VICTOR CARLO TASCHERI
CARLOS ALCIDES PORTO

FATORES SOCIOECÔMICOS ASSOCIADOS À COBERTURA VACINAL INFANTIL EM MUNICÍPIOS BRASILEIROS NO PERÍODO DE 2011-2021: UMA ABORDAGEM EM CIÊNCIA DE DADOS

PROJETO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
APRESENTADO AO DEPARTAMENTO DE ENGENHARIA INDUSTRIAL
DA PUC-RIO, COMO PARTE DOS REQUISITOS PARA OBTENÇÃO
DO TÍTULO DE ENGENHEIRO DE PRODUÇÃO

Orientador: Leonardo dos Santos Lourenço Bastos
Co-orientador: Igor Tona Peres

Departamento de Engenharia Industrial
Rio de Janeiro, 16 de junho de 2023.

Resumo

A queda na adesão vacinal no Brasil tem se tornado cada vez mais evidente, gerando diversas preocupações para as autoridades de saúde. Com a diminuição da cobertura vacinal, enfermidades previamente erradicadas no país estão em risco de ressurgirem, o que se aplica ao conjunto de vacinas analisadas neste estudo. Diante dessa ameaça iminente, é crucial compreender os aspectos socioeconômicos relacionados à baixa adesão vacinal. Diversas hipóteses podem ser determinantes para diminuição nas taxas de imunização contra doenças. Uma delas sugere que municípios com desvantagens socioeconômicas, como baixo índice de desenvolvimento e escassa cobertura de atenção primária à saúde, possam estar relacionados a uma menor taxa de cobertura vacinal. Além disso, outra hipótese sugere que municípios com alta cobertura de planos de saúde podem apresentar uma aparente baixa cobertura, quando analisados os dados disponibilizados, visto que esses dados provêm predominantemente de sistemas do setor público, como o DATASUS. Portanto, este trabalho teve como objetivo investigar a conexão entre a cobertura vacinal deficiente para o grupo de vacinas analisadas e os indicadores socioeconômicos e de desenvolvimento nas diferentes regiões e municípios do Brasil. Para tanto, empregaram-se métodos de análise descritiva dos dados e modelagem de machine learning não supervisionado (clusterização). Por meio das análises realizadas, constatou-se que aparentemente há uma associação entre certos fatores socioeconômicos, como o Índice de Desenvolvimento Humano (IDH) e o Índice de Gini, e a diminuição da cobertura vacinal no país, evidenciando que municípios e regiões menos desenvolvidos apresentam uma proporção menor de indivíduos imunizados.

Palavras-Chave

Cobertura vacinal, Clusterização, Fatores socioeconômicos, Machine learning

Abstract

The decline in vaccination adherence in Brazil has become increasingly evident, raising multiple concerns for health authorities. With the decrease in vaccination coverage, previously eradicated diseases in the country are at risk of resurging, which applies to the set of vaccines analyzed in this study. Faced with this imminent threat, it is crucial to understand the socioeconomic aspects related to low vaccination adherence. Several hypotheses can be determinants for the decrease in immunization rates against diseases. One of them suggests that municipalities with socioeconomic disadvantages, such as low development index and scarce coverage of primary healthcare, may be associated with a lower vaccination coverage rate. Additionally, another hypothesis suggests that municipalities with high coverage of private health plans may exhibit an apparent low coverage when analyzing the available data, as these data predominantly come from public sector systems, such as DATASUS. Therefore, this study aimed to investigate the connection between deficient vaccine coverage for the analyzed group of vaccines and the socioeconomic and development indicators in different regions and municipalities of Brazil. Descriptive data analysis methods and unsupervised machine learning modeling (clustering) were employed for this purpose. Through the conducted analyses, it was found that there apparently is an association between certain socioeconomic factors, such as the Human Development Index (HDI) and the Gini Index, and the decrease in vaccine coverage in the country, highlighting that less developed municipalities and regions have a lower proportion of immunized individuals.

Keywords

Vaccination coverage, Clustering, Socioeconomic factors, Machine learning

Sumário

Lista de Figuras

Lista de tabelas

1. Introdução

A compreensão da evolução da cobertura vacinal é crucial para o sistema de saúde pública no Brasil, que desde 1973 conta com o Programa Nacional de Imunizações (PNI) para a proteção da população contra as doenças imunopreveníveis (UNASUS, 2023). O PNI tem sido fundamental para o controle e erradicação de várias doenças infecciosas, mas nas últimas décadas, o país enfrentou desafios significativos em relação à cobertura vacinal. Entre 2010 e 2021, o país registrou quedas na cobertura vacinal, o que pode ter consequências graves para a saúde pública (UNASUS, 2023).

Além disso, dados quantitativos e evidências têm confirmado a preocupante queda na cobertura vacinal no Brasil ao longo dos anos, por exemplo a cobertura vacinal da poliomielite 2015 era de 95% e chegou pela primeira vez abaixo dos 80% em 2021 (SILVA, 2022). De acordo com o Ministério da Saúde, entre 2010 e 2021, observou-se uma diminuição nas taxas de vacinação em diferentes faixas etárias e para diversas vacinas do calendário básico de imunização (BRASIL, 2023)., colocando em risco os esforços de erradicação alcançados anteriormente (BRASIL, 2023). Esses dados quantitativos destacam a necessidade urgente de investigar e compreender as causas subjacentes à queda na adesão vacinal no país.

Pesquisas passadas fornecem evidências substanciais de que os fatores socioeconômicos desempenham um papel importante na taxa de cobertura vacinal. Além disso, a falta de informações adequadas sobre a importância das vacinas e a disseminação de desinformação também têm sido identificadas como fatores contribuintes para a queda na cobertura vacinal (OLIVEIRA, 2020).

Portanto, compreender essas relações e identificar os fatores que impactam negativamente a adesão vacinal é fundamental para o desenvolvimento de estratégias eficazes de saúde pública voltadas para o aumento da cobertura vacinal no Brasil.

O trabalho teve como objetivo investigar se os aspectos socioeconômicos estão relacionados à baixa adesão vacinal. Para isso, a monografia está estruturada em cinco capítulos, sendo este o primeiro. Os demais capítulos abordarão: a evolução da cobertura vacinal no Brasil; os principais desafios e fatores que afetaram a cobertura vacinal no país; e, por fim, a relação entre os indicadores socioeconômicos e a cobertura vacinal no Brasil. Ao final deste trabalho, espera-se contribuir para o entendimento das causas e consequências da queda na cobertura vacinal no Brasil e fornecer informações relevantes

para a formulação de políticas públicas de saúde que visem a melhorar a cobertura vacinal do país (UNASUS, 2023).

2. Referencial Teórico

2.1 Contexto histórico da vacinação no Brasil

A vacinação é uma das principais estratégias de saúde pública em todo o mundo, contribuindo significativamente para a prevenção de doenças infecciosas e para a redução da morbidade e mortalidade (FIOCRUZ, 2023). No Brasil, a vacinação tem uma história de mais de um século, com importantes marcos e desafios ao longo do tempo. A primeira campanha de vacinação em massa no Brasil ocorreu em 1904, com a introdução da vacina contra a varíola, seguida pela implementação da vacinação obrigatória no mesmo ano, o que gerou a "Revolta da Vacina", um conflito social ocorrido no Rio de Janeiro no início do século XX (FIOCRUZ, 2023).

Ao longo do século XX, o Brasil introduziu diversas vacinas no calendário nacional de vacinação. Nos anos 70, iniciou-se a produção nacional de vacinas, com destaque para a vacina contra a febre amarela, produzida pelo Instituto Oswaldo Cruz. Nos anos 80, o Ministério da Saúde criou o Programa Nacional de Imunizações (PNI), que teve como objetivo organizar ações de vacinação em todo o país (BRASIL, 2023). O PNI estabeleceu metas de cobertura vacinal para diversas vacinas e implementou campanhas de vacinação em massa, como a bem-sucedida erradicação da poliomielite em 1994. A partir dos anos 90, o Brasil se tornou um importante produtor de vacinas, com a introdução de vacinas contra a meningite C, hepatite B e outras doenças (UNASUS, 2023). A incorporação da vacina contra o *Haemophilus influenzae* tipo b (Hib) em 1999 e a vacina pneumocócica conjugada em 2010 representaram avanços importantes na prevenção de doenças infecciosas (BRASIL, 2023).

No século XXI, a vacinação continua sendo uma prioridade para a saúde pública brasileira, com a introdução de novas vacinas, como a vacina contra o HPV em 2014 e a vacina contra a febre amarela em áreas urbanas em 2017 (UNASUS, 2023; BRASIL, 2023). No entanto, o Brasil enfrenta desafios na manutenção das coberturas vacinais, com surtos de doenças preveníveis por vacinação, como o sarampo e a poliomielite (NUNES, 2021). Esses desafios ressaltam a importância contínua da vacinação como estratégia de prevenção e a necessidade de aprimorar as políticas de saúde pública.

O PNI, coordenado pelo Ministério da Saúde através do Departamento de Imunizações e Doenças Transmissíveis (DEIDT), adota diferentes estratégias para a coordenação das campanhas de vacinação, de acordo com as características de cada vacina (UNASUS, 2023). Campanhas nacionais de vacinação são realizadas para vacinas como a poliomielite e o sarampo, visando atingir um amplo público-alvo em um curto período de tempo. Essas campanhas são intensificadas em faixas etárias específicas, como crianças e adolescentes, devido à maior vulnerabilidade a essas doenças. Por outro lado, para vacinas como a hepatite B e a febre amarela, são adotadas estratégias de vacinação rotineira, integradas aos serviços de saúde e disponíveis ao longo do ano para grupos de risco determinados (UNASUS, 2023).

Graças às ações do PNI, o Brasil alcançou resultados impressionantes no combate a doenças imunopreveníveis. O país foi certificado como livre da poliomielite em 1994 e tem mantido a erradicação da doença desde então (UNASUS, 2023). Além disso, o programa desempenhou um papel fundamental na redução dos casos de sarampo no país. No entanto, recentemente, ocorreram surtos da doença em algumas regiões, o que ressalta a importância contínua da vacinação para prevenir o ressurgimento dessas enfermidades (BRASIL, 2023).

No contexto da pandemia de COVID-19, o PNI desempenhou um papel crucial na coordenação da vacinação contra o coronavírus. O programa trabalhou em conjunto com os estados e municípios para a distribuição e administração das vacinas contra a COVID-19, seguindo um plano nacional de imunização. Apesar dos desafios logísticos e da alta demanda, o PNI conseguiu avançar significativamente na vacinação da população, contribuindo para a redução dos casos graves da doença e a proteção da saúde pública (BRASIL, 2023).

Esses feitos do PNI destacam a importância desse programa para a saúde pública brasileira. No entanto, é fundamental que o programa continue a ser monitorado e aprimorado, com investimentos contínuos, a fim de garantir a manutenção da cobertura vacinal adequada e a prevenção de doenças imunopreveníveis e emergentes. O PNI demonstrou sua eficácia ao longo das décadas, mas é necessário estar atento aos desafios atuais e futuros, garantindo que a vacinação seja acessível a todos os segmentos da população e que as estratégias sejam adaptadas às necessidades epidemiológicas do país (UNASUS, 2023). Através do PNI e do compromisso contínuo com a imunização, o

Brasil pode continuar a proteger a saúde da população e a alcançar conquistas significativas no controle e prevenção de doenças infecciosas.

