revista principia

REVISTA DE DIVULGAÇÃO CIENTÍFICA E TECNOLÓGICA DO INSTITUTO FEDERAL DE CIÊNCIA, EDUCAÇÃO E TECNOLOGIA DA PARAÍBA

Vol. 61, N. 3

DOI: http://dx.doi.org/10.18265/1517-0306a2022id7168

ARTIGO ORIGINAL

SUBMETIDO 08/08/2022

APROVADO 23/10/2022

PUBLICADO ON-LINE 07/11/2022

PUBLICADO 10/07/2024

EDITOR ASSOCIADO
Gilberto Reynoso Meza

Técnicas de agrupamento aplicadas aos indicadores de Crescimento Verde da OCDE

Matheus Santos Dias [1]

Amauri Ornellas da Silva [2]

Bruno Samways dos Santos [3]*

Rafael Henrique Palma Lima [4]

[1] matheusdias.1995@alunos.utfpr.edu.br
[2] amauris@alunos.utfpr.edu.br
[3] brunosantos@utfpr.edu.br
[4] rafaelhlima@utfpr.edu.br

Departamento Acadêmico de Engenharia de Produção, Universidade Tecnológica Federal do Paraná (UTFPR), Campus Londrina, Brasil

*Autor para correspondência.

RESUMO: A Organização para a Cooperação e Desenvolvimento Econômico (OCDE) publica anualmente dados sobre os indicadores de Crescimento Verde de todos os países. Em geral, essa base é discutida na literatura usando estatísticas descritivas, as quais fornecem uma visão geral sobre o desempenho sustentável dos países. No entanto, não há trabalhos que relatem a aplicação de técnicas de agrupamento associadas aos algoritmos de mineração de dados com o intuito de encontrar fatores que explicam as semelhanças e diferenças entre os países avaliados por esses indicadores. Por essa razão, este trabalho relata a aplicação de técnicas de agrupamento *k-means* e clusterização hierárquica para encontrar grupos de países com desempenhos semelhantes com relação aos indicadores sustentáveis e demográficos avaliados pela OCDE. Para essa aplicação, foram usados os dados do ano de 2019 considerando todos os países de forma individual, excluindo os dados sobre blocos econômicos. Após a limpeza e preparação dos dados, 153 países e 15 indicadores foram avaliados, resultando em 5 grupos de países. Alguns grupos apresentaram características dominantes entre os países nele incluídos. O *cluster* 3 foi o maior grupo, englobando 96 países subdesenvolvidos ou em desenvolvimento, com economia agroexportadora. O *cluster* 0 agrupou países com grande crescimento populacional, e o *cluster* 1 destacou países com altas taxas de mortalidade por exposição ao radônio. Por fim, o *cluster* 2 teve como destaque as variáveis demográficas referentes a idade e gênero, e o *cluster* 4 agrupou países com baixas taxas de exposição a poluição decorrente de materiais particulados.

Palavras-chave: agrupamento; crescimento verde; mineração de dados; OCDE; sustentabilidade.

Clustering techniques applied to the OECD Green Growth indicators

ABSTRACT: *The Organization for Economic Co-operation and Development (OECD) annually publishes data about Green Growth*

ISSN (on-line): 2447-9187

indicators for all countries. In general, this basis is discussed in the literature using descriptive statistics, which provide an overview of the sustainable performance of the surveyed countries. However, there are no published researches that report the application of clustering techniques associated with data mining algorithms in order to find factors that explain the similarities and differences between countries assessed by these indicators. This paper reports on the application of k-means clustering techniques and hierarchical clustering to find clusters of countries with similar performance in relation to sustainable and demographic indicators assessed by OECD. For this application, we used data from 2019 considering all countries individually, excluding data on economic blocks. After cleaning and preparing the data, 153 countries and 15 indicators were evaluated, resulting in 5 clusters of countries. Some clusters showed dominant characteristics among the countries contained in them. Cluster 3 was the largest cluster with 96 underdeveloped or developing countries, whose economy is based on agricultural exports. Cluster 0 grouped countries with high population growth and Cluster 1 highlighted countries with high mortality rates due to exposure to radon. Finally, Cluster 2 highlighted demographic variables related to age and gender whereas Cluster 4 grouped countries with low rates of exposure to pollution caused by particulate material.

Keywords: *clustering; data mining; green growth; OECD; sustainability.*

1 Introdução

Em um mundo cada vez mais informatizado, os dados se tornaram componentes essenciais a ponto de serem comumente anunciados como "o novo petróleo" por executivos e comunicadores, que fazem referência à frase original "*data is the new oil*" criada em 2006 por Clive Humby, um matemático londrino especializado em ciência de dados (Arthur, 2013). Essa expressão vem sendo repetida diversas vezes desde então, tanto para defender a nova *commodity* quanto para apontar seus perigos (Bhageshpur, 2019).

Avanços nas áreas de aprendizado de máquina, internet das coisas e robótica facilitam o cotidiano do ser humano (Ferreira, 2021), no entanto, sem uma sistematização que extraia informações úteis, os dados são de pouca valia. E é isso o que o processo de descoberta de conhecimento em bases de dados (do inglês *Knowledge Discovery in Databases* – KDD) propõe. Os algoritmos de reconhecimento de padrões, que fazem parte do processo KDD, são capazes de realizar inúmeras tarefas, visando a diferentes objetivos, entre elas a regressão, a classificação, o agrupamento, a associação e a visualização de dados (Larose; Larose, 2014).

O agrupamento de dados, também conhecido como clusterização, é uma categoria de técnicas de aprendizado não supervisionado, pois utiliza dados não rotulados para o reconhecimento de padrões (Zengin *et al*., 2011), o que permite descobrir estruturas ocultas em dados, das quais não se sabe a resposta certa antecipadamente. O objetivo dessas técnicas é encontrar padrões e formar agrupamentos naturais de dados de forma que os itens no mesmo grupo sejam mais semelhantes entre si do que àqueles de grupos diferentes (Raschka, 2015).

As aplicações de técnicas de mineração de dados podem ser úteis em diversas áreas, como detecção de fraudes em cartão de crédito, previsão de valores na área financeira, diagnóstico médico, desenvolvimento de produtos, sumarização de texto, entre outras (Bramer, 2016). Contudo, de acordo com levantamento realizado por Zhang *et al*. (2021), cerca de 47% das publicações sobre mineração de dados se concentram na área de ciência da computação e dão ênfase ao desenvolvimento de novos algoritmos. Isso mostra a importância da realização de trabalhos que apliquem as técnicas de mineração de dados para efetivamente descobrir novos conhecimentos em outras áreas, sobretudo sobre temas relacionados à sustentabilidade.

Sustentabilidade foi definida como o desenvolvimento que supre a necessidade do presente sem impedir as gerações futuras de suprir suas necessidades (Zhu; Hua, 2017), popularizando-se na economia a partir de 1987, quando foi apresentada na Comissão Mundial sobre Meio Ambiente e Desenvolvimento (Brown *et al*., 1987) como um padrão de crescimento factível e socialmente aceito. Desde então, milhares de iniciativas que procuram desenvolvimento equilibrado com sustentabilidade têm surgido em diferentes frentes, como tomada de decisão, administração, políticas e também pesquisa e análise.

Recentemente, o termo *green growth* (ou "crescimento verde") tem ganhado cada vez mais notoriedade, pois reflete os efeitos das políticas de novos acordos ou indústrias "verdes", como elas podem ser reajustadas com o tempo, e como produzir uma relação de mútuo benefício, ou seja, para a economia e para o meio ambiente. Investimentos significativos têm sido feitos por agências governamentais americanas, da Coreia do Sul, da China e da União Europeia. Entretanto, afirma-se que mesmo com a grande quantidade de indicadores e políticas voltadas ao *green growth*, a sua exploração por métodos de aprendizado de máquina tem sido muito limitada, considerando que esses métodos já estão bastante consolidados para outras áreas do conhecimento (Herman; Shenk, 2021).

A Organização para a Cooperação e Desenvolvimento Econômico (OCDE) tem monitorado, ao longo do tempo, diversos indicadores de sustentabilidade de países, os quais são publicados anualmente e mostram o progresso em direção às metas de sustentabilidade. Um estudo aprofundado desses indicadores pode revelar características de países que possuem melhor desempenho em aspectos "verdes", sendo possível sugerir políticas estratégicas quanto à promulgação do desenvolvimento sustentável (Mahmood *et al*., 2022).

Muitos trabalhos publicados recentemente têm explorado técnicas quantitativas para avaliar dados coletados pela OCDE ou de nações pertencentes à organização. Com relação a esses estudos, quando se realiza uma busca pelos termos "*clustering*" e "OECD" (sigla para *Organization for Economic Co-operation and Development*) em bases como *ScienceDirect* e *Scopus*, menos de 100 trabalhos são encontrados, sendo que, a partir de 2017, esse número se torna ainda mais reduzido.

Analisando as pesquisas que utilizaram técnicas de agrupamento como meio principal de análise, tendo como objeto de estudo os países ou regiões pertencentes à OCDE ou investigando diretrizes da OCDE, muitos autores fizeram uso dos algoritmos -*means* (Cai *et al*., 2022; Fujii; Iwata; Managi, 2017; Lamb; Minx, 2020; Linden; Ray, 2017; Menna; Walsh, 2019), agrupamento hierárquico (Ahlborn; Schweickert, 2019; Lalloué *et al*., 2019; Marti; Puertas, 2021; Proksch *et al*., 2019; Radu *et al*., 2018) ou as duas técnicas, para efeitos de comparação ou complementação (Ariaans; Linden; Wendt, 2021; Reibling; Ariaans; Wendt, 2019). Outros trabalhos utilizaram técnicas de agrupamento diferentes do

Quadro 1 ▼
Síntese dos trabalhos correlatos.
Fonte: dados da pesquisa

k-means e do agrupamento hierárquico para atender objetivos bem distintos, como resumido no Quadro 1.

Autor(es) (ano)	Objetivo	Técnicas principais
Fujii, Iwata e Managi (2017)	Identificar fatores determinantes que afetam as emissões de CO_2 e intensidade baseada no tipo de cidade, usando um conjunto de dados de áreas metropolitanas de 276 cidades	*k-means clustering*; modelo de regressão de efeito fixo
Linden e Ray (2017)	Explorar relações entre expectativa de vida e gastos com a saúde pública e privada, a partir de métodos de séries temporais de dados em painel para 34 países	*k-means clustering*; função de resposta ao impulso
Parker e Liddle (2017a)	Desenvolver um maior entendimento das dinâmicas da produção energética a partir de 33 países	Clusterização (método de Phillips e Sul, 2007); métodos de convergência *sigma* e *gamma*
Parker e Liddle (2017b)	Analisar a produtividade energética em 61 países da OCDE	Clusterização (método de Phillips e Sul, 2007); análise dos fatores de impacto na clusterização
Yan *et al*. (2017)	Investigar o desenvolvimento de tendências de tecnologias de baixa emissão de carbono a partir de 72 economias no período de 1990 a 2012 e de economias da OCDE no período de 1960 a 2012	Modelo fatorial não linear variante no tempo; clusterização (método de Phillips e Sul, 2007)
Radu *et al*. (2018)	Analisar a diferença entre custo laboral e salário líquido correlacionado com a taxa de desemprego e emprego. A análise foi realizada em 41 países da OCDE	Agrupamento hierárquico
Ahlborn e Schweickert (2019)	Identificar sistemas econômicos em países em desenvolvimento por uma abordagem de *clusters* macroeconômicos	Agrupamento hierárquico; *fuzzy c-means*; PCA
Lalloué *et al*. (2019)	Examinar a variabilidade da performance de hospitais dentro e entre países, usando dados de mortalidade por infarto agudo do miocárdio. Foram analisados mais de 1.000 hospitais, incluindo 10 membros da OCDE	Regressão linear univariada e multivariada; agrupamento hierárquico
Menna e Walsh (2019)	Aplicar um *framework* conceitual à indústria de vinhos global para identificar as estratégias de comercialização para pequenas e médias empresas em 22 membros da OCDE	*k-means clustering*
Proksch *et al*. (2019)	Agrupar países da OCDE utilizando saídas da inovação na área de serviços à saúde	Agrupamento hierárquico; análise discriminante; análise de correlação
Reibling, Ariaans e Wendt (2019)	Propor um *framework* a partir de indicadores para sistemas de serviços à saúde de acordo com os sistemas propostos pela OCDE	*k-means clustering*; agrupamento hierárquico
Ruggeri e Corsi (2019)	Representar o mercado global de cana-de-açúcar, focando na rede *Fairtrade* (FT) de produtores	Estatística descritiva e *two-step cluster*
Lamb e Minx (2020)	Explorar as razões para a resposta limitada à descarbonização acordada no Acordo de Paris. Os resultados incluíram análise específica de países da OCDE	Correlação de Spearman; PCA; *k-means clustering*; análise de tendências

