http://journal.unoeste.br/index.php/ce/index
DOI: 10.5747/ce.2025.v17.e434
ISSN on-line 2178-8332

Colloquium Exactarum

Submetido: 13/08/2024 Revisado: 08/09/2025 Aceito: 24/09/2025

APLICAÇÃO DO MÉTODO K-MEANS NA ANÁLISE DE COEFICIENTES EPIDEMIOLÓGICOS DA COVID-19 EM UMA MACRORREGIÃO DO SUL DA BAHIA

Application of the K-Means Method in the Analysis of Epidemiological Coefficients of COVID-19 in a Macroregion of Southern Bahia

Dely da Silva Lima Neto[1]; Francisco Bruno Souza Oliveira[2]; Martha Ximena Torres Delgado[3]

Universidade Estadual de Santa Cruz – UESC, Ilhéus, BA.
e-mail: dslima@uesc.br, fbsoliveira@uesc.br, mxtd@uesc.br

RESUMO – A pandemia de COVID-19 atingiu o Brasil em 2020, com impactos distintos em cada região. No Núcleo Regional de Saúde Sul (NRS Sul) da Bahia, as cidades vivenciaram a pandemia de formas diversas. Aplicou-se uma análise de cluster (k-means) para agrupar municípios segundo semelhanças em coeficientes epidemiológicos (incidência, prevalência e letalidade), em três períodos da pandemia. Os resultados evidenciaram padrões específicos de comportamento e heterogeneidade regional, destacando municípios com maior letalidade em áreas de menor acesso à saúde. A abordagem contribui para decisões estratégicas e planejamento de políticas públicas.
Palavras-chave: K-means; COVID-19; Coeficientes epidemiológicos.

ABSTRACT – The COVID-19 pandemic hit Brazil in 2020, with heterogeneous impacts across regions. In the Southern Health Region (NRS Sul) of Bahia, cities experienced the crisis differently. A k-means cluster analysis grouped municipalities by epidemiological coefficients (incidence, prevalence, lethality) across three periods. Results revealed specific pandemic patterns and regional heterogeneity, with higher lethality in municipalities with limited healthcare access. This approach provides evidence to support strategic decision-making and public policy planning.
Keywords: K-means; COVID-19; Epidemiological coefficients.

1. INTRODUÇÃO

A pandemia da COVID-19, causada pelo SARS-CoV-2, emergiu no final de 2019 e rapidamente se espalhou, resultando em milhões de casos e óbitos (WHO, 2020). No Brasil, o primeiro caso foi confirmado em fevereiro de 2020, e o país logo se tornou um dos epicentros, enfrentando desafios devido a desigualdades socioeconômicas e diferenças regionais na resposta (Oliveira *et al.*, 2021).

De acordo com um estudo conduzido por Oliveira *et al.* (2021), que analisou a evolução da pandemia de COVID-19 no Brasil, observou-se uma tendência preocupante de aumento significativo no número de casos confirmados e óbitos. O estudo destacou a importância de implementar medidas de contenção efetivas, como o distanciamento social, o uso de máscaras e a ampliação da capacidade de testagem, como forma de controlar a disseminação do vírus e reduzir o impacto da doença na saúde pública do país. A COVID-19 chegou ao Brasil no início de 2020 e, desde então, o país ainda convive com o reflexo dos resultados da doença.

Na Bahia, impactos significativos exigiram análises regionais. Estudos anteriores (Santos *et al.*, 2022) destacaram a necessidade de fortalecer testagem, rastreamento e distribuição equitativa de recursos de saúde. Nesse contexto, este trabalho analisa a evolução da COVID-19 no NRS Sul da Bahia, utilizando o algoritmo k-means para agrupar municípios de acordo com coeficientes epidemiológicos, permitindo compreender padrões espaço-temporais da pandemia e subsidiar políticas locais.

A análise epidemiológica desempenha um papel fundamental na compreensão e no controle de doenças em uma determinada região. Com o avanço da tecnologia e a disponibilidade de grandes volumes de dados, técnicas de aprendizado de máquina têm sido amplamente exploradas para auxiliar nesse processo. Entre essas técnicas, destaca-se o algoritmo *K-means*, que tem sido aplicado com sucesso em diversas áreas, incluindo a epidemiologia. Neste contexto, objetiva-se analisar o impacto da COVID-19 no Brasil, com enfoque na realidade em cidades do sul da Bahia. Foi aplicado o *K-means* para analisar dados epidemiológicos da COVID-19, considerando variáveis como número de casos confirmados, taxa de mortalidade e taxa de transmissão. Além disso, este estudo concentrou-se especificamente no sul da Bahia, uma sub-região com características próprias e um contexto socioeconômico particular. A aplicação desta técnica de *clustering* nos dados do sul da Bahia permitiu uma análise mais localizada e direcionada, fornecendo informações relevantes para o enfrentamento da COVID-19 nessa área específica.

Embora estudos prévios tenham aplicado técnicas de *clustering* na análise da COVID-19 (Alves *et al.*, 2020; Muhtasim; Masud, 2023), este trabalho avança ao: Analisar padrões espaço-temporais em três fases distintas da pandemia (inicial, intermediária e de transição); Focar especificamente em municípios do NRS Sul da Bahia, considerando suas particularidades socioeconômicas; Integrar dados epidemiológicos normalizados (incidência, prevalência e letalidade) com georreferenciamento e Oferecer subsídios para políticas públicas locais baseadas em evidências

Este artigo está organizado da seguinte forma: a seção 2 (Trabalhos Correlatos) apresenta estudos prévios sobre aplicações de técnicas de *clustering* na análise epidemiológica da COVID-19. A seção 3 (Materiais e Métodos) detalha a base de dados utilizada, o processamento dos coeficientes epidemiológicos e a implementação do algoritmo *K-means*. Na seção 4 (Resultados e Discussões), são analisados os padrões de agrupamento identificados em cada período temporal estudado. Por fim, a seção 5 (Considerações Finais) sintetiza as principais contribuições do trabalho e suas implicações para o enfrentamento da pandemia na região estudada.

2. TRABALHOS CORRELATOS

Alves *et al.* (2020) aplicaram o método k-means para agrupar as Unidades Federativas brasileiras com base em incidência, prevalência e letalidade da COVID-19. Esses coeficientes

permitem avaliar risco de contágio, proporção de doentes e gravidade da doença. O estudo foi dividido em três períodos (início, fase intermediária e transição) e identificou padrões distintos de propagação, evidenciando a heterogeneidade da pandemia no país. Os resultados contribuem para a compreensão da situação epidemiológica e para o planejamento de estratégias de enfrentamento.

Adicionalmente, Muhtasim e Masud (2023) também utilizaram o *K-means* como técnica de aprendizagem não supervisionada em seus estudos para identificar os aspectos cruciais de casos confirmados, casos de morte, casos de recuperação, total de casos por milhão e total de mortes por milhão em dados da Covid-19 em uma combinação de duzentos e trinta países. Os autores calcularam as taxas de incidência e mortalidade em uma análise de série temporal (o período de estudo começou desde o primeiro dia de um caso de Covid-19 até 22 de setembro de 2022) para identificar o comportamento do vírus nas diferentes nações. Ainda na metodologia do trabalho, também utilizaram técnicas de aprendizado de máquina, incluindo árvores de decisão, florestas aleatórias (Random forests), além de abordagens de agrupamento hierárquico. O estudo utilizou técnicas de clusterização para agrupar países com base em seus dados de COVID-19 e identificou três clusters significativos. O cluster 0 incluiu países com altas taxas de recuperação e baixas taxas de mortalidade, o cluster 1 incluiu países com baixas taxas de mortalidade e recuperação, e o cluster 2 incluiu países com altas taxas de mortalidade e baixas taxas de recuperação. O estudo também descobriu que a prevalência da doença tem uma correlação substancial com o COVID-19, mas os indicadores de saúde ambiental têm uma correlação menor. Além disso, o estudo mostrou que há uma correlação significativa entre o número total de mortes e pacientes classificados como críticos. Os resultados do estudo podem ser usados pelos governantes para tomar decisões informadas sobre o controle da epidemia.