2.2 Ciclo de vida de um projeto de Ciência de Dados

A manipulação de dados se tornou uma prática de extrema importância para extrair informações e auxiliar na tomada de decisões. A quantidade cada vez maior de dados armazenados e em circulação demanda análises que permitam uma extração maior de conhecimento e *insights* das bases de dados. Com esse objetivo, a Ciência de Dados, uma área que compreende campos interdisciplinares como a ciência da computação, matemática, estatística e *Machine Learning*, busca transformar números dispersos em informações mais evidentes e visualmente acessíveis (BICHLER, 2017; CAO, 2016).

Para alcançar esse propósito, há diversos *frameworks* que propõem orientar o desenvolvimento de métodos por meio de técnicas de mineração de dados, estabelecendo padrões que dão sentido aos dados brutos. Um exemplo popular é o *CRoss Industry Standard Process for Data Mining* (CRISP-DM), uma metodologia que abrange um plano completo para conduzir um projeto de Ciência de Dados (MANRESA, 2020). O CRISP-DM é caracterizado por um ciclo de seis fases que interagem entre si, conforme apresentado na Figura 1 (BICHLER, 2017).

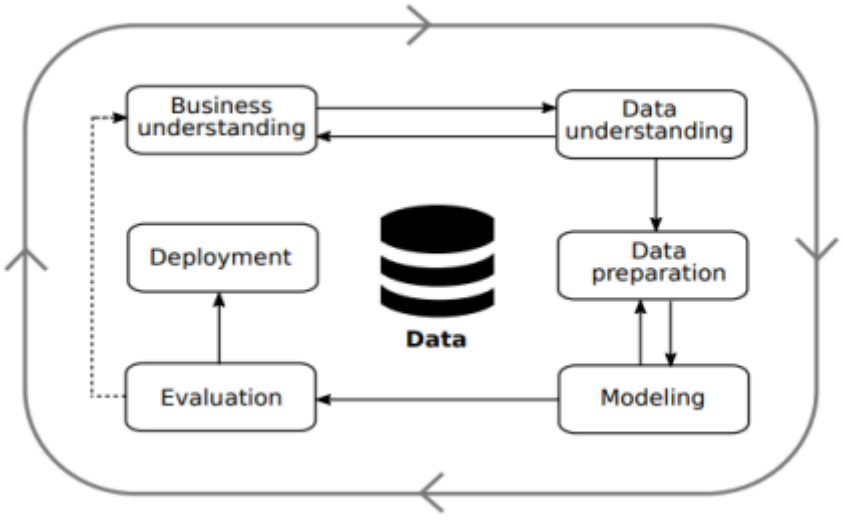

Figura 1: Ciclo de vida de um projeto de ciência de dados

Fonte: Manresa (2020)

A primeira fase (*Business understanding*) aborda a compreensão do negócio e a percepção do problema, incluindo a determinação dos objetivos do negócio, a avaliação do contexto e o objetivo da mineração de dados. Já a segunda fase (*Data understanding*) concentra-se no entendimento dos dados, incluindo a coleta de dados, bases de dados disponíveis, a percepção de suas características e sua qualidade. A terceira fase (*Data preparation*) envolve a preparação dos dados, que consiste em sua seleção, limpeza, construção e formatação, a fim de atender aos requisitos necessários na fase de modelagem (BICHLER, 2017). A quarta fase (*Modeling*) é composta pela escolha das técnicas de modelagem, criação e avaliação de modelos, e tem uma ligação com a fase anterior devido à exigência de diferentes procedimentos na formatação e preparação dos dados (CAO, 2016). A quinta fase é a avaliação (*Evaluation*), onde os resultados são interpretados e avaliados após a seleção do melhor modelo, seguidos por uma revisão do processo de aprendizagem e a determinação dos próximos passos. Por fim, a última etapa do ciclo é a implementação do modelo (*Deployment*), onde a solução é introduzida no contexto prático e é feita uma devolutiva de sua implementação (MANRESA, 2020).

2.3 Técnicas para modelagem e análise de Dados

Quando se trata de análise e modelagens de dados, há diversas técnicas e recursos diferentes que podem ser utilizados. Para os propósitos desse trabalho e os dados que foram adquiridos, foram utilizadas as técnicas de análise descritiva, clusterização e regressão linear. Nesta seção, iremos falar mais sobre esses métodos utilizados para o trabalho.

2.3.1 Análise descritiva

Para se obter bons resultados no estudo de um conjunto de dados, é essencial compreender o comportamento dos mesmos. Cada conjunto de informações possui um comportamento único e singular, que pode ser revelado através de técnicas de análise descritiva. Esse procedimento inicial permite identificar padrões e tendências nos dados, além de destacar características peculiares do conjunto de informações. Para realizar essa análise, são empregadas técnicas de síntese da informação e ferramentas de visualização (COSTA e LEO, 2020).

O primeiro passo na análise descritiva é identificar as variáveis presentes na base de dados. Essas variáveis podem ser de dois tipos: qualitativas ou quantitativas. As variáveis qualitativas descrevem características do objeto observado por meio de uma categoria, em vez de um valor numérico. Elas podem ser classificadas como nominais ou ordinais. Já as variáveis quantitativas expressam características mensuráveis, que podem ser representadas por valores numéricos. Existem dois tipos de variáveis quantitativas: as contínuas e as discretas (COSTA e LEO, 2020).

A análise gráfica também é uma ferramenta importante na análise de dados. Ela permite uma visualização mais clara dos dados, facilitando comparações e identificação de padrões. Dentre os diferentes tipos de gráficos existentes, os gráficos de linhas, colunas e dispersão são alguns dos mais utilizados na representação de um conjunto de dados (MAIÇARA ,2022). Os gráficos de linhas são ideais para a representação de séries temporais, mostrando como a variável de interesse se comporta ao longo do tempo. Já os gráficos de dispersão são utilizados para fazer comparações e entender a relação entre duas variáveis quantitativas de um indivíduo do conjunto de dados. Por fim, os gráficos de colunas fazem uma comparação entre variáveis quantitativas e qualitativas, mostrando a quantidade de um atributo (REIS e REIS, 2002).

Além dos gráficos, também foram utilizadas matrizes de correlação entre as variáveis, o coeficiente de correlação utilizado para no cálculo foi o coeficiente de correlação de Pearson (ρ), definido pela seguinte fórmula:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}.\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov(X,Y)}{\sqrt{var(X).var(Y)}} \quad (1)$$

Nessa expressão (1), $x_1 \ldots x_n$ e $y_1 \ldots y_n$ são os valores medidos em ambas as variáveis, enquanto $\bar{x}$ e $\bar{y}$ são as médias aritméticas calculadas para x e y, respectivamente.

A correlação entre duas variáveis é definida por um coeficiente ρ que pode variar em um intervalo [-1, 1]. Quanto mais próximo o valor for de -1 ou 1, mais forte é a correlação negativa ou positiva entre as variáveis, respectivamente. Em contrapartida, valores mais próximos de 0 indicam uma correlação mais fraca, ou inexistente, entre as

variáveis. Calcular e compreender os coeficientes de correlação nos ajuda a entender a relação entre as variáveis e, consequentemente, a interpretar os resultados da análise de dados

2.3.2 Clusterização

A técnica de clusterização de dados consiste na divisão de um conjunto heterogêneo de dados em subgrupos chamados *clusters*, onde os objetos possuem semelhanças entre si. Essa técnica realiza o agrupamento dos dados com base em suas similaridades, sem a necessidade de uma classificação prévia das informações. O objetivo principal é criar grupos com objetos altamente similares, porém com uma ampla distinção entre os *clusters* (CASSIANO, 2014; SILVA, 2022).

Encontrar a melhor forma de decompor os dados em grupos é uma tarefa desafiadora, por isso a clusterização conta com diversos métodos para realizar esse agrupamento. Cada método possui características específicas, alguns são mais adequados para grandes conjuntos de dados, enquanto outros funcionam melhor com menos informações. Além disso, alguns exigem a definição prévia da quantidade de *clusters*, enquanto outros não possuem essa restrição (DONI, 2004; SILVA, 2022).

Entre os diferentes métodos existentes, dois se destacam: o método hierárquico e o método particional ou não hierárquico. Além das diferenças entre eles, os métodos de *clusterização* também apresentam técnicas específicas. Na Tabela 1, podem ser visualizados dois algoritmos de cada um desses métodos mencionados acima (SILVA, 2022).

Tabela 1: Descrição de técnicas de clusterização

Método	Técnica	Descrição
Hierárquico	Aninhamento Aglomerativo (ANGES)	Clustering aglomerativo é uma técnica na qual nós únicos são mesclados até que todo o conjunto de dados seja um único cluster. Este procedimento constrói um dendrograma no qual uma hierarquia é calculada em relação ao número de clusters. O procedimento de mesclagem depende da função de ligação usada para estimar a distância entre dois clusters e dos critérios para combinar dois clusters calculando sua dissimilaridade. Existem quatro populares funções de ligação: - Ligação simples: considera a distância mínima entre os componentes de dois clusters; - Ligação completa: inversamente, funde dois clusters que possuem a distância máxima entre eles; - Ligação média: une dois clusters com a menor distância média; - Distância da ala: considera a variância mínima dentro do cluster.
	Análise Divisória (DIANA)	Ao contrário do AGNES, o DIANA começa com todo o conjunto de dados como um único cluster. Em seguida, ele se divide iterativamente em clusters até que todos os pontos de dados sejam clusters considerando a distância mínima entre os pontos de dados dentro de um cluster. Ele também calcula uma hierarquia que serve como base para a escolha do melhor número de clusters.
Particionais	K-Means	Um número inicial de k clusters deve ser definido. Ele seleciona k pontos aleatoriamente como centróides e atribui aos clusters os pontos de dados mais próximos do respectivo centróide. No K-Means, o centróide corresponde à média das coordenadas dos pontos dentro do mesmo cluster. Assim, em cada iteração, os centróides são recalculados e os pontos de dados são reatribuídos aos clusters para minimizar a distância média total com relação ao centróide.
	K-Medoids	Um número inicial de k clusters deve ser definido. Ele seleciona k pontos de dados como centróides aleatoriamente e atribui aos clusters os pontos de dados mais próximos do respectivo centróide. Em K-Medoids, os centróides são pontos de dados. Assim, em cada iteração os centróides são recalculados e os pontos de dados são reatribuídos aos clusters para minimizar a distância média total com relação ao centróide.