continua

continuação

López, Arce e Jiang (2020)	Usar um *framework* da OCDE *Inter-Country Input-Output* (OCDE-ICIO) para capturar a existência de grupos de indústria ou setores na transmissão da emissão de carbono, da China ao resto do mundo, por meio das importações	Análise de redes complexas; análise de caminho estruturado (*structural path analysis*)
Ariaans, Linden e Wendt (2021)	Propor uma nova tipologia de sistemas de saúde de cuidados a longo prazo em um conjunto de dados da OCDE	*k-means clustering*; agrupamento hierárquico
Jun, Yoo e Lee (2021)	Entender como a declaração da Organização Mundial da Saúde (OMS) sobre a pandemia afetou a preocupação e comportamento do público a partir de dados do *Google Trends*, encontrados em 27 países da OCDE	DBSCAN; análise de intervenção
Marti e Puertas (2021)	Identificar as práticas mais sustentáveis focadas no tratamento de resíduos em 41 países da OCDE e/ou União Europeia	Agrupamento hierárquico; tabelas de contingência; teste de Kruskal-Wallis
Cai *et al*. (2022)	Desenvolver um novo conjunto de dados de emissão de CO_2 em alta resolução como meio de quantificação das emissões urbanas baseadas em limite de densidade populacional. O estudo utiliza definições estabelecidas pela OCDE	*k-means clustering*; modelo de regressão de efeito fixo
Eva *et al*. (2022)	Investigar a desigualdade regional a partir do crescimento econômico em dados de países da OCDE	Várias técnicas de agrupamento espacial
Hasse e Lajaunie (2022)	Reexaminar o poder preditivo do *spread* de rendimento entre países ao longo do tempo em 13 países da OCDE no período de 1975 a 2019	Modelo univariado adaptado; clusterização regional e estatística; *quantile-regression-based clustering* adaptado

Legenda: DBSCAN (*Density-based spatial clustering of applications with noise*); PCA (*Principal Component Analysis*)

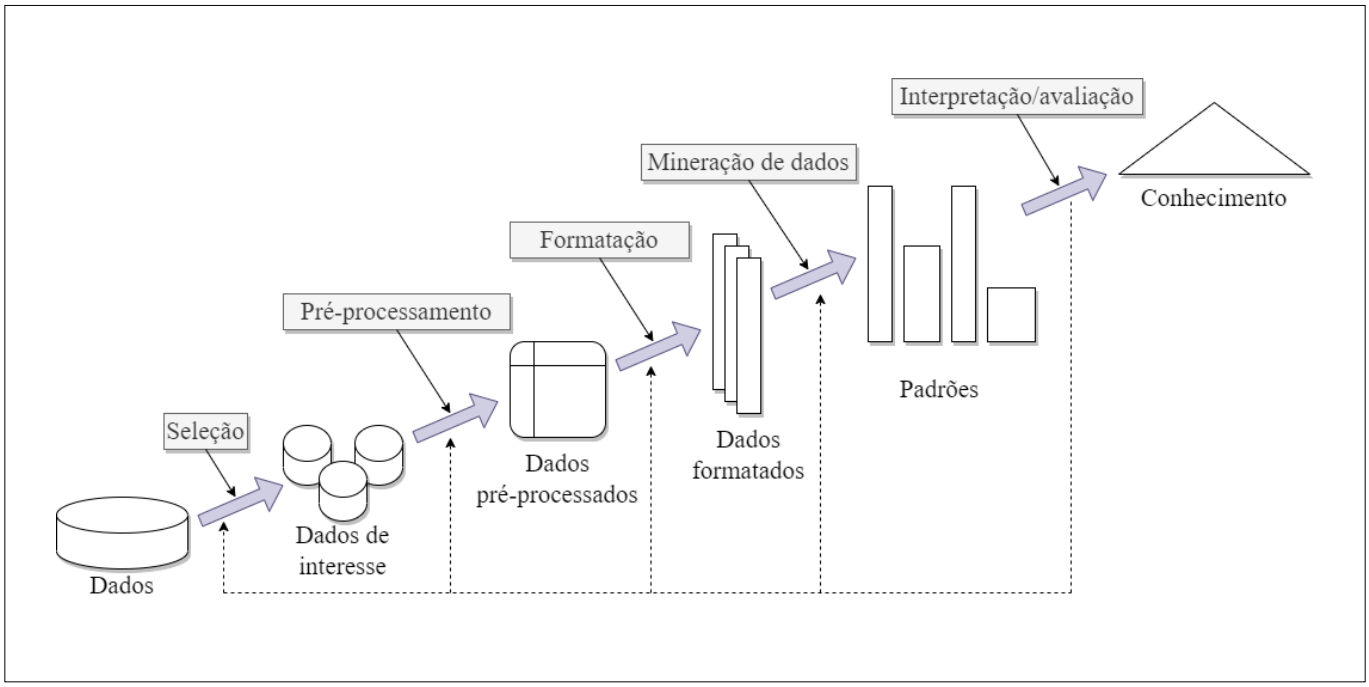
Nota-se no Quadro 1 que técnicas como -*means* e agrupamento hierárquico são aplicadas em dados relacionados em algum grau com a OCDE, incluindo a técnica estatística de correlação. Entretanto, nota-se que não há trabalhos relacionados aos dados de indicadores no contexto do "*Green Growth*" relativos à OCDE. Considerando a necessidade de mais aplicações práticas de técnicas de clusterização para extração de conhecimento de bases de dados reais (Zhang *et al*., 2021) e a inexistência de trabalhos que aplicam algoritmos de clusterização aos indicadores de Crescimento Verde da OCDE, este trabalho relata a aplicação dos algoritmos de agrupamento *k-means* e agrupamento hierárquico à base de dados de indicadores de Crescimento Verde da OCDE, considerando os resultados publicados no ano de 2019. Após o agrupamento, foram feitas análises utilizando mapas e *boxplots*, analisando-se as características e indicadores que diferenciam esses *clusters*, discutindo-os em seguida.

Após esta seção introdutória, o restante do artigo está organizado da seguinte maneira: a seção 2 discorre sobre a fundamentação teórica, com o embasamento dos conceitos sobre o KDD, os métodos de agrupamento e a métrica para avaliação do número de *clusters*. A seção 3 caracteriza a base de dados utilizada, bem como descreve a metodologia e as ferramentas utilizadas para a obtenção dos resultados dos algoritmos de clusterização. A seção 4 apresenta e discute os resultados da aplicação da metodologia proposta, tanto da etapa de agrupamento como da visualização dos resultados (grupos) formados. Por fim, a seção 5 apresenta as conclusões e direções para pesquisas futuras.

2 Fundamentação teórica

Figura 1 ▼
O processo de descoberta de conhecimento em bases de dados.
Fonte: adaptado e traduzido de Fayyad, Piatetsky-Shapiro e Smyth (1996)

A mineração de dados é o processo de descoberta de conhecimento através da análise de grandes volumes de dados, também conhecida como extração de conhecimento ou análise de padrões e muitas vezes utilizada como sinônimo para o *Knowledge Discovery from Data* (KDD) – em português, Descoberta de Conhecimento a partir de Dados –, apesar de serem coisas distintas e de a mineração de dados ser uma etapa do método de KDD (Han; Kamber; Pei, 2011). As etapas do KDD estão ilustradas na Figura 1.



O KDD pode ser dividido em cinco etapas. Inicialmente deve-se selecionar os dados de interesse, aqueles que são relevantes para a análise. Em seguida é feito o pré-processamento para limpar os ruídos do conjunto de dados. Na sequência é realizada a transformação dos dados já processados, com procedimentos como a normalização ou a padronização, por exemplo, que contribuem para um melhor desempenho do algoritmo de mineração. Aplica-se então o algoritmo de mineração de dados. Por fim, os resultados obtidos através do algoritmo são interpretados e avaliados.

Os algoritmos de *machine learning* (ou aprendizado de máquina, em português) aplicados para extrair padrões de bases de dados são diversos e atendem a necessidades específicas. Esses métodos podem ser subdivididos genericamente em métodos supervisionados e não supervisionados. Os métodos supervisionados são aqueles em que existe um rótulo para os dados analisados, e o objetivo da tarefa de mineração de dados é classificar novos dados não rotulados a partir do que foi aprendido com os dados rotulados. Já nos métodos não supervisionados, esses rótulos não existem e o objetivo da tarefa é agrupar esses dados de forma coerente (Grus, 2016).

As técnicas de clusterização, ou agrupamento de dados, são de aprendizado não supervisionado e consistem no agrupamento de objetos que são semelhantes entre si e dissemelhantes dos objetos pertencentes aos demais *clusters*. O particionamento de muitos pontos de dados em menos grupos ajuda a resumir os dados e a

entendê-los (Aggarwal, 2015). O agrupamento de dados pode ainda ser usado como uma etapa de pré-processamento para outros algoritmos. Existe uma grande variedade de técnicas de agrupamento de dados, sendo que os diferentes modelos podem funcionar melhor em diferentes cenários e tipos de dados. Vários pesquisadores têm se dedicado ao desenvolvimento de novos algoritmos de clusterização, cada um com estratégias distintas para agrupar registros de uma base de dados.

2.1 Método *k-means*

O algoritmo *k-means* é fundamentado no trabalho de MacQueen (1967) e tem como objetivo subdividir uma base de dados em *k* grupos. Conforme explicam Jin e Han (2011), a ideia básica do algoritmo é partir de um agrupamento inicial não ótimo e iterativamente redefinir os centroides dos grupos e realocar os pontos até que um critério de convergência seja atingido. Trata-se de um algoritmo computacionalmente muito eficiente em comparação com outros algoritmos de agrupamento, sendo amplamente utilizado na literatura especializada. Uma das principais razões de sua popularidade é a facilidade de implementação, o que reduz a barreira para sua adoção em diversas aplicações práticas (Raschka, 2015).