Ainda no contexto de utilizar o *K-means* na análise de dados epidemiológicos da Covid-19, Gohari *et al.* (2022) utilizaram uma abordagem que inclui três etapas para agrupar tendências de incidência e mortalidade específicas por Covid-19 de cada um dos países dos duzentos e dezesseis selecionados para estudos. O período selecionado para estudo dos dados de novos casos diários e mortalidade da Covid-19 foi entre 3 de janeiro de 2020 à 28 de agosto de 2021. O método extrai as características essenciais de uma trajetória e as considera em um algoritmo de agrupamento *K-means*. Na primeira etapa, dividiram as linhas de tendência em medidas que descrevem diferentes aspectos de uma tendência. Em seguida, aplicaram um método de redução de dimensão às medidas e, finalmente, um algoritmo *K-means* as agrupa. Realizaram esta abordagem para taxas de incidência e mortalidade separadamente e, finalmente, as compararam. Os autores identificaram três padrões distintos de incidência e mortalidade por COVID-19, que eles chamaram de "clusters". O primeiro cluster incluiu países com baixa incidência e mortalidade, o segundo cluster incluiu países com alta incidência e mortalidade, e o terceiro cluster incluiu países com alta incidência e baixa mortalidade. Os autores também descobriram que fatores como a idade média da população, a densidade populacional e o acesso a cuidados de saúde afetaram a disseminação e o impacto da pandemia em diferentes países.

Embora os trabalhos revisados compartilhem similaridades metodológicas: Alves *et al.* (2020) aplicaram *K-means* apenas em nível estadual, enquanto este estudo opera em escala municipal; Muhtasim e Masud (2023) focaram em comparações entre países, sem análise temporal estratificada; Gohari *et al.* (2022) não consideraram variáveis geoespaciais, diferentemente da abordagem aqui proposta. A presente pesquisa inova ao combinar: segmentação cronológica em três fases pandêmicas, análise integrada de coeficientes epidemiológicos normalizados e mapeamento georreferenciado dos clusters identificados

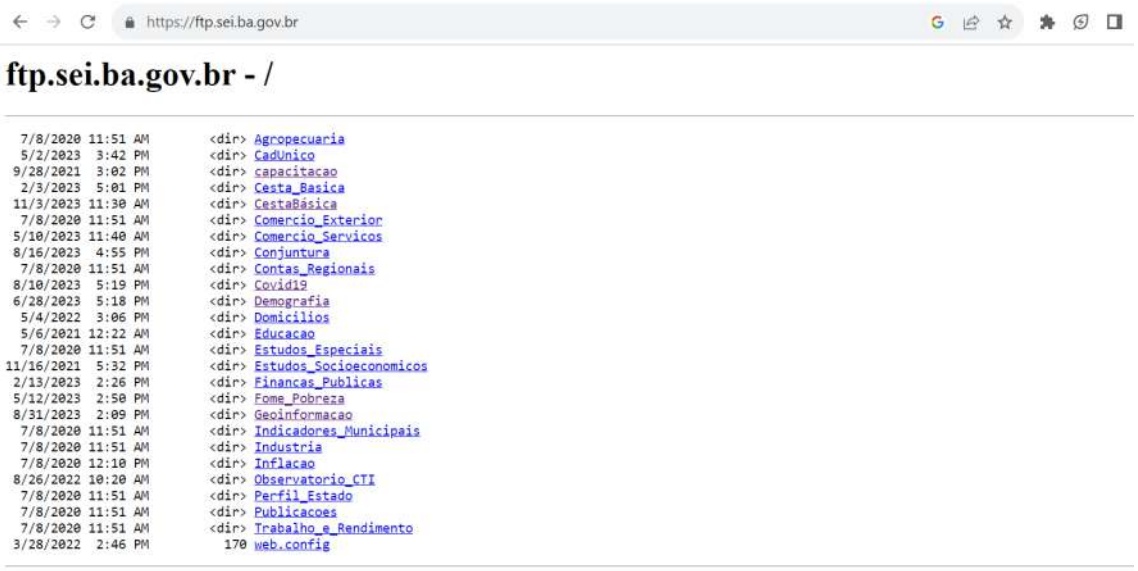
3. MATERIAIS E MÉTODOS

3.1 Base de Dados

O portal ftp.sei.ba.gov.br (Figura 1) é um endereço de FTP que pertence ao governo do estado da Bahia. FTP significa "*File Transfer Protocol*" e é um protocolo utilizado para transferir arquivos pela *internet*. Nesse caso, o endereço ftp.sei.ba.gov.br indica que é possível acessar arquivos do governo da Bahia através desse servidor FTP. Alguns exemplos de dados comuns encontrados no servidor:

1. Documentos oficiais: Arquivos em formato PDF, DOC, XLS, entre outros, contendo informações sobre políticas, regulamentos, relatórios, etc.;
2. Dados estatísticos: Conjuntos de dados em formato CSV ou outros formatos utilizados para análise e pesquisa;
3. Publicações e materiais educacionais: E-books, manuais, guias e outros materiais informativos relacionados a áreas específicas, como saúde, educação, meio ambiente, entre outros.

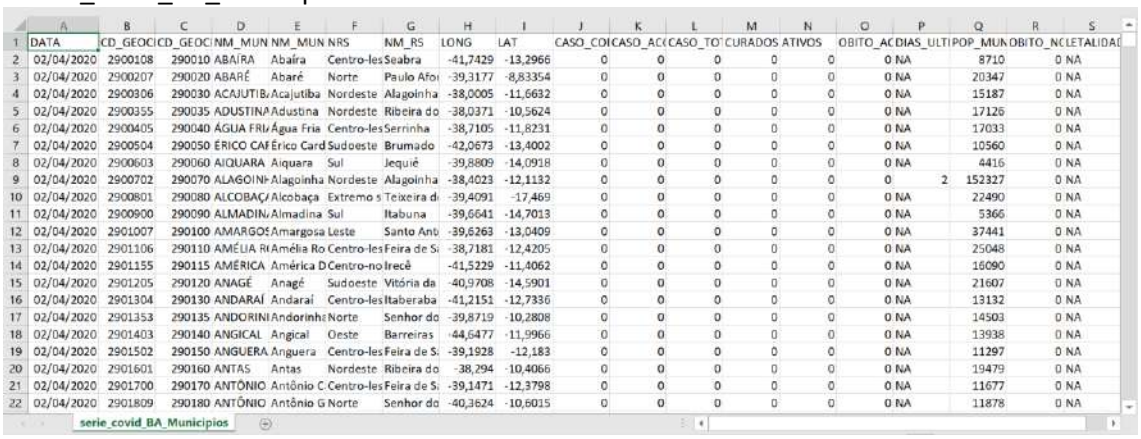
Figura 1. Portal ftp.sei.ba.gov.br.



Fonte: https://ftp.sei.ba.gov.br/.

Dentre os diferentes tipos de dados encontrados no servidor, foi utilizado para esta pesquisa a base de dados serie_covid_BA_Municipios.csv (Figura 2), arquivo este que é atualizado quase que diariamente no repositório com dados epidemiológicos da COVID-19 de todas as cidades da Bahia. O arquivo apresenta as seguintes informações por colunas: data, código de geografia do município, código de geocodificação municipal de 6 dígitos, nome do município, núcleo regional de saúde, nome da região de saúde, longitude, latitude, casos confirmados, casos acompanhados, casos totais, curados, ativos, óbitos acumulados, dias do último caso, população, óbitos novos, letalidade e prevalência.