Fontes: SILVA, 2022

Apesar da diversidade de métodos, o K-Means é um método tradicional amplamente utilizado devido à sua simplicidade(SILVA, 2022). Após a definição da quantidade de grupos e a atribuição dos centróides de cada grupo, o algoritmo calcula a distância entre os centróides e os demais pontos, atribuindo os dados ao cluster mais próximo (SILVA, 2022). Para isso, o método segue a seguinte formulação matemática (DABBURA, 2018):

$$J = \sum_{i=1}^{m} \sum_{k=1}^{m} w_{ik} \ \left\| x^i - \mu_k \right\|^2 \qquad (2)$$

Nessa expressão 2, w_{ik} = 1 se o ponto x^i pertencer ao cluster k; caso contrário, w_{ik}= 0. Além disso, μ_k representa o centróide do cluster de x^i.

Além dos diferentes tipos de clusterização, existem diversas métricas de avaliação de clusters para determinar a quantidade ideal de clusters a ser utilizada. O objetivo geral ao analisar essas métricas é verificar se os agrupamentos são coerentes e se os membros de um cluster são mais semelhantes entre si ou a membros de outros clusters (SILVA, 2022). Algumas das métricas de avaliação mais comuns incluem o coeficiente de silhueta, o índice Calinski Harabasz e o índice Davies-Bouldin.

O coeficiente de silhueta de um cluster é calculado como a média da distância intra-cluster e a distância média para o cluster mais próximo de todos os membros do conjunto. Com um intervalo de variação de -1 a 1, um valor alto indica que os clusters resultantes são densos e bem separados (MEHTA, 2022). O coeficiente de silhueta é definido para cada amostra e composto por duas pontuações, sendo calculado da seguinte forma (ROUSSEAUW, 1986).

$$s = \frac{b - a}{max(a, b)} \quad (3)$$

Nessa fórmula (3), a representa a distância média entre uma amostra e os demais pontos do mesmo cluster, e b é a distância média entre uma amostra e todos os pontos do cluster mais próximo.

O índice de Calinski Harabasz, também conhecido como critério de razão da variância, é calculado como a razão entre a soma da dispersão entre os clusters e a dispersão dentro de cada cluster para todos os clusters. Quanto maior o valor do índice, melhor é o agrupamento e mais separados estão os clusters. Para um conjunto de dados E com tamanho n_E agrupado em k clusters, a pontuação de Calinski-Harabasz s é dada pela seguinte formulação (CALÍNSKI e HATABASZ, 1974).

$$s = \frac{tr(B_k)}{tr(W_k)} . \frac{n_E - k}{k - 1} \quad (4)$$

Onde:

$$w_k = \sum_{q=1}^{k} \sum_{x \in c_q} w_{ik} \ (x - c_q)(x - c_q)^T \quad (5)$$

$$B_k = \sum_{q=1}^{k} n_q \ (c_q - c_E)(c_q - c_E)^T \quad (6)$$

Nessas expressões 5 e 6, c_q representa o conjunto de pontos no cluster q, c_E é o centro de E e n_q é o número de pontos no cluster q $tr(.)$ corresponde ao traço da matriz.

Por fim, o índice de Davies-Bouldin mede a similaridade média entre cada cluster Ci, para i = 1, ..., k, e seu cluster mais similar Cj. Essa medida de similaridade leva em consideração as distâncias dentro dos clusters e entre os clusters, e é representada por R_{ij}. Um valor menor do índice é preferível, indicando baixas medidas de dispersão intra-grupo e grandes distâncias intergrupos. O valor mínimo é zero, que representa uma divisão ideal dos membros (HALKIDI, BATISTAKIS e VAZIRGIANNIS, 2001). O índice de Davies-Bouldin é definido da seguinte forma:

$$DB = \frac{1}{k} \sum_{i=1}^{k} max \ R_{ij} \quad (i \neq j) \qquad (7)$$

Onde:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \qquad (8)$$

Na expressão fornecida acima 8, s_i representa a média das distâncias entre cada ponto pertencente ao cluster i e seu centróide, enquanto d_{ij} representa a distância entre os centróides dos clusters i e j.

2.3.3 Regressão Linear

Regressão Linear é um conjunto de técnicas estatísticas que visa criar modelos para descrever as relações entre variáveis. O objetivo principal da regressão linear é modelar uma variável contínua Y como uma função matemática de uma ou mais variáveis X. Quando há apenas uma variável explicativa X para explicar Y, chamamos de Regressão

Linear Simples. Já quando temos mais de uma variável explicativa, é denominada Regressão Linear Múltipla (WEISBERG, 2014).

Além disso, é importante ressaltar a inferência dos coeficientes de regressão, conhecidos como betas β. A interpretação dos betas é crucial para entender o efeito das variáveis explicativas no modelo. Esses coeficientes nos fornecem informações sobre o impacto que uma unidade de mudança em uma variável explicativa tem na variável resposta Y, mantendo todas as outras variáveis constantes (WEISBERG, 2014).

Para realizar inferência sobre os betas, é comum utilizar testes de hipótese. Esses testes são a base para o cálculo do p-valor e do intervalo de confiança. O p-valor nos indica a probabilidade de obter um efeito igual ou mais extremo do que o observado, assumindo que a hipótese nula seja verdadeira. Se o p-valor for menor que um nível de significância pré-definido, podemos rejeitar a hipótese nula e concluir que há uma relação significativa entre a variável explicativa e a variável resposta. O intervalo de confiança fornece uma faixa de valores plausíveis para o coeficiente de regressão, com base na estimativa amostral e no nível de confiança selecionado (WEISBERG, 2014). A equação matemática da Regressão Linear é generalizada da seguinte forma:

$$Y = X\beta + \varepsilon \qquad (9)$$

Onde

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad (10)$$

Y é um vetor n x 1 cujos componentes corresponde às n respostas. X é uma matriz de dimensão n x (p+1) denominada matriz do modelo. ε é um vetor de dimensão n x 1 cujos componentes são os erros, a parte Y que o modelo de regressão não consegue explicar. β é um vetor (p+1) x 1 cujos elementos são os coeficientes de regressão.

3. Materiais e Métodos

O objetivo desta pesquisa consiste em realizar análises e identificar os múltiplos elementos relacionados à diminuição da taxa de cobertura vacinal de um determinado

conjunto de vacinas no território brasileiro. A queda observada pode ser influenciada por uma série de fatores, tais como a restrição de acesso às vacinas, a disparidade de desenvolvimento entre distintas regiões do país, bem como as diferenças socioeconômicas que permeiam os estados e municípios brasileiros. Essas são apenas algumas das diversas questões que podem estar contribuindo para este cenário atual.

Com o intuito de alcançar esse propósito, foram conduzidas análises dos dados, embasando-se no ciclo de vida característico do campo da Ciência de Dados. Essa abordagem metodológica segue uma sequência de etapas estruturadas, as quais foram minuciosamente abordadas neste capítulo, com o intuito de fornecer um panorama completo das atividades realizadas.

3.1 Compreensão do problema

Nos últimos anos, tem sido observada uma diminuição progressiva na taxa de cobertura vacinal no Brasil, o que acarreta uma redução no número de pessoas imunizadas (SILVA, 2022). Essa queda levanta sérias preocupações em relação ao ressurgimento de doenças que antes estavam controladas ou até mesmo erradicadas no país. Diante desse panorama, torna-se imprescindível realizar uma análise abrangente, tanto em nível nacional quanto municipal, a fim de identificar os possíveis fatores associados a essa redução.

Diante desse contexto, o presente estudo busca responder a duas indagações fundamentais: (i) Qual tem sido o panorama atual da cobertura vacinal no Brasil e em suas cinco regiões? e (ii) Há uma relação evidente entre características socioeconômicas e a diminuição ou aumento da cobertura vacinal? Com o intuito de elucidar essas questões, foram conduzidos três conjuntos de análises distintos. O primeiro bloco de análises tem como objetivo avaliar a progressão da cobertura vacinal no país e em seus estados. Já os dois blocos subsequentes visam identificar possíveis associações entre fatores socioeconômicos e a cobertura vacinal no nível dos municípios.

3.2 Entendimento dos dados

Para tais análises, os dados foram obtidos da base de dados do Instituto de Estudos para Políticas de Saúde (IEPS, 2023), denominada IEPS Data, que é um repositório de indicadores socioeconômicos, e de saúde do Brasil, o qual agrega dados de saúde e socioeconômicos de diferentes fontes, tais como, e-Gestor, PNI, e TabNet/DATASUS e

Censo Populacional. Foram obtidos os dados para oito principais vacinas administradas em território nacional e por município entre os anos de 2010 e 2021 (IEPS, 2023).

Tabela 2 - Variáveis presentes nas bases de dados retiradas do IEPS Data

Variável	Nome dos Indicadores	Bloco	Fonte
cob_ab	**Cobertura da Atenção Básica (%)**	Atenção Primária	e-Gestor
cob_vac_bcg	**Cobertura Vacinal de BCG (%)**	Atenção Primária	PNI, TabNet/DATASUS
cob_vac_hepb	**Cobertura Vacinal de Hepatite B em crianças de até 30 dias (%)**	Atenção Primária	PNI, TabNet/DATASUS
cob_vac_menin	**Cobertura Vacinal de Meningocócica C (%)**	Atenção Primária	PNI, TabNet/DATASUS
cob_vac_penta	**Cobertura Vacinal de Pentavalente (%)**	Atenção Primária	PNI, TabNet/DATASUS
cob_vac_pneumo	**Cobertura Vacinal de Pneumocócica (%)**	Atenção Primária	PNI, TabNet/DATASUS
cob_vac_polio	**Cobertura Vacinal de Poliomielite (%)**	Atenção Primária	PNI, TabNet/DATASUS
cob_vac_rota	**Cobertura Vacinal de Rotavírus Humano (%)**	Atenção Primária	PNI, TabNet/DATASUS
cob_vac_tvd1	**Cobertura Vacinal de Tríplice Viral (1ª Dose) (%)**	Atenção Primária	PNI, TabNet/DATASUS

Fonte: IEPS (2023).

Os Indicadores socioeconômicos escolhidos foram a Cobertura de Atenção Básica, Cobertura da rede privada de saúde, o índice de desenvolvimento humano (IDH) médio, o percentual de saneamento básico adequado, além do índice Gini (IEPS, 2023). O índice Gini é uma medida estatística de desigualdade de renda ou riqueza em uma determinada população, que varia de zero (0) a um (1), onde zero representa a completa igualdade (todos possuem a mesma renda ou riqueza) e um representa a completa desigualdade (uma única pessoa detém toda a renda ou riqueza da população). Quanto mais próximo de 1 for o valor do índice Gini, maior será a desigualdade na distribuição de renda ou riqueza na população.

3.3 Tratamento dos dados

Após coletar as informações, deu-se início à manipulação dos dados, iniciando pelo primeiro passo que consiste na limpeza e formatação dos dados. Todas as bases de dados abrangem o período de 2010 a 2021. A primeira base manipulada continha informações de cobertura vacinal por ano para as vacinas da tabela 2 em cada uma das regiões e estados brasileiros. Nessa base, não foi necessário realizar nenhuma limpeza, pois não havia variáveis com valores faltantes.