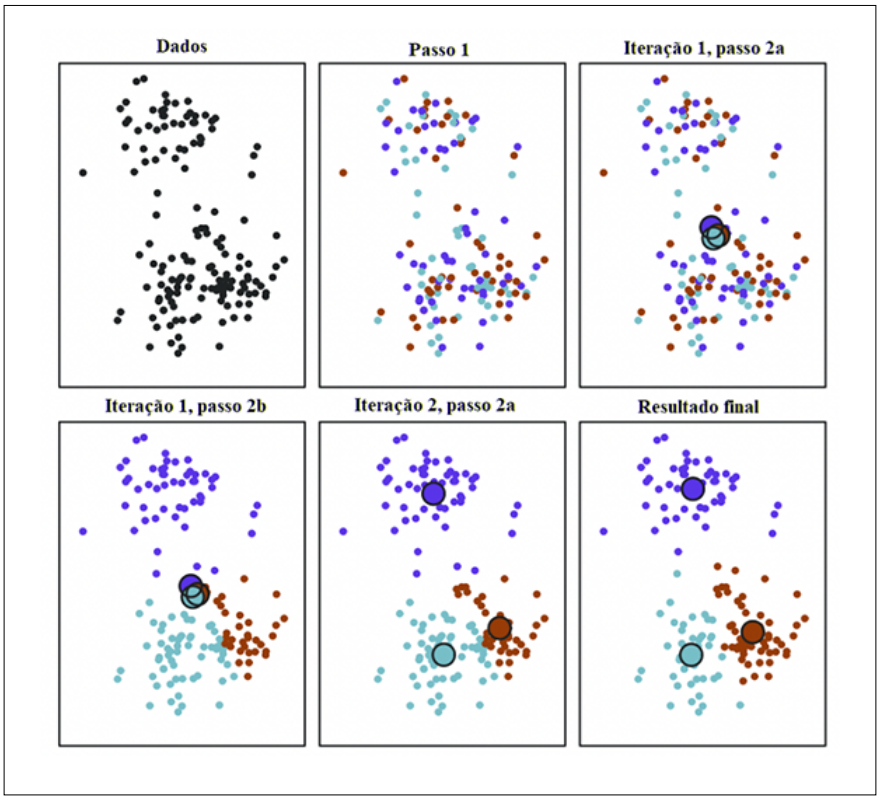
Uma das principais características do algoritmo *k-means* é a realização de agrupamentos exclusivos, ou seja, cada objeto é designado a apenas um *cluster*. Além disso, o algoritmo pertence à categoria de agrupamento baseado em protótipo, o que significa que cada grupo é representado por um protótipo, que pode ser o centroide (média) de pontos semelhantes com atributos contínuos ou o medoide (o ponto mais representativo ou de ocorrência mais frequente), no caso de atributos categóricos (Raschka, 2015). A soma dos quadrados das distâncias euclidianas dos pontos de dados aos seus representantes mais próximos é a métrica mais comumente usada para quantificar a função objetivo do agrupamento (Aggarwal, 2015). Esse algoritmo consegue ter um bom desempenho quando aplicado em pequenos conjuntos de dados (Sreedhar; Kasiviswanath; Reddy, 2017).

O procedimento iterativo do método *k-means* pode ser resumido pelas quatro etapas a seguir (Jin; Han, 2011; Raschka, 2015):

- Etapa 1: escolha aleatoriamente *k* centroides dos pontos de amostra como centros iniciais do grupo;
- Etapa 2: atribua cada amostra ao centroide mais próximo;
- Etapa 3: mova os centroides para o centro das amostras que foram atribuídas a ele;
- Etapa 4: repita as etapas 2 e 3 até que a atribuição do *cluster* não mude ou até que uma tolerância definida pelo usuário ou um número máximo de iterações seja alcançado.

Conforme explica Raschka (2015), o *k-means* é mais efetivo em identificar grupos de forma esférica. Porém, uma de suas desvantagens é a necessidade de especificar previamente o número de grupos, identificado pelo parâmetro *k*. Uma escolha inadequada para *k* pode resultar em baixo desempenho. A Figura 2 (próxima página) ilustra as iterações desse método.

Figura 2 ▶
Iterações do método de agrupamento *k-means*.
Fonte: adaptado James et al. (2017)



2.2 Método do agrupamento hierárquico

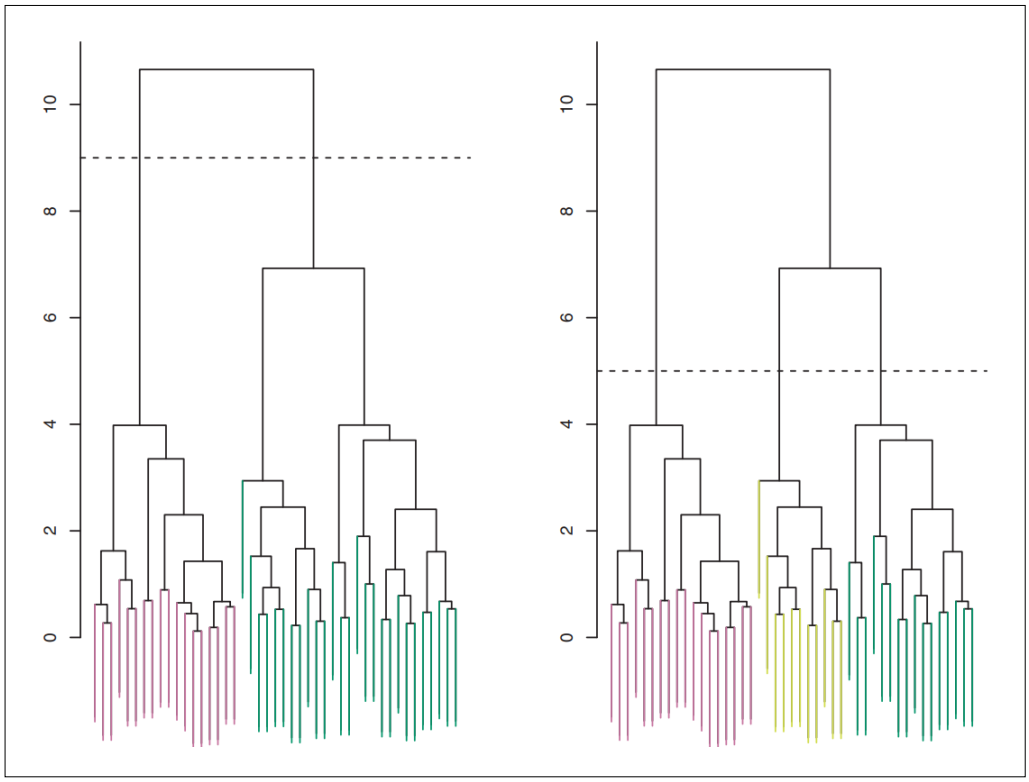
Uma das principais ferramentas de análise de *cluster* usadas é o agrupamento hierárquico, uma técnica recursiva na qual os objetos são unidos sequencialmente em *clusters* (aglomerativo) ou divididos uns dos outros em subgrupos (divisivo) (Govender; Sivakumar, 2020). O agrupamento aglomerativo começa com cada objeto separado e encontra os dois objetos mais próximos um do outro. Esses dois objetos são unidos (aglomerados) para formar um *cluster*, que agora é tratado como um novo objeto. As distâncias entre o novo objeto e os outros objetos são calculadas e o processo é repetido juntando os dois objetos mais próximos. Esse algoritmo se repete até que todos os objetos sejam unidos (James *et al*., 2017).

Em resumo, o procedimento iterativo pode ser dividido nas seguintes etapas:

- Etapa 1: calcule a matriz de distância entre todas as amostras;
- Etapa 2: represente cada ponto de dados como um *cluster* único;
- Etapa 3: mescle os dois *clusters* mais próximos com base na distância dos membros mais dissimilares (distantes);
- Etapa 4: atualize a matriz de similaridade;
- Etapa 5: repita as etapas 2 a 4 até que um único *cluster* permaneça.

O resultado final desse processo de agrupamento hierárquico pode ser visto em um dendrograma, como na Figura 3.

Figura 3 ▶
Representação de diferentes números de *clusters* por dendrograma.
Fonte: adaptado James et al. (2017)



Para o caso da Figura 3, o dendrograma à esquerda possui um corte na altura com valor 9, formando apenas dois *clusters*, um roxo e um verde. Já o dendrograma à direita representa um corte na altura com o valor 5, possuindo assim um *cluster* a mais, representado na cor amarela. Um dendrograma é uma árvore binária, ou seja, cada nó é subdividido em dois ramos. No entanto, o posicionamento dos *clusters* não corresponde à sua localização física no diagrama original. Todos os objetos originais são colocados no mesmo nível (a parte inferior do diagrama), como nós de folha. A raiz da árvore é mostrada no topo do diagrama. É um *cluster* que contém todos os objetos. Os outros nós mostram *clusters* menores que foram gerados conforme o processo prosseguia (Bramer, 2016).

2.3 Coeficiente da silhueta

Para o aprendizado não supervisionado, não há rótulos definidos para cada uma das instâncias. Logo, a avaliação deve ser conduzida utilizando como referência os próprios grupos que o modelo acaba criando (Wang *et al*., 2021). Para isso, pode ser utilizada uma métrica conhecida como coeficiente da silhueta. Esse coeficiente é uma medida de similaridade que verifica o quão perto um dado objeto (ou instância) está do grupo ao qual pertence, comparado aos outros grupos. O coeficiente pode ser calculado a partir da Equação 1:

$$\text{Coeficiente da silhueta} = \left(\frac{b-a}{max\,(a,b)} \right) \quad (1)$$

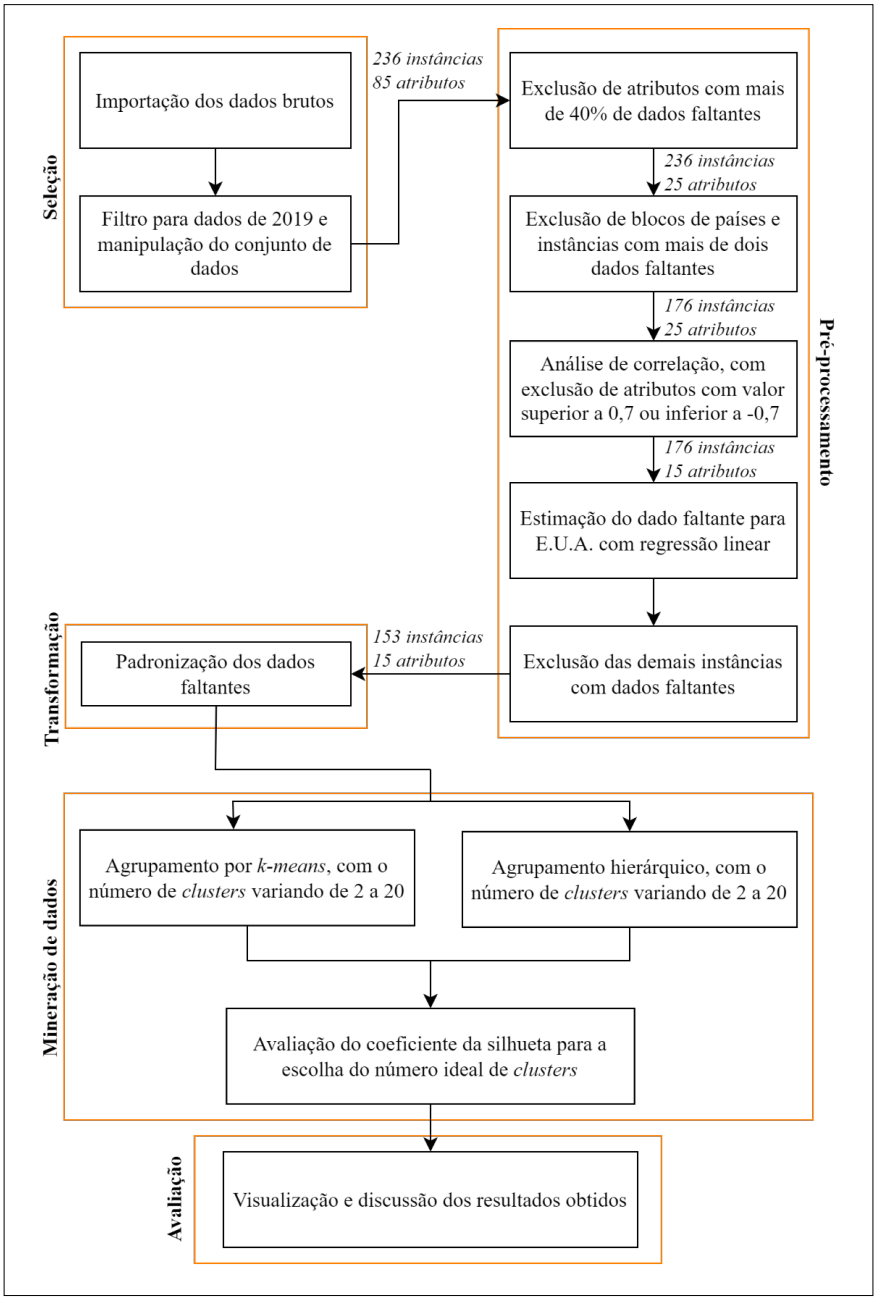
Nessa equação, o valor *a* é a distância média entre a amostra avaliada e todos os outros pontos do mesmo *cluster*, enquanto *b* é a distância média entre a amostra e todos os pontos do outro *cluster* (Wang *et al*., 2021). Quando há mais de dois *clusters*, utiliza-se para o cálculo de *b* o grupo mais próximo.