Figura 2. Serie_covid_BA_Municipios.csv.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	DATA	CD_GEOCI	CD_GEOCI	NM_MUN	NM_MUN	NRS	NM_RS	LONG	LAT	CASO_CO	CASO_AC	CASO_TO	CURADOS	ATIVOS	OBITO_AC	DIAS_ULTI	POP_MUN	OBITO_NC	LETALIDAI
2	02/04/2020	2900108	290010	ABAÍRA	Abaíra	Centro-les	Seabra	-41,7429	-13,2966	0	0	0	0	0	0	NA	8710	0	NA
3	02/04/2020	2900207	290020	ABARÉ	Abaré	Norte	Paulo Afo	-39,3177	-8,83354	0	0	0	0	0	0	NA	20347	0	NA
4	02/04/2020	2900306	290030	ACAJUTIB	Acajutiba	Nordeste	Alagoinha	-38,0005	-11,6632	0	0	0	0	0	0	NA	15187	0	NA
5	02/04/2020	2900355	290035	ADUSTINA	Adustina	Nordeste	Ribeira do	-38,0371	-10,5624	0	0	0	0	0	0	NA	17126	0	NA
6	02/04/2020	2900405	290040	ÁGUA FRI	Água Fria	Centro-les	Serrinha	-38,7105	-11,8231	0	0	0	0	0	0	NA	17033	0	NA
7	02/04/2020	2900504	290050	ÉRICO CA	Érico Card	Sudoeste	Brumado	-42,0673	-13,4002	0	0	0	0	0	0	NA	10560	0	NA
8	02/04/2020	2900603	290060	AIQUARA	Aiquara	Sul	Jequié	-39,8809	-14,0918	0	0	0	0	0	0	NA	4416	0	NA
9	02/04/2020	2900702	290070	ALAGOINH	Alagoinha	Nordeste	Alagoinha	-38,4023	-12,1132	0	0	0	0	0	0	2	152327	0	NA
10	02/04/2020	2900801	290080	ALCOBAÇA	Alcobaça	Extremo s	Teixeira d	-39,4091	-17,469	0	0	0	0	0	0	NA	22490	0	NA
11	02/04/2020	2900900	290090	ALMADINA	Almadina	Sul	Itabuna	-39,6641	-14,7013	0	0	0	0	0	0	NA	5366	0	NA
12	02/04/2020	2901007	290100	AMARGOS	Amargosa	Leste	Santo Ant	-39,6263	-13,0409	0	0	0	0	0	0	NA	37441	0	NA
13	02/04/2020	2901106	290110	AMÉLIA RI	Amélia Ro	Centro-les	Feira de S	-38,7181	-12,4205	0	0	0	0	0	0	NA	25048	0	NA
14	02/04/2020	2901155	290115	AMÉRICA	América D	Centro-no	Irecê	-41,5229	-11,4062	0	0	0	0	0	0	NA	16090	0	NA
15	02/04/2020	2901205	290120	ANAGÉ	Anagé	Sudoeste	Vitória da	-40,9708	-14,5901	0	0	0	0	0	0	NA	21607	0	NA
16	02/04/2020	2901304	290130	ANDARAÍ	Andaraí	Centro-les	Itaberaba	-41,2151	-12,7336	0	0	0	0	0	0	NA	13132	0	NA
17	02/04/2020	2901353	290135	ANDORINI	Andorinha	Norte	Senhor do	-39,8719	-10,2808	0	0	0	0	0	0	NA	14503	0	NA
18	02/04/2020	2901403	290140	ANGICAL	Angical	Oeste	Barreiras	-44,6477	-11,9966	0	0	0	0	0	0	NA	13938	0	NA
19	02/04/2020	2901502	290150	ANGUERA	Anguera	Centro-les	Feira de S	-39,1928	-12,183	0	0	0	0	0	0	NA	11297	0	NA
20	02/04/2020	2901601	290160	ANTAS	Antas	Nordeste	Ribeira do	-38,294	-10,4066	0	0	0	0	0	0	NA	19479	0	NA
21	02/04/2020	2901700	290170	ANTÔNIO	Antônio C	Centro-les	Feira de S	-39,1471	-12,3798	0	0	0	0	0	0	NA	11677	0	NA
22	02/04/2020	2901809	290180	ANTÔNIO	Antônio G	Norte	Senhor do	-40,3624	-10,6015	0	0	0	0	0	0	NA	11878	0	NA

serie_covid_BA_Municipios

Fonte: https://ftp.sei.ba.gov.br/Covid19/.

Os municípios da Bahia são agrupados por Núcleo Regional de Saúde (NRS) que são divididos em 9 macrorregiões de saúde: Macrorregião Norte, Macrorregião Nordeste, Macrorregião Leste, Macrorregião Sul, Macrorregião Extremo Sul, Macrorregião Sudoeste, Macrorregião Oeste, Macrorregião Centro Norte e Macrorregião Centro Leste. O NRS Sul, com 68 cidades (Aiquara (0), Almadina (1), Apuarema (2), Arataca (3), Aurelino Leal (4), Barra do Rocha (5), Barro Preto (6), Boa Nova (7), Brejões (8), Buerarema (9), Cairu (10), Camacan (11), Camamu (12), Canavieiras (13), Coaraci (14), Cravolândia (15), Dário Meira (16), Floresta Azul (17), Gandu (18), Gongogi (19), Ibicaraí (20), Ibirapitanga (21), Ibirataia (22), Igrapiúna (23), Ilhéus (24), Ipiaú (25), Irajuba (26), Iramaia (27), Itabuna (28), Itacaré (29), Itagi (30), Itagibá (31), Itaju do Colônia (32), Itajuípe (33), Itamari (34), Itapé (35), Itapitanga (36), Itaquara (37), Itiruçu (38), Ituberá (39), Jaguaquara (40), Jequié (41), Jitaúna (42), Jussari (43), Lafaiete Coutinho (44), Lajedo do Tabocal (45), Manoel Vitorino (46), Maracás (47), Maraú (48), Mascote (49), Nilo Peçanha (50), Nova Ibiá (51), Nova Itarana (52), Pau Brasil (53), Piraí do Norte (54), Planaltino (55), Santa Cruz da Vitória (56), Santa Inês (57), Santa Luzia (58), São José da Vitória (59), Taperoá (60), Teolândia (61), Ubaitaba (62), Ubatã (63), Uma (64), Uruçuca (65), Valença (66), Wenceslau Guimarães (67)) foi escolhido para pesquisa.

As informações contidas foram separadas em três períodos seguindo uma ordem cronológica de contágio da doença como fase inicial: (02-04-2020 a 31-12-2020), fase intermediária: (01-01-2021 a 30-06-2021) e fase de transição da doença (01-07-2023 a 01-12-2023).

Os períodos selecionados para análise foram escolhidos por serem representativos das diferentes fases da pandemia de COVID-19.

O período de 02-04-2020 a 31-12-2020 (período 1) corresponde à primeira onda da pandemia, quando a doença ainda era pouco conhecida e não havia vacinas ou tratamentos eficazes. Nesse período, o número de casos e mortes aumentou rapidamente, e os sistemas de saúde de todo o mundo foram sobrecarregados.

O período de 01-01-2021 a 30-06-2021 (período 2) corresponde à fase de contenção ou período de estabilização, quando a doença voltou a se espalhar rapidamente após uma breve redução no número de casos no final de 2020. Nesse período, as vacinas começaram a ser distribuídas em alguns países, mas ainda não estavam amplamente disponíveis. No início deste período, as primeiras doses da vacina começam a ser aplicadas no estado da Bahia. As medidas de distanciamento social e restrição de atividades foram reintroduzidas em muitos países, período em que se pode analisar como as cidades vão se comportar perante ao vírus.

