetores de médias dos grupos que foram formados naquele passo.

Busca-se o maior valor de Pseudo F, ou seja, aquele que estaria relacionado com a menor probabilidade de significância do teste, resultando assim na partição com maior

heterogeneidade entre grupos (MINGOTI, 2005). Timm (2002) destaca que o índice Pseudo F é o mais útil para identificar o número de clusters.

4 RESULTADOS

Os dados do presente estudo referem-se à média da quantidade de pessoas compondo a PEA por mês (obtido dentro de cada cidade e cada sexo) é a média de todos os meses entre 2005 e 2015. Primeiramente foi realizado um estudo descritivo dos dados para cada sexo (Tabelas 1 e 2).

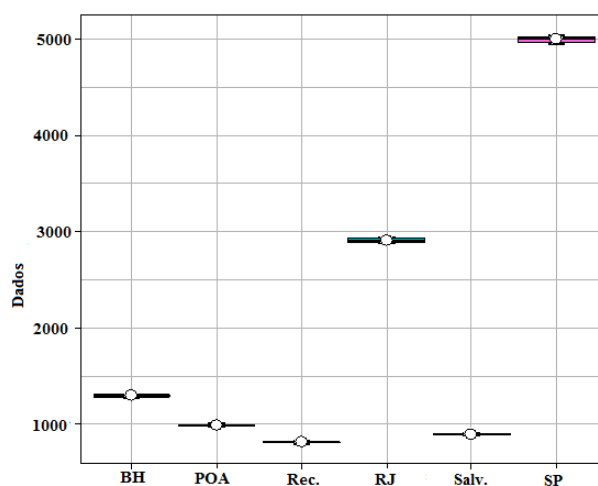
Tabela 1: Resumo descritivo dos dados referentes à média do número de pessoas compondo a PEA (em milhares) do sexo masculino em cada cidade, no período de 2005 a 2015.

Estatísticas	Belo Horizonte	Porto Alegre	Recife	Rio de Janeiro	Salvador	São Paulo
Mínimo	1272,09	978,09	800,82	2882,45	880,27	4947,82
1º Quartil	1279,93	981,27	807,66	2888,09	883,68	4968,20
Média	1294,0	988,4	812,2	2909,4	891,5	4996,65
Mediana	1296,14	985,41	810,73	2902,00	891,91	4997,18
3º Quartil	1305,62	996,41	817,64	2931,58	900,00	5026,64
Máximo	1307,50	1006,60	824,20	2944,36	901,45	5038,60

Tabela 2: Resumo descritivo dos dados referentes à média do número de pessoas compondo a PEA (em milhares) do sexo feminino em cada cidade, no período de 2005 a 2015.

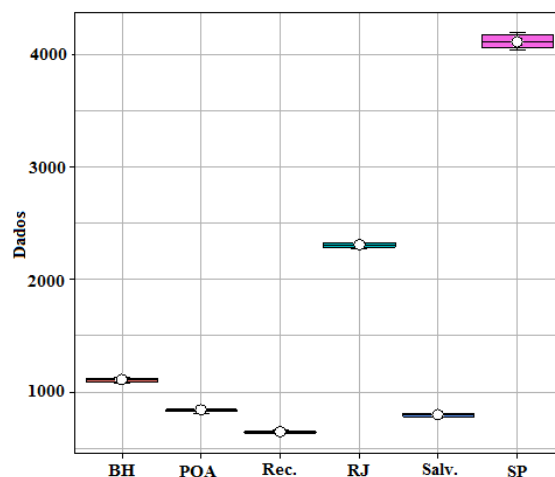
Estatísticas	Belo Horizonte	Porto Alegre	Recife	Rio de Janeiro	Salvador	São Paulo
Mínimo	1077,73	812,82	631,91	2268,18	774,36	4041,27
1º Quartil	1088,68	823,84	633,48	2286,30	781,32	4059,05
Média	1104,85	836,68	644,20	2306,04	793,63	4111,48
Mediana	1105,68	839,91	643,32	2306,32	792,09	4106,59
3º Quartil	1121,14	849,02	653,70	2328,58	809,02	4176,61
Máximo	1131,80	851,70	663,90	2331,60	810,70	4192,82

Os gráficos Boxplot (Figuras 4 e 5) evidenciam uma baixa variabilidade dos dados. Os resíduos são normalmente distribuídos e independentes, como pode ser observado pelos gráficos *QQ-Plot*, Histograma e plotagem dos resíduos (Figuras 6, 7 e 8). Foram detectados *outliers*, como era esperado, devido ao fato de São Paulo ter médias da PEA com valores muito maiores e destoantes das demais regiões metropolitanas.



*BH corresponde a Belo Horizonte (MG), Rec. a Recife (PE), Salv. a Salvador (BA), POA a Porto Alegre (RS), RJ a Rio de Janeiro (RJ) e SP a São Paulo (SP).