A interpretação do coeficiente da silhueta leva em conta a coesão do grupo e a separação entre os grupos em uma análise de *cluster* (Duarte *et al*., 2021). O intervalo que essa métrica pode assumir varia de -1 a 1, com um valor positivo, próximo a 1, indicando que a instância pertence ao *cluster* mais compacto e está a uma distância razoável de qualquer outro *cluster* (Igual; Seguí, 2017).

3 Materiais e métodos

As etapas executadas para a realização da presente pesquisa foram divididas em cinco fases principais, sendo elas: seleção, pré-processamento, transformação, mineração e avaliação. É importante ressaltar que todas essas fases estão previstas no processo do KDD, descrito na seção 2 deste artigo. As atividades de cada uma das etapas estão organizadas de acordo com a Figura 4.

Figura 4 ▶
Fluxograma das atividades realizadas neste estudo.
Fonte: elaborado pelos autores



A descrição do conjunto de dados utilizado bem como o pré-processamento e a análise exploratória estão na sequência desta seção.

3.1 Descrição do conjunto de dados

Este trabalho utiliza o conjunto de dados chamado "*Green Growth*" publicado pela OCDE[1], que contém indicadores selecionados para monitorar o progresso em direção ao crescimento verde, apoiando a formulação de políticas públicas, e fornecer esses dados a todos os interessados de forma geral. O conjunto sintetiza dados e indicadores em uma ampla gama de domínios, com os seus indicadores selecionados de acordo com critérios bem especificados e incorporados em uma estrutura conceitual, que é distribuída em torno de quatro eixos para capturar as características principais do crescimento verde. Esses eixos são: "produtividade ambiental e de recursos", "base de ativos naturais", "dimensão ambiental da qualidade de vida" e "oportunidades econômicas e respostas de políticas públicas", além de indicadores de contexto socioeconômico dos países (OECD, 2020).

[1] Disponível em: https://www.oecd.org/greengrowth/green-growth-indicators. Acesso em: 2 jan. 2024..

A base originalmente possuía dados desde o ano de 1990, no entanto, para a tarefa de agrupamento, optou-se por um corte transversal para o ano de 2019. Após a importação dos dados e aplicação do filtro para 2019, um conjunto de dados brutos foi obtido com 236 instâncias (países, blocos econômicos, fóruns internacionais ou agrupamentos geográficos) e 85 atributos (indicadores). A partir desse conjunto, foi realizado o pré-processamento, incluindo uma posterior análise exploratória. O conjunto de dados explorado e pré-processado ficou no formato mostrado na Tabela 1.

Tabela 1 ▼
Exemplo do conjunto parcial de dados explorado.
Fonte: dados da pesquisa

País	AGRGDP_PC	BIRD_FARM	CO2_AIRTRACAP	TEMPCHANGE5180
Equador	9,662	NaN	13,65	1,308
Egito	NaN	NaN	31,099	0,949
El Salvador	5,605	NaN	8,821	1,052
Guiné Equatorial	2,532	NaN	63,409	1,603
Eritreia	NaN	NaN	4,783	0,964
Estônia	2,865	58,923	0,270	2,213

Nota: as células com a identificação "NaN", significando "*Not a Number*", é uma representação da biblioteca *numpy* para a identificação de dados ausentes em um vetor (*array*).

3.2 Pré-processamento e análise exploratória de dados

Na etapa de pré-processamento foi identificada uma alta quantidade de dados faltantes. Ao todo havia 12.189 valores faltantes, o que corresponde a mais de 60% dos dados. Foi adotado como critério a exclusão dos atributos com menos de 60% de dados válidos, dessa forma, o conjunto passou de 85 para 25 atributos. Das instâncias resultantes, foram excluídas todas aquelas que não eram referentes aos países de forma individual – ou seja, blocos econômicos inteiros – e as que possuíam mais de dois dados faltantes dos 25 atributos restantes, o que fez com que o número de instâncias diminuísse de 236 para 176.

Depois disso, fez-se a análise de correlações entre os atributos a partir do método de correlação de Pearson. Esse método faz a medição da dependência linear entre um par de

variáveis (Raschka, 2015) e é muito aplicado no contexto do aprendizado de máquina, pois busca-se retirar variáveis altamente correlacionadas entre si por possuírem informações redundantes (Vettoretti; Di Camillo, 2021). Para o caso deste artigo, sempre que dois atributos apresentaram um coeficiente de correlação superior a 0,7 ou inferior a -0,7, um dos dois foi excluído. Ao final, o número de atributos caiu de 25 para 15.

Nesse ponto constatou-se que os Estados Unidos da América (EUA) ainda possuíam um dado faltante para o indicador "Emissões de dióxido de carbono (CO_2) de transporte aéreo por unidade do Produto Interno Bruto (PIB)". Um dos critérios adotados nessa etapa de pré-processamento foi garantir a inclusão de países com grande importância econômica ou política no cenário global, optando-se assim pela estimação desse dado para os EUA. A série histórica de 2013 a 2018 desse país foi levantada para esse indicador e, a partir desses dados, foi feita a regressão linear para estimar o dado faltante. Posteriormente, todas as demais instâncias que ainda tinham dados faltantes foram excluídas, resultando em um conjunto com dimensão final de 153 instâncias e 15 atributos.

Quadro 2 ▼
Indicadores utilizados para a tarefa de agrupamento.
Fonte: dados da pesquisa

Os atributos remanescentes podem ser observados no Quadro 2. A descrição detalhada de cada indicador e informações sobre como foram calculados podem ser acessadas no *website* https://www.oecd-ilibrary.org/, na opção "*Dataset Green Growth Indicators*".

Indicador	Eixo conceitual
Emissões de CO_2 de transporte aéreo por unidade do PIB	Produtividade ambiental e de recursos
Mudança da temperatura anual da superfície, desde 1951 a 1980	Base de ativos naturais
Mortalidade por exposição ao ozônio	Dimensão ambiental da qualidade de vida
Mortalidade por exposição ao chumbo	
Mortalidade por exposição a PM2.5	
Exposição média da população a PM2.5	
Porcentagem da população exposta a mais de 10 µg/m³ de PM2.5	
Mortalidade por exposição ao radônio residencial	
PIB real, índice 2000 = 100	Contexto socioeconômico
PIB real per capita	
População	
Densidade populacional, habitantes por km²	
População por faixa etária 15 a 64 anos, % do total	
Rede de migração	
Mulheres, % da população total	

Nota-se, a partir dos indicadores listados no Quadro 2, que os dados utilizados nas tarefas de agrupamento estão balanceados entre contexto ambiental e contexto socioeconômico. Esses dados apresentaram uma alta quantidade de *outliers*, e, por conta disso, inferiu-se que a normalização poderia não mostrar a real variabilidade dos dados. Optou-se, portanto, pela padronização dos dados pelo escore *z*, ou seja, transformar os dados para que a média geral seja igual a 0 (zero) e o desvio-padrão igual a 1 (um), evitando que o algoritmo ficasse enviesado pelas variáveis com maior ordem de grandeza.

Para a realização deste trabalho foi utilizada a linguagem de programação Python em sua versão 3.9, o ambiente de desenvolvimento integrado (IDE, *Integrated Development Environment*) Jupyter Notebook, o formato de troca de dados *JavaScript Object Notation* (JSON) e as bibliotecas *Pandas*, *Matplotlib*, *Seaborn*, *NumPy*, *Scikit-learn*, *SciPy* e *Plotly*. Para mais informações sobre as etapas metodológicas do trabalho, incluindo o código desenvolvido em linguagem Python, é possível acessar o diretório do *GitHub*, disponível em https://github.com/diassmatheus/ClusterizacaoDadosOCDE.

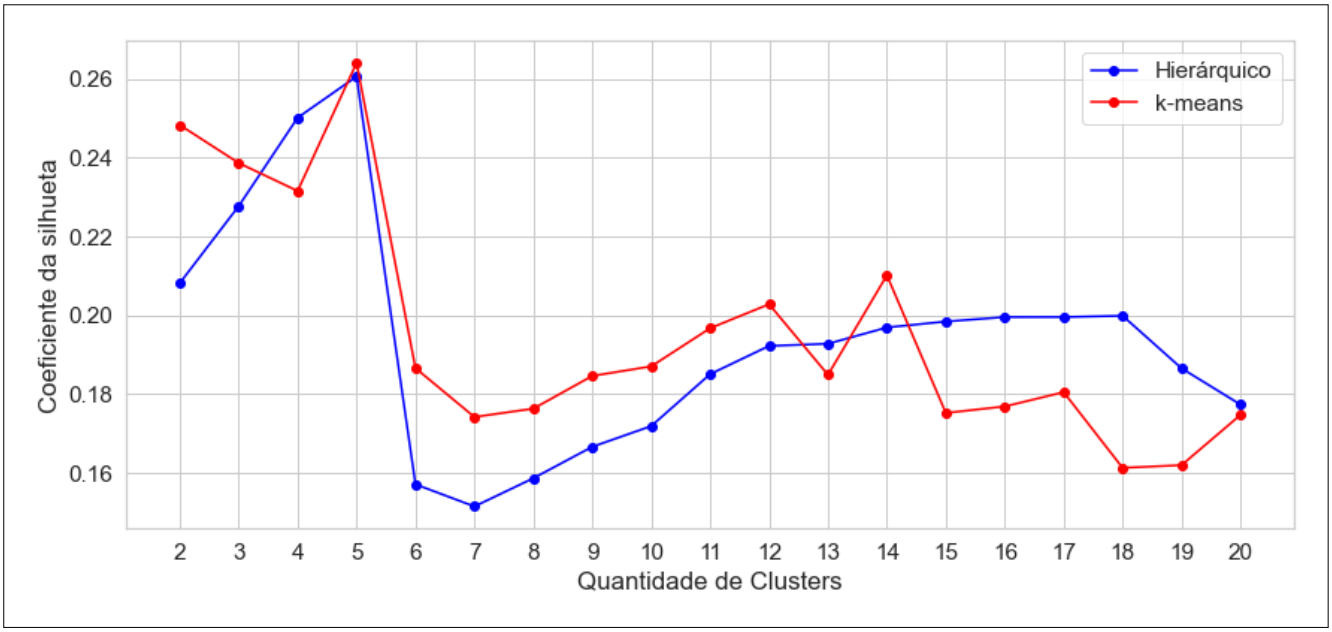
4 Resultados e discussão

Nesta seção, foi verificado o melhor número de *clusters* para cada uma das técnicas, realizou-se uma análise visual da distribuição geográfica dos países e seus respectivos grupos, foram identificados os principais atributos para os *clusters* encontrados e, por fim, uma discussão aprofundada dos resultados encontrados foi descrita. Todas as análises realizadas obedeceram às etapas descritas na Figura 4, totalizando 153 países e cinco atributos para a etapa de clusterização.

4.1 Análise do número de *clusters*

Como mostrado na Figura 4, após a análise exploratória e o pré-processamento do conjunto de dados, foi realizada a tarefa de agrupamento pelos métodos *k-means* e agrupamento hierárquico (etapa de mineração de dados). Como citado na seção 2, uma das limitações do método *k-means* é ter que definir o número de *clusters* (grupo) antecipadamente. Nesse contexto, foi gerado o coeficiente da silhueta para ambos os métodos com o número de *clusters* variando de 2 a 20 (vide Figura 5).