O período de 01-07-2023 a 01-12-2023 (período 3) corresponde à fase de transição para normalidade, o Brasil passou por um período de recuperação gradual e adaptação às mudanças resultantes da pandemia. Com a implementação bem-sucedida de estratégias de saúde pública e a

disseminação da vacinação, as taxas de infecção diminuíram significativamente. As restrições foram relaxadas progressivamente, permitindo a retomada das atividades econômicas e sociais. No entanto, a vigilância contínua e a preparação para possíveis ressurgimentos da doença permaneceram essenciais para garantir uma transição segura para a normalidade. Nesse período, a maioria das pessoas já tinha sido vacinada ou infectada com o vírus, o que reduziu a gravidade dos casos.

A escolha desses períodos permite analisar a evolução dos coeficientes epidemiológicos da COVID-19 ao longo do tempo, em diferentes fases da pandemia. A SESAB (Secretaria de Saúde do Estado da Bahia) não informou dados epidemiológicos (casos confirmados, casos acompanhados, casos totais) em seus boletins entre 01/07/2021 a 30/06/2023, o que impossibilitou o processamento dos coeficientes epidemiológicos.

Posteriormente, foi calculado os coeficientes de incidência, prevalência e letalidade dos municípios do NRS Sul pertencentes a cada um dos grupos em cada período considerado e, finalmente, utilizou-se esses coeficientes calculados como variáveis para obter os clusters dos municípios em cada período estudado, o que possibilita entender como ocorreu o avanço da COVID-19 em cada conjunto encontrado.

3.2 Processamento dos Coeficientes Epidemiológicos

Para a construção dos coeficientes epidemiológicos, foram analisados os seguintes dados: casos confirmados, casos acompanhados, óbitos acumulados e tamanho da população. É importante destacar que a coleta e a divulgação de dados epidemiológicos precisos e atualizados são fundamentais para o monitoramento e a resposta efetiva à pandemia de COVID-19. No entanto, em muitos casos, a coleta e a divulgação de dados envolvem múltiplos atores, incluindo autoridades locais de saúde, departamentos estaduais de saúde e agências nacionais de saúde.

Atrasos na divulgação de dados podem ocorrer por diversos motivos, como problemas técnicos, preocupações com a qualidade dos dados ou desafios administrativos. Para lidar com essa questão, é importante estabelecer diretrizes claras e protocolos para a coleta e a divulgação de dados, bem como fornecer recursos e suporte adequados às autoridades locais de saúde. Além disso, a transparência e a responsabilidade no processo de divulgação de dados podem ajudar a construir confiança e garantir que dados precisos e atualizados estejam disponíveis para a tomada de decisões e o desenvolvimento de políticas.

3.2.1 Coeficiente de Incidência

O coeficiente de incidência é a razão entre o número de casos novos de uma doença que ocorre em uma comunidade, em um intervalo de tempo determinado, e a população exposta ao risco de adquirir essa doença no mesmo período (Menezes, 2001).

A coluna CASO_CONF (Soma de Casos Confirmados) já presente na base de dados apresenta a soma dos casos totais de cada município por período selecionado.

A coluna POP_MUN (Tamanho População) já presente na base de dados apresenta o tamanho da população do município.

A construção do coeficiente de incidência (INCIDENCIA) da COVID-19 no NRS é dada pela equação 1:

$$Incidência = \left(\frac{CasosConfirmados}{TamanhodaPopulação}\right) x100 \qquad (1)$$

3.2.2 Coeficiente de Prevalência

Para Rothman e Greeland (2008), o coeficiente de prevalência é uma variação dessa medida que leva em consideração a duração da doença ou condição. É definido como o número de casos de uma doença ou condição em uma população dividido pelo tempo total de pessoa em risco de desenvolver a doença ou condição.

O coeficiente de prevalência é uma medida útil em estudos epidemiológicos porque fornece uma estimativa mais precisa da carga de uma doença ou condição em uma população do que a taxa de prevalência sozinha. Ao levar em consideração a duração da doença ou condição, o coeficiente de prevalência pode ajudar os pesquisadores a entender melhor o impacto da doença ou condição na população ao longo do tempo.

A coluna CASO_TOTAL (Soma de Casos Totais) já presente na base de dados apresenta a soma dos casos confirmados de cada município por período selecionado.

A coluna POP_MUN (Tamanho População) já presente na base de dados apresenta o tamanho da população do município.

A construção do coeficiente de prevalência (PREVALENCIA) da COVID-19 no NRS é dada pela equação 2:

$$Prevalência = \left(\frac{CasosTotais}{TamanhodaPopulação}\right) x100 \qquad (2)$$

3.2.3 Coeficiente de Letalidade

Em um estudo elaborado por de Souza *et al.* (2020), eles explicam que no contexto da pandemia de COVID-19, o coeficiente de letalidade tem sido amplamente utilizado para avaliar a gravidade da doença em diferentes populações e comparar o impacto da pandemia em diferentes países. Por exemplo, em um estudo que analisou o impacto da COVID-19 na população do Brasil, os autores calcularam o coeficiente de letalidade para estimar o risco de morte associado à doença.

A coluna OBITO_ACUM (Soma de Óbitos Acumulados) já presente na base de dados apresenta a soma dos óbitos acumulados de cada município por período selecionado.

A coluna CASO_TOTAL (Soma de Casos Totais) já presente na base de dados apresenta a soma dos casos confirmados de cada município por período selecionado.

A construção do coeficiente de letalidade (LETALIDADE) da COVID-19 no NRS é dada pela equação 3:

$$Letalidade = \left(\frac{ÓbitosAcumulados}{CasosTotais}\right) x100 \qquad (3)$$

3.2.4 Normalização dos Dados

Para o processamento dos coeficientes epidemiológicos, foi necessário a normalização dos dados. Os autores Hastie *et al.* (2009) discutem a importância de ajustar a escala e distribuição dos dados antes de aplicar modelos de aprendizado de máquina. Eles explicam que a normalização de dados permite que diferentes atributos sejam comparáveis em uma mesma escala, evitando problemas como a dominância de características com maior magnitude. A normalização de dados envolve a aplicação de transformações nos valores dos atributos de um conjunto de dados, de modo a colocá-los em uma escala comum.

Para normalizar as colunas INCIDÊNCIA, PREVALENCIA e LETALIDADE, identificaram-se os valores mínimo (min) e o valor máximo (max) de cada coluna. A cada valor da coluna, aplicou-se a fórmula de normalização min-max, conforme apresentado na equação 4:

$$ValorNormalizado = \frac{ValorOriginal - Mínimo}{Máximo - Mínimo} \qquad (4)$$

Essa técnica ajusta os valores dos atributos para um intervalo específico, geralmente entre 0 e 1. Isso é feito através de uma transformação linear que preserva a forma dos dados originais.

Após a normalização dos dados, foram criadas as colunas: INCIDENCIA_NORM (Coeficiente de Incidência Normalizado), PREVALENCIA_NORM (Coeficiente de Prevalência Normalizado) e LETALIDADE_NORM (Coeficiente de Letalidade Normalizado).

3.3 Bibliotecas em Python

As bibliotecas em Python consistem em pacotes de código que expandem as funcionalidades da linguagem. A plataforma oferece bibliotecas especializadas para diversas

aplicações computacionais, incluindo processamento de dados, análise estatística e manipulação de imagens.