Uma segunda base de dados com os dados de variáveis socioeconômicas e de cobertura vacinal por município foi analisada. Para esses dados, realizou-se um tratamento devido à presença de valores faltantes. A primeira etapa de limpeza ocorreu neste base utilizando o filtro disponível na ferramenta Microsoft Excel. Ao filtrar os valores em branco das variáveis, foi possível identificar quais municípios possuíam dados incompletos. Em seguida, esses municípios, juntamente com seus dados correspondentes, foram removidos da base, de modo a manter apenas aqueles que continham informações completas de todas as variáveis nos anos de interesse.

Para realizar as análises de municípios, também foi necessário tratar dados incompletos de algumas vacinas. Em alguns casos, não havia dados disponíveis para determinadas vacinas em alguns estados nos anos de 2010 a 2014. Por exemplo, em determinado período, não havia dados disponíveis sobre a cobertura vacinal de determinada vacina em um estado específico. Além disso, foram identificados casos em que os dados apresentavam valores que se destacavam, indicando situações incomuns. Por exemplo, para determinados estados, as coberturas vacinais de algumas vacinas foram registradas com valores muito baixos. Para esses casos específicos, os dados foram tratados como *outliers*, e foram eliminados a fim de evitar impactos nos resultados finais da pesquisa.

Após a limpeza dos dados, foi possível iniciar o cálculo dos indicadores para a análise. Foram realizados dois procedimentos: o cálculo da média anual das variáveis analisadas e a determinação da variação anual de cada variável. Para obter esse resultado, foi calculada a diferença entre o valor do ano em análise e o valor do ano anterior, dividida pelo valor do ano anterior. Essas análises foram iniciadas no ano de 2011. O primeiro procedimento de formatação foi aplicado a todas as bases de dados, enquanto o segundo procedimento foi aplicado exclusivamente às bases contendo informações por estado.

No desenvolvimento das análises do bloco 2, os dados foram tratados excluindo todos os municípios que apresentavam algum dado faltante da cobertura vacinal de determinada vacina. Por exemplo, municípios que não possuíam um dos indicadores selecionados foram removidos da análise. Além disso, municípios com indicadores nulos ou muito próximos de zero em cobertura vacinal, também foram excluídos.

3.4 Modelagem

Nas primeiras análises realizadas, buscamos aprofundar a compreensão em nível macro, começando com o panorama do país como um todo e, em seguida, explorando regionalmente estados e municípios. Com o objetivo de analisar a cobertura vacinal no Brasil de forma mais detalhada, dividimos as coberturas vacinais em dois blocos distintos.

No Bloco 1, examinamos as principais vacinas do Programa Nacional de Imunizações (PNI), levando em consideração as diferenças regionais e municipais. Analisamos as vacinas mais relevantes do PNI presentes na base de dados, incluindo BCG, Hepatite B, Poliomielite, Pneumocócica, Pentavalente, Meningocócica C, Tríplice Viral (Sarampo, Rubéola e Caxumba) e Rotavírus Humano. Avaliamos as coberturas vacinais em cada região do país e nos principais municípios.

Na análise do Bloco 2, optamos por analisar as vacinas tríplice viral, rotavírus, meningite, pneumocócica e poliomielite. Escolhemos essas cinco vacinas por possuírem dados mais completos e coerentes no IEPS Data, o que proporcionou maior estabilidade em nossa análise. Além disso, todas essas vacinas são administradas na mesma faixa etária, facilitando a comparação das coberturas vacinais. Selecionamos apenas os municípios que possuíam dados completos. No entanto, decidimos não prosseguir com a análise da vacina BCG devido ao fato de ela ser um outlier, com um comportamento muito diferente das outras vacinas como será observado nos resultados.

No Bloco 3, utilizamos a técnica de clusterização na base de dados de municípios para identificar cidades com características socioeconômicas parecidas. A clusterização é uma técnica de análise de dados que permite agrupar objetos com características semelhantes em clusters ou grupos. Nesse caso, aplicamos a técnica de clusterização particional com o algoritmo k-medoids, uma técnica que possibilita utilizar pontos de dados específicos, chamados de medoids, como representantes dos clusters. Para definir o número de clusters, utilizamos a métrica Silhouette, a abordagem foi calcular o

silhouette para a quantia de clusters, escolhendo a quantia com o número que maximiza a métrica.

Com as separações feitas pela clusterização, podemos ter grupos com características socioeconômicas semelhantes divididos, permitindo que seja feito um estudo sobre os fatores que podem estar por trás da vacinação em diferentes níveis de performance dos indicadores selecionados. A regressão linear por sua vez nos permite estudar as relações entre as variáveis, permitindo ter uma visão precisa do peso que cada fator socioeconômico possuí sobre a cobertura vacinal, a nível municipal. Neste contexto, foram conduzidas análises para os dados gerais, seguidas por análises separadas para cada um dos dois clusters identificados, com o objetivo de verificar se o comportamento das variáveis se mantém constante. As análises foram conduzidas utilizando a cobertura vacinal como variável resposta Y, e os indicadores socioeconômicos como variáveis explicativas X. Comparamos então os coeficientes e valores p das variáveis explicativas X. Para uma comparação melhor entre os coeficientes, foram multiplicados os valores das variáveis IDHM e GINI por 100. Essas técnicas combinadas nos permitem maior compreensão dos dados, e nos irão auxiliar a chegar em uma conclusão sobre o que afeta as taxas de vacinação.

4. Análise e discussões de resultados

4.1. Análise 1: Evolução temporal da cobertura vacinal da infância no Brasil

Nas primeiras análises realizadas, o objetivo foi abordar o cenário nacional em um nível macro, observando e analisando dados apenas do Brasil como um todo. Inicialmente, elaborou-se gráfico com a média da cobertura vacinal de todas as vacinas que se possuíam dados no Brasil, ano a ano, a fim de analisar a tendência macro de cada período e verificar se havia uma tendência nos dados. Para tanto, foram elaborados dois gráficos, um separando as diferentes vacinas, e um com a média geral de todas as vacinas e um intervalo de confiança.

Analisando os gráficos de cobertura vacinal da infância no Brasil no período de 2010-2021, podemos destacar alguns padrões. Em particular, nos cinco primeiros anos observados, a cobertura vacinal se manteve alta, em torno de 95%, com valores próximos de 100% em 2013 com alta confiabilidade, esse pico se repetiu em 2015, seguido por uma ligeira queda. Outro destaque é o ano de 2018, em que após um período de ascensão na cobertura vacinal, houve uma queda na cobertura vacinal que permaneceu nos anos seguintes, atingindo o menor valor em 2021, o último ano analisado. Complementando a

análise com o gráfico das coberturas vacinais, podemos observar que o comportamento da maioria das vacinas é semelhante, especialmente a queda nas taxas a partir de 2018.

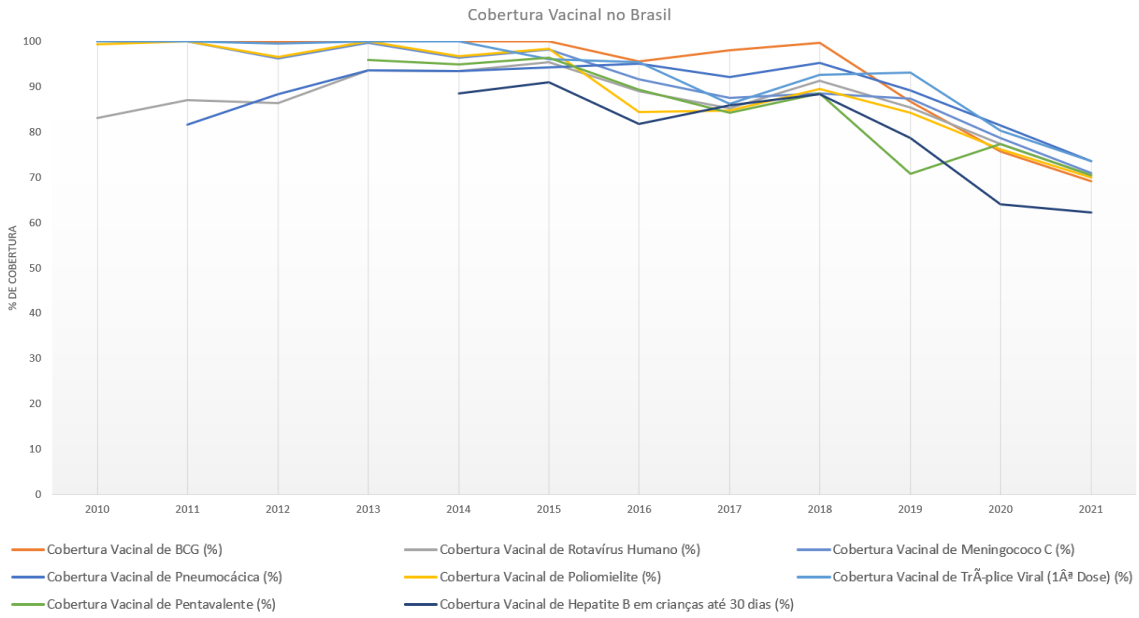

Figura 2: Evolução da cobertura vacinal da infância no Brasil por ano no período de 2010-2021, separação por vacina.

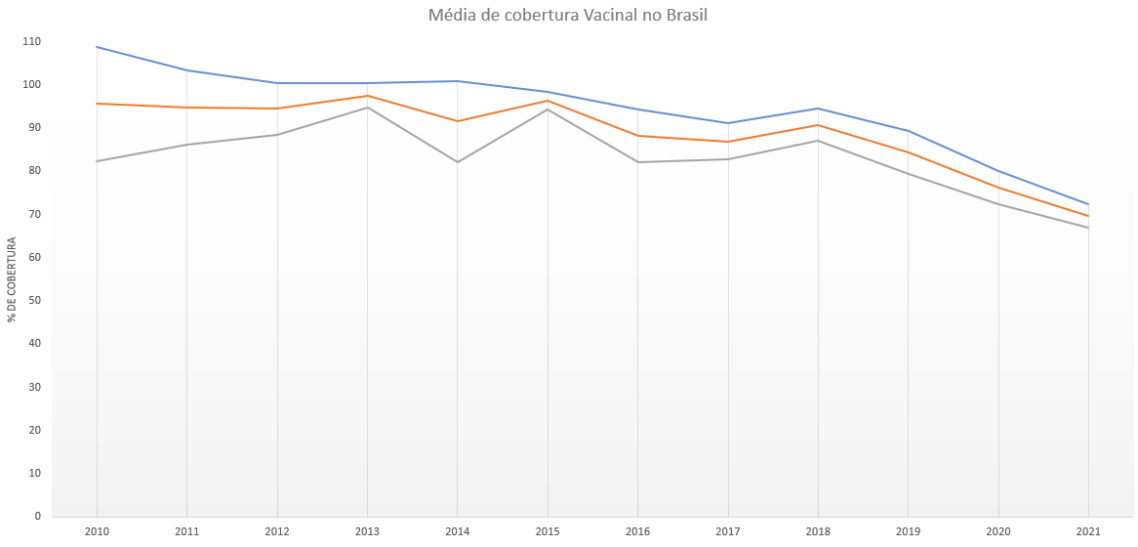

Figura 3: Evolução da média de cobertura vacinal da infância no Brasil por ano no período de 2010-2021, média de todas as vacinas.