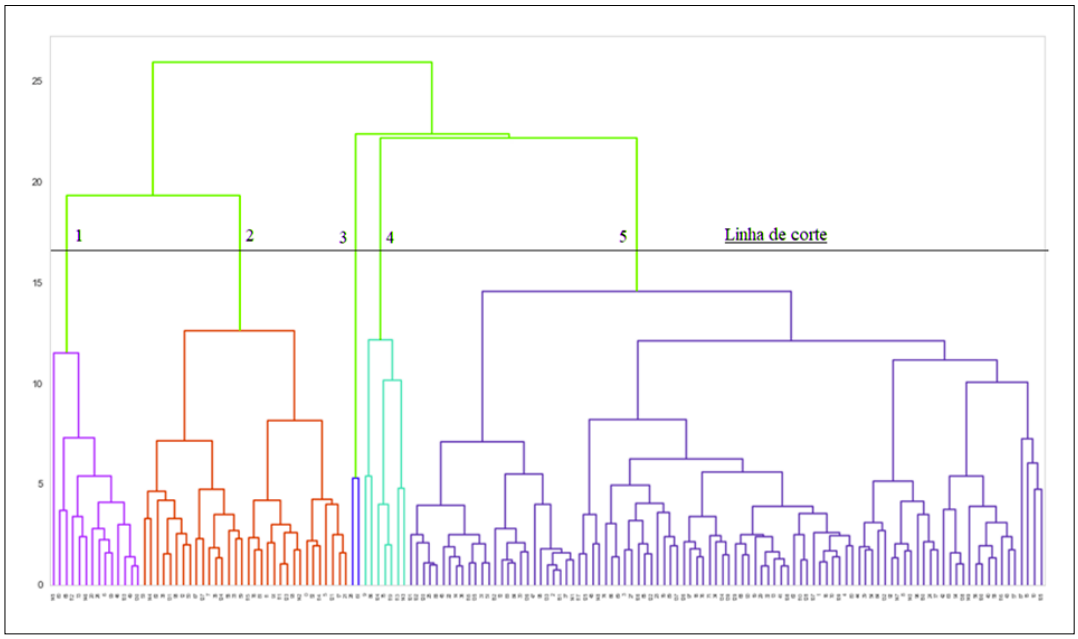
Figura 5 ▼
Coeficiente da silhueta para os métodos de agrupamento.
Fonte: dados da pesquisa



Constatou-se que o número ideal de *clusters* para esse conjunto de dados em ambos os métodos é igual a cinco. Assim, todas as discussões foram realizadas em cima dos cinco grupos encontrados pelos métodos.

Figura 6 ▼
Dendrograma da clusterização por agrupamento hierárquico.
Fonte: dados da pesquisa

Também foi criado o dendrograma da clusterização por meio do agrupamento hierárquico (Figura 6). O nível de similaridade é medido ao longo do eixo vertical, em que se avalia onde há o maior espaçamento entre a junção de dois *clusters*. As diferentes observações são listadas ao longo do eixo horizontal.

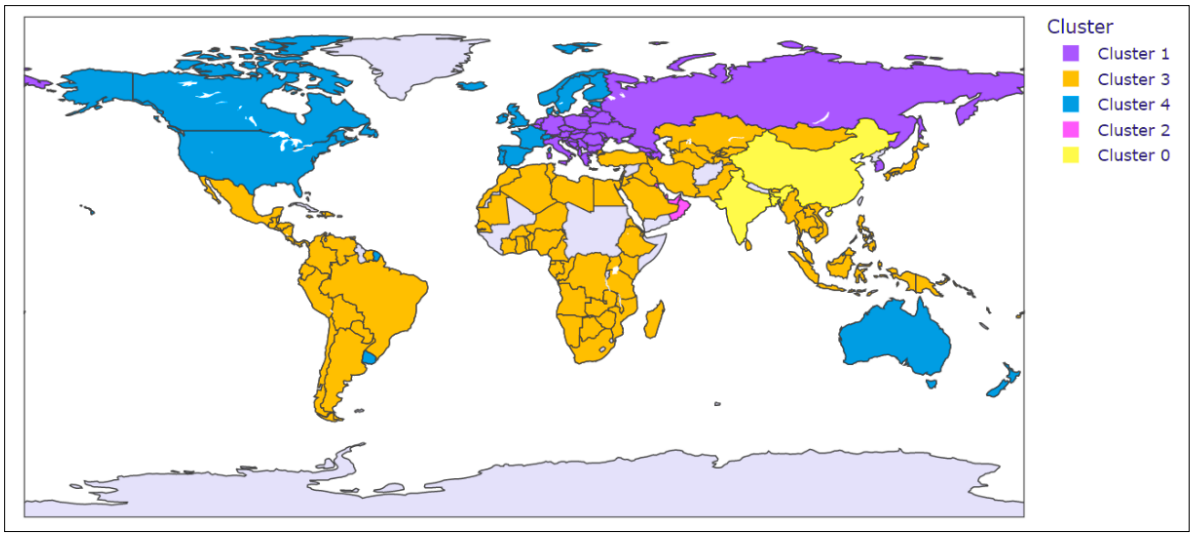


O dendrograma da Figura 6 confirma o que foi observado através dos coeficientes de silhueta – que o melhor número de *clusters* é cinco –, o que é mostrado pela linha de corte entre as junções com cinco grupos e com quatro grupos.

4.2 Distribuição geográfica dos *clusters*

Figura 7 ▼
Clusters obtidos por *k-means* identificados no mapa-múndi.
Fonte: dados da pesquisa

Para melhor visualização dos resultados obtidos através desses algoritmos de clusterização, foi ilustrada no mapa-múndi a distribuição dos cinco *clusters* obtidos pelos dois métodos. A Figura 7 mostra a distribuição dos grupos formados pelo método *k-means*.

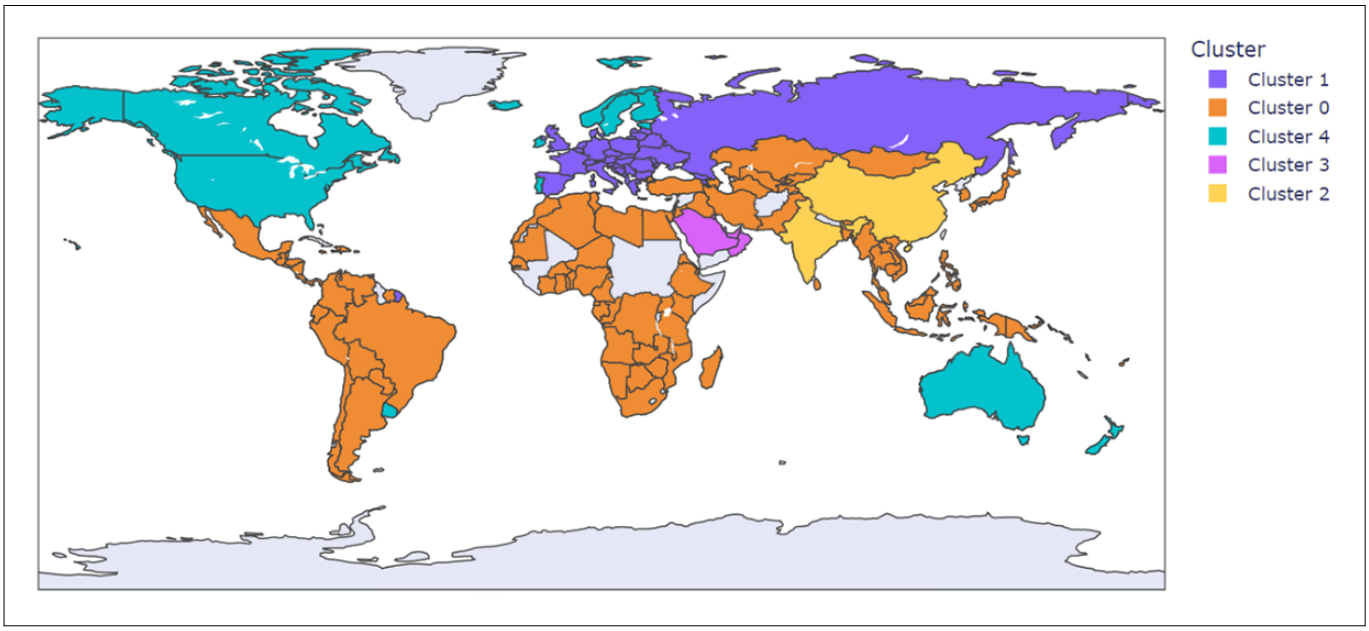


Nota-se pela Figura 7 que os países ficaram divididos da seguinte forma:

- *Cluster* "0" (Amarelo): países asiáticos com grande população, como a Índia e a China;
- *Cluster* "1" (Roxo): países do Leste Europeu e a Rússia;
- *Cluster* "2" (Rosa): nesse grupo, tem-se países do Oriente Médio, como o Bahrein, Emirados Árabes Unidos, Qatar, Kuwait e Omã, além das Maldivas, que possui uma proximidade geográfica;
- *Cluster* "3" (Laranja): agrupou muitos países, sendo a maioria deles pertences à América Latina e à África, com economias subdesenvolvidas ou em desenvolvimento;
- *Cluster* "4" (Azul): agrupou países desenvolvidos, como os EUA, Canadá, Austrália e países da Europa Ocidental.

Figura 8 ▼
Clusters obtidos por agrupamento hierárquico.
Fonte: dados da pesquisa

A Figura 8 mostra a distribuição dos grupos formados pelo método agrupamento hierárquico.



Já por esse agrupamento, apesar de algumas mudanças para países da Europa e do Oriente Médio, a maioria dos países foi agrupado da mesma forma que no método anterior, só mudando a coloração devido aos identificadores dos *clusters* quando implementado cada algoritmo (vide Figura 8).

Uma vez que não houve muitas diferenças entre os países dos *clusters* encontrados pelos métodos *k-means* e hierárquico, optou-se, para fins de análise e discussão, por explorar as características intraclusters e interclusters encontradas pelo método *k-means*.

4.3 Análise e discussão de atributos dos clusters encontrados

Uma aferição possível a partir da distribuição dos *clusters* no mapa-múndi, e que se confirma a partir de um gráfico de frequências, é o desbalanceamento da quantidade de países por *cluster*. A quantidade de países em cada grupo pode ser observada na Figura 9 (próxima página).