Figura 4: Gráfico Boxplot da média do número de pessoas (em milhares) do sexo masculino que compunham a PEA entre os anos de 2005 a 2015, em relação às regiões metropolitanas em estudo.



*BH corresponde a Belo Horizonte (MG), Rec. a Recife (PE), Salv. a Salvador (BA), POA a Porto Alegre (RS), RJ a Rio de Janeiro (RJ) e SP a São Paulo (SP).

Figura 5: Gráfico Boxplot da média do número de pessoas (em milhares) do sexo feminino que compunham a PEA entre os anos de 2005 a 2015, em relação às regiões metropolitanas em estudo.

Foi realizada a verificação das pressuposições da ANOVA, verificadas nas Figuras 6, 7, 8 e 9 e também no resultado do Teste de Bartlett, o qual apresentou uma estatística de valor igual a 159,98, com 5 graus de liberdade e valor-p menor que 0,0001.

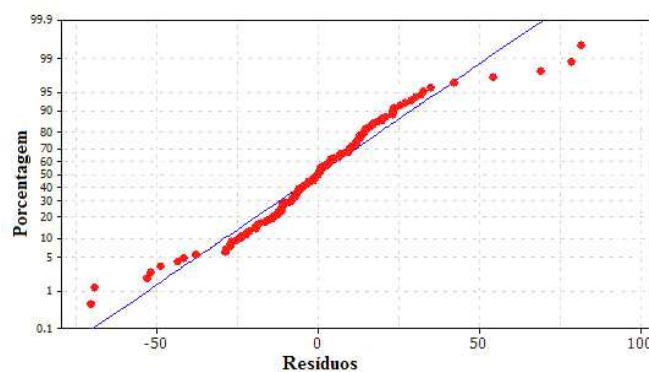


Figura 6: Gráfico da Probabilidade Normal dos resíduos.

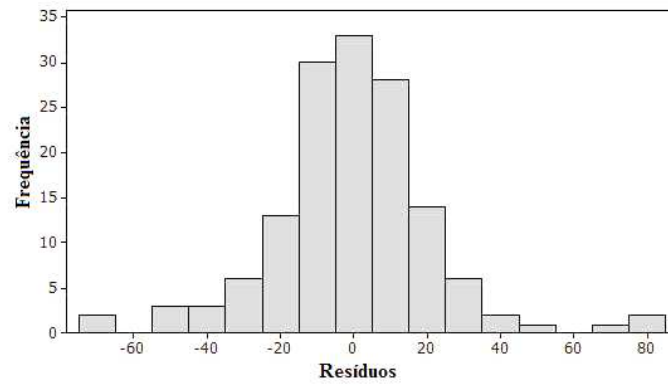


Figura 7: Gráfico Histograma dos resíduos.

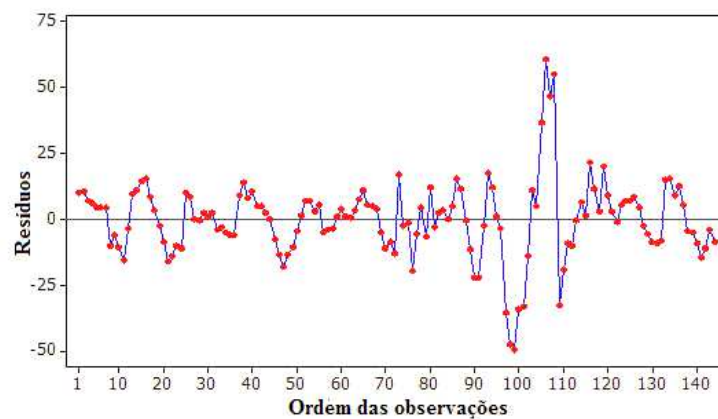


Figura 8: Gráfico dos Resíduos versus os dados ordenados.

A homogeneidade das variâncias não foi atendida, como observado na Figura 9 e no resultado do Teste de Bartlett, devido à presença destes *outliers*, porém, como as demais suposições da Análise de Variância (normalidade e independência dos resíduos) foram válidas, e devido à robustez do teste ANOVA, seguiu-se com a análise dos dados.

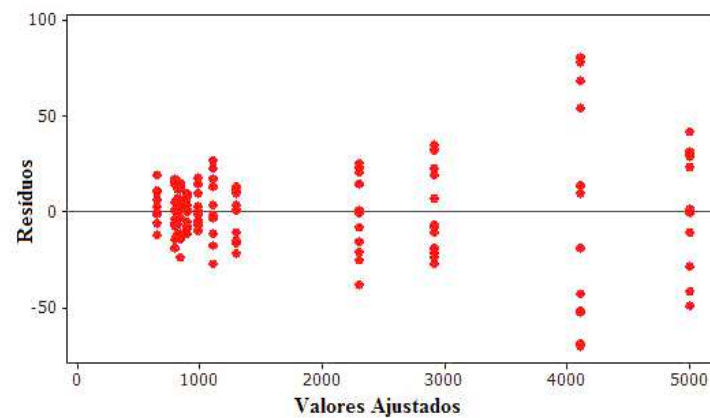


Figura 9: Gráfico dos Resíduos versus valores ajustados.

