Research, Society and Development, v. 9, n. 2, e60922067, 2020
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v9i2.2067

Comparação de métodos de agrupamentos hierárquicos aglomerativos em indicadores de sustentabilidade em municípios do estado do Pará

Comparison of cluster methods agglomerative hierarchical in sustainability indicators in municipalities of Pará state

Comparación de los métodos de agrupación jerárquica aglomerativa en los indicadores de sostenibilidad de los municipios del estado de Pará

Recebido: 11/11/2019 | Revisado: 11/11/2019 | Aceito: 12/11/2019 | Publicado: 13/11/2019

Diêgo Lima Crispim
ORCID: https://orcid.org/0000-0003-1491-2636
Universidade Federal do Pará, Brasil
E-mail: dlimacrispim@gmail.com
Lindemberg Lima Fernandes
ORCID: https://orcid.org/0000-0003-1806-4670
Universidade Federal do Pará, Brasil
E-mail: lberge@ufpa.br
David Figueiredo Ferreira Filho
ORCID: https://orcid.org/0000-0002-5890-3515
Universidade Federal do Pará, Brasil
E-mail: davydferreira@gmail.com
Bruna Roberta Pereira Lira
ORCID: https://orcid.org/0000-0003-4152-7709
Universidade Federal do Pará, Brasil
E-mail: bruna.lira@itec.ufpa.br

Resumo

Este estudo teve como objetivo comparar o desempenho dos métodos de agrupamento hierárquico aglomerativo utilizando um conjunto de dados composto por diversos indicadores de sustentabilidade referentes aos municípios do estado do Pará. Assim como, definir a quantidade de agrupamentos iniciais a serem constituídos pela utilização dos índices de validade. Para seleção dos indicadores, foi feito um check-list de estudos científicos de abrangência nacional, regional e local que abordam a temática da sustentabilidade. Posteriormente, foi realizado a padronização dos indicadores, devido às unidades e escalas de

medidas diferentes, não interferindo no resultado e possuindo pesos semelhantes no cômputo do coeficiente de similaridade. A medida de dissimilaridade empregada foi a distância euclidiana, e para determinar o método de agrupamento hierárquico foi utilizado o coeficiente aglomerativo (CA). Para estabelecer o número de agrupamento inicial foram empregados índices de validação. O método aglomerativo com melhor desempenho quanto ao (CA) foi de Ward com 0,94, indicando uma melhor força e qualidade entre as técnicas aglomerativos. Os índices de validação Davies Bouldin (DB), Dunn (D) e Silhouette (SIL), indicaram que a quantidade ideal de agrupamentos iniciais a ser formado são 2, todavia o índice PBM constatou que a formação ideal é com 4 grupos. Com relação aos municípios maior homogeneidade, verificou-se que na composição com 2 grupos, as observações mais similares foram m105(Salinópolis) e m109(Santa Izabel do Pará), seguido das observações m102 (Rio Maria) e m144 (Xinguara), todas inseridas no grupo 1.

Palavras-chave: Método de Ward; Distância euclidiana; Índice de validação; Dendrograma.

Abstract

This study aimed to compare the performance of hierarchical agglomerative clustering methods using a data set composed of several sustainability indicators referring to the municipalities of the state of Pará. As well as determining the number of initial groups to be formed by applying of validity indexes. For the selection of indicators, a check-list of national, regional and local scientific studies addressing the theme of sustainability was carried out. Subsequently, the indicators were standardized due to the units and scales of different measures, not interfering in the result and having similar weights in the calculation of the coefficient of similarity. The measure of dissimilarity used was the Euclidean distance, and to determine the hierarchical grouping method was used the agglomerative coefficient (AC). Validation indexes were used to establish the initial grouping number. The agglomerative method with the best performance regarding the (AL) was Ward with 0.94, indicating a better strength and quality among the agglomerative techniques. The Davies Bouldin (DB), Dunn (D) and Silhouette (SIL) validation indexes indicated that the ideal amount of initial clusters to be formed is 2, however the PBM index found that the ideal formation is with 4 groups. Regarding the municipalities with greater homogeneity, it was found that in the composition with 2 groups, the most similar observations were m105 (Salinópolis) and m109 (Santa Izabel do Pará), followed by the observations m102 (Rio Maria) and m144 (Xinguara), all inserted in group 1.

Keywords: Ward's method; Euclidean distance; Validation index; Dendrogram.

Resumen

El objetivo de este estudio fue comparar el desempeño de los métodos de agrupación aglomerativa jerárquica utilizando un conjunto de datos compuesto por varios indicadores de sostenibilidad referidos a los municipios del estado de Pará. Además de determinar el número de grupos iniciales que se formarán mediante la aplicación de índices de validación. Para la selección de los indicadores, se llevó a cabo una lista de control de estudios científicos nacionales, regionales y locales sobre el tema de la sostenibilidad. Posteriormente, los indicadores se estandarizaron debido a las unidades y escalas de las diferentes medidas, no interfiriendo en el resultado y teniendo pesos similares en el cálculo del coeficiente de similitud. La medida de disimilitud utilizada fue la distancia euclídea, y para determinar el método de agrupación jerárquica se utilizó el coeficiente aglomerativo (CA). Se utilizaron índices de validación para establecer el número de agrupación inicial. El método aglomerativo con el mejor rendimiento respecto al (AL) fue Ward con 0,94, lo que indica una mayor resistencia y calidad entre las técnicas aglomerativas. Los índices de validación de Davies Bouldin (DB), Dunn (D) y Silhouette (SIL) indicaron que la cantidad ideal de clusters iniciales a formar es 2, sin embargo el índice PBM encontró que la formación ideal es con 4 grupos. En cuanto a los municipios con mayor homogeneidad, se encontró que en la composición con 2 grupos, las observaciones más similares fueron m105 (Salinópolis) y m109 (Santa Izabel do Pará), seguidas por las observaciones m102 (Río María) y m144 (Xinguara), todas insertadas en el grupo 1.

Palabras clave: El método de Ward; Distancia euclidiana; Índice de validación; Dendrograma.