Figura 9 ▶
Quantidade de países por *cluster*.
Fonte: dados da pesquisa

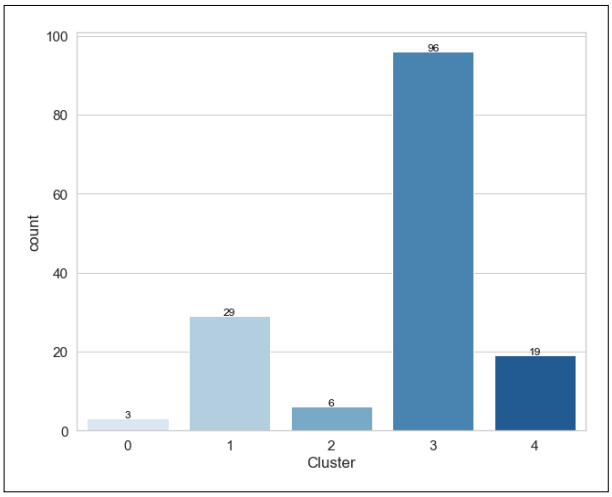
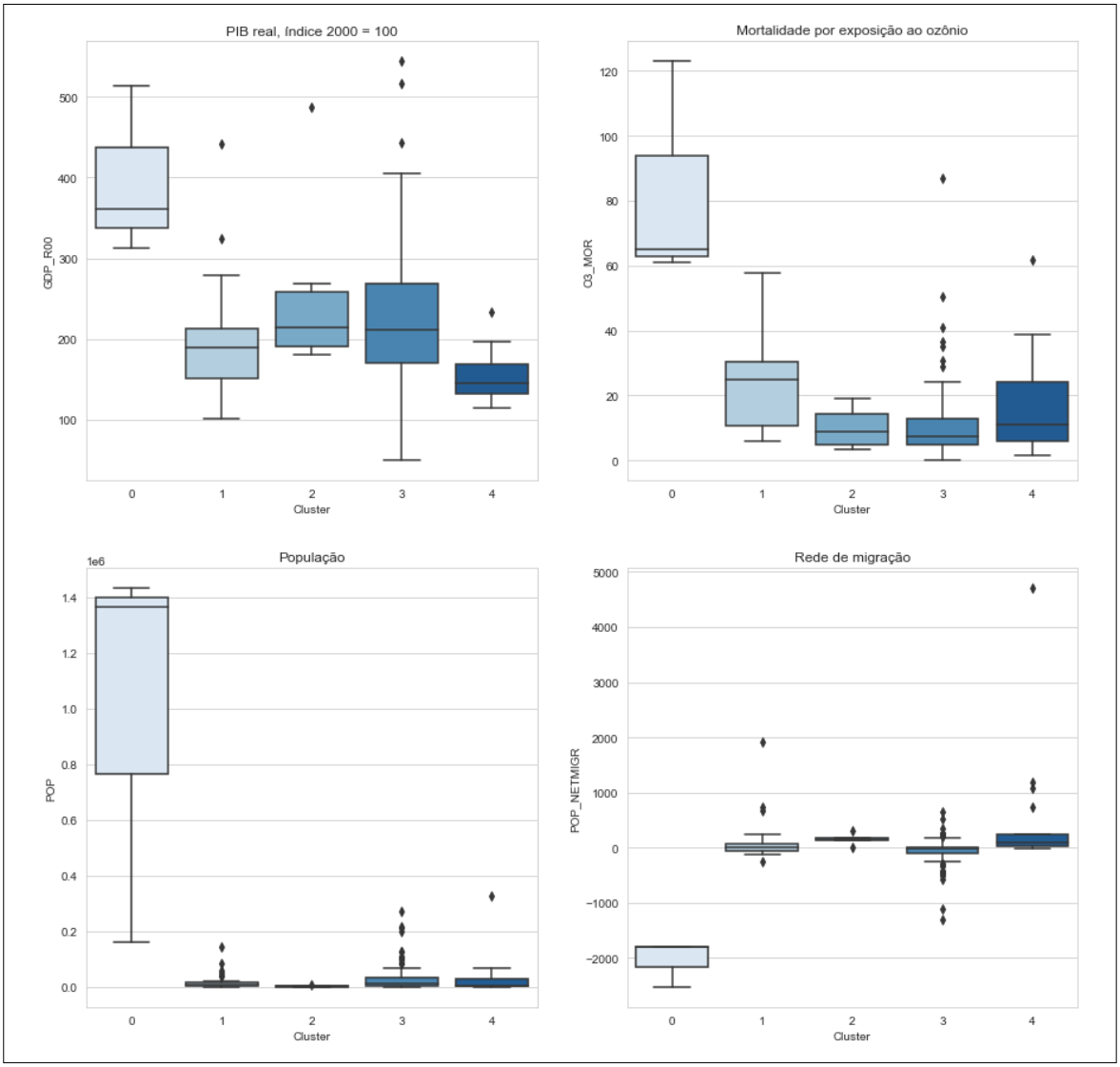


Figura 10 ▼
Boxplots de indicadores de interesse para o *cluster* "0".
Fonte: dados da pesquisa

O *cluster* "0", que possui a menor quantidade de países, é composto por China, Índia e Bangladesh. Esses países asiáticos, além de apresentarem similaridades demográficas, possuem uma alta taxa de crescimento econômico. Algumas comparações podem ser observadas na Figura 10.



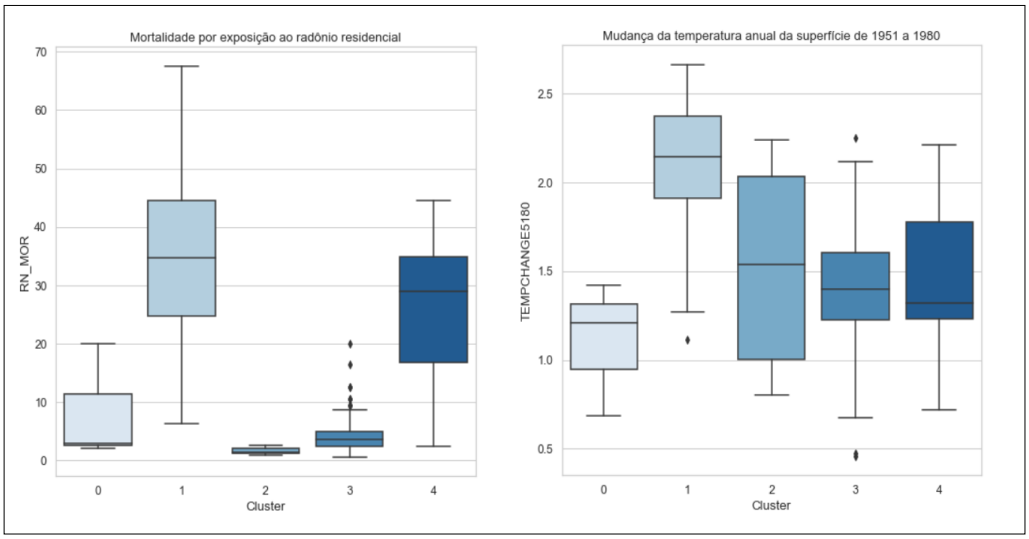
Bangladesh, que já foi considerado o país menos desenvolvido do mundo em termos econômicos *per capita*, devido a um recente crescimento econômico, associado aos avanços em educação e saúde pública e a um menor índice de vulnerabilidade, deve sair, até 2024, do selo de Países Menos Desenvolvidos (PMD) da Organização das Nações Unidas (ONU) (O País [...], 2020).

Os países do *cluster* "0" apresentam uma alta mortalidade por exposição ao ozônio a nível do solo quando comparados com os outros *clusters*. Os padrões de qualidade do ar são diferentes para diferentes nações. Por exemplo, para a Organização Mundial da Saúde (OMS), o limite aceitável de exposição ao ozônio é de 100 µg/m³, enquanto tem-se 120 µg/m³ para a União Europeia, 140 µg/m³ para o padrão da *US National Ambient Air Quality Standard* e 160 µg/m³ para a *Chinese Ambient Air Quality*. Esses padrões são importantes, pois a exposição prolongada a determinados níveis de ozônio interfere em mecanismos fisiopatológicos, também podendo levar à mortalidade ou morbidades (Vicedo-Cabrera *et al*., 2020). Porém, em uma pesquisa em nível global realizada por Vicedo-Cabrera *et al*. (2020) sobre a mortalidade a partir da exposição ao ozônio em diversas cidades, a China foi um dos países analisados com um risco relativo de morte de 0,55% por cada aumento de 10 µg/m³, similar a países como Austrália, República Tcheca, França, Alemanha, Itália, Japão, Coreia do Sul, Suécia, Suíça e Estados Unidos. Entretanto, os autores destacam que o número de cidades de alguns países, incluindo a China, pode não ter sido representativo, o que levaria o estudo a não revelar os verdadeiros impactos. Os autores também sugerem que fatores relacionados às características populacionais devem ser investigados.

A rede de migração – que representa o número de imigrantes menos o número de emigrantes – nesse *cluster* é menor que nos demais, o que indica uma maior emigração nesses países. A emigração é primeiramente condicionada por oportunidades econômicas; nesse contexto, uma grande porcentagem de trabalhadores asiáticos é nascida na Índia, e esses buscam oportunidades de melhores condições de trabalho nos Emirados Árabes Unidos (muito em função do acordo entre os dois países realizado em 2018 para o desenvolvimento de habilidades e qualificação), EUA e Arábia Saudita (IOM, 2021). Também há uma pesquisa relevante sobre o encorajamento da China e da Índia para repatriação de empreendedores, uma vez que estes podem contribuir para o desenvolvimento econômico local (Zweig; Tsai; Singh, 2021). Esse estudo constatou que a China é mais ativa nesse encorajamento, com os seus empresários tendo uma boa visão sobre o seu Estado, enquanto empreendedores indianos veem o Estado local como predatório (Zweig; Tsai; Singh, 2021).

O *cluster* "1" é composto por 29 países, sendo a maioria deles ex-membros da extinta URSS (União das Repúblicas Socialistas Soviéticas), com algumas exceções, como Grécia e Coreia do Sul, por exemplo. Grande parte desses países enfrentaram no passado uma realidade ambiental deteriorada, uma vez que a URSS e os países da Europa Oriental priorizaram o desenvolvimento econômico por meio de um processo de industrialização a todo custo e agricultura intensiva em detrimento de políticas de conservação ambiental (Mnatsakanian, 2000). A Figura 11 traz algumas comparações de interesse para o *cluster* "1".

Figura 11 ▶
Boxplots de indicadores de interesse para o *cluster* "1".
Fonte: dados da pesquisa

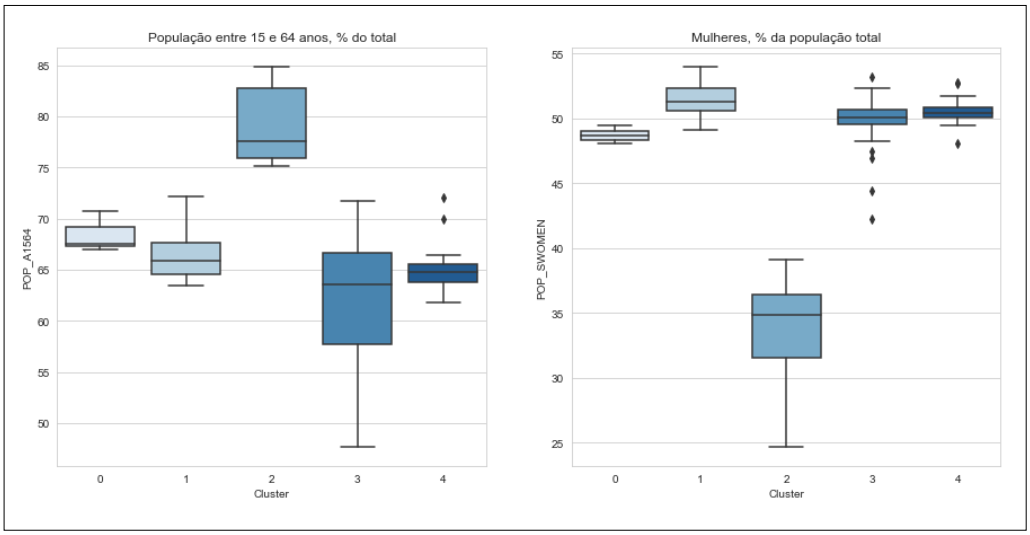


Analisando-se a Figura 11, é possível perceber que o *cluster* "1" possui maior mediana de mortalidade por exposição ao radônio residencial e maior mediana de mudança de temperatura da superfície entre 1951 e 1980, fato que ocorre em diversos outros países pelo mundo. Uma pesquisa publicada recentemente na *Earth's Future* indicou que o aumento da temperatura da superfície ocorrido entre 1971 e 2000 merece maior atenção na região da Escandinávia e em alguns países do Leste Europeu, devido às suas maiores médias de aumento de temperatura em relação ao resto do mundo, principalmente com o aquecimento ocorrido em temporadas de inverno (Jacob *et al*., 2018).

De acordo com a OMS, o gás inerte natural conhecido como radônio pode ser encontrado em grandes concentrações em ambientes fechados, como casa e locais de trabalho, sendo um dos principais causadores do câncer de pulmão (WHO, 2021). O *boxplot* do *cluster* 1 sugere uma quantidade de países com taxas de mortalidade por exposição ao radônio residencial, porém, oito países pertencentes a esse grupo possuem os maiores valores e, quando comparados ao monitoramento de radiação natural do *European Commission Joint Research Centre* (EC, 2019), Hungria, Sérvia, República Tcheca, Eslovênia e Áustria, de fato, possuem altas concentrações de radônio, variando de 0 a 20 Bq/m³ (ou Becquerel por m³), até 200 a 500 Bq/m³ em alguns pontos da Áustria e da República Tcheca. Isso fica mais evidente ainda quando se verifica o mapa da concentração de urânio no solo, que nesses países, em diversos territórios, é de até 5,0 a 7,6 U (mg/kg), bem acima de outras regiões da Europa. Isso é importante devido a o radônio surgir de forma natural a partir do decaimento do urânio (EC, 2019).