Fonte: Autores, 2023

No próximo passo, realizamos uma análise individual de cada região do país: Norte, Nordeste, Centro-Oeste, Sul e Sudeste. Ao analisar a tendência de cada região,

pudemos observar que, no geral, há uma diferença significativa na cobertura vacinal entre as regiões.

No entanto, notamos que o comportamento não é homogêneo para todos os tipos de vacina, uma vez que, em algumas vacinas, a diferença na cobertura entre Sul-Sudeste e Norte-Nordeste é notável, apresentando maior divergência na cobertura vacinal. Um caso que podemos destacar dessa discrepância é a cobertura inicial da vacina de Rotavírus, que inicia próxima de 90% nas regiões Sul e Sudeste, enquanto ela se encontra próxima de 70% na região Norte. Isso nos leva a crer que existem outros fatores externos que afetam a cobertura. É importante ressaltar que essas observações foram feitas ao analisar gráficos de cada região de forma separada.

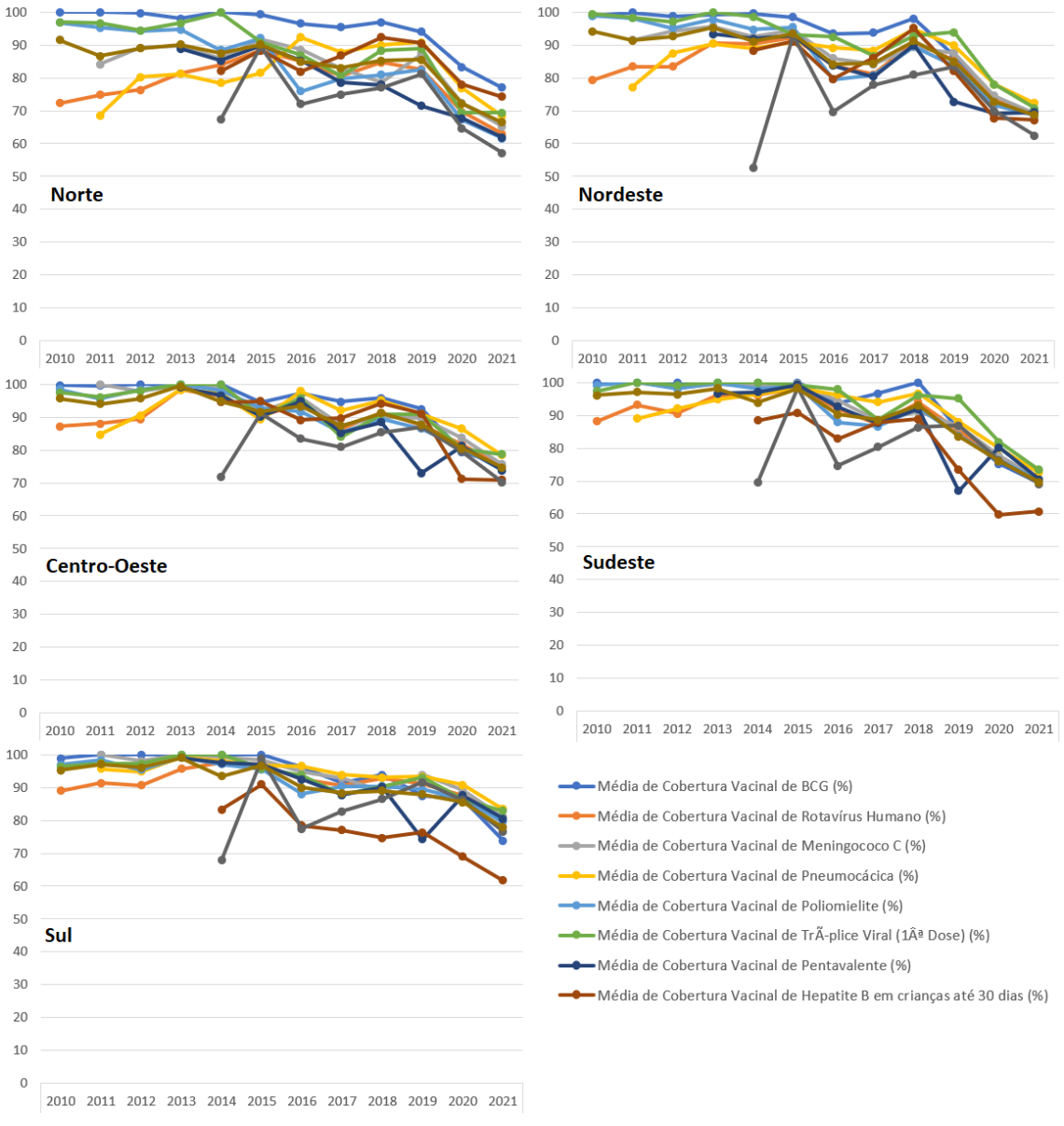

Figura 4: Evolução da cobertura vacinal da infância nas regiões do Brasil por ano no período de 2010-2021, separação por vacina.

Fonte: Autores, 2023

Neste bloco demarcamos 5 gráficos diferentes para as 5 regiões, sendo 1 para cada região. Observando esse painel, podemos notar que há diferenças na cobertura vacinal das diferentes vacinas entre as regiões. Em particular, podemos notar que todas as regiões apresentam queda na cobertura vacinal, com diferenças nas taxas de queda nas diferentes vacinas entre as diferentes regiões. Observando com cuidado, podemos ver que a região Sul teve os melhores desempenhos nas taxas vacinais.

Em sequência, foi iniciada a análise por estados. Os 27 estados diferentes foram agrupados em ordem de média de cobertura vacinal, dos estados, três deles foram destacados: o que apresentou a melhor cobertura vacinal (Santa Catarina - SC), o estado com a pior cobertura vacinal (Pará - PA) e o estado que se mostrou a mediana entre os demais (Goiás - GO).

Tabela 3: Evolução da cobertura vacinal anual da infância por ano estratificado pelos Estados brasileiros no período de 2010-2021. Em destaque de cobertura vacinal: Santa Catarina - SC (verde, maior média), Goiás - GO (amarelo, mediana) e Pará - PA (vermelho, menor média)

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Média
SC	98,04	99,42	99,40	98,66	96,23	99,92	95,41	90,80	91,52	90,03	86,98	79,72	93,84
CE	96,90	94,84	95,06	98,14	97,57	99,62	98,99	98,49	98,95	91,23	82,70	70,08	93,55
RO	96,18	96,54	97,83	97,93	96,37	100,00	98,18	95,86	95,94	92,15	81,18	73,09	93,44
MS	96,80	95,78	97,92	100,00	97,35	100,00	96,90	94,16	97,17	96,06	77,99	70,88	93,42
MG	98,24	99,37	98,51	99,87	93,96	98,45	89,75	88,65	97,11	88,46	86,07	75,73	92,85
ES	97,69	99,51	99,06	98,95	96,74	98,19	90,16	84,06	92,90	87,17	83,21	77,61	92,10
PR	96,72	97,66	97,20	99,81	92,69	98,62	88,41	89,33	89,20	87,80	84,80	78,67	91,74
TO	94,04	95,28	92,59	95,57	93,12	96,65	88,65	88,91	92,24	91,00	85,77	81,21	91,25
MT	95,49	93,55	96,00	97,57	94,66	99,45	92,62	86,67	92,04	86,14	81,48	76,03	90,98
PE	95,82	97,08	97,62	97,60	91,40	97,96	92,73	88,07	95,72	87,94	73,94	68,60	90,37
SP	95,19	96,76	95,97	98,50	93,56	98,22	87,46	89,55	91,88	84,77	79,15	70,78	90,15
DF	93,90	87,91	91,17	100,00	95,54	73,77	100,00	86,88	89,08	86,63	84,29	78,92	89,01
SE	94,62	97,62	94,86	98,27	91,49	94,85	85,47	84,85	93,17	83,78	72,21	73,31	88,71
GO	96,63	98,72	97,71	100,00	91,81	93,61	83,71	81,68	86,58	82,01	78,17	72,89	88,63
RS	91,15	94,30	91,49	98,65	91,25	91,57	86,30	84,78	86,18	85,67	84,72	75,04	88,43
AL	93,34	84,38	89,06	93,21	92,21	94,34	87,38	90,37	97,72	89,59	72,30	72,26	88,01
PB	94,27	91,65	89,13	96,80	89,06	92,51	87,64	87,34	94,86	93,41	72,40	66,65	87,98
RR	90,72	89,07	85,59	83,95	87,92	94,29	89,81	93,73	90,40	84,35	79,30	58,44	85,63
RN	93,11	91,61	93,30	92,09	92,48	93,30	78,65	69,10	90,05	83,61	74,64	72,95	85,41
RJ	93,47	92,47	91,93	95,22	91,18	97,74	93,60	92,07	90,05	73,74	55,74	54,60	85,15
AM	87,79	81,77	87,15	88,32	92,12	95,12	82,02	81,04	84,82	87,39	75,44	71,96	84,58
PI	93,68	93,13	94,03	93,38	82,74	83,76	76,12	81,55	86,82	81,99	73,09	71,79	84,34
BA	92,02	89,19	90,75	92,99	91,28	92,50	75,64	79,27	79,35	76,31	72,38	61,06	82,73
MA	93,21	83,18	90,06	95,05	91,75	92,72	74,67	80,12	84,62	77,35	61,04	61,28	82,09
AC	90,91	83,05	85,48	84,59	70,90	81,63	78,99	78,48	83,44	86,33	65,71	62,87	79,36
AP	89,09	80,38	86,53	89,18	84,67	88,07	89,18	72,43	76,90	81,07	54,93	58,21	79,22
PA	91,72	79,90	88,23	91,03	87,20	75,21	67,79	70,62	72,60	76,08	62,73	59,26	76,86
Média	94,10	92,00	93,10	95,38	91,38	93,41	87,27	85,51	89,68	85,63	75,64	70,14	87,77

Fonte: Autores (2023)

Para ajudar a visualizar a evolução da cobertura vacinal em cada estado, foi feito a seguir um painel com histogramas para cada estado, com destaques para os estados SC, GO e PA.