1. Introdução

Em geral a humanidade faz o uso dos recursos naturais de forma desorganizada e avassaladora, sem respeitar os limites de capacidade de suporte. Nesse contexto, a partir da metade do século XX, os indivíduos começaram a dar relevância pelas questões relacionadas aos meios de produção utilizados para o desenvolvimento econômico e os insumos empregados nas atividades industriais. Assim, nesse período, inicia-se o surgimento do conceito de sustentabilidade associado a definição de limites ao desenvolvimento (LACERDA; CÂNDIDO, 2013).

A temática que envolve a sustentabilidade está em evidência devido à necessidade de

buscar novas formas de organização do processo de produção que favoreçam o prosseguimento da capacidade de estruturas dos recursos naturais, atendendo as necessidades das atuais e futuras gerações, e assegurando o bem-estar e qualidade de vida (CRISPIM et al., 2019; MACEDO et al., 2016).

A sustentabilidade passou a ser um tema bastante discutido e estudado no meio acadêmico, porém, ainda apresenta limitações quanto a determinação ou análise com aplicação de métodos adequados que proporcionem uma avaliação para uma abrangência nacional, regional ou local, assim como dificuldades na utilização de metodologias distintas e subjetivas inerentes aos aspectos sociais, econômicos e ambientais (CARVALHO et al., 2011; ROHAN; BRANCO; SOARES, 2018).

Marzall e Almeida (2000) destacam vários modelos de mensuração da sustentabilidade, a citar dois, o caso do WRI (Hammond et al., 1995) e de Clain (1997), nos quais, foram desenvolvidos com o objetivo de identificar características específicas de aplicabilidade, nos quais, normalmente, permitem a construção de indicadores, representando um determinado conjunto de dados.

Assim, os indicadores são ferramentas importantes, que desde a Rio-92, tem ganhado destaques frente a decisões sobre tendências ambientais globais, regionais e locais, de tal forma, que representam uma tentativa de quantificar e auxiliar na tomada de decisões, análises de desempenhos e cálculos de impactos das atividades antrópicas (REZENDE et al., 2017; VIEIRA, 2019).

Na espacialidade municipal, a utilização dos indicadores pelos agentes públicos e políticos é fundamental para mostrar ações necessárias a serem efetuadas, bem como observar e evidenciar políticas públicas não bem-sucedidas e efetivas, visando uma sustentabilidade em nível local, posteriormente regional, buscando atingir um grau estável de sustentabilidade (REZENDE et al., 2017).

Com base em um conjunto de indicadores, aplicam-se técnicas de análises de agrupamentos, com a finalidade de delinear uma análise, bem como verificar as semelhanças entre as mesmas, assim como perceber as distinções, e uma das técnicas empregadas são através de métodos hierárquicos aglomerativos, dos quais buscam aglomerar os dados mais semelhantes, possuindo, em geral, características mais próximas ou em comum (BEM; GIACOMINI; WAISMANN, 2014).

Portanto, o presente estudo teve como objetivo comparar o desempenho de métodos de agrupamento hierárquico aglomerativo empregando um grupo de dados formados por vários indicadores de sustentabilidade relacionados aos municípios do estado do Pará. Bem como,

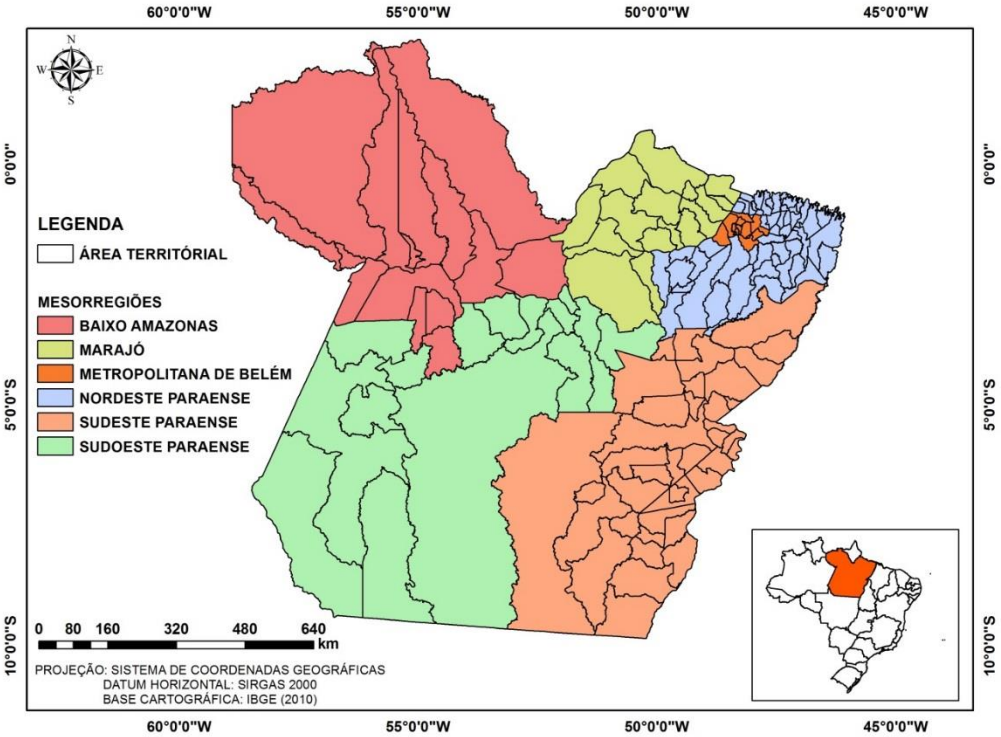
determinar o número de grupos iniciais a serem formados pela aplicação de índices de validade.

2. Metodologia

2.1 Área de estudo

O estado do Pará está localizado na região Norte do Brasil, sendo composto por um total de 144 municípios (Figura 01), distribuídos em seis mesorregiões: Baixo Amazonas; Marajó; Metropolitana de Belém; Nordeste Paraense; Sudoeste Paraense e Sudeste Paraense (IBGE, 2018). Esse estado é o nono mais habitado do Brasil (CHAVES et al., 2017), e o mais povoado da região Norte (BRABO, 2016), com uma população no último censo demográfico de 7.581.051 de pessoas (LOUREIRO et al., 2014; IBGE, 2010) e uma densidade demográfica de 6,07 hab./km² (IBGE, 2010).

Figura 01. Mapa de Localização do estado do Pará



Fonte: Elaborado pelo autor (2019).