O *cluster* "2" é composto por seis países, sendo eles Catar, Emirados Árabes Unidos, Bahrein, Kuwait, Omã e Maldivas. Com exceção das Maldivas, todos os países desse *cluster* são do Oriente Médio. De acordo com os indicadores analisados, os maiores pontos de similaridade entre esses países são de cunho demográfico, como mostrado na Figura 12.

Figura 12 ▶
Boxplots de indicadores de interesse para o *cluster* "2".
Fonte: dados da pesquisa

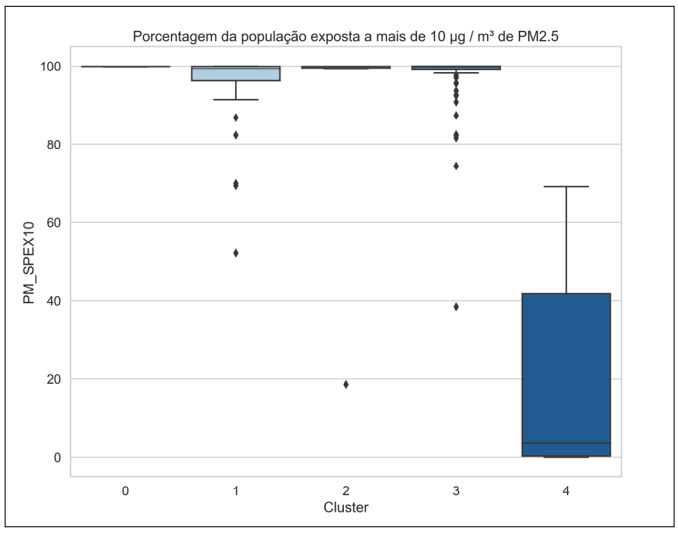


Como é possível observar na Figura 12, os países do *cluster* "2" possuem o percentual de sua população com faixa etária entre 15 e 64 anos maior que os demais países e uma menor porcentagem de mulheres.

O *cluster* "3" foi o que agrupou a maior quantidade de países, totalizando 96. Majoritariamente, esses países são do que estudos pós-coloniais e transnacionais chamam de "Sul Global". Este é composto, no geral, por economias subdesenvolvidas ou em desenvolvimento, agroexportadoras, com baixa emissão de carbono e que têm uma estrutura social e econômica com grandes desigualdades em padrões de vida, expectativa de vida ou acesso a recursos. Os dados desse *cluster* para os indicadores utilizados não possibilitaram encontrar informações de destaque; o fato de o *cluster* possuir uma quantidade de instâncias significativamente maior que os demais pode ser uma razão.

Por fim, o *cluster* "4" agrupou 19 países, entre eles EUA, Canadá, Austrália e países da Europa Ocidental. Países ricos, com acesso a tecnologias avançadas, sistemas políticos estáveis e alta expectativa de vida. A Figura 13 traz uma comparação entre o *cluster* "4" e os demais.

Figura 13 ▶
Boxplots de indicadores de interesse para o *cluster* "4".
Fonte: dados da pesquisa



Como ilustrado na Figura 13, um ponto fundamental proveniente da análise dos dados desse *cluster* foi a baixa taxa de exposição a material particulado com concentrações

anuais que excedem 10 µg/m³. O material particulado identificado por PM10 pode ser originado de diferentes fontes, como locais de construção, aterros sanitários e agricultura, incêndios florestais, entre outros, e essa faixa de material também inclui o chamado material particulado fino (PM2.5), encontrado na combustão da gasolina e do óleo diesel, por exemplo (CARB, 2022).

Pensando em uma escala global, em 1999 foi criado o Protocolo de Gotemburgo, o primeiro tratado sobre redução e controle de múltiplos poluentes do ar e suas fontes. Esse protocolo incluiu diversos países, como Bulgária, Canadá, EUA, Suíça, Holanda, Portugal, Finlândia etc., no intuito de se obter metas diretas de redução de emissões (IISD, 2019). Porém, um estudo recente afirmou que suas diretrizes devem ser revistas, com objetivos mais ambiciosos e específicos para as várias regiões do mundo (Sokhi *et al*., 2021).

Um outro estudo de 2017 sobre os países em desenvolvimento constatou que vários deles, principalmente na região asiática, possuem muitos problemas com a poluição do ar, incluindo o PM2.5 e o PM10, pois muitos deles experimentaram industrialização desenfreada, urbanização e desenvolvimento dos transportes, favorecendo a poluição do ar. Vários outros países em desenvolvimento têm sérios problemas na qualidade do ar pelo uso de combustíveis sólidos e biomassa para atender as necessidades energéticas locais, enquanto países desenvolvidos, como vários do *cluster* "4", têm seus programas de industrialização completos após vários anos de implantação (Mannucci; Franchini, 2017).

5 Conclusão

Este estudo aplicou os algoritmos *k-means* e agrupamento hierárquico para identificar grupos de países similares a partir dos dados de indicadores de crescimento verde da OCDE, definindo o número ideal de *clusters* através dos coeficientes da silhueta.

Devido principalmente ao alto número de dados faltantes, a maioria dos indicadores levados em consideração na etapa de mineração dos dados foram do eixo conceitual dimensão ambiental da qualidade de vida e do contexto socioeconômico.

Considerando que uma das finalidades das tarefas de agrupamento é descobrir estruturas ocultas nos dados, pode-se afirmar que os resultados deste trabalho foram satisfatórios, desde o cálculo da melhor quantidade de grupos até o agrupamento em si. Os agrupamentos possibilitaram mostrar, em alguns indicadores, os principais destaques por visualização via *boxplot*, como a baixa taxa de exposição a partículas com concentrações anuais que excedem 10 µg/m³ para grande parte dos países desenvolvidos (*cluster* "4"), enquanto características demográficas foram o que definiu o *cluster* "2". Os países considerados ricos dentro da OCDE, como Reino Unido, Alemanha e Dinamarca, são declarados líderes do clima, pois têm aplicado com sucesso ações para reduzir emissões na última década (Lamb; Minx, 2020). Essas nações são consideradas desenvolvidas, já finalizaram programas de industrialização e hoje buscam otimizar as suas emissões, mostrando que o *cluster* 4 corrobora a declaração anterior.

Também se observou altas taxas de mortalidade por exposição ao radônio residencial em países do Leste Europeu ou da antiga União Soviética, enquanto o *cluster* "0", com apenas três países asiáticos (China, Índia e Bangladesh), agrupa países com grande crescimento populacional, e também industrial em várias de suas regiões. Entretanto, algumas interpretações merecem uma maior atenção: por exemplo, a mortalidade por câncer de pulmão aumenta com a expectativa de vida. Nesse contexto, apesar da grande população da Índia e da sua alta exposição ao radônio, os EUA apresentam um índice maior de mortalidade por exposição ao radônio do que a Índia (Gaskin *et al*., 2018).

Vale destacar que esta pesquisa se baseou nos dados publicados em 2019, o que proporcionou a formação de agrupamentos relevantes à realidade mundial naquele ano. Durante a realização deste trabalho, os dados referentes a 2020 e 2021 não haviam sido disponibilizados ou encontravam-se incompletos. Portanto, trabalhos futuros precisam continuar a análise dos grupos formados a partir de dados mais recentes, assim que se tornarem disponíveis. Após a publicação dos dados de 2020 a 2022, também será possível avaliar o impacto do período da pandemia de COVID-19 sobre os indicadores verdes e a formação dos grupos de países.

Trabalhos futuros também podem considerar os dados históricos e sua evolução temporal, uma vez que a base possui dados a partir de 1990, permitindo a identificação de tendências futuras. Para isso, novas pesquisas podem explorar a base de dados usando não apenas técnicas de agrupamento, mas também algoritmos de regressão ou classificação com o intuito de extrair novos tipos de conhecimento.

Como limitação do trabalho, pode-se citar o baixo número de atributos dos eixos conceituais "produtividade ambiental e de recursos" e "base de ativos naturais" e a ausência de atributos do eixo conceitual "oportunidades econômicas e respostas de políticas públicas". Sugere-se estudar também os indicadores de outros países que ficaram fora do estudo a partir de anos anteriores ou anos futuros, possibilitando uma visão mais generalizada de cada indicador no escopo global, mostrando as tendências em cada um dos indicadores disponibilizados pela OCDE.

Financiamento

Esta pesquisa não recebeu nenhum financiamento externo.

Conflito de interesses

Os autores declaram não haver conflito de interesses.

Referências

AGGARWAL, C. C. **Data mining**: the textbook. 1. ed. Cham: Springer International, 2015. 734 p. DOI: https://dx.doi.org/10.1007/978-3-319-14142-8.

AHLBORN, M.; SCHWEICKERT, R. Economic systems in developing countries: a macro cluster approach. **Economic Systems**, v. 43, n. 3-4, 100692, 2019. DOI: https://doi.org/10.1016/j.ecosys.2019.100692.

ARIAANS, M.; LINDEN, P.; WENDT, C. Worlds of long-term care: a typology of OECD countries. **Health Policy**, v. 125, n. 5, p. 609-617, 2021. DOI: https://doi.org/10.1016/j.healthpol.2021.02.009.

ARTHUR, C. Tech giants may be huge, but nothing matches big data. **The Guardian**, 23 ago. 2013. Disponível em: https://www.theguardian.com/technology/2013/aug/23/tech-giants-data. Acesso em: 29 out. 2022.

BHAGESHPUR, K. Data is the new oil -- and that's a good thing. **Forbes**, 15 nov. 2019. Disponível em: https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/. Acesso em: 29 out. 2022.

BRAMER, M. **Principles of Data Mining**. 3. ed. London: Springer-Verlag London, 2016. 526 p. ISBN 978-1-4471-7307-6. DOI: https://doi.org/10.1007/978-1-4471-7307-6.

BROWN, B. J.; HANSON, M. E.; LIVERMAN, D. M.; MERIDETH JUNIOR, R. W. Global sustainability: toward definition. **Environmental Management**, v. 11, n. 6, p. 713-719, 1987. DOI: https://doi.org/10.1007/BF01867238.

CAI, B.; LIU, H.; ZHANG, X.; PAN, H.; ZHAO, M.; ZHENG, T.; NIE, J.; DU, M.; DHAKAL, S. High-resolution accounting of urban emissions in China. **Applied Energy**, v. 325, 119896, 2022. DOI: https://doi.org/10.1016/j.apenergy.2022.119896.

CARB – CALIFORNIA AIR RESOURCES BOARD. **Inhalable particulate matter and health (PM2.5 and PM10)**. [2022]. Disponível em: https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health. Acesso em: 27 mar. 2022.

DUARTE, J.; VIEIRA, L. W.; MARQUES, A. D.; SCHNEIDER, P. S.; PUMI, G.; PRASS, T. S. Increasing power plant efficiency with clustering methods and Variable Importance Index assessment. **Energy and AI**, v. 5, 100084, 2021. DOI: https://doi.org/10.1016/j.egyai.2021.100084.

EC – EUROPEAN COMMISSION. Joint Research Centre. **European Atlas of Natural Radiation**. Luxemburgo: Publications Office of the European Union, 2019. DOI: https://dx.doi.org/10.2760/46388.

EVA, M.; CEHAN, A.; CORODESCU-ROȘCA, E.; BOURDIN, S. Spatial patterns of regional inequalities: empirical evidence from a large panel of countries. **Applied Geography**, v. 140, 102638, 2022. DOI: https://doi.org/10.1016/j.apgeog.2022.102638.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-53, 1996. DOI: https://doi.org/10.1609/aimag.v17i3.1230.

FERREIRA, J. M. Análise de pesquisas sobre o impacto das tecnologias modernas e as transformações no mundo do trabalho (2013 – 2020). **Future Studies Research Journal: Trends and Strategies [FSRJ]**, v. 13, n. 3, p. 435-462, 2021. DOI: https://doi.org/10.24023/FutureJournal/2175-5825/2021.v13i3.593.

FUJII, H.; IWATA, K.; MANAGI, S. How do urban characteristics affect climate change mitigation policies? **Journal of Cleaner Production**, v. 168, p. 271-278, 2017. DOI: https://doi.org/10.1016/j.jclepro.2017.08.221.