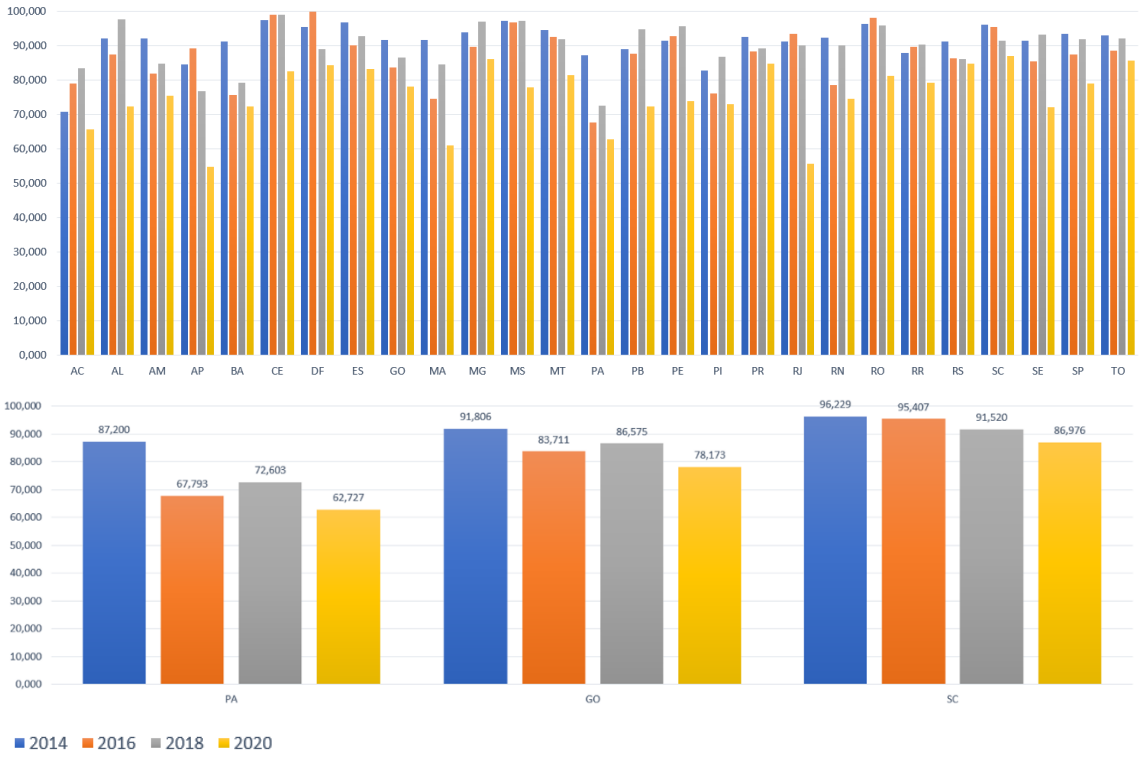

Figura 5: Evolução da cobertura vacinal anual da infância por ano estratificado pelos Estados brasileiros nos anos de 2014, 2016, 2018 e 2020. Em destaque as coberturas vacinais dos estados PA, GO e SC.

Fonte: Autores, 2023

Neste gráfico conseguimos novamente observar as tendências, e verificar que, de fato, os pontos de inflexão ocorreram nos anos de 2016 e 2018 para a maioria dos estados. Alguns casos como o de SC não apresentaram esse ponto de inflexão, porém, um comportamento consistente em todos os estados observados é a queda na cobertura vacinal a partir de 2018.

Observamos que, para regiões e Estados brasileiros, em 2016, a cobertura vacinal apresentava uma tendência ascendente. Porém, em 2018, houve uma queda, comportamento que se repetiu em grande parte dos estados. Além disso, após 2018, houve uma redução na cobertura vacinal para a maioria dos estados. Esse comportamento de queda então se manteve e se acentuou nos anos seguintes, sem sinais de parada.

4.2 Análise 2: Cobertura vacinal e fatores socioeconômicos entre municípios brasileiros

Na segunda parte de nossa análise, nós decidimos optar pelas vacinas tríplice viral, rotavírus, meningite, pneumocócica, poliomielite e BCG. Selecionamos estas vacinas pelo motivo dos dados serem mais completos na base de dados presente no IEPS data, dessa forma conseguimos ter mais estabilidade em nossa análise. Além disso, todas essas vacinas escolhidas são aplicadas em uma mesma faixa etária, exceto a BCG, o que facilita a comparação em suas coberturas.

A partir desse bloco, iniciou-se o uso dos indicadores socioeconômicos mencionados na metodologia. Utilizamos os dados tratados da forma elaborada na metodologia, retirando os outliers e os municípios com dados faltantes. Feito o tratamento de dados, chegamos em um total de 4935 municípios, tendo representatividade em todas as cinco regiões abordando diversas características socioeconômicas diferentes.

4.2.1 Análise 2: Tabelas de correlação

Para avaliar a relação de variáveis socioeconômicas e a cobertura vacinal média dos municípios, além de possíveis influências na taxa de vacinação ao longo do tempo, calculamos o intervalo de correlação de Pearson. Ao observar a tabela conseguimos observar uma forte ligação entre as tendências de cada vacina por indicador. Representamos o coeficiente de correlação com a variável ρ e ilustramos uma tabela para caracterizar as diferentes forças de correlação encontradas.

Tabela 4: Correlação entre as vacinas e indicadores

	vac_bcg	*vac_rota*	*vac_menin*	*vac_pneumo*	*vac_polio*	*vac_tvd1*	*Média Vac*	*cob_ab*	*cob_priv*	*idhm*	*san_adeq*	*Gini*
vac_bcg	1,000											
vac_rota	0,347	1,000										
vac_menin	0,340	0,964	1,000									
vac_pneumo	0,355	0,965	0,964	1,000								
vac_polio	0,333	0,956	0,959	0,943	1,000							
vac_tvd1	0,315	0,905	0,914	0,905	0,912	1,000						
Média Vac	0,599	0,944	0,944	0,943	0,937	0,907	1,000					
cob_ab	-0,104	0,163	0,169	0,148	0,175	0,150	0,110	1,000				
cob_priv	0,202	0,099	0,073	0,068	0,075	0,082	0,131	-0,427	1,000			
idhm	0,320	0,238	0,202	0,174	0,209	0,179	0,273	-0,286	0,702	1,000		
san_adeq	0,110	0,106	0,079	0,057	0,078	0,097	0,107	-0,296	0,609	0,605	1,000	
Gini	-0,057	-0,297	-0,270	-0,226	-0,281	-0,245	-0,249	-0,058	-0,244	-0,397	-0,324	1,000

Valor de ρ (+ ou -)	Interpretação
0.00 a 0.09	Correlação bem fraca
0.10 a 0.19	Correlação fraca
0.20 a 0.59	Correlação moderada
0.60 a 0.89	Correlação forte
0.90 a 1.00	Correlação muito forte

Fonte: Autores, 2023

Podemos observar que a correlação da cobertura vacinal entre a vacina BCG e as demais vacinas é apenas moderada, ao invés de muito forte. Em decorrer disso, optamos por não incluir a vacina BCG em nossas futuras análises considerando a taxa média de vacinação. As vacinas Rotavírus, Meningococo, pneumocácica, póliomelite e tríplice viral todas possuem correlação muito forte, com ρ acima de 0,9 para todos os pares entre essas vacinas. Dentre os indicadores socioeconômicos, os que possuem maior correlação com a cobertura vacinal são o IDHm, com um valor ρ de 0,273 e o índice Gini, com um valor ρ de -0,249.

4.2.2 Análise 2: Gráficos de dispersão

Seguindo as análises de correlação utilizando a média das cinco vacinas de alta correlação, traçamos gráficos gráfico de dispersão comparando as taxas médias de vacinação com as variáveis socioeconômicas de interesse.

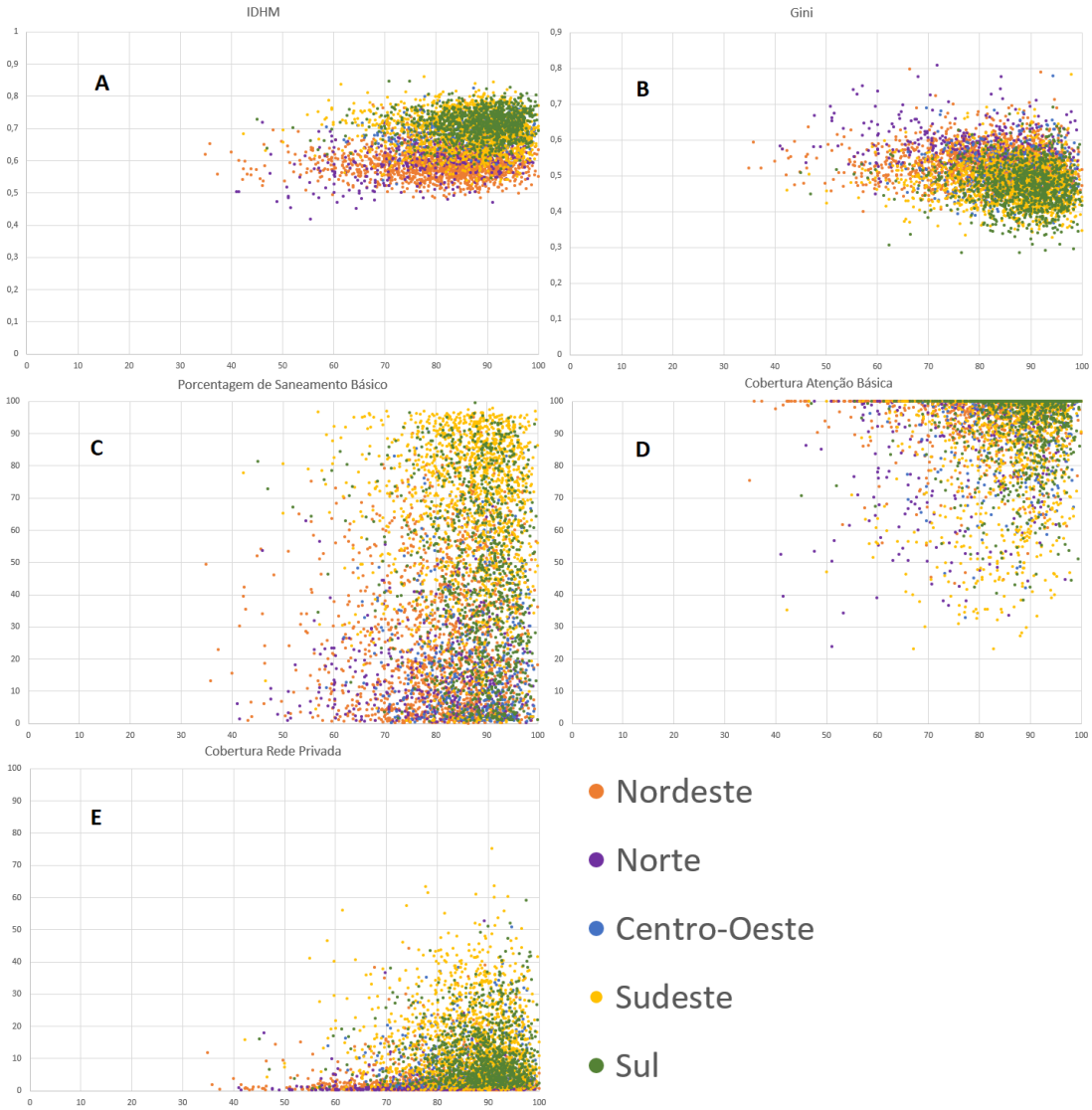

Figura 6: Painel de gráficos de dispersão: Correlação entre os 5 indicadores sócio econômicos e a média de cobertura vacinal para as 5 regiões do Brasil. Os gráficos a correlação da cobertura vacinal com os seguintes indicadores: A: IDHm, B: índice Gini, C: Porcentagem de Saneamento básico, D: Cobertura de atenção básica, E: Cobertura de rede privada de saúde.