O Pará possui uma área de unidade territorial de 1.247.955,238 km² (IBGE, 2018), consistindo no segundo maior estado em extensão territorial do Brasil (BRABO, 2016;

BRABO, 2014; CHAVES et al., 2017), com aproximadamente 14,65% (8.515.767,049 km²), correspondendo cerca de 29,73% do território da Amazônia brasileira (4.196.943,00 km²) (CORDEIRO; ARBAGE; SCHWARTZ, 2017).

De acordo com a classificação climática de Köppen, são caracterizados 3 subtipos climáticos para o estado do Pará: "Af", "Am", "Aw", tais subtipos relacionam-se ao clima tropical chuvoso, particularizando-se por ter temperaturas médias mensais geralmente maiores que 18 °C, se distinguem pelo volume de chuva média mensal e anual (MENEZES et al., 2015). Além disto, possui alta variedade de recursos naturais, por exemplo, os recursos hídricos (MENEZES et al., 2015).

No setor econômico se destaca por ser um dos maiores produtores de estoque pesqueiro no país (BRABO et al. 2016). Também, se destaca no setor agropecuário, no extrativismo vegetal, na exploração de recursos minerais e na indústria do turismo (BRABO, 2014), assim como atividades econômicas ligadas a geração de energia (GOMES; ANDRADE, 2011). Segundo o IBGE (2017), o rendimento nominal mensal domiciliar per capita da população residente do estado do Pará no ano de 2017 foi de R$ 715,00. Também, possui o Produto Interno Bruto (PIB) mais elevado da região Norte do Brasil (ALBUQUERQUE et al., 2016; BRABO, 2016).

2.2 Seleção dos indicadores

A preparação de uma estrutura para fazer a separação e constituição de indicadores de sustentabilidade é essencial para mensurar o cenário de um determinado local quanto ao desenvolvimento sustentável e a prováveis ações ou soluções a serem tomadas (FRAINER et al., 2017). Assim, para seleção das componentes e indicadores de sustentabilidade utilizados na análise de agrupamento, foi feito um check-list de estudos nacionais, regionais e locais que versam sobre a questão da sustentabilidade.

Os fatores empregados para escolha dos indicadores foram suas características quanto a espacialidade municipal, existência e acessibilidade de base de informações, clareza e simplicidade (JUWANA et al., 2012). Para a análise multivariada foram empregados dados secundários disponíveis na página do Programa das Nações Unidas para o Desenvolvimento (PNUD, 2010). Assim, os indicadores que foram selecionados, são usualmente utilizados no Barômetro da sustentabilidade, Índice de Desenvolvimento Sustentável para Municípios (IDSM) e Índice de Desenvolvimento Sustentável para Municípios Participativos (IDSMP) (Quadro 01).

Os indicadores foram empregados para uma análise multivariada, especialmente a técnica de análise de agrupamento, com o propósito de agrupar observações (municípios) em uma pequena quantidade de grupos similares formados pelas semelhanças entre si e diferenças com os demais. Desse modo, para realização dos cálculos matemáticos, foi usado o programa estatístico (R) versão 3.6.1, que consiste em um software livre, de linguagem e um ambiente computacional para cálculos estatísticos e gráficos.

Quadro 01. Componentes e indicadores empregados para análise de agrupamento segundo dados do PNUD (2010)

Componente	Indicador	Referência
Social	Esperança de vida ao nascer	Macedo et al. (2016)
	Expectativa de anos de estudo	Rezende et al. (2017)
	IDHM	Jeunon e Santos (2014)
	Mortalidade Infantil	Cardoso, Toledo e Vieira (2016)
	Mortalidade até 5 anos de idade	Cardoso, Toledo e Vieira (2016)
	População total	Macedo et al. (2016)
Econômico	Índice de Theil	Jeunon e Santos (2014)
	Índice de GINI	Cardoso, Toledo e Vieira (2016)
	Porcentagem de extremamente pobres	Cardoso, Toledo e Vieira (2016)
	Renda per capita	Macedo et al. (2016)
Habitação	Porcentagem da população em domicílios com banheiro e água encanada	Sousa, Santos e Sousa (2016)
	Porcentagem de pessoas em domicílios com energia elétrica	Cardoso, Toledo e Vieira (2016)
Saneamento Ambiental	Porcentagem da população em domicílios com água encanada	Sousa, Santos e Sousa (2016)
	Porcentagem da população em domicílios com coleta de lixo	Cardoso, Toledo e Vieira (2016)
	Porcentagem de pessoas em domicílios com abastecimento de água e esgotamento sanitário inadequados	Rezende et al. (2017)

Fonte: Dados da pesquisa (2019).

2.3 Análise de agrupamento

A análise de agrupamento é uma técnica multivariada que se constitui em separar observações em grupos, fundamentando-se nas particularidades que essas observações possuem (NEVES; VANZELLA, 2017), formando componentes semelhantes e homogêneas entre si em um dado grupo, porém diferentes das observações de outros agrupamentos (RODRIGUES; FACHEL; PASSUELLO, 2012).

Os métodos de agrupamento podem ser divididos em duas classificações, que são as técnicas hierárquicas e não hierárquicas. Segundo Comunello et al. (2013), os métodos hierárquicos distinguem dos não hierárquicos, por não constituírem um número específico de agrupamentos, porém, geram grupos por uma sequência crescente de divisões ou ligações contínuas de grupos.

Segundo Seidel et al. (2008), a técnica hierárquica fundamenta-se na definição de uma hierarquia ou composição em forma de árvore. Assim, o agrupamento hierárquico liga as amostras por suas associações, formando uma reprodução gráfica chamada de dendrograma, no qual as observações similares, baseada nos indicadores empregados no estudo, são reunidos ou agrupados entre si.

Os métodos hierárquicos são compostos por duas classes naturais de algoritmos para formação dos agrupamentos, por exemplo, agrupamento aglomerativo (agglomerative clustering) e agrupamento por divisão (divisive clustering) (MINGOTI, 2013). Nesse estudo, foram empregadas inicialmente 4 técnicas hierárquicas aglomerativas (Single Linkage, Average Linkage, Complete Linkage e Ward) para determinar municípios homogêneos, selecionando o método com maior valor apresentado no Coeficiente Aglomerativo (CA).