GASKIN, J.; COYLE, D.; WHYTE, J.; KREWKSI, D. Global estimate of lung cancer mortality attributable to residential radon. **Environmental Health Perspectives**, v. 126, n. 5, 057009, 2018. DOI: https://doi.org/10.1289/EHP2503.

GOVENDER, P.; SIVAKUMAR, V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: a review (1980-2019). **Atmospheric Pollution Research**, v. 11, n. 1, p. 40-56, 2020. DOI: https://doi.org/10.1016/j.apr.2019.09.009.

GRUS, J. **K-means and hierarchical clustering with Python**. Sebastopol: O'Reilly Media, 2016.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3. ed. Burlington, MA, EUA: Morgan Kaufmann Publishers, 2011. 703 p.

HASSE, J.-B.; LAJAUNIE, Q. Does the yield curve signal recessions? New evidence from an international panel data analysis. **The Quarterly Review of Economics and Finance**, v. 84, p. 9-22, 2022. DOI: https://doi.org/10.1016/j.qref.2022.01.001.

HERMAN, K. S.; SHENK, J. Pattern discovery for climate and environmental policy indicators. **Environmental Science & Policy**, v. 120, p. 89-98, 2021. DOI: https://doi.org/10.1016/j.envsci.2021.02.003.

IGUAL, L.; SEGUÍ, S. **Introduction to data science**: a Python approach to concepts, techniques and applications. Cham: Springer International Publishing, 2017. (Undergraduate Topics in Computer Science). DOI: https://doi.org/10.1007/978-3-319-50017-1.

IISD – INTERNATIONAL INSTITUTE FOR SUSTAINABLE DEVELOPMENT. Protocol regulating fine particulate matter enters into force. **SDG Knowledge Hub**, 15 out. 2019. Disponível em: https://sdg.iisd.org/news/air-pollution-protocol-regulating-fine-particulate-matter-enters-into-force/. Acesso em: 27 mar. 2022.

IOM – INTERNATIONAL ORGANIZATION FOR MIGRATION. **Spotlight on labour migration in Asia**: A factor analysis study. Geneva: IOM, 2021. Disponível em: https://impact.economist.com/perspectives/sites/default/files/spotlight_on_labour_migration_in_asia_7th_december_1.pdf. Acesso em: 7 nov. 2022.

JACOB, D.; KOTOVA, L.; TEICHMANN, C.; SOBOLOWSKI, S. P.; VAUTARD, R.; DONNELLY C.; KOUTROULIS, A. G.; GRILLAKIS, M. G.; TSANIS, I. K.; DAMM, A.; SAKALLI, A.; VAN VLIET, M. T. H. Climate impacts in Europe Under +1.5°C Global Warming. **Earth's Future**, v. 6, n. 2, p. 264-285, 2018. DOI: https://doi.org/10.1002/2017EF000710.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning**: with applications in R. 8. ed. New York: Springer, 2017. DOI: https://dx.doi.org/10.1007/978-1-4614-7138-7.

JIN, X.; HAN, J. K-means clustering. *In*: SAMMUT, C.; WEBB, G. I. (ed.). **Encyclopedia of machine learning**. 1. ed. Boston: Springer, 2011. p. 563-564. DOI: https://doi.org/10.1007/978-0-387-30164-8_425.

JUN, S.-P.; YOO, H. S.; LEE, J.-S. The impact of the pandemic declaration on public awareness and behavior: focusing on COVID-19 google searches. **Technological Forecasting and Social Change**, v. 166, 120592, 2021. DOI: https://doi.org/10.1016/j.techfore.2021.120592.

LALLOUÉ, B.; PADGET, M.; BROWNWOOD, I.; MINVIELLE, E.; KLAZINGA, N. Does size matter? The impact of caseload and expertise concentration on AMI 30-day mortality: a comparison across 10 OECD countries. **Health Policy**, v. 123, n. 5, p. 441-448, 2019. DOI: https://dx.doi.org/10.1016/j.healthpol.2019.03.007.

LAMB, W. F.; MINX, J. C. The political economy of national climate policy: architectures of constraint and a typology of countries. **Energy Research & Social Science**, v. 64, 101429, 2020. DOI: https://doi.org/10.1016/j.erss.2020.101429.

LAROSE, D. T.; LAROSE, C. D. **Discovering knowledge in data**: an introduction to Data Mining. 1. ed. Hoboken: Wiley, 2014. DOI: https://dx.doi.org/10.1002/9781118874059.

LINDEN, M.; RAY, D. Life expectancy effects of public and private health expenditures in OECD countries 1970–2012: panel time series approach. **Economic Analysis and Policy**, v. 56, p. 101-113, 2017. DOI: https://doi.org/10.1016/j.eap.2017.06.005.

LÓPEZ, L. A.; ARCE, G.; JIANG, X. Mapping China's flows of emissions in the world's carbon footprint: a network approach of production layers. **Energy Economics**, v. 87, 104739, 2020. DOI: https://doi.org/10.1016/j.eneco.2020.104739.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. *In*: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 5., Berkeley, 1967. **Proceedings** […]. Berkeley: University of California, 1967. v. 1, p. 281-297. Disponível em: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Some-methods-for-classification-and-analysis-of-multivariate-observations/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992. Acesso em: 15 out. 2022.

MAHMOOD, N.; ZHAO, Y.; LOU, Q.; GENG, J. Role of environmental regulations and eco-innovation in energy structure transition for green growth: evidence from OECD. **Technological Forecasting and Social Change**, v. 183, 121890, 2022. DOI: https://doi.org/10.1016/j.techfore.2022.121890.

MANNUCCI, P. M.; FRANCHINI, M. Health effects of ambient air pollution in developing countries. **International Journal of Environmental Research and Public Health**, v. 14, n. 9, p. 1048, 2017. DOI: https://dx.doi.org/10.3390/ijerph14091048.

MARTI, L.; PUERTAS, R. Influence of environmental policies on waste treatment. **Waste Management**, v. 126, p. 191-200, 2021. DOI: https://doi.org/10.1016/j.wasman.2021.03.009.

MENNA, A.; WALSH, P. R. Assessing environments of commercialization of innovation for SMEs in the global wine industry: a market dynamics approach. **Wine Economics and Policy**, v. 8, n. 2, p. 191-202, 2019. DOI: https://doi.org/10.1016/j.wep.2019.10.001.

MNATSAKANIAN, R. Environmental disaster in Eastern Europe. **Le Monde Diplomatique**, jul. 2000. Disponível em: https://mondediplo.com/2000/07/19envidisaster. Acesso em: 24 mar. 2022.

OECD – THE ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. **OECD Green Growth Indicators**. n. March, p. 68, 2018. Disponível em: https://stats.oecd.org/wbos/fileview2.aspx?IDFile=0eddc076-a4f9-4a2b-8e86-4190c8523b59. Acesso em: 30 dez. 2020.

O PAÍS que já foi o "menos desenvolvido" e hoje supera a China em crescimento. **BBC News Brasil**, 5 mar. 2020. Disponível em: https://www.bbc.com/portuguese/geral-51614697. Acesso em: 24 mar. 2022.

PARKER, S.; LIDDLE, B. Analysing energy productivity dynamics in the OECD manufacturing sector. **Energy Economics**, v. 67, p. 91-97, 2017a. DOI: https://doi.org/10.1016/j.eneco.2017.07.016.

PARKER, S.; LIDDLE, B. Economy-wide and manufacturing energy productivity transition paths and club convergence for OECD and non-OECD countries. **Energy Economics**, v. 62, p. 338-346, 2017b. DOI: https://doi.org/10.1016/j.eneco.2016.07.018.

PROKSCH, D; BUSCH-CASLER, J.; HABERSTROH, M. M.; PINKWART, A. National health innovation systems: clustering the OECD countries by innovative output in healthcare using a multi-indicator approach. **Research Policy**, v. 48, n. 1, p. 169-179, 2019. DOI: https://doi.org/10.1016/j.respol.2018.08.004.

RADU, C. F.; FENIŞER, C.; SCHEBESCH, K. B.; FENIŞER, F.; DOBREA, F. M. Study of the tax wedge in EU and other OECD Countries, using cluster analysis. **Procedia - Social and Behavioral Sciences**, v. 238, p. 687-696, 2018. DOI: https://doi.org/10.1016/j.sbspro.2018.04.051.

RASCHKA, S. **Python machine learning**. 1. ed. Birmingham: Packt Publishing, 2015. ISBN: 978-1-78355-513-0.

REIBLING, N.; ARIAANS, M.; WENDT, C. Worlds of healthcare: a healthcare system typology of OECD countries. **Health Policy**, v. 123, n. 7, p. 611-620, 2019. DOI: https://doi.org/10.1016/j.healthpol.2019.05.001.

RUGGERI, G.; CORSI, S. An analysis of the Fairtrade cane sugar small producer organizations network. **Journal of Cleaner Production**, v. 240, 118191, 2019. DOI: https://doi.org/10.1016/j.jclepro.2019.118191.

SOKHI, R. S.; SINGH, V.; QUEROL, X.; FINARDI, S.; TARGINO, A. C.; ANDRADE, M. F. *et al*. A global observational analysis to understand changes in air quality during exceptionally low anthropogenic emission conditions. **Environment International**, v. 157, 106818, 2021. DOI: https://doi.org/10.1016/j.envint.2021.106818.

SREEDHAR, C.; KASIVISWANATH, N.; REDDY, P. C. Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop. **Journal of Big Data**, v. 4, n. 27, 2017. DOI: https://doi.org/10.1186/s40537-017-0087-2.

VETTORETTI, M.; DI CAMILLO, B. A variable ranking method for machine learning models with correlated features: in-silico validation and application for diabetes prediction. **Applied Sciences (Switzerland)**, v. 11, n. 16, 7740, 2021. DOI: https://doi.org/10.3390/app11167740.

VICEDO-CABRERA, A. M.; SERA, F.; LIU, C.; ARMSTRONG, B.; MILOJEVIC, A.; GUO, Y. *et al*. Short term association between ozone and mortality: global two stage time series study in 406 locations in 20 countries. **The BMJ**, v. 368, m108, 2020. DOI: https://doi.org/10.1136/bmj.m108.

WANG, L.; WANG, S.; YUAN Z.; PENG, L. Analyzing potential tourist behavior using PCA and modified affinity propagation clustering based on Baidu index: taking Beijing city as an example. **Data Science and Management**, v. 2, p. 12-19, 2021. DOI: https://doi.org/10.1016/j.dsm.2021.05.001.

WHO – WORLD HEALTH ORGANIZATION. **Radon and health**. 2021. Disponível em: https://www.who.int/news-room/fact-sheets/detail/radon-and-health. Acesso em: 24 mar. 2022.

YAN, Z.; DU, K.; YANG, Z.; DENG, M. Convergence or divergence? Understanding the global development trend of low-carbon technologies. **Energy Policy**, v. 109, p. 499-509, 2017. DOI: https://doi.org/10.1016/j.enpol.2017.07.024.

ZENGIN, K.; ESGI, N.; ERGINER, E.; AKSOY, M. E. A sample study on applying data mining research techniques in educational science: developing a more meaning of data. **Procedia - Social and Behavioral Sciences**, v. 15, p. 4028-4032, 2011. DOI: https://doi.org/10.1016/j.sbspro.2011.04.408.

ZHANG, J. Z.; SRIVASTASA, P. R.; SHARMA, D.; EACHEMPATI, P. Big data analytics and machine learning: a retrospective overview and bibliometric analysis. **Expert Systems with Applications**, v. 184, 115561, 2021. DOI: https://doi.org/10.1016/j.eswa.2021.115561.

ZHU, J.; HUA, W. Visualizing the knowledge domain of sustainable development research between 1987 and 2015: a bibliometric analysis. **Scientometrics**, v. 110, n. 2, p. 893-914, 2017. DOI: https://doi.org/10.1007/s11192-016-2187-8.

ZWEIG, D.; TSAI, K. S.; SINGH, A. D. Reverse entrepreneurial migration in China and India: the role of the state. **World Development**, v. 138, 105192, 2021. DOI: https://doi.org/10.1016/j.worlddev.2020.105192.