Para este estudo, foram utilizadas as seguintes bibliotecas:

- *folium*: uma biblioteca em Python que permite a criação de visualizações interativas de dados geoespaciais em mapas. Ela é construída sobre a biblioteca *JavaScript Leaflet.js* e oferece uma interface simples e intuitiva para criar mapas interativos com camadas personalizadas.
- *matplotlib.pyplot*: uma das principais ferramentas para visualização de dados em *Python*. Com ela, é possível criar gráficos e visualizações de alta qualidade para análise exploratória de dados, apresentação de resultados e comunicação de *insights*.
- *mpl_toolkits.mplot3d:* é um conjunto de ferramentas para plotagem de gráficos tridimensionais no *Matplotlib*. Ela fornece uma variedade de funções e classes para criar gráficos 3D de dados científicos e de engenharia. A classe utilizada neste trabalho foi *Axes3D*.
- *pandas*: ferramenta poderosa para análise de dados, fornecendo estruturas de dados flexíveis e eficientes para manipulação e processamento de dados. Com ela, é possível realizar tarefas como leitura de arquivos CSV, limpeza e transformação de dados, além de operações avançadas como agrupamento e pivoteamento de dados.
- *pandas.plotting:* fornece uma variedade de funções para plotar dados em um *DataFrame* do *Pandas*. Foi usada para criar uma tabela visual dentro de um gráfico *matplotlib*. Especificamente, a tabela é criada usando a função *table* da submódulo *plotting* do *Pandas*.
- *sklearn.cluster*: biblioteca voltada para o aprendizado de máquina. Oferece uma variedade de algoritmos de *clustering*, como *k-means* e DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), que permitem agrupar dados com base em suas características e padrões.

3.4 O método *k-means*

Neste trabalho, o método *k-means* foi aplicado no estudo para analisar os coeficientes epidemiológicos da COVID-19 na macrorregião sul da Bahia. O objetivo é identificar padrões de agrupamento nos dados e entender como as cidades se comportam em relação à pandemia.

Na literatura, existem técnicas de agrupamentos que consideram múltiplas variáveis como condicionantes, conhecida como análise de agrupamentos para dados multivariados (FÁVERO e BELFIORE 2019; FERREIRA 2018), cujo objetivo é dividir as observações em grupos que são homogêneos internamente e heterogêneos externamente.

De acordo com Fontana e Naldi (2009), *K-means* utiliza o conceito de centroides como protótipos representativos dos grupos, onde o centroide representa o centro de um grupo, sendo calculado pela média de todos os objetos do grupo.

Segundo Makih e Tarigan (2021), a análise de cluster *K-Means* é um método de análise de cluster não hierárquica que busca particionar objetos existentes em um ou mais clusters ou grupos de objetos com base em suas características, de modo que objetos que têm as mesmas características sejam agrupados no mesmo cluster e objetos que têm características diferentes sejam agrupados em outros clusters. A finalidade do agrupamento é minimizar a função objetiva definida no processo de agrupamento, que basicamente tenta minimizar as variações dentro de um cluster e maximizar a variação entre clusters.

Após a normalização dos dados, foi necessário definir o número de clusters (K), número este definido através do método elbow com análise do ponto de inflexão no gráfico da inércia, retornando k=3 como número ótimo e os dados a serem agrupados, os coeficientes

epidemiológicos processados da base de dados. Para implementar o algoritmo usando a "scikit-learn", foi necessário importar a classe *K-Means* da biblioteca.

A fórmula principal do *k-means* envolve a minimização da soma dos quadrados das distâncias entre os pontos de dados e os centroides dos clusters.

A minimização é alcançada iterativamente em dois passos principais:

Atribuição de Clusters (Responsabilidade):

- Para cada ponto de dados *xi*, calcular a distância entre *xi* e todos os centroides.
- Atribuir *xi* ao cluster cujo centroide é o mais próximo.
- Isso é geralmente feito usando a distância euclidiana, mas outras métricas podem ser utilizadas (Manhattan, Minkowski e Mahalanobis)

A fórmula da distância euclidiana entre dois pontos *xi* e *cj* no espaço n-dimensional é dada pela equação 5:

$$D(xi, cj) = \sqrt{\sum_{i=1}^{n}(xi_i - cj_i)^2} \qquad (5)$$

Atualização dos centroides

- Calcular os novos centroides para cada cluster, que é a média dos pontos atribuídos a esse cluster.

A fórmula para o centroide *cj* em um cluster *j* é dada pela equação 6:

$$cj = \frac{1}{|Cj|}\sum_{x_i \in C_j} x_i \qquad (6)$$

Onde:

- cj: é o vetor centróide do cluster j
- C_j: é o conjunto de pontos pertencentes ao cluster j
- $\left|C_j\right|$: é o número de pontos em C_j
- x_i : são os vetores (pontos) pertencentes ao cluster

Esses dois passos são repetidos iterativamente até que a atribuição de clusters não mude significativamente ou um número máximo de iterações seja atingido. O *K-Means* é uma combinação da fórmula de distância euclidiana para atribuição de clusters e a fórmula para a atualização dos centroides.

A variabilidade nos resultados do *K-Means* pode ocorrer devido à natureza estocástica do algoritmo. O *K-Means* depende de inicializações aleatórias dos centroides, o que pode levar a diferentes agrupamentos em diferentes execuções, especialmente se os dados possuírem uma estrutura complexa ou distribuição irregular. Nesta fase, o algoritmo de agrupamento a cada vez que era executado gerava uma nova combinação de clusters.

Para garantir uma certa consistência nos resultados, definimos manualmente os centroides iniciais usando o parâmetro init na inicialização do *K-Means*. Uma opção comum nestes casos é usar o método *k-means*++ para inicialização, que tenta posicionar os centroides de maneira mais estratégica. O algoritmo *K-Means* foi aplicado com 3 clusters usando inicialização *k-means*++ para uma convergência mais rápida e um estado randômico fixo para reprodutibilidade, resolvendo o problema citado no parágrafo anterior.

Os resultados do *k-means* foram adicionados de volta ao DataFrame original, atribuindo a cada cidade um rótulo de cluster. Em seguida, um mapa interativo foi criado utilizando a biblioteca folium, centrado na Bahia, para cada coeficiente epidemiológico nos três períodos selecionados. De forma complementar, a saída dos resultados também pode ser visualizada através de tabelas e gráficos.

As cidades foram plotadas no mapa com cores diferentes de acordo com seus grupos. Foram utilizadas cores como azul para o cluster 1, verde para o cluster 2 e vermelho para o cluster 3 para representar os diferentes grupos identificados pelo *k-means*.

O mapa resultante foi salvo em um arquivo HTML (HyperText Markup Language) chamado "mapa_bahia_clusters.html", permitindo a visualização interativa dos clusters e sua distribuição geográfica na macrorregião sul da Bahia. Adicionalmente, foi inserido uma legenda de identificação no mapa com a média de cada cluster com os seus respectivos coeficientes epidemiológicos.

4. RESULTADOS E DISCUSSÕES

Nesta seção, são discutidas as descobertas e insights derivados da clusterização dos coeficientes epidemiológicos, fornecendo uma compreensão mais aprofundada sobre a distribuição e comportamento da doença nas cidades investigadas.

4.1. O método *elbow* (método do cotovelo) para definição do número ótimo de *clusters*

O método Elbow é uma técnica amplamente utilizada para determinar o número ideal de agrupamentos em um conjunto de dados. Esta abordagem envolve traçar o número de clusters em relação à soma dos quadrados das distâncias dos pontos de dados em relação aos seus centroides correspondentes. O gráfico resultante geralmente exibe uma curva semelhante a um "cotovelo", e o ponto de inflexão desse cotovelo é considerado o número ideal de clusters. Este marco indica o ponto em que adicionar mais agrupamentos não resulta em uma redução significativa na soma dos quadrados das distâncias.

Primeiramente, para determinar o número ideal de clusters, dados foram carregados do arquivo CSV usando a biblioteca pandas do Python e as linhas com valores ausentes foram removidas. Em seguida, foram selecionadas as colunas relevantes (Coeficiente de Incidência Normalizado, Coeficiente de Prevalência Normalizado e Coeficiente de Letalidade Normalizado) para o algoritmo *K-Means*.

Para testar diferentes números de clusters e calcular a inércia, foi utilizada uma lista vazia para armazenar os valores. A inércia, medida da variabilidade dos dados dentro de um cluster, serve para avaliar a qualidade do agrupamento. O objetivo do *K-Means* é minimizar a inércia, agrupando os dados em conjuntos homogêneos.