2.4 Padronização dos indicadores

Fávero et al. (2009), recomendam que os dados utilizados para análise de agrupamento necessitam que sejam padronizados, visto que, possuem escalas e unidades de medidas distintas, tendo em vista que podem modificar a estrutura do agrupamento. Assim, a padronização dos indicadores empregados nesse processo pode atenuar esse problema (HAIR et al., 2009).

De acordo com Fávero et al. (2009), a padronização das unidades de medidas dos indicadores proporcionará que esses possuam pesos semelhantes no cálculo do coeficiente de similaridade. Assim, para tal processo foi empregado a técnica denominada range, no qual faz com que o indicador tenha uma variação entre 0 a 1 (Equação 1):

$$Y = \frac{X\ - mínimo}{amplitude\ total} \quad (1)$$

Em que: Y é o valor obtido através do método range; X é o valor da variável observada.

2.5 Distância euclidiana

A distância euclidiana possui especificidades métricas, assim como é bastante empregada para variáveis divisivas ou separativas e medidas em uma escala de intervalo (CRISPIM et al., 2019; EVERITT; DUNN, 2010). Assim, essa medida é constantemente utilizada quando os indicadores ou variáveis são completamente quantitativos (CRISPIM et al., 2019; SILVA et al., 2018).

Para determinar a distância entre as observações objeto de estudo, foi aplicada a distância euclidiana. Assim, na Equação 1 é apresentada a solução geral da medida da distância euclidiana (Dx,y) entre o componente x e o componente y, em um espaço n-dimensional, estabelecida por:

$$D_{X,y} = \sqrt{\sum_{k=1}^{n} \left(P_{kx} - P_{ky}\right)^2} \quad (2)$$

Em que: Dx,y é a distância entre os componentes x e y; Pkx consiste no valor do indicador Pk para o componente (x); Pky é o valor do componente para a observação (y); executa-se a soma para todas as variáveis (p) consideradas.

2. 6 Coeficiente Aglomerativo

Para analisar as soluções de grupos originados pelas técnicas hierárquicas foram visualizados os dendrograma, que retrata a estrutura dos grupos de dados. Logo, a altura em que dois agrupamentos se unem no dendrograma refere-se à distância entre esses dois grupos (ROTH et al., 2016).

Nesse estudo, foi utilizado uma métrica para analisar a qualidade do dendrograma e, por seguinte, a força da estrutura dos clusters nos dados. Assim, o critério aplicado foi Coeficiente Aglomerativo (CA), conforme metodologia sugerida por Roth et al. (2016) e

Abson et al. (2014). Essa métrica verifica a precisão da composição da estrutura de agrupamento do conjunto de dados (SHARMA et al., 2018).

Conforme esse coeficiente, para cada observação i, $d_{(i)}$ é sua dissimilaridade com relação ao primeiro grupo que foi introduzido, separado pela dissimilaridade na fase final do algoritmo (ROTH et al., 2016; ABSON et al., 2014). Assim, o CA é determinado pela seguinte expressão:

$$CA = \frac{1}{n}\sum_{i=1}^{n}(1 - d(i)) \quad (3)$$

Em que: n é a quantidade total de observações do grupo de dados.

O (CA) assume valores que varia de 0 a 1, em que valores próximos de 1 indicam para uma estrutura de cluster satisfatória para o conjunto de dados. No programa R versão 3.6.1., o (CA) é computado utilizando a rotina Agnes no cluster de pacotes (ROTH et al., 2016).

2.7 Índices de validação

Para evitar prováveis indagações com relação às subjetividades na escolha dos grupos iniciais, foram utilizados índices de validação com o propósito de verificar os resultados formados pelos algoritmos de agrupamento (STARCZEWSKI, 2017; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002). Assim, esses métodos de análises dos resultados na composição dos grupos apresentam quatro particularidades fundamentais:

1. verificar se há uma estrutura não aleatória nos dados com o objetivo de evitar tendência nos resultados;
2. definir o número de agrupamentos iniciais;
3. avaliar como um resultado de um agrupamento se organiza a um dado conjunto de dados, sendo essa a única informação disponível;
4. analisar como bem localizados estão os municípios dentro dos grupos conforme as partições obtidas segundo outras fontes de dados.

Neste estudo, foram aplicados os índices de validade Davies Bouldin (DB), Dunn (D), Silhouette (SIL) e PBM (Quadro 02). Esses índices possuem especificidades diferentes, visto

que, o menor valor obtido para o grupo de dados no índice Davies Bouldin (DB) aponta um particionamento apropriado de um conjunto de dados, enquanto nos índices Dunn (D), Silhouette (SIL) e PBM, o maior valor alcançado no índice indica o melhor particionamento de um certo conjunto de dados (STARCZEWSKI, 2017; PARCHURE; GEDAM, 2019).

Quadro 02. Índices de validação para definição do número inicial de agrupamentos

Índice de validade	**Equação**		**Fonte**
Davies Bouldin (DB)	$DB = \frac{1}{K}\sum_{i=1}^{K} Ri,qt$	(3)	STARCZEWSKI (2017)
Dunn (D)	$Vd = \min_{1\le s\le K}\{\min_{1\le t\le K, t\ne s}\{\frac{\delta i\{Cs, Ci)}{\max_{1\le k\le K}\Delta j(Ck)}\}\}$	(4)	STARCZEWSKI (2017)
Silhouette (SIL)	$\bar{s}(k) = \frac{1}{n}\sum_{i=1}^{n} s(i)$	(5)	SILVA et al. (2018)
PBM	$PBM(k) = (\frac{1}{k} * \frac{E1}{Ek} * Dk)^2$	(6)	STARCZEWSKI (2017)

Fonte: Adaptado de STARCZEWSKI (2017); PARCHURE; GEDAM (2019).

3. Resultados e Discussão

Inicialmente foi realizado a determinação do número inicial de grupos, em que os índices de validação apresentaram a composição de dois agrupamentos distintos, no qual os métodos Davies Bouldin (DB), Dunn (D) e Silhouette (SIL) estabeleceram que a quantidade apropriada de grupos a serem constituídos são 2 (Tabela 01). Já o índice de validação PBM indicou uma formação com 4 agrupamentos. Assim, nesse estudo, considerou-se às duas proposições distintas de grupos iniciais sugeridos pelos índices de validação.

Tabela 01: Determinação do número ideal de agrupamentos por meio de índices de validação

Grupo	**Davies Bouldin (DB)**	**Dunn (D)**	**Silhouette (SIL)**	**PBM**
2	1,18	0,09	0,33	10,98
3	1,21	0,08	0,24	9,98
4	1,34	0,08	0,24	11,65
5	1,25	0,08	0,21	8,18
6	1,24	0,08	0,20	8,61
7	1,35	0,08	0,19	6,58
8	1,41	0,08	0,18	5,45
9	1,46	0,08	0,16	4,56
10	1,34	0,08	0,15	3,97

Fonte: Dados da pesquisa (2019).

O (CA) entre os métodos de agrupamento, obtidos com base na matriz da distância euclidiana, alternou entre 0,58 (Single Linkage/Ligação Simples) e 0,94 (Ward). Assim, pode-se inferir que a estrutura de agrupamento hierárquico de Ward obteve a melhor força e qualidade entre os métodos aglomerativos empregados (Tabela 02). Assim, os resultados indicam que essa técnica é a mais apropriada para ser empregada na base de dados relacionadas aos indicadores de sustentabilidade para o estado do Pará.

Tabela 02: Definição do melhor método de agrupamento para agrupar as variáveis com base no Coeficiente Aglomerativo (CA)

Método de agrupamento	Coeficiente Aglomerativo (CA)
Single Linkage	0,58
Average Linkage	0,78
Complete Linkage	0,86
Ward	0,94

Fonte: Dados da pesquisa (2019).

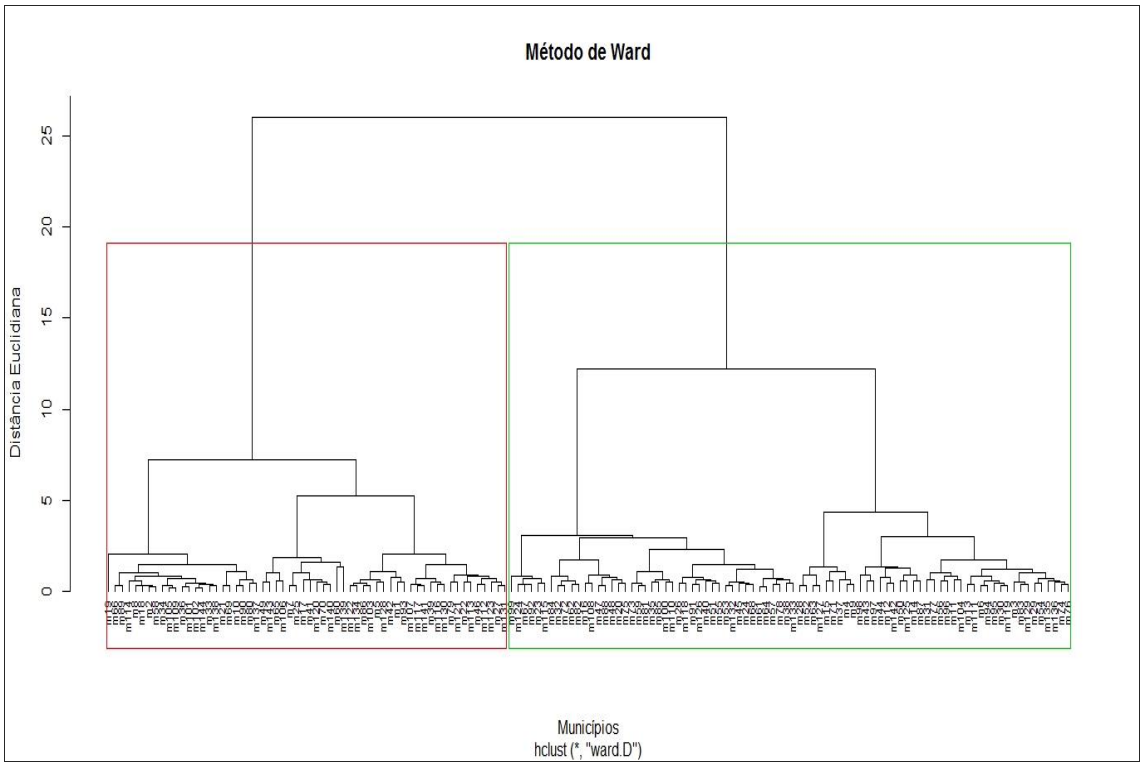
O resultado do CA (0,94) alcançado nesse estudo para o método de Ward indica uma ordenação de agrupamento forte nos conjuntos de dados. Além disso, esse desempenho do CA obtido nessa pesquisa foi superior ao obtido por Bertussi (2008), em que obteve um CA=0,826. Ademais, o desempenho do coeficiente Aglomerativo (CA) nesse trabalho foi quase similar ao alcançado na pesquisa de Abson et al. (2014), em que alcançaram um (CA) = 0, 95, e um pouco inferior ao resultado obtido por Roth et al. (2016), com (CA) = 0,99.

Em seguida, foi aplicado a técnica de análise de agrupamento hierárquico para um conjunto de dados concernentes aos municípios do estado do Pará, com n= 144 (observações) e p= 15 (indicadores), sendo os dados inseridos e computados no Programa R versão 3.6.1. Posteriormente, foi empregado o método hierárquico Aglomerativo (Ward), usando como medida de dissimilaridade a distância Euclidiana no grupo de indicadores padronizados. Assim, o resultado desse procedimento pode ser observado no diagrama de árvore (Figura 02).

Pode-se inferir no dendrograma ilustrado na Figura 02, que o agrupamento 1, foi constituído por 60 observações (municípios), o que representa 41,67% do total de municípios, enquanto o segundo grupo foi composto por 84 observações, correspondendo a 58,83% do universo total da amostra. Desse modo, percebe-se que o grupo 2, agrupou o maior número de

municípios.

Figura 02. Dendrograma com dois agrupamentos gerado pelo método hierárquico de Ward e distância euclidiana



Nota: os municípios do estado do Pará foram numerados conforme sua ordem alfabética e abreviados com a letra m.
Fonte: Dados da pesquisa (2019).

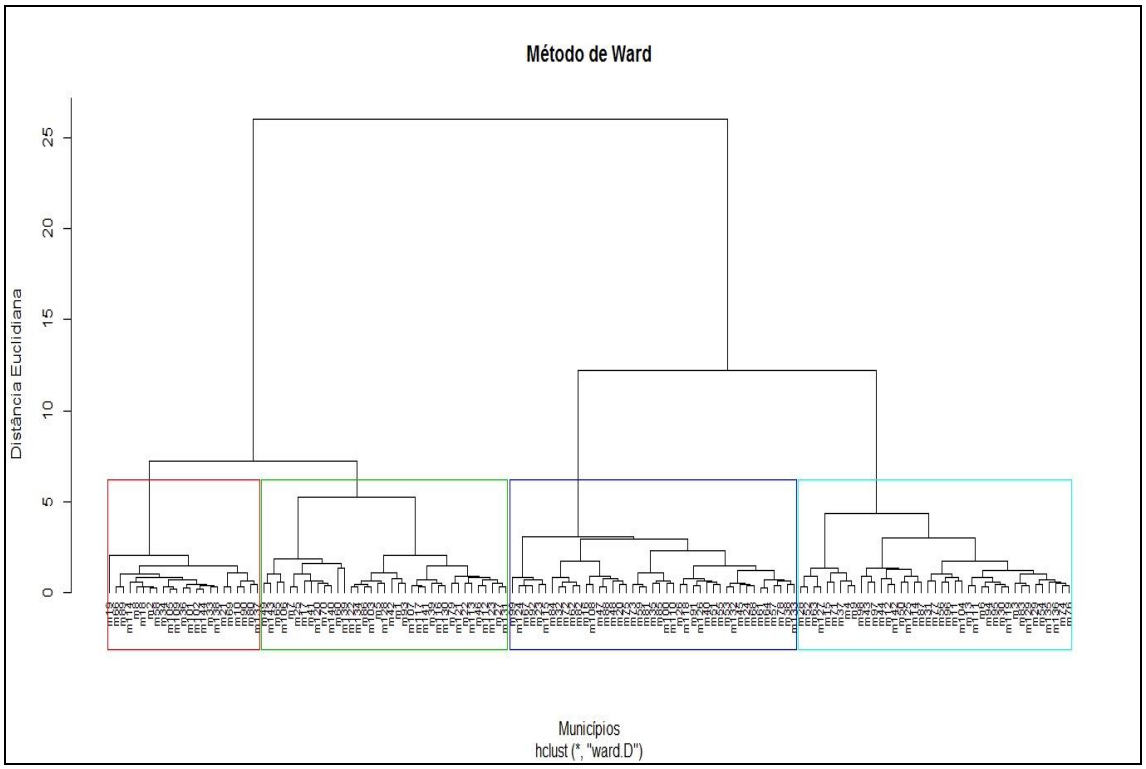
Quanto aos municípios mais similares ou distintos, pode-se depreender que a observação denominada m19 (Belém) apresenta maior dissimilaridade com os outros municípios estudados no grupo 1, visto que, não possui vizinhança para uma dada altura no eixo da distância euclidiana. Além disso, constata-se no grupo 2, que a observação m99 (Primavera) indica também maior dissimilaridade para os demais municípios que integram esse grupo. Logo, esses municípios apresentam dispersão com os demais dentro dos seus respectivos agrupamentos.

Os municípios que apresentaram serem mais similares, segundo a análise, são m105(Salinópolis) e m109(Santa Izabel do Pará), posteriores das observações m102 (Rio Maria) e m144 (Xinguara), respectivamente, do grupo 1. Enquanto, no grupo 2, os municípios mais semelhantes foram o m61 (Jacundá) e m64 (Mãe do Rio). Deste modo, presume-se que esses apresentam maior homogeneidade entre si quanto aos desempenhos dos indicadores de sustentabilidade para o período analisado e características diferentes dos demais.

Quando se avalia outro grau de similaridade para uma menor distancia euclidiana conforme apresentado no dendrograma apresentado na Figura 03, observa-se a formação de 4

grupos. Assim, constata-se que no primeiro agrupamento foram considerados 37 municípios, com uma representatividade de 25,7% do total de observações; para o segundo, foram sugeridos 23 municípios, com uma representação de 15,9% do universo de observações; ao passo que o grupo 3 foi composto por 41 municípios, indicando um percentual de 28,5% do total; e, por fim, o quarto agrupamento, foi constituído por 43 municípios, retratando 29,9% do total de municípios.

Figura 03. Dendrograma com 4 agrupamentos formado pelo método hierárquico de Ward e distância euclidiana



Nota: os municípios do estado do Pará foram numerados conforme sua ordem alfabética e abreviados com a letra m.
Fonte: Dados da pesquisa (2019).

No tocante aos municípios mais semelhantes ou que não apresentaram muita similaridade conforme apresentado no dendrograma com 4 grupos, constata-se um comportamento próximo aos resultados obtidos para o dendrograma com 2 duas formações de agrupamento, onde o município de Belém-PA intitulado de m19 apresentou uma maior dispersão quando comparado com as outras observações do seu grupo, já no grupo 4 pode-se inferir que a observação denominada m99 (Primavera) também teve um comportamento disperso dentro do seu agrupamento.

Quanto aos municípios que apresentaram semelhanças, pode-se perceber que as observações m122 (São Geraldo do Araguaia) e m134 (Tomé-Açu) no grupo 1, possuem similaridades, assim como os municípios denominados m28 (Breves) e m52 (Gurupá), que

estão localizados na mesorregião do Marajó, possuem similaridades nos desempenhos dos indicadores de sustentabilidade, e pertencem ao grupo 3.

4. Conclusão

O resultado alcançado quanto ao coeficiente aglomerativo (CA) demonstrou que dentre os métodos hierárquicos aglomerativos o que apresentou um melhor desempenho foi a técnica de Ward com (CA) = 0,94, indicando uma ordenação de agrupamento forte nos conjuntos de dados.

A utilização dos índices de validação possibilitou a definição da quantidade de agrupamentos iniciais, evitando questionamento quanto a subjetividade na seleção dos grupos iniciais. Assim, os índices Davies Bouldin (DB), Dunn (D) e Silhouette (SIL) determinaram que o número adequado de grupos a serem compostos são 2, porém o índice de validação PBM demonstrou uma composição com 4 grupos.

No tocante aos municípios mais similares ou dissimilares nos dendrograma para as formações com 2 e 4 grupos, constatou-se que na composição com 2 agrupamentos, as observações mais semelhantes foram m105(Salinópolis) e m109(Santa Izabel do Pará), posteriores das observações m102 (Rio Maria) e m144 (Xinguara), todas inseridas no grupo 1.

Quanto ao dendrograma com 4 formações de agrupamentos, observou-se que os municípios denominados m122 (São Geraldo do Araguaia) e m134 (Tomé-Açu) inseridos no grupo 1, apresentam similaridades e homogeneidade, bem como as observações denominadas m28 (Breves) e m52 (Gurupá), integrantes no grupo 3. Essas últimas estão localizadas na mesorregião do Marajó e apresentam desempenhos similares nos indicadores de sustentabilidade, sendo uma possível explicação para suas ligações entre si.

Uma proposição de estudos futuros será comparar métodos de agrupamentos não hierárquicos (k-means e Fuzzy C-Means) e hierárquico (Ligação Simples, Ligação Média, Ligação Completa, Centroide e Ward), empregando a medida de dissimilaridade (distância euclidiana), com o propósito de analisar possíveis similaridades ou dissimilaridades entre os municípios do estado do Pará, empregando dados de indicadores de sustentabilidade.
Posteriormente, os agrupamentos formados com aplicação do método selecionado serão espacializados por um software de Sistema de Informação Geográfica (SIG) para facilitar na leitura e realização de inferências a serem realizadas com aplicação dos métodos de agrupamentos.

Para verificar os dendrograma que serão gerados e as matrizes de dissimilaridade, será

aplicado o coeficiente de correlação cofenética (CCC), com a finalidade de constatar se esses demonstrarão um ajuste apropriado. Logo, a técnica hierárquica e não hierárquica que será selecionada vai ser a que obter o maior valor de (CCC).

Agradecimentos

Congratulamos a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão de bolsa acadêmica do Processo de n.º 1848167. Ao Programa de Pós-Graduação em Engenharia Civil da Universidade Federal do Pará pelo apoio para execução do estudo.

Referências

Abson, D. J., von Wehrden, H., Baumgärtner, S., Fischer, J., Hanspach, J., Härdtle, W., & Walmsley, D. (2014). Ecosystem services as a boundary object for sustainability. *Ecological Economics*, 103, 29-37.

Albuquerque, N. C., Portal, L. C., Nogueira, L. M. V., & Rodrigues, I. L. A. (2016). Busca ativa de hanseníase por meio de educação em saúde entre populações ribeirinhas. *Rev Enferm UFPE*, 10(7), 2634-40.

Bem, J. S. D., Giacomini, N. M. R., & Waismann, M. (2015). Utilização da técnica da análise de clusters ao emprego da indústria criativa entre 2000 e 2010: estudo da Região do Consinos, RS. *Interações (Campo Grande)*, 16(1), 27-41

Bertussi, G. L. (2008). *Análise do processo de convergência de renda na América Latina e no leste Asiático.* 2008. 90 f. Dissertação (Mestrado do Centro de Desenvolvimento e Planejamento Regional) - Faculdade de Ciências Econômicas, Universidade Federal de Minas Gerais, 2008.

Brabo, M. F. (2014). Piscicultura no Estado do Pará: situação atual e perspectivas. *Acta Fish. Aquat. Res.*, 2(1), 1-7.

Brabo, M. F.; Pereira; L. F. S.; Santana; J. V. M.; Campelo; D. A. V.; Veras; G. C. (2016). Cenário atual da produção de pescado no mundo, no Brasil e no estado do Pará: ênfase na aquicultura. *Acta Fish*, 4(2), 50-58.

Cardoso, A.S.; Toledo, P.M.; Vieira, I.C.G. (2016). Barômetro da sustentabilidade aplicado ao município de Moju, estado do Pará. *Revista Brasileira de Gestão e Desenvolvimento Regional*, 12(1), 234-263.

Carvalho, J. R. M., Curi, W. F., de Araújo Carvalho, E. K. M., & Curi, R. C. (2011). Proposta e validação de indicadores hidroambientais para bacias hidrográficas: estudo de caso na sub-bacia do alto curso do rio Paraíba, PB. *Sociedade & Natureza*, 23(2), 295-310.

Chaves, E. C., Costa, S. V., Flores, R. L. D. R., & Neves, E. O. S. D. (2017). Índice de carência social e hanseníase no Estado do Pará em 2013: análise espacial. Epidemiologia e Serviços de Saúde, 26, 807-816.

Clain, N. (1997). Les indicateurs de développement durable en agriculture, aspects écologiques et environnementaux. *Paris: Université de Paris 7.*

Comunello, É., de Araújo, L. B., Sentelhas, P. C., Araújo, M. F. C., dos Santos Dias, C. T., & Fietz, C. R. (2013). O uso da análise de cluster no estudo de características pluviométricas. *Sigmae*, 2(3), 29-37.

Cordeiro, I. M. C. C., Arbage, M. J. C., & Schwartz, G. (2017). *Nordeste do Pará: configuração atual e aspectos identitários.* Embrapa Amazônia Oriental-Capítulo em livro científico (ALICE), 19-58.

Crispim, D. L.; Fernandes, L. L.; Albuquerque, R. L. De O. (2019). Aplicação de técnica estatística multivariada em indicadores de sustentabilidade nos municípios do Marajó-PA. *Revista Principia - Divulgação Científica e Tecnológica do IFPB*, (46), 145-154.

Fávero, L. P. L., Belfiore, P. P., Silva, F. L. D., & Chan, B. L. (2009). Análise de dados: modelagem multivariada para tomada de decisões.

Frainer, D. M., Souza, C. C. D., Reis Neto, J. F., & Castelão, R. A. (2017). Uma aplicação do Índice de Desenvolvimento Sustentável aos municípios do estado de Mato Grosso do Sul. *Interações (Campo Grande)*, 18(2), 145-156.

Gomes, S.C.; Andrade, L.C. (2011). *Análise espacial do crescimento econômico dos municípios paraenses no período 2002- 2006.*

Instituto Brasileiro de Geografia e Estatística. (2017). *IBGE divulga o rendimento domiciliar per capita 2017.* Rio de Janeiro: IBGE. Disponível em:https://agenciadenoticias.ibge.gov.br/agencia-sala-deimprensa/2013agencia denoticias/releases/20154-ibge-divulga-o-rendimento-domiciliar-per-capita2017.html. Acesso em: 16 abr. 2018.