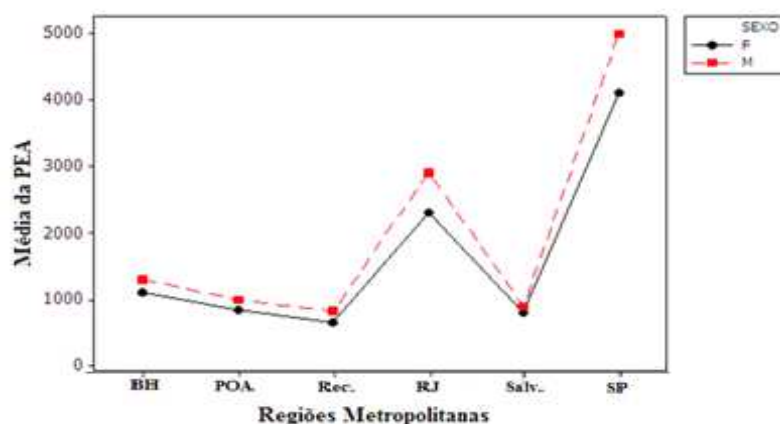
Observou-se a existência de variações entre um mês e outro, verificando-se que todos os tratamentos (combinação de sexo com cidade) se encontram dentro de cada mês, os blocos foram considerados como sendo os meses, os anos foram considerados como repetições (onze), e dois fatores (sexo e cidade, dois e seis níveis, respectivamente). Assim, foi aplicado o Delineamento Aleatorizado em Blocos (também conhecido como DBC) em esquema Fatorial, e, na ANOVA (Tabela 3), foi obtida significância estatística ao nível de 1% para os fatores sexo, cidade e a interação entre estes, indicando que a resposta de um fator depende da presença ou ausência do outro. Neste caso, a alternativa utilizada foi estudar o comportamento de um fator dentro de cada nível de outro fator, considerando o fator de blocagem.

Tabela 3: ANOVA referente à média do número de pessoas compondo a PEA (em milhares) no período de 2005 a 2015.

Fator de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Estatística F	Valor-p
Mês	11	41466	3770	14,61	Menor que 0,0001
Sexo	1	4389720	4389720	17011,51	Menor que 0,0001
Cidade	5	274816762	54963352	212999,78	Menor que 0,0001
Sexo*Cidade	5	3074895	614979	2383,23	Menor que 0,0001
Erro	121	31223	258		
Total	143	282354067			

Aplicado o Teste Tukey, as médias entre os sexos na PEA diferiram para todas as regiões metropolitanas ao nível de 5% de significância, exceto entre a interação de Recife no sexo masculino e Salvador no sexo feminino.

Percebe-se que a quantidade média de pessoas compondo a PEA em relação às regiões metropolitanas é maior para o sexo masculino, o que é evidenciado pelo gráfico da interação entre o sexo e as regiões metropolitanas (Figura 10).