Por fim, plotou-se a inércia em relação aos clusters testados. O objetivo era encontrar o número ideal que minimizasse a inércia. Esse número foi determinado pelo ponto de inflexão no gráfico, onde adicionar mais clusters não reduzia a inércia significativamente. Um algoritmo desenvolvido em Python utilizando as bibliotecas: “pandas”, “matplotlib.pyplot” e “sklearn.cluster”. KMeans realizou o agrupamento *K-Means* em um conjunto de dados. Uma função nomeada encontrar_numero_clusters_ideal retorna o índice do ponto de cotovelo no gráfico que representa o número ideal de clusters, testando diferentes números de clusters e calculando a inércia para cada um.

Uma função nomeada *gerar_tabela_inercias* gera uma tabela de valores de inércia para diferentes números de *clusters*, ela chama a função *encontrar_numero_clusters_ideal* e cria um *DataFrame* para armazenar os resultados.

O código utiliza o método *inertia_* fornecido pela implementação do algoritmo *K-Means* na biblioteca “scikit-learn” para calcular a inércia. Este método é dado pela equação 7, que calcula a soma das distâncias quadráticas das amostras para os centroides mais próximos.

$$I = \sum_{i=1}^{n} min_{\mu_{j \in \mu}} \vee x_i - \mu_j \vee^2 \qquad (7)$$

Onde:

- I é inércia
- $\sum_{i=1}^{n}$ Indica a soma sobre todas as amostras $x_i$no conjunto de dados
- $min_{\mu_{j \in \mu}}$ Denota o valor mínimo dentre as distâncias das amostras para os centroides $\mu_j$pertencentes ao conjunto de todos os centroides μ
- $\vee x_i - \mu_j \vee^2$ Refere-se à distância euclidiana ao quadrado entre a amostra x_i e o centroide μ_j

As Figuras 3, 4 e 5 apresentam a relação entre o número de *clusters* e os respectivos valores de inércia para os três períodos analisados. A inércia, neste contexto, representa a soma das distâncias quadradas de cada ponto ao centroide de seu *cluster*, servindo como medida de coesão interna dos grupos. Valores mais baixos indicam *clusters* mais compactos e bem definidos, onde os pontos estão mais próximos de seus respectivos centroides. No método elbow utilizado, procura-se o ponto onde a redução da inércia começa a diminuir significativamente (o "cotovelo" da curva), que representa o número ideal de *clusters* para os dados. Esta análise permite verificar como a estrutura de agrupamentos varia entre os diferentes períodos da pandemia.

Figura 3. Número de clusters e inércia – Período 1 (02-04-2020 a 31-12-2020).

Número de clusters	Inércia
1	7.875983
2	4.555579
3	2.946883
4	2.223393
5	1.809834
6	1.501277
7	1.269472
8	1.097884
9	0.928343
10	0.765083

Fonte: Os autores.

Figura 4. Número de clusters e inércia – Período 2 (01-01-2021 a 30-06-2021).

Número de clusters	Inércia
1	7.469620
2	3.917677
3	2.420343
4	1.822562
5	1.413741
6	1.181577
7	1.007944
8	0.846966
9	0.715529
10	0.635912

Fonte: Os autores.

Figura 5. Número de clusters e inércia – Período 3 (01-07-2023 a 01-12-2023).

Número de clusters	Inércia
1	6.990064
2	3.334758
3	2.166772
4	1.009424
5	0.655969
6	0.377293
7	0.262903
8	0.201185
9	0.154538
10	0.119676

Fonte: Os autores.

As tabelas mostram que, à medida que o número de *clusters* aumenta, a inércia diminui. Isso significa que, à medida que mais *clusters* são criados, eles se tornam mais compactos e separados uns dos outros.

4.2. Descrição dos clusters em cada série temporal

Após a aplicação do método *k-means* em cada um dos três períodos mencionados, os municípios foram agrupados. Para resumir a situação do surgimento de novos casos, dos casos acumulados e dos óbitos em cada um desses grupos, foram calculadas as médias dos coeficientes correspondentes a cada município.

4.2.1 Período 1

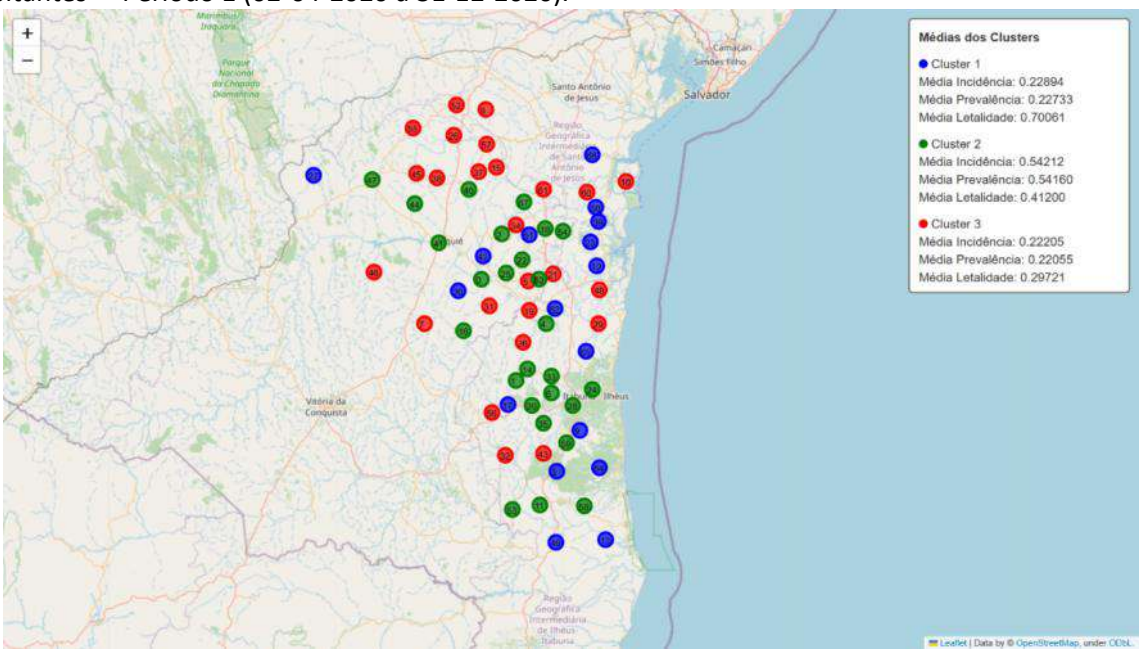
O período de 02 de abril de 2020 a 31 de dezembro de 2020 na pandemia da COVID-19 também pode ser chamado de "Fase de Continuidade e Adaptação". Durante esse intervalo, o mundo continuou a lidar com os desafios impostos pelo novo coronavírus, mas houve uma mudança gradual no foco e nas estratégias em comparação com as fases iniciais da pandemia. O período pode se destacar pelas seguintes características:

- Persistência da Pandemia: Nesse período, a pandemia continuou a afetar comunidades globalmente. Houve uma persistência da transmissão do vírus, com surtos ocorrendo em diferentes regiões.
- Adaptação a Novas Realidades: Sociedades e sistemas de saúde começaram a se adaptar às novas realidades impostas pela COVID-19. Foram implementadas medidas de mitigação, como distanciamento social, uso de máscaras e higiene intensificada, que se tornaram parte integrante da vida cotidiana.
- Decreto toque de recolher: O Governo da Bahia decretou toque de recolher em 33 cidades para frear o avanço do novo coronavírus nesses locais entre 04 de agosto de 2020 a 11 de agosto de 2020. O toque de recolher valia das 19h às 5h. Dos municípios selecionados, três deles fazem parte do NRS Sul: Almadina (1), Ibirataia (22) e Jitaúna (42). Nessas cidades, ficou autorizado, durante o intervalo do toque de recolher, somente o funcionamento dos serviços essenciais e, em especial, as atividades relacionadas ao enfrentamento da pandemia, o transporte e o serviço de entrega de medicamentos e demais insumos necessários para manutenção das atividades de saúde, as obras em hospitais e a construção de unidades de saúde.

A análise de cluster revelou diferentes padrões de incidência, prevalência e letalidade da COVID-19 nos municípios do Núcleo Regional de Saúde Sul (NRS Sul) da Bahia. O cluster 2, com maior taxa de ocorrência da doença e alta proporção de casos, apresentou incidência e prevalência elevadas, porém letalidade intermediária, sugerindo um número significativo de casos, mas também uma parcela considerável de óbitos. O cluster 1, com menor incidência e prevalência, mas alta letalidade, indica uma proporção elevada de casos fatais. O cluster 3 apresentou os menores riscos. Fatores como a movimentação de pessoas entre as cidades, a infraestrutura densamente povoada, o uso intenso de transporte público, as condições socioeconômicas desfavoráveis, as dificuldades de implementação de medidas de distanciamento e o acesso limitado a recursos de saúde em algumas áreas urbanas podem ter contribuído para a propagação diferenciada da pandemia no território baiano.

A Figura 6 representa o mapa obtido da clusterização com a média dos coeficientes epidemiológicos, onde essas médias foram avaliadas para cada 100 habitantes no período inicial de contágio da COVID-19 no Brasil (período 1). Adicionalmente, foi inserido nas plotagens do mapa a identificação do número do índice de cada município conforme Tabela 1.

Figura 6. Identificação de clusters no mapa da Bahia com médias dos coeficientes epidemiológicos por cada 100 habitantes – Período 1 (02-04-2020 a 31-12-2020).



Fonte: Os autores.

4.2.2 Período 2

As medidas de contenção, como o distanciamento social e o uso de máscaras, continuaram sendo implementadas em todo o mundo para tentar controlar a propagação do vírus. O período de 01 de janeiro de 2021 a 30 de junho de 2021 considerado fase de transição, também foi marcado por alguns eventos importantes, como:

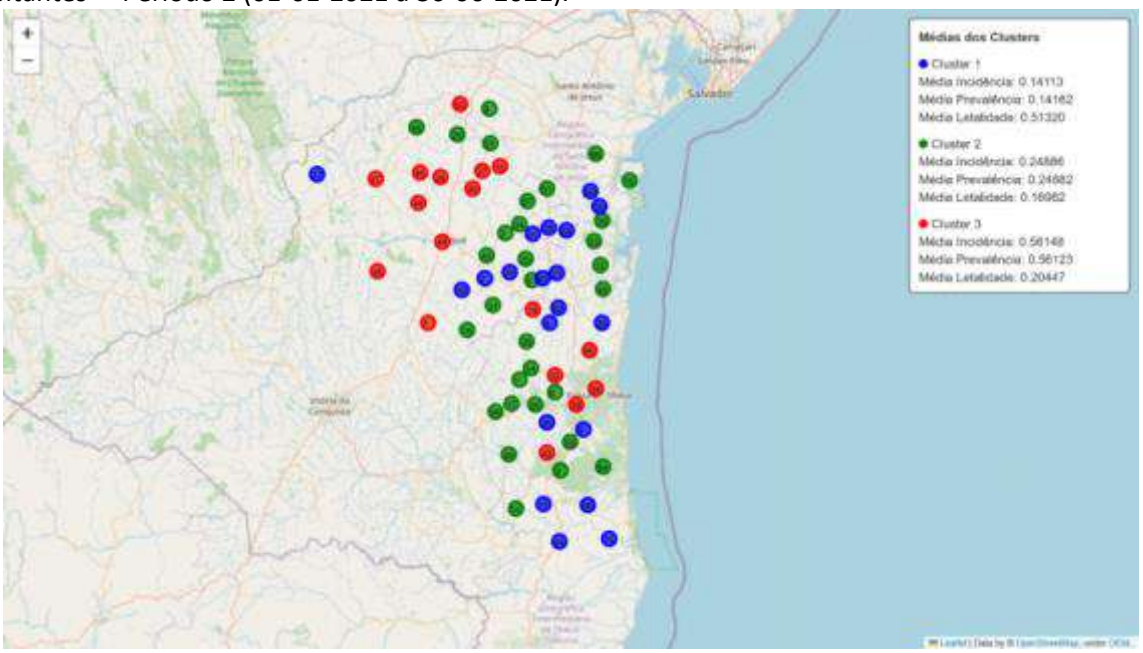
- A aprovação da vacina contra a COVID-19 pela Agência Nacional de Vigilância Sanitária (Anvisa);
- O início da vacinação contra a COVID-19 na Bahia em 19 de janeiro de 2021.

Durante o período analisado, observou-se uma mudança nos padrões dos clusters. O cluster 1, composto principalmente por municípios com acesso limitado a recursos de saúde, apresentou baixa incidência e prevalência da doença, mas alta letalidade, indicando maior risco de mortalidade. Fatores como diagnóstico tardio, falta de leitos e equipamentos, desigualdades socioeconômicas, barreiras culturais/educacionais e coexistência de doenças crônicas podem ter contribuído para o aumento da letalidade nesse cluster.

Já o cluster 3, com as cidades mais populosas de Itabuna (28) e Ilhéus (24), apresentou os maiores riscos de novos casos e casos acumulados, porém uma letalidade moderada. Por outro lado, o cluster 2 mostrou um risco intermediário de novos casos e casos acumulados, com baixa letalidade, destacando-se as cidades de Itamari (34), Irajuba (25) e Maracás (47).

A Figura 7 representa o mapa obtido da clusterização com a média dos coeficientes epidemiológicos.

Figura 7. Identificação de clusters no mapa da Bahia com médias dos coeficientes epidemiológicos por cada 100 habitantes – Período 2 (01-01-2021 a 30-06-2021).



Fonte: Os autores.

4.2.3 Período 3

No período de 01 de julho de 2023 a 01 de dezembro de 2023 o número de casos e mortes estava diminuindo. No entanto, a partir de setembro de 2023, o número de casos começou a aumentar novamente. Essa alta pode ser explicada por alguns fatores, como:

- A diminuição da cobertura vacinal;
- A diminuição da adesão às medidas de controle da pandemia.

Neste período de transição para a normalidade, observou-se uma heterogeneidade entre os municípios do Núcleo Regional de Saúde (NRS) Sul da Bahia em relação aos padrões de incidência, prevalência e letalidade da COVID-19. O cluster 3, composto por municípios como Mascote (49), Lajedo do Tabocal (45), Ubatã (63) e Wenceslau Guimarães (67), apresentou os maiores riscos de óbitos acumulados, com elevada letalidade (média de 0,56944) e baixa incidência (média de 0,13709) e prevalência (média de 0,13709).

