UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE GRADUAÇÃO EM BACHARELADO EM ESTATÍSTICA

GABRIEL GRACIANO DE MENDONÇA

ESTUDO DO ÍNDICE DE DESENVOLVIMENTO DA EDUCAÇÃO BÁSICA DO ESTADO DA PARAÍBA ATRAVÉS DE ALGORITMOS NÃO SUPERVISIONADOS E ESTATÍSTICA ESPACIAL

CAMPINA GRANDE - PB
2023

GABRIEL GRACIANO DE MENDONÇA

ESTUDO DO ÍNDICE DE DESENVOLVIMENTO DA EDUCAÇÃO BÁSICA DO ESTADO DA PARAÍBA ATRAVÉS DE ALGORITMOS NÃO SUPERVISIONADOS E ESTATÍSTICA ESPACIAL

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Estatística.

Orientador: Prof.Dr. