

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

WANDER RODRIGUES DE OLIVEIRA

**ESTUDO DA POPULAÇÃO ECONOMICAMENTE ATIVA NO BRASIL VIA
ANÁLISE DE AGRUPAMENTOS**

UBERLÂNDIA

2019

WANDER RODRIGUES DE OLIVEIRA

**ESTUDO DA POPULAÇÃO ECONOMICAMENTE ATIVA NO BRASIL VIA
ANÁLISE DE AGRUPAMENTOS**

Artigo Científico apresentado ao Curso de Graduação em Administração, da Universidade Federal de Uberlândia, como exigência parcial para a obtenção do título de Bacharel.

Orientador Prof. Dr. Luciano Ferreira Carvalho.

UBERLÂNDIA

2019

Sumário

RESUMO	4
ABSTRACT	5
1 INTRODUÇÃO	6
2 REFERENCIAL TEÓRICO	8
3 METODOLOGIA	13
4 RESULTADOS E DISCUSSÃO	15
5 CONSIDERAÇÕES FINAIS.....	18
REFERÊNCIAS	20

RESUMO

A pesquisa realizada se encontra em um campo ainda pouco estudado, tendo-se analisado com suporte em métodos estatísticos, mais especificamente por meio da análise de agrupamentos, dados da População Economicamente Ativa – PEA da série consolidada, a qual se baseia em informações de sexo, idade, qualificação, distribuição pelos setores econômicos e pelas atividades rurais e urbanas para confeccionar os números apresentados que são indicadores essenciais no mundo moderno. Em base do supramencionado método é possível distinguir padrões que interligam grupos vistos sob o mesmo prisma, ou seja, torna viável estigmatizar os dados a fim de se obter modelos que indicam a similaridade existente dentro do espectro de variáveis selecionadas para o filtro. Doutro modo, pode-se, valendo-se deste sistema, estruturar um agrupamento homogêneo que apresenta características análogas em diferentes graus. Com fundamentação primordialmente estatística, pôde-se averiguar a correlação existente entre o PEA e o PIB estaduais, indicando um padrão de encadeamento deste para com aquele em um movimento de notável similitude, findando na formação de três grupos de enquadramento que apresentaram mesma ordenação com base nos dois indicadores. Desta forma, a pesquisa permite reestruturar a utilização da bruta informação para transformá-la em utensílio basilar, quando acionadas mediante análise fundada em outras variáveis complementares, ou para compreensão do cenário macro-socioeconômico deste país de proporções continentais, o que é de grande valor na gestão estratégica de empresas que buscam enquadramento em nichos diversos, ou para o setor público no desenvolvimento de políticas públicas direcionais.

Palavras-chave: População Economicamente Ativa. Produto Interno Bruto. Análise de Agrupamentos.

ABSTRACT

The research is in a field still little studied, having been analyzed supported by statistical methods, more specifically through cluster analysis, data from the Economically Active Population - EAP of the consolidated series, which is based on information on gender, age, qualification, distribution by economic sectors and rural and urban activities to make the numbers presented that are essential indicators in the modern world. Based on the aforementioned method, it is possible to distinguish patterns that interconnect groups viewed from the same perspective, what means, it is feasible to stigmatize the data in order to obtain models that indicate the similarity within the spectrum of variables selected for the filter. Otherwise, using this system, one can structure a homogeneous grouping that has similar characteristics to varying degrees. Based on a primarily statistical basis, it was possible to verify the correlation between the EAP and the state GDP, indicating a chaining pattern of this with that in a movement of remarkable similarity, ending in the formation of three framing groups that presented the same order based on both indicators. This time, the research allows to restructure the use of raw information to transform it into a basic tool to elaborate new proposals, when triggered by analysis based on other complementary variables, or to understand the macro-socioeconomic scenario of this country of continental proportions, which is of great value in the strategic management of companies that seek framing in different niches, or for the public sector in the development of directional public policies.

Keywords: Economically Active Population. Gross Domestic Product. Cluster Analysis.

1 INTRODUÇÃO

O Brasil é um país com extensões continentais e possui 27 estados. Administrar todos esses estados e conseguir qualidade de vida, manter a economia aquecida e uma boa distribuição de renda é um papel do governo. Conduzir esses fatores para gerar no fim algo positivo para a sociedade é facilitado quando em meio a tanta heterogeneidade, conseguimos ter uma visão homogênea. A Análise de Agrupamento, quando somada ao estudo da População Economicamente Ativa (PEA), pode gerar informações de comportamento de certo grupo, corroborando assim como dito acima, promovendo uma visão ampla e talvez mais assertiva na resolução de problemas sociais e econômicos, que podem estar relacionados a alterações na PEA. A População Economicamente Ativa (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.