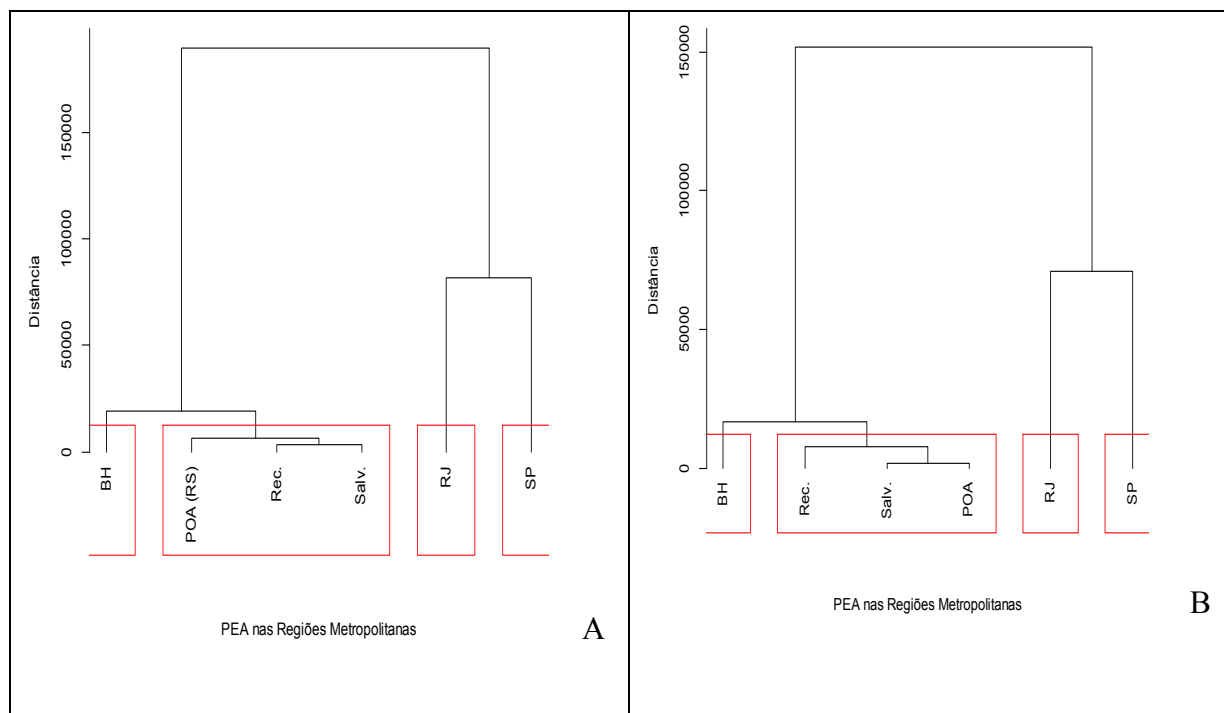


*BH corresponde a Belo Horizonte (MG), Rec. a Recife (PE), Salv. a Salvador (BA), POA a Porto Alegre (RS), RJ a Rio de Janeiro (RJ) e SP a São Paulo (SP).

Figura 10: Gráfico da interação entre os sexos da média do número de pessoas (em milhares) que compunham a PEA entre os anos de 2005 a 2015, em relação às regiões metropolitanas em estudo.

Foi realizada a Análise de Agrupamentos, sendo utilizado o Método de Ward, pois apresentou valores do Coeficiente de Correlação Cofenética próximos de 1 para ambos os sexos, (sendo igual a 0,82 para o sexo feminino e 0,85 para o sexo masculino), indicando, portanto, boa consistência dos agrupamentos e pouca distorção após a formação dos grupos das regiões metropolitanas. No critério Pseudo F, foi observada compatibilidade entre os resultados dos grupos formados pelos dendrogramas com a análise preliminar descritiva realizada, e, portanto, seguiu-se com a formação dos *clusters*.

No dendrograma resultante da Análise de Agrupamentos (Figura 11), observa-se que, tanto para o sexo masculino, quanto para o feminino, a subdivisão das cidades dentro dos grupos manteve a classificação em relação à média obtida na análise descritiva (Tabelas 2 e 3).



*BH corresponde a Belo Horizonte (MG), Rec. a Recife (PE), Salv. a Salvador (BA), POA a Porto Alegre (RS), RJ a Rio de Janeiro (RJ) e SP a São Paulo (SP).

Figura 11: Dendrograma com 4 grupos distintos referente ao agrupamento das Regiões Metropolitanas em relação ao período de 2005 a 2015: **(A)** para os dados do sexo masculino e **(B)** do sexo feminino.

Foi possível observar que o padrão de classificação referiu-se ao fato de São Paulo, Rio de Janeiro e Belo Horizonte terem formado grupos isolados (cada um sendo o seu próprio grupo), onde São Paulo registrou uma média de 5 milhões e 4,1 milhões para os sexos masculino e feminino, respectivamente, enquanto que, no Rio de Janeiro, a média foi de 3 milhões e 2,3 milhões, para os sexos masculino e feminino, respectivamente. Já em relação a Belo Horizonte, este número caiu mais que a metade, sendo, em média de 1,2 milhão e 1 milhão, respectivamente, para os sexos masculino e feminino, enquanto que as demais regiões metropolitanas apresentaram valores da PEA abaixo de 1 milhão de pessoas, para ambos os sexos.

Tabela 4: Composição dos agrupamentos formados pela análise de *cluster* para os sexos masculino e feminino.

Grupo 1	Grupo 2	Grupo 3	Grupo 4
Recife (PE)	Belo Horizonte (BH)	Rio de Janeiro (RJ)	São Paulo (SP)
Salvador (BA)			
Porto Alegre (RS)			

Como os grupos formados tiveram a mesma composição para ambos os sexos, seguiu-se a análise descritiva dos dados e também os valores-p obtidos através do Teste *t*-Student (Tabela 5). Para a comparação das médias da PEA das Regiões Metropolitanas do Grupo 1 em cada sexo, realizou-se a ANOVA, sendo observados os resultados nas Tabelas 6 e 7.

Tabela 5: Análise descritiva dos agrupamentos das regiões metropolitanas em relação aos anos para os dados do sexo masculino e do sexo feminino, com 4 grupos distintos (em milhares) e Teste *t* para a comparação dos valores das médias obtidos pelos agrupamentos entre os sexos.

	Masculino		Feminino		Teste <i>t</i> - <i>Student</i>
	Média	Desvio Padrão	Média	Desvio Padrão	Valor-p
Grupo 1	10602,97	950,94	8957,515	1005,92	$5,299 \times 10^{-7}$
Grupo 2	15290,18	993,80	13052,73	1049,39	$3,95 \times 10^{-4}$
Grupo 3	34379,45	1395,14	27248,82	1649,80	$3,709 \times 10^{-8}$
Grupo 4	59044,64	2769,92	48576,82	3374,02	$2,581 \times 10^{-6}$

Tabela 6: ANOVA para comparação das médias da PEA das Regiões Metropolitanas do Grupo 1, para os dados do sexo masculino.

	<i>Graus de Liberdade</i>	<i>Soma de Quadrados</i>	<i>Quadrado Médio</i>	<i>Estatística F</i>	<i>Valor-P</i>
Grupos	3	13950893,21	4650297,74	597,48	0,0017
Resíduos	2	15566,45	7783,23		

Assim, ao nível de 1% de significância, foi obtido que, para ambos os sexos, mesmo estando em um único agrupamento, as médias das regiões metropolitanas de Recife, Salvador e Porto Alegre são estatisticamente diferentes entre si.

Tabela 7: ANOVA para comparação das médias da PEA das Regiões Metropolitanas do Grupo 1, para os dados do sexo feminino.

	<i>Graus de Liberdade</i>	<i>Soma de Quadrados</i>	<i>Quadrado Médio</i>	<i>Estatística F</i>	<i>Valor-P</i>
Grupos	3	1282906043,58	427635347,86	41906,57	Menor que 0,0001
Resíduos	2	20408,99	10204,49		

5 DISCUSSÕES

A composição dos agrupamentos da PEA nas regiões metropolitanas em relação aos anos analisados foi similar para ambos os sexos quando considerados quatro grupos distintos, porém, a média da quantidade de pessoas segundo os sexos compondo a PEA são diferentes em todos os agrupamentos, sendo maior para o sexo masculino, como mostram os resultados do Teste t de Student. Foi observado que a diferença média entre os sexos na composição da PEA tem caído ao passar dos anos, mas o presente estudo confirma que tal diferença ainda é significativa. Os resultados obtidos reforçam o que foi afirmado e citado neste trabalho por Hirata (1998) e também vão de acordo com o Gender Gap Index (2015), ambos apresentados anteriormente.

Dados do IBGE revelam que entre 2003 e 2010, a PEA feminina cresceu 17,3%, enquanto a PEA masculina cresceu 9,7%. A participação das mulheres na PEA passou de 44,4% em 2003, para 46,1%, em 2011. No mesmo período, os homens cresceram sua participação na população ocupada, passando de 43,0% para 45,4%.

A PEA acompanha o registro populacional das regiões metropolitanas, exceto para Salvador e Belo Horizonte, uma vez que o número de habitantes para Salvador é maior do que o de Belo Horizonte, porém, a PEA de Belo Horizonte é superior à de Salvador, o que pode ser justificado pelo fato do estado da Bahia ser o que mais recebe benefícios no programa Bolsa Família, se comparado aos outros estados brasileiros. Como também é o caso do estado de Pernambuco, com a terceira maior participação do Bolsa Família, comparado aos outros estados, justificando o fato de Recife, no presente

estudo, possuir o menor registro na PEA em ambos os sexos. Porto Alegre e Recife, por outro lado, possuem o número de habitantes aproximadamente iguais, no entanto, a PEA de Recife é ainda menor que a de Porto Alegre, o que pode ser justificado pela forte presença do Bolsa Família, conforme explicado anteriormente.

São Paulo é a cidade mais populosa do país, possuindo, automaticamente, a maior PEA, aliada à migração de pessoas vindas de outras cidades e até mesmo outros estados rumo a São Paulo. A partir da década de 40, quando o país iniciou o seu processo de industrialização, houve um acelerado crescimento urbano provocado pela mecanização do campo, o que prossegue acontecendo pelo avanço tecnológico, ocasionando perda de postos de trabalho e promovendo um grande fluxo de trabalhadores para os grandes centros urbanos, como é o caso de São Paulo, Rio de Janeiro e Belo Horizonte. Segundo dados do IBGE, o Rio de Janeiro possui uma população total com cerca da metade da de São Paulo, sendo a segunda cidade mais populosa do país, justificando também que a PEA decaia nesta mesma proporção.

Vale ressaltar que não foram considerados e nem comparados os salários entre os sexos, uma variável que é plausível ser estudada e analisada posteriormente. Sugere-se, para estudos futuros, a aplicação dos dados em medidas repetidas e também em Regressão, sendo as variáveis categóricas transformadas em variáveis *Dummy*.

6 CONCLUSÃO

O presente estudo obteve resultados que evidenciaram a existência de diferença significativa entre a quantidade média de pessoas compondo a PEA em relação aos sexos entre os anos de 2005 e 2015, nas seis regiões metropolitanas consideradas, confirmando a desigualdade entre os sexos, apesar da evolução nítida da participação feminina no mercado de trabalho. Esta diferença encontrada entre a composição da PEA em relação aos sexos é observada desde a inserção da mulher no mercado de trabalho. Apesar disto, nota-se que o comportamento da PEA feminina em relação aos agrupamentos e nas estatísticas descritivas se assemelhou ao comportamento da PEA masculina quanto às regiões metropolitanas.

REFERÊNCIAS BIBLIOGRÁFICAS:

- ALVES, J. E. D. **O crescimento da PEA e a redução do hiato de gênero nas taxas de atividade no mercado de trabalho**. Rio de Janeiro: UFRJ/Aparte, 2013.
- ALTTIMAN, C. N.; COSTA, S. G. **Revolução Feminina: as conquistas da mulher no Século XX**. Trabalho de conclusão de curso (Monografia). Faculdade Eça de Queiroz, Jandira (SP), 2009.
- AQUILINI, G. H.; COSTA, P. L. **O sobre-desemprego e a inatividade das mulheres na metrópole paulista**. Estudos Avançados, São Paulo, v.17, n.49, 2003.
- BEVILACQUA, S. **Segmentação a partir das bases específicas não observáveis do produto e o emprego conjunto de técnicas multivariadas: o estudo de caso do veículo multiuso**. Revista Eletrônica de Administração, v. 19, n. 2, 2013.
- BRUSCHINI, M. C. A. **Trabalho e gênero no Brasil nos últimos dez anos**. Cadernos de Pesquisa, São Paulo, v.37, n.132, Set. / Dec. 2007.
- BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. F. **Introdução à Análise de Agrupamentos**. 9º Simpósio Brasileiro de Probabilidade e Estatística. São Paulo: IME – USP, 1990.
- CATELA, E. Y. S. **Mudança estrutural e crescimento dos municípios fluminenses: uma análise da mobilidade relativa no período 1998-2008**. Econ. Apl. [online], vol.17, n.2, 2013.
- CHIERICE JUNIOR, N.; LANDIM, P. M. B. **Análise da chuva e vazão na Bacia Hidrográfica do Rio Pardo**. Geociências (São Paulo. Impresso), v. 33, 2014.
- COMUNELLO, E.; ARAÚJO, L. B.; SENTELHAS, P. C.; ARAÚJO, M. F. C., DIAS, C. T. S.; FIETZ, C. R. **O uso da análise de cluster no estudo de características pluviométricas**. Sigmae, vol. 2, n.3, 2013.
- CRUZ, C. D.; REGAZZI, A. J. **Métodos biométricos aplicados ao melhoramento genético**. Viçosa, MG: Universidade Federal de Viçosa. 1994.
- DEVORE, J. L. **Probabilidade e estatística: para engenharia e ciências**. 1ª Ed. São Paulo: Pioneira Thomson Learning, 2006.
- DIAS, A. **Seleção multivariada e identidade de modelos não lineares para o crescimento e acúmulo de nutrientes em frutos de mangueira**. Dissertação (Doutorado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras. 2014.

DILLON, W.; GOLDSTEIN, M. **Multivariate Analysis**. New York: John Wiley & Sons, 1994.

DOWNING, D.; CLARK, J. **Estatística Aplicada**. 1ª Ed. São Paulo: Saraiva, 2000.

ESPÍNDOLA, G. **A trajetória do poder da mulher: do lar ao mercado de trabalho**. 2008. Disponível em: <http://www.slideshare.net/eudelucy/a-trajetria-do-poder-da-mulher-do-lar-ao-mercado-de-trabalho>. Acesso em: 15 de março de 2016.

FEE. **Banco de Dados Estatísticos**. Disponível em: <http://www.fee.tche.br>

FELIX, F. N. **Aplicando bootstrap para determinação de intervalos de confiança para o número de grupos no procedimento hierárquico aglomerativo de Ward**. Dissertação de Mestrado. Belo Horizonte: Departamento de Estatística/UFMG, 2004.

FONSECA, J.S.; MARTINS, G.A. **Curso de Estatística**. 6ª Ed. São Paulo: Atlas, 1996.

FREITAS, E. **População Economicamente Ativa no Brasil**. Disponível em: <http://mundoeducacao.bol.uol.com.br/geografia/populacao-economicamente-ativa-brasil.htm>> Acesso em 19/03/2016.

FREITAS, J. A. B. **Análise de Cluster da Lisozima**. Relatório Científico apresentado ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) por conta da conclusão de projeto de Iniciação Científica. Universidade de São Paulo. 2006.

FREUND, J.E.; SIMON, G.A. **Estatística Aplicada**. 9ª Ed. Porto Alegre: Bookman, 2000.

Gender Gap Index, Disponível em: <http://reports.weforum.org/global-gender-gap-report2015/economies/#economy=BRA>>. Acesso em 18 de março de 2016.

HAIR, J. F. Jr.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Multivariate data analysis with readings**. 4. ed. . Black, W. C. New Jersey: Prentice Hall, 1995.

HAIR, J. F. Jr.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise Multivariada de Dados**. Trad. Adonai S. Sant'Anna e Anselmo C. Neto. 5ª ed. Ed Bookman, Porto Alegre, 2005.

HIRATA, H. **Reestruturação produtiva, trabalho e relações de gênero**. Revista Latino-americana de Estudos do Trabalho, v.4, n.7, 1998.

HOFFMANN, R. **A dinâmica da modernização da agricultura em 157 microrregiões homogêneas do Brasil**. Revista de Economia e Sociologia Rural, Brasília, v. 30, out. / dez. 1992.

HOFFMANN, R.; LEONE, E. T. **Participação da mulher no mercado de trabalho e desigualdade da renda domiciliar per capita no Brasil: 1981-2002**. Nova Economia, 2004.

- HOFFMANN, M. B. P.; MENDONÇA, S. E. A. **O Mercado de Trabalho na Região metropolitana de São Paulo**. Estud. av. vol.17 no.47 São Paulo Jan./Apr. 2003
- JOHNSON, R.; WICHERN, D. **Applied multivariate statistical analysis**. 3^a ed. New Jersey: Prentice Hall, 1992.
- KELLER FILHO, T.; ASSAD, E. D.; LIMA, P. R. S. R. **Regiões pluviometricamente homogêneas no Brasil**. Pesquisa Agropecuária Brasileira, Brasília, v.40, 2005.
- KOETZ, A. L.; LACERDA, D. P.; MOREIRA, G.; KLIPPED, M.; VACCARO, G. L. R. **Uma análise interpretativa do produto interno bruto versus população economicamente ativa dos municípios gaúchos utilizando a análise de Cluster**. In: SEGET 2004 - Simpósio Sobre Excelência em Gestão e Tecnologia, 2004, Resende, RJ. Anais do SEGET 2004 - Simpósio Sobre Excelência em Gestão e Tecnologia, 2004. v. 01.
- LAMEIRAS, M.A.P. **Efeitos da População Economicamente Ativa sobre a Taxa de Desemprego**. Carta de Conjuntura. Rio de Janeiro. 2013.
- LANDIM, P. M. B. **Análise Estatística de Dados Geológicos Multivariados**. Departamento de Geologia Aplicada. Laboratório de Geomatemática. UNESP/Campus de Rio Claro. Texto Didático 03. 2000.
- LATTIN, J.; AVRITSCHER, H.; CARROLL, J. D.; GREEN, P. E. **Análise de Dados Multivariados**. Cengage, 2011.
- LAVINAS, L.; Cordilha, A. C. **Assimetrias de Gênero no Mercado de Trabalho: rumos da formalização**. Colóquio internacional: Travail, Careet Politiques Sociales: Débat Brésil - France, 2014.
- LEONE, E. T.; BALTAR, P. **A Mulher na Recuperação Recente do Mercado de Trabalho Brasileiro**. Revista Brasileira de Estudos Populacionais. São Paulo, v.25, n.2, 2008.
- LINDEN, R. **Técnicas de Agrupamento**. Revista Sistemas de Informação da FMSA, n.4, 2009.
- MANTOVANI L. S. D. **O processo de escolha de serviço de telefonia celular dos jovens em Curitiba**. Dissertação de Mestrado. UFPR, 2006.
- MARTINS, M. R. F. O.; PEDRO, S.; ROSA, S. **Escolha do número de grupos e validação da solução em análise classificatória: da teoria à prática**. Instituto Superior de Estatística e Gestão de Informação. Universidade Nova de Lisboa. Campolide, Lisboa. 2004.