Fonte: Autores (2023)

Ao analisar o gráfico de dispersão para as 5 variáveis socioeconômicas, podemos observar alguns padrões. No primeiro gráfico (Cobertura Vacinal x IDHm), é possível perceber uma correlação positiva, além disso, podemos perceber os pontos em verde (região Sul) estão no geral mais acima e a direita em relação aos outros. No segundo gráfico, podemos observar uma correlação leve entre o índice gini e a cobertura vacinal, com os pontos em verde, representando a região Sul estando no geral mais abaixo e à direita. Em contrapartida, o terceiro gráfico, representando a porcentagem de saneamento

básico adequada, teve os pontos mais dispersos, indicando uma correlação mais fraca com a taxa de vacinação.

4.3 Análise 3: Grupos Socioeconômicos e associação com cobertura vacinal nos municípios brasileiros.

A terceira parte da nossa análise apresentará os resultados da utilização da clusterização e regressão linear para análise de dos dados socioeconômicos e a relação deles com a cobertura vacinal.

4.3.1 Análise de Clusterização

A aplicação das técnicas de clusterização e regressão linear possui como objetivo identificar padrões de relacionamentos entre variáveis. Nesse caso, o objetivo da nossa clusterização foi separar os municípios do Brasil em diferentes grupos, levando em conta as características socioeconômicas selecionadas. Para definir o número ideal de clusters, buscamos a maior pontuação pela métrica Silhueta.

Tabela 5: Avaliação dos resultados de clusterização pela métrica Silhueta

Número de Clusters (k)	Silhueta Média K-Means	Silhueta Média K-Medoids
2	0,545	0,542
3	0,450	0,432
4	0,447	0,437
5	0,462	0,384
6	0,422	0,390
7	0,377	0,383
8	0,387	0,325
9	0,371	0,381
10	0,371	0,351
11	0,351	0,330
12	0,353	0,330
13	0,361	0,298
14	0,363	0,272
15	0,366	0,310
16	0,375	0,320
17	0,357	0,319
18	0,366	0,318
19	0,366	0,293
20	0,336	0,300
21	0,346	0,303
22	0,338	0,290
23	0,331	0,302
24	0,318	0,305
25	0,320	0,293
26	0,324	0,302
27	0,330	0,300

Fonte: Autores (2023)

A partir da tabela ilustrada acima, podemos concluir que o melhor número de clusters para análise é dois, pois essa quantia apresentou o maior silhouette. Para caracterizar os clusters, criamos uma tabela comparativa entre os dois, para visualizar a distribuição de municípios em cada um.

Tabela 6: Distribuição dos municípios analisados em cada cluster por região

	Cluster 1		Cluster 2	
Total	**2357**	100,00%	**2578**	100,00%
Centro-Oeste	65	2,76%	364	14,12%
Nordeste	572	24,27%	1043	40,46%
Norte	20	0,85%	375	14,55%
Sudeste	1207	51,21%	289	11,21%
Sul	493	20,92%	507	19,67%

Fonte: Autores, 2023

Ambos os clusters possuem número próximos de municípios, porém podemos observar que a principal diferença entre ambos os clusters é que o cluster 2 possui uma quantidade significativamente maior de municípios das regiões Norte, Nordeste e Centro-Oeste enquanto o Cluster 1 possuí uma proporção maior de municípios no Sudeste.

Em seguida, traçamos o perfil de cada variável de interesse para ambos os clusters, para pode comparar os dois diretamente.

Tabela 7: Características socioeconômicas para o cluster 1

Cluster1	Média Vacinas	Cobertura AB	Cobertura Privada	IDHM	%San. Adeq.	Gini
Média	87,1	90,8	12,3	0,683	59,4	0,491
Desvio Padrão	9,2	15,1	11,8	0,069	27,2	0,060
Mediana	89,3	99,4	8,6	0,697	64,3	0,492
1º Quartil	82,4	87,8	2,7	0,633	46,6	0,450
3º Quartil	93,8	100,0	18,6	0,735	81,8	0,530

Fonte: Autores (2023)

Tabela 8: Características socioeconômicas para o cluster 2

Cluster2	Média Vacinas	Cobertura AB	Cobertura Privada	IDHM	%San. Adeq.	Gini
Média	85,5	94,9	4,8	0,635	23,2	0,519
Desvio Padrão	10,6	10,5	6,7	0,066	20,3	0,064
Mediana	87,8	100,0	2,1	0,630	18,3	0,518
1º Quartil	80,6	95,3	0,9	0,584	6,9	0,477
3º Quartil	93,3	100,0	6,1	0,686	34,4	0,560

Fonte: Autores (2023)

Na comparação dos valores, podemos destacar alguns pontos. A variável que apresentou maior diferença entre ambos os clusters foi a cobertura de saneamento básico, com mais de 36% de diferença no valor médio de ambos os clusters. Comparando diretamente os valores da Média de vacinação, a diferença de média é de apenas 1,63%, sugerindo que, apesar das diferenças notáveis nos indicadores socioeconômicos observados, não há uma correlação tão forte deles com a cobertura vacinal.

Para observar melhor o comportamento de ambos os clusters em torno da variável de vacinação, traçamos um gráfico com a evolução de taxas de vacinação entre ambos os clusters nos últimos anos. Podemos observar que o cluster 1 é o mais favorecido em quesito de cobertura vacinal, enquanto o cluster 2 é o menos favorecido. Porém, ambos os clusters apresentam comportamentos semelhantes de ascensão e queda ao longo dos anos, entre 2016 e 2021.

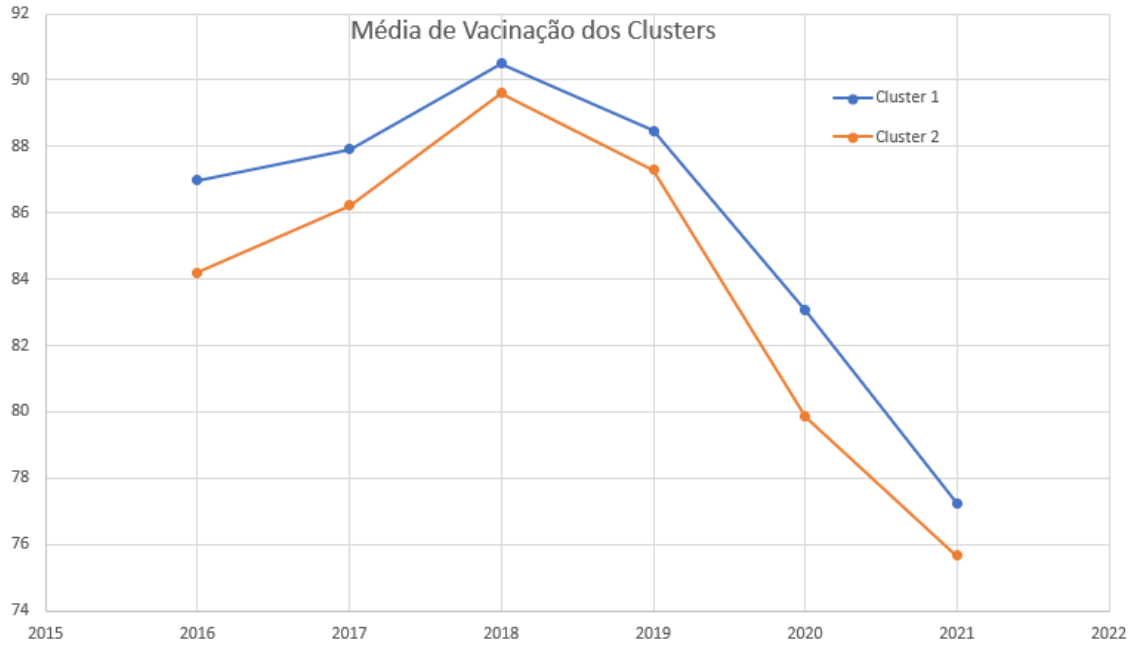

Figura 7: Evolução da média de cobertura vacinal entre os clusters de municípios.

Fonte: Autores, 2023

Com o resultado dessa média de vacinação e dos clusters, chegamos novamente à conclusão de que municípios com piores performances socioeconômicas possuem uma menor taxa média de vacinação. Porém apesar da grande diferença de indicadores socioeconômicos como o de saneamento básico, a diferença entre as taxas de vacinação se manteve consistentemente apenas entre 2% e 3% comparando os dois clusters ao longo dos anos.

4.3.2 Análise de regressão linear

Para estudar melhor a influência de cada variável socioeconômica escolhida nas taxas de vacinação, realizou-se uma análise de regressão linear múltipla utilizando as variáveis de interesse. Nesse caso, a variável dependente (Y) foi a cobertura média de vacinação, e as variáveis independentes (X) foram as seguintes: Índice de Desenvolvimento Humano Municipal (IDHM), índice de Gini, percentagem de saneamento adequado, cobertura da rede privada e cobertura da atenção básica.

Tabela 9: Análise de regressão linear múltipla geral dos dados socioeconômicos dos municípios.

Geral	Coeficiente	Intervalo de confiança (95%)		P-valor
Intercepto	67,150	61,711	72,589	0,000
Gini	-0,295	-0,341	-0,249	0,000
IDHM	0,309	0,252	0,365	0,000
% San. Adeq.	-0,015	-0,027	-0,003	0,012
Cobertura Privada	-0,005	-0,045	0,035	0,810
Cobertura AB	0,155	0,132	0,178	0,000

Fonte: Autores (2023)

Com base na tabela 9, resultante da análise da regressão linear múltipla, podemos chegar a algumas conclusões: As variáveis independentes Gini, IDH e cobertura da atenção básica foram as que retornaram os maiores coeficientes de regressão, além dos menores valores P. Isso significa que elas têm o impacto mais significativo na variável dependente, a cobertura vacinal. Outro aspecto interessante é o percentual de saneamento básico, que não apresentou uma associação relevante considerando todos os outros fatores.

Em contraste, as variáveis de saneamento adequado e cobertura de rede privado não apresentaram coeficientes de regressão significativos. Essas variáveis também apresentaram os maiores valores p, sugerindo não ter um impacto significativo em nossa variável dependente , a cobertura vacinal.

Tabela 10: Análise de regressão linear múltipla dos dados socioeconômicos dos municípios pertencentes ao Cluster 1

Cluster 1	Coeficiente	Intervalo de confiança (95%)		P-valor
Intercepto	74,435	66,188	82,681	0,000
Gini	-0,307	-0,374	-0,241	0,000
IDHM	0,230	0,133	0,326	0,000
% San. Adeq.	-0,024	-0,055	0,007	0,130
Cobertura Privada	0,030	-0,017	0,077	0,216
Cobertura Atenção Básica	0,143	0,116	0,171	0,000

Fonte: Autores (2023)

Tabela 11: Análise de regressão linear múltipla dos dados socioeconômicos dos municípios pertencentes ao Cluster 2

Cluster 2	Coeficiente	Intervalo de confiança (95%)		P-valor
Intercepto	61,220	53,382	69,059	0,000
Gini	-0,284	-0,348	-0,219	0,000
IDHM	0,366	0,290	0,441	0,000
% San. Adeq.	-0,002	-0,028	0,026	0,914
Cobertura Privada	-0,076	-0,161	0,009	0,081
Cobertura Atenção Básica	0,174	0,134	0,214	0,000

Fonte: Autores (2023)

Com essa comparação, podemos observar que os indicadores sócio econômicos possuem maior influência sobre a cobertura vacinal no cluster 2, especialmente o IDHm, refletindo maiores desigualdades entre esses municípios.

Iniciando as análises com o IDHm, em ambos os clusters houve uma associação positiva, com um P-valor <0,05. No caso do cluster 2, a magnitude de associação foi maior quando comparado ao cluster 1, indicando que nesse grupo especialmente, o desenvolvimento socioeconômico é um marcador na cobertura vacinal. O índice Gini foi outra variável que apresentou forte associação com a cobertura vacinal, negativa nesse caso, indicando que as desigualdades socioeconômicas estão ligadas com uma menor cobertura vacinal. A variável de cobertura de atenção básica também apresentou correlação significativa com a cobertura vacinal, indicando a importância dela no acesso às vacinas para a população, especialmente para os municípios do cluster 2. Em contrapartida, as variáveis de cobertura de saúde privada e porcentagem de saneamento adequado apresentaram pouquíssima associação com a cobertura vacinal, com P-valor > 0,05 para essas variáveis em ambos os clusters e valores de coeficiente próximos de 0.

5. Conclusão

O objetivo desse trabalho foi entender o comportamento da cobertura vacinal de diversas vacinas no Brasil e verificar se há uma correlação forte da cobertura vacinal com fatores socioeconômicos.

Nos últimos 10 anos, de 2011 até 2021, a cobertura vacinal da maioria das vacinas do Brasil caiu significativamente, com algumas regiões apresentando quedas de 30% na taxa de algumas vacinas, e o país no geral apresentando uma queda de aproximadamente 15% na vacinação geral nos últimos 4 anos. Percebemos que as regiões mais vulneráveis com condições socioeconômicas menos favorecidas, como a Norte e Nordeste, são especialmente afetadas pelas baixas na cobertura vacinal.

Pela clusterização feita dos municípios analisados, foi possível notar uma clara divisão socioeconômica entre diversos municípios brasileiros, enquanto um cluster possuiu a maioria com municípios da região Sudeste e apresentou melhores resultados, o outro apresentou maioria de municípios das regiões Norte e Nordeste e apresentou resultados significativamente pior. Isso explicou uma parte da cobertura vacinal, porém, não foram fatores tão influentes quanto o esperado. Apesar do cluster com a maior taxa de cobertura vacinal ter indicadores socioeconômicos significativamente melhores, as variações nas quedas de vacinação se apresentaram semelhantes para ambos, com pontos de inflexão e quedas e subidas nos mesmos períodos.

Com esse estudo foi possível identificar uma associação entre fatores socioeconômicos e a cobertura vacinal, em particular com fatores como o idhm, índice Gini e cobertura de atenção básica. A taxa menor de vacinação em municípios menos favorecidos indica que esses deveriam ter mais atenção do sistema de saúde público, e que essa atenção foi sendo reduzida ao longo dos anos. Apesar de aplicar um modelo de regressão linear para estudar melhor a influência de cada variável socioeconômica em cima das taxas de vacinação, e de usar múltiplas vacinas para traçar os perfis de vacinação dos municípios, chegamos à conclusão que fatores socioeconômicos possuem influência nas taxas de vacinação dos diferentes municípios, porém não são o principal fator que explique a queda nas taxas de vacinação como um todo. A dificuldade de obtenção de bases de dados completas para outras variáveis pertinentes, como dados de campanhas e políticas públicas de vacinação durante o período de 2010 até 2019, limita o escopo do

estudo, porém, esses dados seriam instrumentais para termos respostas mais definitivas para responder o que aconteceu com as taxas de vacinação nos anos recentes.

Esse estudo possuiu limitações em utilizar dados numéricos de indicadores socioeconômicos e coberturas vacinais catalogados e disponíveis. Apesar de ter sido possível chegar a uma análise satisfatória com os dados disponíveis, acreditamos que há outros fatores importantes que colaboraram para a queda da cobertura vacinal que não couberam no escopo desse trabalho. Para trabalhos futuros, recomendamos a busca por dados ligados a políticas públicas de saúde e campanhas de vacinação, que vão além da cobertura de atenção básica.

6. Bibliografia

BASTOS, L. Analysis of Performance in Intensive Care Units. Dissertação (Mestrado). Departamento de Engenharia Industrial. Pontifícia Universidade Católica do Rio de Janeiro, 2018. Disponível em: <www.maxwell.vrac.puc-rio.br/35727/35727.PDF>. Acesso em: 16 maio 2023.

BASTOS, L., Aguilar, S., RACHE, B., MAÇAIRA, P., BAIÃO, F., CERBINO-NETO, J., . . . BOZZA, F. A.. Primary Healthcare Protects Vulnerable Populations from Inequity in COVID-19 Vaccination: An Ecological Analysis of Nationwide Data from Brazil. The Lancet Regional Health - Americas, 14, 100335, 2022. doi: https://doi.org/10.1016/j.lana.2022.100335. Acesso em: 11 maio 2023.

BICHLER, M.; HEINZL, A.; VAN DER AALST, W. M. P. Business Analytics and Data Science: Once Again? Business & Information Systems Engineering, v. 59, n. 2, p. 77-79, 2017. Disponível em: https://doi.org/10.1007/s12599-016-0461-1. Acesso em: 11 maio 2023.

BRASIL. Ministério da Saúde. Programa Nacional de Imunizações - Vacinação. Disponível em: https://www.gov.br/saude/pt-br/acesso-a-informacao/acoes-e-programas/programa-nacional-de-imunizacoes-vacinacao. Acesso em: 03 mai. 2023.

SILVA, C. , Análise de fatores socioeconômicos associados à queda de cobertura vacinal de poliomielite no Brasil por meio de aprendizado não-supervisionado. Trabalho de Conclusão do Curso de Engenharia de produção. Departamento de Engenharia Industrial. Pontifícia Universidade Católica do Rio de Janeiro, 2022.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. Communications in Statistics, v. 3, n. 1, p. 1-27, 1974. DOI: 10.1080/03610927408827101.

CAO, L. Data Science and Analytics: A New Era. International Journal of Data Science and Analytics, 2016. 1(1), 1-2. doi: 10.1007/s41060-016-0006-1

CASSIANO, K. M. Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de Suas Componentes Baseada em Densidade. Dissertação (Doutorado). Departamento de Engenharia Elétrica. Pontifícia Universidade Católica do Rio de Janeiro, 2014. Disponível em: https://www.maxwell.vrac.puc-rio.br/colecao.php?strSecao=resultado&nrSeq=24787@1. Acesso em: 16 maio 2023.

Costa, A., & Leo, A. Estatística Descritiva: Principais Conceitos Data Science e Direito, 2020. Disponível em https://dsd.arcos.org.br/estatistica-descritiva-principais-conceitos/. Acessado em: 11 de Maio de 2023.

DABBURA, I. K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Disponível em: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. Acesso em: 16 maio 2023.

DONI, M. V. Análise de Cluster: Métodos Hierárquicos e de Particionamento. Trabalho de Conclusão de Curso (Bacharel em Sistemas de Informação). Universidade Presbiteriana Mackenzie, 2004. Disponível em: http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF. Acesso em: 16 maio 2023.

FIOCRUZ. Cinco dias de fúria: Revolta da Vacina envolveu muito mais do que insatisfação com a vacinação. Disponível em: https://portal.fiocruz.br/noticia/cinco-dias-de-furia-revolta-da-vacina-envolveu-muito-mais-do-que-insatisfacao-com-vacinacao#:~:text=No%20in%C3%ADcio%20de%20novembro%20de,Cultural%20do%20Minist%C3%A9rio%20da%20Sa%C3%BAde. Acessado em: 19 de Maio de 2023

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On Clustering Validation Techniques. Journal of Intelligent Information Systems, v. 17, p. 107-145, 2001. Disponível em: https://link.springer.com/article/10.1023/A:1012801612483. Acesso em: 16 maio 2023.

IEPS. Instituto de Estudos para Políticas de Saúde. IEPS Data. Disponível em https://iepsdata.org.br/. Acessado em: 19 de Maio de 2023.

Maçaira, P. B., L.; Aguilar, S. & Peres, I.. Inferência Estatística com R (1ª edição). Rio de Janeiro, Brasil, 2022.

Manresa, A. P.. Machine Learning to Predict High-Cost Hospitalizations. Dissertação Mestrado. Departamento de Engenharia Industrial. Pontifícia Universidade Católica do Rio de Janeiro, 2022. Disponível em 49137.PDF&ved=2ahUKEwjC1c7iyPb6AhWdA7kGHamfBM4QFnoECBMQAQ&usg=AOvVaw1HuWe8r5upztgyAljV4UjS (puc-rio.br). Acessado em: 11 de Maio de 2023.

MEHTA, S. A Tutorial on Various Clustering Evaluation Metrics. Disponível em: https://analyticsindiamag.com/a-tutorial-on-various-clustering-evaluation-metrics/. Acesso em: 16 maio 2023.

Nunes, L. Cobertura Vacinal no Brasil 2020. Instituto de Estudos para Políticas de Saúde, 2021. Disponível em https://ieps.org.br/wp-content/uploads/2021/05/Panorama_IEPS_01.pdf Acessado em: 10 de Maio de 2023.

Reis, E. A., & Reis, I. A. Análise Descritiva de Dados. Relatório Técnico do Departamento de Estatística da UFMG 1, 2002. Disponível em http://www.est.ufmg.br/portal/arquivos/rts/rte0202.pdf Acessado em: 11 de Maio de 2023.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, v. 20, p. 53-65, 1987. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/0377042787901257. Acesso em: 16 maio 2023.

WEISBERG, S Applied Linear Regression., 2014 Disponível em: https://www.stat.purdue.edu/~qfsong/teaching/525/book/Weisberg-Applied-Linear-Regression-Wiley.pdf . Acesso em: 4 junho 2023.

SCIKIT-LEARN. Clustering Performance Evaluation. Disponível em: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation. Acesso em: 16 maio 2023.

UNASUS. PNI: Entenda Como Funciona um dos Maiores Programas de Vacinação do Mundo. Disponível em https://www.unasus.gov.br/noticia/pni-entenda-como-funciona-um-dos-maiores-programas-de-vacinacao-do-mundo. Acessado em, 05 de mai. de 2023.

OLIVEIRA, W, DUARTE, E, FRANÇA, G, GARCIA, L. Como o Brasil pode deter a COVID-19, 2020. Disponível em: https://doi.org/10.5123/S1679-49742020000200023. Acesso em: 30 maio 2023.