Não há um método consolidado em todo o mundo para definir aqueles que fazem parte da PEA. Por exemplo, nos países subdesenvolvidos, o índice inclui os indivíduos que possuem entre 10 e 60 anos, já nos países desenvolvidos geralmente considera-se apenas aquele que possui mais de 15 anos de (SILVA, 2014). Assim, a parte da população que está desempregada e que não busca empregos, como crianças menores que 10 anos, estudantes que não trabalham, donas de casa que exercem apenas funções domésticas não remuneradas, entre outros, é incluída naquilo que se denomina por População Economicamente Inativa. No mundo capitalista, saber qual melhor lugar para se instalar uma empresa, qual estado se encontra seu público-alvo, se esse público possui a renda necessária que encaixa no seu produto, aonde conseguir melhores incentivos, pessoas qualificadas para trabalhar, dentre outros fatores que devem ser analisados na concepção e/ou expansão de uma empresa. A PEA revela então um dado importante nessa análise, quanto de capital humano tem disponível, e de certa forma, quando de possíveis potenciais consumidores (considerando que quanto maior a PEA, maior o consumo por pessoas e mais aquecida será a economia local) existe em determinado local, e quando comparado e agrupado entre os estados, pode servir de suporte estratégico.

No Brasil, de acordo com Instituto Brasileiro de Geografia e Estatística - IBGE, a PEA brasileira compreende 51% da população (IBGE, 2010), apesar de esse número não

considerar aqueles que não trabalham com contrato formal ou carteira assinada. Além disso, é de grande relevância um estudo relacionando a PEA diferenciado a mesma entre os sexos masculino e feminino nos últimos anos, interligando essa informação aos vinte e sete estados da federação, visto que é um tema pouco estudado, visando buscar um resultado que interprete o cenário dos estados brasileiros e assim gerando um resultado que sirva de respaldo para decisões estratégicas empresariais e até para políticas públicas.

Neste sentido, a Análise Multivariada pode ser de grande importância na análise de dados deste tipo uma vez que corresponde a um conjunto de métodos e técnicas que analisam simultaneamente todas as variáveis na interpretação teórica do conjunto de dados. Dentro desta análise, tem-se a técnica de Análise de Agrupamentos, que é classificada como técnica de interdependência por se tratar de uma análise simultânea de todas as variáveis em estudo, na tentativa de se encontrar uma estrutura subjacente ao conjunto inteiro de variáveis, objetivo do presente estudo.

A Análise de Agrupamentos é uma metodologia que consiste em uma técnica estatística que possibilita a criação de agrupamentos de itens diversos, de acordo com as semelhanças apresentadas por esses itens em relação a algum critério de seleção, determinado previamente pelo analista/pesquisador (HAIR JR. *et al.*, 2005). O objetivo da ferramenta é classificar um pequeno número de grupos que tenham a característica de serem homogêneos internamente, heterogêneos entre si e mutuamente excludentes (HAIR JR. *et al.*, 2005). A técnica multivariada de análise de agrupamento (Análise de Clusters - AC) é uma técnica que tem como objetivo básico descobrir os agrupamentos naturais das variáveis, onde estes são feitos com base nas similaridades ou dissimilaridades (caracterizadas por diversas formas de cálculo de “distâncias”). É uma maneira de se obter grupos homogêneos, por um esquema que possibilite reunir os dados em questão em um determinado número de grupos, de modo que exista grande homogeneidade dentro de cada grupo e heterogeneidade entre eles (JONHSON E WICHERN, 1992; CRUZ E CARNEIRO, 2003).

O objetivo deste estudo consiste em agrupar os estados brasileiros e o Distrito Federal por similaridades quanto ao PEA considerando sexo masculino e feminino, interpretando a homogeneidade e heterogeneidade entre os grupos, possibilitando assim um instrumento estratégico para o campo empresarial, suporte para tomadas de decisões de onde instalar filiais de empresas, observando capital humano, desenvolvimento econômico, dentre outros fatores.

A pesquisa justifica-se pela utilidade de se ter informação, que nada mais é do que conhecimento estruturado de uma determinada forma para que possa ser compreendido por um receptor, que, por sua vez, absorve este conhecimento e o interpreta. Ademais, a pesquisa serve de referencial para análises de abordagem estratégica que aperfeiçoam a gestão de recursos empresariais e públicos. Isso porque os dados extraídos do PEA quando observados sob o prisma da Análise de Agrupamentos, possibilita uma melhor perspectiva sobre os componentes multivariados que estão ligados às mais diversas regiões do país, o que interliga informações de cunho puramente econômico com aspectos não-econômicos que impactam aquele de modo bastante efetivo.

Por fim, a presente pesquisa se coloca como instrumento auxiliar na elaboração de planos de abordagem em caráter crítico às operações de expansão e recondicionamento empresarial ou governamental, permitindo a observação de critérios esparsos no que tange aos seus reflexos na funcionalidade orgânica local, sendo isso de enorme interesse tanto para planos de introdução mercadológica quanto para direcionamento de políticas públicas alternativas voltadas para o desenvolvimento local diferenciado.

Este artigo está dividido em quatro partes além desta introdução. No tópico dois foi desenvolvida uma revisão bibliográfica com os principais conceitos sobre o tema. Em seguida, no capítulo três, foi desenvolvida a metodologia da pesquisa. Na parte quatro foram apresentados os resultados da pesquisa e, por fim, foram tecidas as últimas considerações.

2 REFERENCIAL TEÓRICO

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.

Segundo o IBGE (2010) 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. As principais dificuldades de operacionalização do conceito de PEA 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.

Conforme Loesch e Hoeltgebaum, (2012), as técnicas hierárquicas, na maioria das vezes, são utilizadas em análises exploratórias dos dados com o intuito de identificar possíveis agrupamentos e o valor provável do número de grupos. Já para o uso de técnicas não hierárquicas é necessário que o valor do número de grupos já esteja pré-especificado pelo pesquisador, e tem como objetivo encontrar diretamente uma participação de “n” elementos em “k” grupos (clusters), de modo que a participação satisfaça dois requisitos básicos: “coesão” interna (ou “semelhança” interna) e isolamento (ou separação) dos clusters formados. As utilizações dos métodos hierárquicos são recomendadas quando não se sabe o número de grupos, e os não-hierárquicos, quando se sabe. Porém, são métodos complementares, pois para usar o método não hierárquico necessariamente deve ser feito o método hierárquico para saber o número de grupos (LOESCH; HOELTGEBAUM, 2012).

Após definidas as variáveis, espera-se que elas apresentem contribuição equivalente na análise de agrupamento, de modo que a distância entre as amostras não seja alterada com a adoção de características com unidades de medidas distintas, num mesmo conjunto de dados, de forma que as variáveis admitidas apresentem poder discriminatório semelhante, não baseado na amplitude de seus valores. Caso uma determinada característica apresente uma maior amplitude em seus valores, em comparação às demais, ela certamente terá um maior peso na análise. Assim, se todos os dados estiverem num mesmo padrão (unidade) de medida,

as variabilidades de cada característica serão homogêneas ou quase homogêneas, podendo neste caso, utilizar os dados originais. Entretanto, a não ocorrência deste mesmo padrão de medida entre as variáveis sugere uma transformação dos dados (HAIR JR. *et al.*, 2005).

Conforme Hair Jr. (2015), existem duas razões para a padronização dos dados. A primeira visa evitar que as unidades escolhidas para mensurar as variáveis afetem arbitrariamente a similaridade entre os elementos (amostras). A segunda é que a padronização faz com que as variáveis contribuam igualmente na avaliação da similaridade entre amostras. O termo dissimilaridade surgiu em função da relação da distância entre dois pontos P e Q, definida como $d(P,Q)$, pois, à medida que ela cresce, diz-se que a divergência entre os pontos (unidades amostrais) P e Q aumenta, ou seja, tornam-se cada vez mais dissimilares. Os valores de distâncias são geralmente obtidos a partir de informações de “n” unidades amostrais, mensurados em relação a “p” caracteres (variáveis). É necessário especificar um coeficiente de semelhança que indique a proximidade entre os indivíduos sendo importante considerar, em todos os casos semelhantes a este, a natureza da variável (discreta, contínua, binária) e a escala de medida (nominal, ordinal, real ou razão).

Fávero *et. al.* (2009) diz que as distâncias são medidas utilizadas para a representação dos pontos na estrutura de similaridade. Esta medida representa o menor espaço entre dois pontos, sendo uma extensão do teorema de Pitágoras para o caso multidimensional. A distância euclidiana mede a dissimilaridade dos dados, quando não existem grandes diferenças na escala e grandeza das variáveis. Porém, caso exista é necessário usar a distância euclidiana padronizada, a qual padroniza os dados na mesma escala. A distância euclidiana média mede a distância média entre dados, ou seja, calcula a distância euclidiana e divide pela raiz do número de variáveis. A "distância euclidiana" preconiza que a "distância entre duas observações (i e j) corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações (i e j), para todas as p variáveis" (FÁVERO *et. al.*, 2009, p. 201), conforme a fórmula $d_{ii'} = \sqrt{\sum_j (Y_{ij} - Y_{i'j})^2}$, considerando Y_{ij} a observação no i-ésimo indivíduo para a j-ésima característica.

A "distância euclidiana quadrada" consiste exatamente na "distância euclidiana", porém sem extração da raiz quadrada: "A Distância Euclidiana Quadrada tem a vantagem de que não é necessário calcular a raiz quadrada, o que acelera o tempo de computação, e é a distância recomendada para o método de agrupamento Ward" (HAIR JR. *et al.*, 2005, p. 394).

Conforme descrito por Johnson e Wichern (1998), cita-se como medida de dissimilaridade a Distância Euclidiana, que é insatisfatória para muitas situações estatísticas. Isso ocorre devido à contribuição de cada coordenada ter o mesmo peso para o cálculo da distância. Segundo Fávero *et al.* (2009, p. 198), "A maior parte das medidas de distância sofre influência das diferentes escalas ou magnitudes das variáveis de similaridade". Dessa forma, com o objetivo de evitar distorções ou vies, causados pelas diferenças de escalas dessas cinco variáveis de similaridade definidas, providenciou-se a padronização delas antes de se executar a *cluster analysis*, transformando-as em escores padrão (também denominados de escores Z), de maneira a apresentar média 0 (zero) e desvio padrão 1 (um), conforme a fórmula $Z_j = \frac{Y_j - \bar{Y}}{\hat{\sigma}_j}$, em que $\hat{\sigma}_j$ é o desvio-padrão e $\bar{Y}$ é a média associados à j-ésima característica.

A métrica de Mahalanobis é usada quando as variáveis estiverem correlacionadas e/ou quando existem repetições dentro das unidades experimentais. Já a métrica de Minkowski é conhecida como “city-block” para “m=1”, e para “m=2” tem-se a distância euclidiana. Esta métrica é menos afetada pela presença de valores discrepantes na amostra se comparado com a distância euclidiana. A métrica de Canberra e Coeficiente de Czekanowski são distâncias que são definidas para variáveis não negativas. A Distância de Mahalanobis é muito utilizada uma vez que considera as diferenças de variação e a presença de correlação, ou seja, é uma distância que depende das variâncias e das covariâncias amostrais. Esta distância tem sua expressão definida por $D_i^2 = (x_i - \mu)' \Sigma^{-1} (x_i - \mu)$, em que x_i é um vetor de observações da população avaliada, μ_i é o centróide da população avaliada, Σ^{-1} é a inversa da matriz de covariância, sendo também outra medida de distância muito utilizada no caso em que as variáveis analisadas sejam correlacionadas entre si. Atualmente, diversas medidas de dissimilaridade são propostas na literatura, principalmente devido ao grande desenvolvimento e utilização das técnicas multivariadas (KHATTREE E NAIK, 2000).

Os métodos de ligação (*linkage methods*) são adequados para analisar tanto o agrupamento de itens quanto de variáveis (o que não é verdadeiro para todos os procedimentos hierárquicos aglomerativos). O método de ligação completa ou do vizinho mais distante tem como vantagem a afirmação de que todos os itens em um grupo estão dentro de alguma distância máxima (ou similaridade mínima) de cada outro. O método centróide calcula valores máximos ou mínimos, e sim um valor médio dos indivíduos que pertencem ao grupo, ou seja, medida central que pode não ser influenciada por valores discrepantes. No método de ligação

única ou do vizinho mais próximo agrupa indivíduos por sua proximidade, portanto os que têm menores distâncias. Existe o método de ligação média ou *Average Linkage* que mede a distância entre dois grupos como a média das distâncias de todos os pares de itens onde um membro de um par pertence a cada grupo. E por fim, o método de Ward que forma grupos de tal modo que a solução resultante tenha a menor soma de quadrados dentro de grupos, ou seja, maximiza a homogeneidade dentro dos grupos diminuindo as diferenças dos indivíduos (CRUZ E CARNEIRO, 2003).

Os Métodos de Otimização diferem, basicamente, dos hierárquicos, pelo fato dos grupos formados serem mutuamente exclusivos, ou seja, independentes. Um Método de Otimização bastante utilizado no Melhoramento Vegetal foi o proposto no Método de Tocher, citado por Rao (1952), que se baseia no seguinte: a partir da matriz de distâncias, genéticos, identifica-se o par mais próximo que irá constituir o primeiro grupo; concomitantemente, obtém-se θ , que é o maior valor do conjunto de menores distâncias entre os acessos. A partir de então, avalia-se a possibilidade de inclusão de outros acessos neste grupo, respeitando o critério da distância média intragrupo ter que ser inferior a quaisquer distâncias intergrupos. Para tanto, procede-se à comparação entre o acréscimo no valor médio da distância dentro do grupo e o máximo permitido, comumente denominado θ . A inclusão de um acesso em um grupo será possível quando:

$$\frac{D_{(ij)k}^2}{n} < \theta$$

Em que:

$D_{(ij)k}^2$ é a distância entre o grupo **ij** e o acesso **k**, obtida pela expressão:

$$D_{(ij)k}^2 = D_{(ik)}^2 + D_{(jk)}^2, \text{ sendo:}$$

$D_{(ik)}^2$ = distância entre os acessos **i** e **k**;

$D_{(jk)}^2$ = distância entre os acessos **j** e **k**; e

n = número de acessos que constitui o grupo original.

Utilizando todas as variáveis disponíveis e, depois de aplicado algum método hierárquico à matriz de distâncias, os agrupamentos são representados de maneira bidimensional através de

um dendrograma (diagrama bidimensional em forma de árvore). Nele estão dispostas linhas ligadas segundo os níveis de similaridade, que agrupará pares de indivíduos ou de variáveis segundo Everitt (1993) e Landim (2001). O dendrograma ilustra as fusões ou partições efetuadas em cada nível sucessivo do processo de agrupamento, no qual o eixo das abscissas representa os indivíduos, e o eixo das ordenadas as distâncias obtidas após a utilização de uma metodologia de agrupamento. Os ramos da árvore fornecem a ordem das $(n-1)$ ligações, em que o primeiro nível representa a primeira ligação, o segundo a segunda ligação, e assim sucessivamente, até que todos se juntem.

Um método é melhor do que outro quando o dendrograma fornece uma imagem menos distorcida da realidade. É possível avaliar o grau de deformação provocado pela construção do dendrograma calculando-se o Coeficiente de Correlação Cofenético - CCC, (VALENTIN, 2000). Desta forma, o menor grau de distorção, será refletido pelo maior coeficiente cofenético, fornecido pela matriz fenética F, na qual seus valores foram obtidos junto à matriz de distâncias inicial e pela matriz cofenética C, sendo estes os valores obtidos junto à matriz final das distâncias. O maior coeficiente cofenético possui a capacidade de evidenciar melhor a estrutura dos dados, isto é, a existência de grupos. Assim, um CCC menor que 0,7 indica inadequação do método de agrupamento. Caso contrário, o método de agrupamento aplicado foi adequado e os agrupamentos são ditos consistentes. A decisão sobre o número total de conglomerados, a constarem na análise, caberá ao pesquisador, pois esta dependerá de cada pesquisa.

3 METODOLOGIA

Foram obtidas as informações sobre a População Economicamente Ativa, pela página de Dados Abertos do Governo Federal, referente ao período de 2010, (IBGE, 2010), que são dados disponibilizados pelo órgão federal de livre acesso e sem restrição de licenças, patentes ou mecanismos de controle, justificando este período por ser o último ano com dados consolidados, visto que o cálculo da PEA é feito por amostras e possui margem de erro. Dessa forma, trabalhar com dados projetados, como nos anos seguintes podem distorcer os resultados.

As informações sobre a População Economicamente Ativa (PEA) são coletadas pelas pesquisas domiciliares produzidas pelo IBGE. 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. Para o estudo será analisado todos os 26 estados do Brasil e o Distrito Federal.

A "distância euclidiana" preconiza que a "distância entre duas observações (i e j) corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações (i e j), para todas as p variáveis" (FÁVERO *et al.*, 2009, p. 201), conforme a fórmula $d_{ij} = \sqrt{\sum_j (Y_{ij} - Y_{ij})^2}$, considerando Y_{ij} a observação no i-ésimo indivíduo para a j-ésima característica.

Segundo Fávero *et al.* (2009, p. 198), "A maior parte das medidas de distância sofre influência das diferentes escalas ou magnitudes das variáveis de similaridade". Dessa forma, com o objetivo de evitar distorções ou viés, causados pelas diferenças de escalas dessas cinco variáveis de similaridade definidas, providenciou-se a padronização delas antes de se executar a *cluster analysis*, transformando-as em escores padrão (também denominados de escores Z), de maneira a apresentar média 0 (zero) e desvio padrão 1 (um), conforme a fórmula $Z_j = \frac{Y_j - \bar{Y}}{\hat{\sigma}_j}$, em que $\hat{\sigma}_j$ é o desvio-padrão e $\bar{Y}$ é a média associados à j-ésima característica. Contudo, com a análise dos dados utilizou-se como instrumento dessa pesquisa a distância euclidiana padronizada como medida de dissimilaridade e posteriormente o agrupamento.

Existem várias formas de representar esta estrutura de agrupamento. O método da ligação média entre grupos (UPGMA) é o mais utilizado, pois tem como vantagem sobre os demais métodos o fato de considerar médias aritméticas das medidas de dissimilaridade, o que evita caracterizar a dissimilaridade por valores extremos entre os indivíduos considerados (CRUZ E CARNEIRO, 2003). Dessa forma, será feito um comparativo entre os métodos UPGMA e Ward (forma grupos de tal modo que a solução resultante tenha a menor soma de quadrados dentro de grupos, ou seja, maximiza a homogeneidade dentro dos grupos diminuindo as diferenças dos indivíduos), sendo a conclusão com base no método que oferecer uma melhor consistência no agrupamento dos dados, isso por meio da análise dos resultados dos Coeficientes Cofenético de cada um.

Portanto, será gerado um dendograma, cujo objetivo é agrupar os estados conforme suas semelhanças, por meio de níveis (EVERITT, 1993; LANDIM, 2001). Ele ilustra as fusões ou partições efetuadas em cada nível sucessivo do processo de agrupamento. Os ramos da árvore fornecem a ordem das $(n-1)$ ligações, em que o primeiro nível representa a primeira ligação, o segundo a segunda ligação, e assim sucessivamente, até que todos se juntem.

4 RESULTADOS E DISCUSSÃO

Aplicando a análise de agrupamento entre os estados brasileiros e o Distrito Federal, alguns apresentaram características similares, compondo o mesmo grupo. Contudo, foi feita análise comparativa entre a População Economicamente Ativa do sexo masculino e a do sexo feminino, pelo método de agrupamento hierárquico UPGMA e Ward.

Analizando o resultado obtido pelo sexo masculino, as formações dos grupos, tanto pelo método do UPGMA, tanto pelo Ward se igualaram, considerando o ponto de corte sendo o maior degrau no dendograma. Além disso, o Coeficiente de Correlação Cofenético obtido pelo método do UPGMA é 0,956 e do Ward 0,862. Como visto, um CCC acima de 0,7 é considerado bom, portanto, os dois instrumentos de agrupamentos podem ser considerados para este estudo. Porém, se for para escolher algum método o UPGMA é o indicado devido ao seu maior CCC, correspondendo a uma melhor conexão entre os dados. O fato dos agrupamentos se igualarem no UPGMA e Ward reforça a consistência dos resultados, já que dois métodos diferentes, com bons CCC resultaram os mesmos grupos.

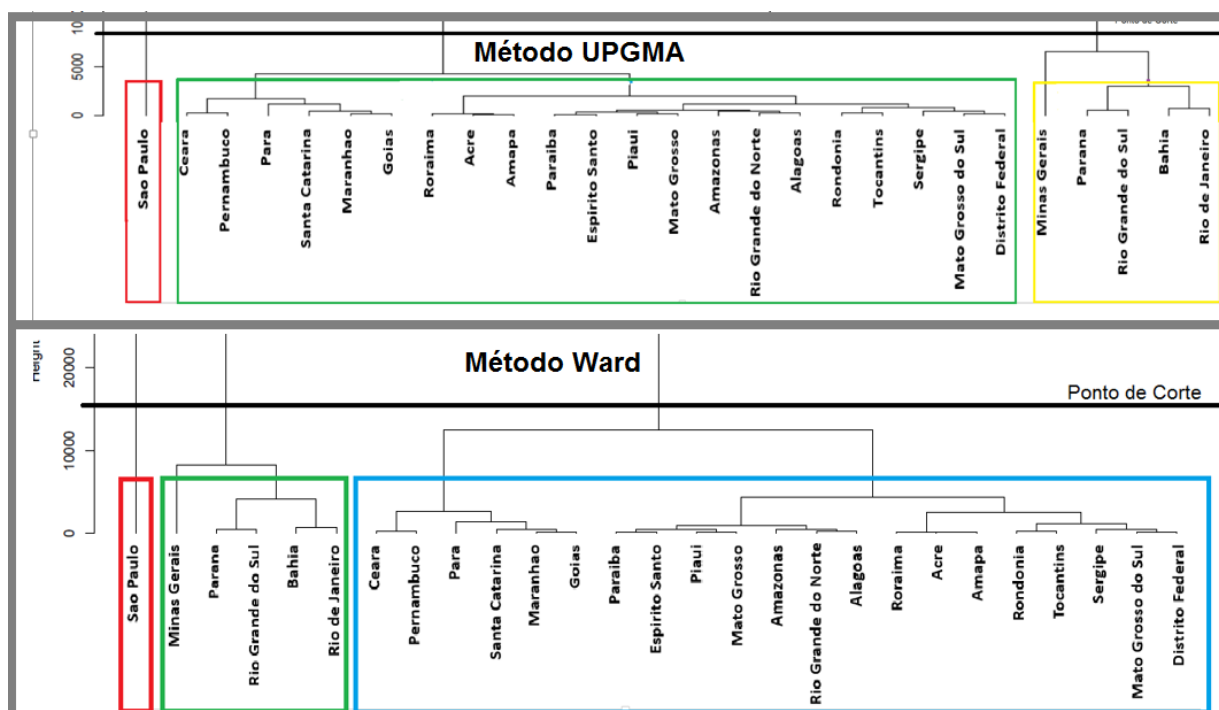


Figura 1 – Dendrograma Método de Agrupamento Ward e UPGMA para o sexo masculino.

Fonte: Próprio autor.

Comparando o sexo feminino com o sexo masculino temos que os grupos formados, por ambos os métodos de agrupamentos foram iguais. Isso revela que independente do estado, o perfil da população economicamente ativa feminino e masculina segue um padrão, ou seja, se um estado se assemelha com o outro no sexo masculino, este é espelhado também o sexo feminino. Contudo, pode verificar que para o sexo feminino a PEA se comporta analogamente a PEA do sexo masculino, ressaltando que a análise de agrupamentos poderia ser feita sem diferir sexos. Na PEA do sexo feminino o CCC foi para UPGMA e Ward: 0,96 e 0,867, respectivamente. Reforçando a análise dita acima, onde até os CCC se diferiram por centésimos.

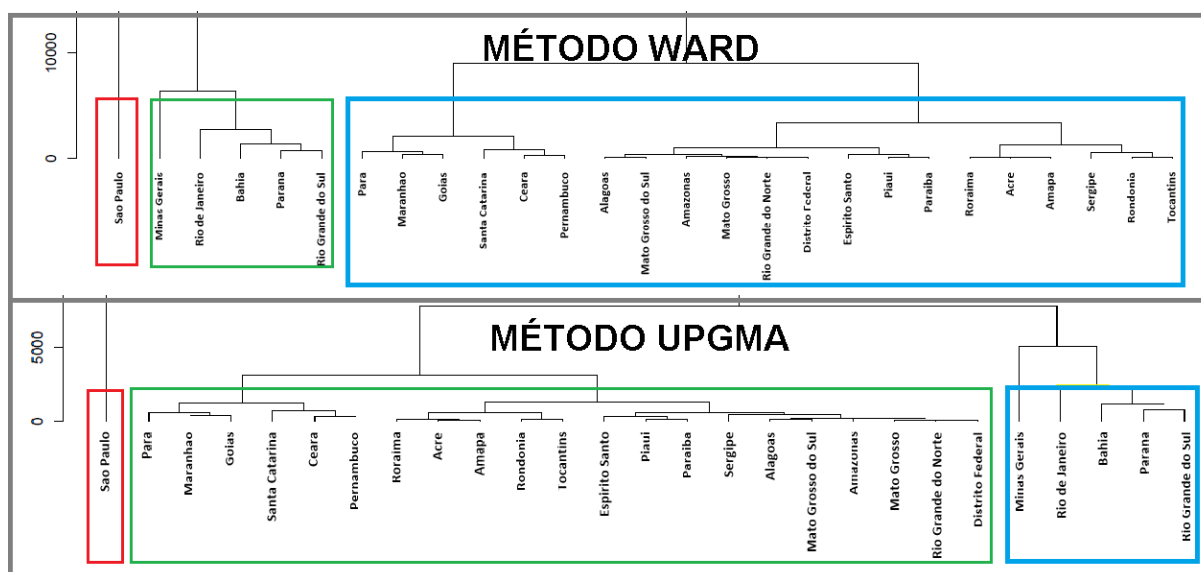


Figura 2 – Dendograma Método de Agrupamento Ward e UPGMA para o sexo feminino.

Fonte: Próprio autor.

São Paulo foi o estado que não se agrupou com nenhum outro, justificando seu número maior de pessoas que se enquadram como PEA, dos dois sexos, sendo o grupo 1. O estado de Minas Geras, Bahia, Rio de Janeiro, Rio Grande do Sul e Paraná formaram o segundo grupo. O restante dos estados e o Distrito Federal formaram o terceiro grupo. Dentro dos grupos, a semelhança entre o número de pessoas da PEA, de cada estado, é muito grande, portanto sendo o grupo 1 com maior número de pessoas que se enquadram como PEA, o segundo grupo estados mais desenvolvidos que o terceiro grupo e que apresentam uma maior quantidade de PEA.

É importante salientar que analisando a lista do Produto Interno Bruto (PIB) dos estados que mais tem participação no PIB nacional, no período de 2002 a 2016, segundo IBGE, se assemelha bastante com a ordem crescente dos estados com maior número de PEA. Sendo, São Paulo o que mais contribui com 32,5% em 2016, Rio de Janeiro, Minas Gerais, Rio Grande do Sul, Paraná, Bahia, com participação de 10,2%; 8,69%; 6,52%; 6,41% e 4,1%, respectivamente, revelando uma possível relação entre o PIB estadual e a PEA estadual.

Tabela 1 – Produto Interno Bruto Brasil 2016.

Estado	PIB (em bilhões)	%	Estado	PIB (em bilhões)	Estado
São Paulo	2.038.005	32,52%	Piauí	41.406	0,66%

Rio de Janeiro	640.186	10,21%	Rondônia	39.451	0,63%
Minas Gerais	544.634	8,69%	Sergipe	38.867	0,62%
Rio Grande do Sul	408.645	6,52%	Tocantins	31.576	0,50%
Paraná	401.662	6,41%	Amapá	14.339	0,23%
Bahia	258.649	4,13%	Acre	13.751	0,22%
Santa Catarina	256.661	4,10%	Roraima	11.011	0,18%
Distrito Federal	235.497	3,76%			
Goiás	181.692	2,90%			
Pernambuco	167.290	2,67%			
Ceará	138.379	2,21%			
Pará	138.068	2,20%			
Mato Grosso	123.834	1,98%			
Espírito Santo	109.227	1,74%			
Mato Grosso do Sul	91.866	1,47%			
Amazonas	89.017	1,42%			
Maranhão	85.286	1,36%			
Rio Grande do Norte	59.661	0,95%			
Paraíba	59.089	0,94%			
Alagoas	49.456	0,79%			

Fonte: IBGE: PIB, 2016.

5 CONSIDERAÇÕES FINAIS

Este artigo teve como objetivo identificar similaridades entre os 26 estados brasileiros e o Distrito Federal, considerando os dados relativos ao PEA e suas variáveis de composição.

Por meio dos resultados apresentados foi possível verificar que os estados que se agruparam possuem quantidades semelhantes de PEA, seja do sexo feminino ou do sexo masculino. A formação de três grupos foi consistente, analisando os resultados o CCC em ambos os métodos foi acima do aceitável, sendo análogo entre os sexos. Isso reflete uma característica entre os estados, em que a PEA feminina e masculina apresenta comportamentos parecidos no período analisado. Além disso, verificou-se uma relação positiva entre o PEA e o PIB, ou seja, os estados com maiores números de pessoas enquadradas como PEA tem um maior Produto Interno Bruto. Revelando que quanto maior a população inserida no mercado de trabalho resulta em uma economia mais ostensiva.

Uma empresa que queira implementar uma nova filial, ou lançar um produto em outro estado, é interessante analisar outros estados em que se agrupam com características bem parecidas

em relação ao PEA que considera: sexo, idade, qualificação, distribuição pelos setores econômicos e pelas atividades rurais e urbanas. Contudo, essa análise serve de instrumento para possíveis decisões estratégicas empresariais, como na área logística. Reforçando que a análise desse estudo agrupa os estados pela sua similaridade, ou seja, por suas semelhanças.

O presente estudo evidência a concentração do PEA e do PIB em apenas 6 estados, 2 grupos que giram a economia do país. Essa concentração pode afetar outros índices importantes para a federação, como por exemplo o Índice de Desenvolvimento Humano (IDH, 2008), que analisa questões relacionada a riqueza, alfabetização, educação, esperança de vida, natalidades, dentre outras. Para o setor governamental é importante identificar fraquezas na economia, estados que precisam de maior atenção, sinais de queda da economia.

Recomenda-se para estudos futuros inserir novas variáveis ao estudo, como salário e etnia, possibilitando verificar se ainda mantem a semelhança entre o sexo feminino e masculino identificado no presente estudo, quanto à formação dos grupos.

Identificar se a migração de pessoas de estados para estados tem haver com a diferença de PEA entre eles, ou seja, se grupos de estados que tem menores índices possui um número maior de evasão de pessoas, e estados agrupados com maiores números de PEA possui menor evasão, ou até mesmo, maior número de migrantes.

REFERÊNCIAS

CRUZ, C. D.; CARNEIRO, P. C. S.; **Modelos Biométricos aplicados ao melhoramento genético**. Viçosa: UFV, v. 2, 585p., 2003.

EVERITT, B.S. 1993. **Cluster analysis**. 3rd ed. London: Heinemann Educational Books, 122p.

FÁVERO, L. P. et al. **Análise de dados: modelagem multivariada para tomada de decisões**. Rio de Janeiro: Campus Elsevier, 2009.

IBGE. **População Economicamente Ativa: Estatísticas Históricas do Brasil**, 2010. Séries Econômicas, Demográficas e Sociais. Disponível em: <https://serieestatisticas.ibge.gov.br/lista_tema.aspx?op=0&de=19&no=7>. Acesso em: 01 de dez. de 2019.

IBGE. **Produto Interno Bruto – PIB**, 2016. Explica – PIB. Disponível em: <<https://www.ibge.gov.br/explica/pib.php>>. Acesso em: 01 de dez. de 2019.

VALENTIN, J. L. **Ecologia numérica: uma introdução à análise multivariada de dados ecológicos**. Interciência, Rio de Janeiro, 2000.

LANDIM, P. M. B. **Geologia Quantitativa: Introdução à análise estatística de dados geológicos multivariados**. Rio Claro - SP, 2001. (Livro em CD-ROM).

KHATTREE, R. and NAIK, D.N. **Multivariate data reduction and discrimination with SAS software**. Cary, SAS Institute, 558p., 2000.

IDH. **Desafios do desenvolvimento: O que é? IDH**. IPEA, 2008. Disponível em: <http://desafios.ipea.gov.br/index.php?option=com_content&view=article&id=2144:catid=28&Itemid=23>. Acesso em: 15 de nov. de 2019.

JOHNSON, R.A. and WICHERN, D.W. **Applied Multivariate Statistical Analysis**. New Jersey-USA: Englewood Cliffs, 642p. 1998.

HAIR JR., J. F. et al. **Análise multivariada de dados**. 5. ed. Tradução Adonai Schlup Sant'Anna e Anselmo Chaves Neto. Porto Alegre: Bookman, 2005.

RAO, C. R.; **An advanced statistical method in biometric research.** New York, Ed. John Wiley e Sons, p.390, 1952.

SILVA, W. S. **População economicamente ativa.** 2014. Disponível em:<
<https://www.infoescola.com/geografia/populacao-economicamente-ativa/>>. Acesso em: 20 de nov. de 2019.