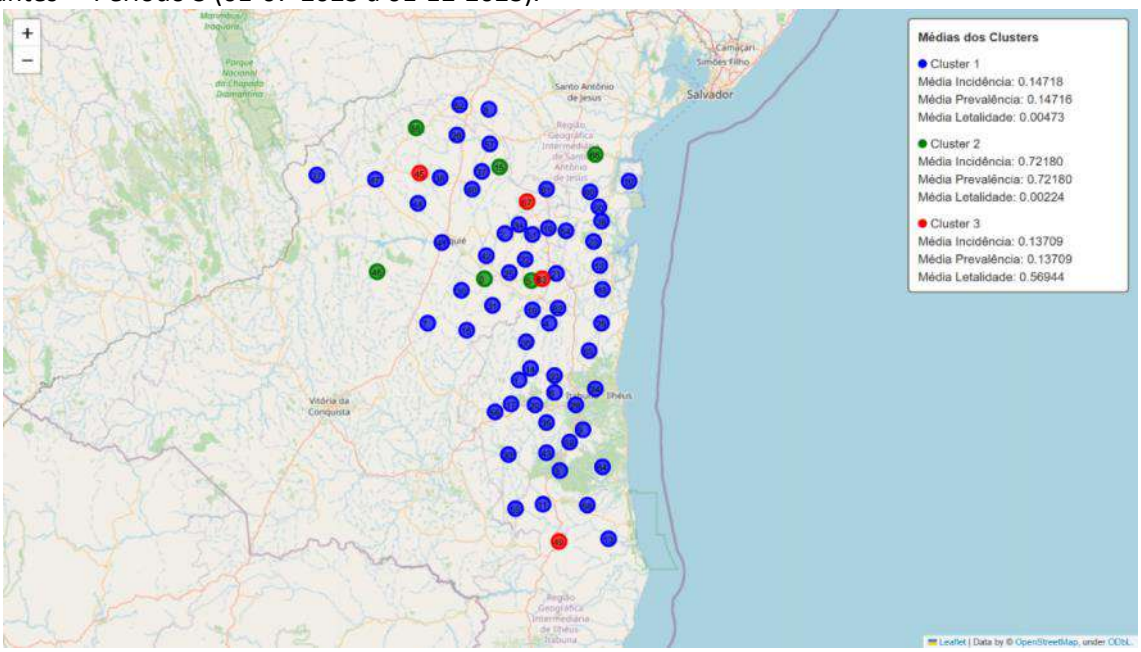
Já o cluster 2, formado por cidades menos populosas como Aiquara (0) e Cravolândia (15), registrou os maiores riscos de novos casos e casos acumulados, com altas incidência (média de 0,72180) e prevalência (média de 0,72180), mas baixa letalidade (média de 0,00224). A cidade de Barra do Rocha (5) se destacou neste cluster com o maior coeficiente de incidência e prevalência, porém sem registro de óbitos.

Por fim, o cluster 1, composto pelas cidades mais populosas de Itabuna (24), Ilhéus (28) e Jequié (41), apresentou o segundo maior risco de novos casos, casos acumulados e óbitos, com incidência (média de 0,14718), prevalência (média de 0,14716) e letalidade (média de 0,00473) moderadas. As cidades de Iramaia (27) e Almadina (1) se destacaram nesse cluster por não registrarem casos confirmados e óbitos acumulados.

Essas diferenças entre os clusters revelam os distintos padrões de propagação e impacto da COVID-19 nos municípios da região, indicando a necessidade de estratégias específicas de acordo com as características de cada localidade.

A Figura 8 representa o mapa obtido da clusterização com a média dos coeficientes epidemiológicos.

Figura 8. Identificação de clusters no mapa da Bahia com médias dos coeficientes epidemiológicos por cada 100 habitantes – Período 3 (01-07-2023 a 01-12-2023).



Fonte: Os autores.

5. CONSIDERAÇÕES FINAIS

O estudo do comportamento viral e a aplicação de técnicas avançadas de análise, como o *clustering*, desempenham um papel crucial na compreensão e gestão eficaz de pandemias. A análise do comportamento do vírus permite identificar padrões de transmissão, mutações e fatores influenciadores, fornecendo insights essenciais para a tomada de decisões em saúde pública. A utilização de técnicas de *clustering*, como o *K-means*, amplifica essa compreensão, ao agrupar regiões ou populações com perfis epidemiológicos semelhantes. Esta abordagem simplifica a interpretação de dados complexos e permite a adaptação de estratégias específicas para cada agrupamento: para municípios do Cluster 1 (alta letalidade e baixa incidência), recomenda-se fortalecer a capacidade diagnóstica e a atenção primária; no Cluster 2 (alta mobilidade urbana), medidas de controle de fluxo populacional se mostram prioritárias; enquanto para o Cluster 3 (grandes centros urbanos), a ampliação de leitos de UTI e campanhas de testagem em massa são estratégias-chave. Tais diferenciais consideram características únicas como: (i) infraestrutura de saúde local, (ii) densidade populacional, e (iii) padrões de mobilidade interurbana identificados nos dados.

A aplicação dessas abordagens analíticas à evolução da COVID-19 em cidades do Sul da Bahia permitiu discernir nuances nas dinâmicas da doença em diferentes fases, identificar áreas de maior risco e orientar a alocação eficiente de recursos. O estudo epidemiológico revelou a presença de cidades mais populosas, como Itabuna e Ilhéus, em clusters junto a municípios menores, possivelmente devido à intensa mobilidade populacional entre centros urbanos e áreas circunvizinhas, facilitando a propagação do vírus.

Em suma, o estudo do comportamento viral e o emprego de técnicas de *clustering* representam ferramentas valiosas para a formulação de políticas preventivas e intervencionistas, visando mitigar os impactos da pandemia e promover respostas mais eficazes em nível local e global.

REFERÊNCIAS

ALVES, H. J. P.; FERNANDES, F. A.; LIMA, K. P.; BATISTA, B. D. O.; FERNANDES, T. J. The COVID-19 pandemic in Brazil: an application of the k-means clustering method. **Research, Society and Development,** [S. l.], v. 9, n. 10, p. e5829109059, 2020. DOI: https://doi.org/10.33448/rsd-v9i10.9059.

DE SOUZA, W.M., BUSS, L.F., CANDIDO, D.D.S., *et al.* Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. **Nature Human Behaviour**, Londres, v. 4, n. 8, p. 856-865, 2020. DOI: https://doi.org/10.1038/s41562-020-0928-4

FÁVERO, L. P.; BELFIORE, P. **Data Science for Business and Decision Making**. Cambridge: Academic Press, 2019.

FERREIRA, D. F. **Estatística multivariada** 3. ed. Lavras: Ed. UFLA, 2018.

FONTANA, A.; NALDI, M. C. **Estudo de comparação de métodos para estimação de números de grupos em problemas de agrupamento de dados**. São Paulo: Universidade de São Paulo, 2009. (Relatório Técnico, n. 340).

GOHARI, K., KAZEMNEJAD, A., SHEIDAEI, A. *et al.* Clustering of countries according to the COVID-19 incidence and mortality rates. **BMC Public Health**, Londres, v. 22, n. 632, 2022. DOI: https://doi.org/10.1186/s12889-022-13086-z.

HASTIE, T.; TIBSHIRANI, R.;FRIEDMAN J. **The Elements of Statistical Learning:** Data Mining, Inference, and Prediction. 2. ed. Nova York: Springer, 2009. DOI: https://doi.org/10.1007/978-0-387-84858-7

MAKIH, A.; TARIGAN, A. Analysis of Covid-19 Case Data in Classification of the Distribution in DKI Jakarta Using Algorithm Method K-Means Clustering. I**nternational Research Journal of Advanced Engineering and Science**, [S. l.], v. 6, n. 2, p. 154-157, 2021.

MENEZES, A. M. B. **Noções básicas de epidemiologia**. Porto Alegre: Artmed, 2001. v. 1. p. 15-30.

MUHTASIM; MASUD, M. A. **C**lustering Countries on COVID-19 Data among Different Waves Using K-Means Clustering. **Journal of Computer and Communications**, [*S. l.*], v. 11, p. 1-14, 2023. DOI: https://doi.org/10.4236/jcc.2023.117001.

OLIVEIRA, R. R. *et al.* Evolução da COVID-19 no Brasil e suas implicações para políticas públicas de saúde. **Revista Panamericana de Salud Pública**, Washington, v. 45, p. e47, 2021. DOI: 10.26633/RPSP.2021.47. Disponível em: https://iris.paho.org/handle/10665.2/53918. Acesso em: 12 maio 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Relatório sobre a situação global da COVID-19.** Genebra: OMS, 2020. Disponível em: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Acesso em: 13 maio 2025.

ROTHMAN, K. J.; GREENLAND, S.; LASH, T. L. **Epidemiologia moderna.** 3. ed. Porto Alegre: Artmed, 2008.

SANTOS, J. A. *et al.* Challenges and Strategies for COVID-19 Control in Bahia, Brazil. **Journal of Public Health Research**, [S. l.], v. 11, n. 4, p. 2430, 2022.