Instituto Brasileiro de Geografia e Estatística. (2010). *Censo Demográfico*. Rio de Janeiro: IBGE.

Instituto Brasileiro de Geografia e Estatística. (2018). *Pará [Internet]*. Disponível em: https://cidades.ibge.gov.br/brasil/pa/panorama. Acesso em: 15 abr. 2018.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados.* Bookman Editora.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Cluster validity methods: part I. *ACM Sigmod Record*, 31(2), 40-45.

Hammond, A., & World Resources Institute. (1995). *Environmental indicators: a systematic approach to measuring and reporting on environmental policy performance in the context of sustainable development* (Vol. 36). Washington, DC: World Resources Institute.

Jeunon, E. E., & Santos, L. M. (2014). Indicadores de Desempenho na Gestão de Projetos Sociais Sustentáveis: Proposição de Modelo para os Centros Vocacionais Tecnológico. Revista *Gestão & Tecnologia*, *14*(2), 225-248.

Juwana, I., Muttil, N., & Perera, B. J. C. (2012). Indicator-based water sustainability assessment—A review. *Science of the Total Environment*, *438*, 357-371.

Lacerda, C. D. S., & Cândido, G. A. (2013). Modelos de indicadores de sustentabilidade para gestão de recursos hídricos. Gestão sustentável dos recursos naturais: uma abordagem participativa [online]. Campina Grande: EDUEPB, 13-30.

Loureiro, R. D., Saraiva, J. M., Saraiva, I., Senna, R. C., & Fredó, A. S. (2014). Estudo dos eventos extremos de precipitação ocorridos em 2009 no estado do Pará. *Rev Bras Meteorol*, *29*, 83-94.

Macedo, L. O. B., Cândido, G. A., de Aguiar Costa, C. G., & da Silva, J. V. F. (2016). Avaliação da sustentabilidade dos municípios do estado de Mato Grosso mediante o emprego do IDSM–Índice de Desenvolvimento Sustentável Para Municípios. *Revista Brasileira de Gestão e Desenvolvimento Regional*, *12*(3), 323-345.

Marzall, K., & Almeida, J. (2000). INDICADORES DE SUSTENTABILIDADE PARA AGROECOSSISTEMAS Estado da arte, limites e potencialidades de uma nova ferramenta para avaliar o desenvolvimento sustentável. *Cadernos de Ciência & Tecnologia*, *17*(1), 41-59.

Menezes, F. P., Fernandes, L. L., & da Rocha, E. J. P. (2015). O uso da estatística para regionalização da precipitação no Estado do Pará, Brasil. *Revista Brasileira de Climatologia*, *16*(11), 64-71.

Mingoti, S. A. (2005). *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Editora UFMG.

Neves, M. R. D. S., & Vanzella, E. (2017). Tempo de máquinas paradas numa indústria têxtil: um estudo por meio de análise de agrupamento. *Revista Mangaio Acadêmico*, *2*(1), 58-65.

Parchure, A. S., & Gedam, S. K. (2019). Homogeneous regionalization via L-moments for Mumbai City, India. Meteorology Hydrology and Water Management. *Research and Operational Applications*, *7*(2), 73-83.

Programa das Nações Unidas para o Desenvolvimento. (2010). *Ranking IDHM Municípios 2010.* Disponível em:http://www.pnud.org.br/atlas/ranking/raking-IDHM-Municipios-2010.aspx. Acesso em: 21 out. 2019.

Rezende, G. B., G. B., Cândido, G. A., Rezende, H. L., & Silva, F. P. (2017). Sustentabilidade de Barra do Garças sob a ótica do índice de desenvolvimento sustentável para municípios. *Desenvolvimento em Questão*, *15*(39), 203-235.

Rodrigues, A., Fachel, J. M. G., & Passuello, A. C. (2014). Estatística espacial e análise de cluster em dados de desastres naturais: mapeamento das inundações no Rio Grande do Sul entre 2003 e 2009. *Revista de Iniciação Científica*, *10*(1), 48-67.

Rohan, U., Branco, R. R., & Soares, C. A. P. (2018). Potencialidades e Limitações dos Instrumentos de Mensuração da Sustentabilidade. *Engenharia Sanitária e Ambiental*, *23*(5), 857-869.

Roth, K. L., Casas, A., Huesca, M., Ustin, S. L., Alsina, M. M., Mathews, S. A., & Whiting, M. L. (2016). Leaf spectral clusters as potential optical leaf functional types within California ecosystems. *Remote Sensing of Environment*, 184, 229-246.

Seidel, E. J., Júnior, F. D. J. M., Ansuj, A. P., & Noal, M. R. C. N. C. (2008). Comparação entre o método Ward e o método K-médias no agrupamento de produtores de leite. *Ciência e Natura*, *30*(1), 07-15.

Sharma, M., Kumari, P., & Rizvi, M. A. (2018, November). An Analysis to Find the Efficient Clustering Algorithm for Identification of User Access Pattern. In 2018 8th *International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 72-77). IEEE.

Silva, M. D. N. A. D., Pessoa, F. C. L., Silveira, R. N. P. D. O., Rocha, G. S., & Mesquita, D. A. (2018). Determination of the Homogeneity and Tendency of Precipitations in the Tapajós River Basin. Revista Brasileira de Meteorologia, *33*(4), 665-675.

Sousa, L.C.R.; Santos; R.B.N.; Sousa, D.S.P. (2016). Desenvolvimento e pobreza multidimensional na Amazônia Legal. *Revista Espacios*, 37(21).

Starczewski, A. (2017). A new validity index for crisp clusters. *Pattern Analysis and Applications*, *20*(3), 687-700.

Vieira, I. C. G. (2019). Abordagens e desafios no uso de indicadores de sustentabilidade no contexto amazônico. *Ciência e Cultura*, *71*(1), 46-50.

Porcentagem de contribuição de cada autor no manuscrito

Diêgo Lima Crispim – 50%

Lindemberg Lima Fernandes – 20%

David Figueiredo Ferreira Filho – 20%

Bruna Roberta Pereira Lira – 10%