- MENEZES FILHO, N. A.; SORZAFAVE, L. **A Evolução da Participação Feminina no Mercado de Trabalho Brasileiro**. Campinas. Anais do XXIV Encontro Nacional de Economia, 2000.
- METZ, J.; MONARD, M. C. **Projeto e Implementação do Módulo de Clustering Hierárquico do Discover**. Instituto de Ciências Matemáticas e de Computação. São Carlos. Agosto de 2006.
- MILLIGAN, G. W.; COOPER, M. C. **An examination of procedures for determining the number of clusters in a data set**. *Psychometrika*, v. 50, n. 2, 1985.
- MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.
- PENA, R. F. A. **"População Economicamente Ativa – PEA"**. *Brasil Escola*. Disponível em <<http://brasilecola.uol.com.br/geografia/populacao-economicamente-ativa-pea.htm>>. Acesso em 19 de março de 2016.
- PENA, R. F. A. **O Conceito de População Economicamente Ativa (PEA)**. Disponível em <<http://mundoeducacao.bol.uol.com.br/geografia/o-conceito-populacao-economicamente-ativa-pea.htm>>. Acesso em 19 de março de 2016.
- PENA, R. F. A. **População Economicamente Ativa – PEA**. Disponível em:<<http://alunosonline.uol.com.br/geografia/populacao-economicamente-ativa-pea.html>> Acesso em 19 de março de 2016.
- PNUD, 2009. Disponível em: <<http://www.pnud.org.br/Noticia.aspx?id=1998>>. Acesso em 18 de março de 2016.
- R version 2.7.1. **The R Foundation for Statistical Computing**. ISBN 3-900051-07-0. 2008.
- REIS, P. R.C.; SILVEIRA, S. F. R.; RODRIGUES, P. E. L. **Impactos da Política Nacional de Irrigação sobre o desenvolvimento socioeconômico da região Norte de Minas Gerais: uma avaliação do Projeto Gorutuba**. *Rev. Adm. Pública*, Ago 2012, vol.46, nº 4, 2012.
- SCHLICKMANN, E.; PIZARRO, D. **A evolução da mulher no trabalho: uma abordagem sob a ótica da liderança**. Artigo apresentado como requisito parcial para obtenção do título de graduação em Administração pela Faculdade Borges de Mendonça, 2012.
- SILVA, C. M.; FISCH, G. F. **Análise do Comportamento da Precipitação em Taubaté - SP na escala de tempo diária em comparação com os eventos de El Niño e La Niña**. *Ambiente e Água – An Interdisciplinary Journal of Applied Science*, 2013.

SILVEIRA, E. R.; DANTAS, M. A. R. **Uma Abordagem de Monitoração de Tráfego de Rede utilizando Lógica Difusa**. 2005.

SIMÕES, F. I. W.; HASHIMOTO, F. **Mulher, mercado de trabalho e as configurações familiares do século XX**. Revista Vozes dos Vales da UFVJM: Publicações Acadêmicas – MG – Brasil – Nº 02 – Ano I – 10/2012.

SOUZA, P. M. **Modernização e mudanças estruturais na agricultura brasileira, 1970 a 1995**. Tese (Doutorado em Economia Rural) - Universidade Federal de Viçosa, Viçosa, 2000.

SPRENT, P. **Applied nonparametric statistical methods**. New York: Chapman Hall, 1993.

TIMM, N. H. **Applied multivariate analysis**. New York: Springer, 2002.

TUKEY, J.W. **The problem of multiple comparisons**. Mimeographs Princeton University, Princeton, N.J., 1953.

UVO, B. C. **Analysis and regionalization of the Northern European winter precipitation based on its relationship with the North Atlantic oscillation**. International Journal of Climatology, v.23, 2003.

WAJNMAN, S.; RTOS-NETO, E. L. G. **Participação Feminina na população economicamente ativa no Brasil: alternativas para projeções de níveis e padrões**. Pesquisa Plano Econômico, Rio de Janeiro. V. 24, n 2. Agosto/1994.

WANGEINHEM, A. von. **Reconhecimento de Padrões: Métodos, Técnicas e Ferramentas para Aprendizado e Classificação de Dados**. Acesso em 16 de abril de 2016. Disponível em: <<https://www.inf.ufsc.br/~patrec/agrupamentos.html>>.

ANEXO A

Programa utilizado no software R para Análise de Agrupamentos Hierárquicos dos dados:

```
# Carregando os dados
rm(list=ls(all.names=TRUE))
dados<-read.table("dados.txt", header=T, sep=" ", dec= ",", row.names=" REGIÃO")
dados<-na.omit(dados)
attach(dados)
names(dados)

#### Agrupamento com Distância Euclidiana:
d<-dist(dados[1:6])
dist<-as.dist(d)

#### Método de Ward
cluster_WARD<-hclust(dist,method="ward.D2") #método de ward
ccccluster_WARD<-cophenetic(cluster_WARD) #coeficiente de correlação cofenético
cor(dist,ccccluster_WARD)

## Dendograma com distância Euclidiana, Método de Ward e 4 grupos
plot(cluster_WARD,hang=-1,main=" ",xlab="Anos",ylab="Distância")
rect.hclust(cluster_WARD,k=4,border="red")

#### Critérios Pseudo F e pseudo t² utilizando distância Euclidiana
# Critério Pseudo F - WARD
#Número mínimo de grupos igual a 2 e número máximo igual a 4

require(NbClust)
pseudof_WARD<-NbClust(dados, distance = "euclidean", min.nc=2, max.nc=4, method
= "ward.D2",index="ch")
pseudof_WARD
```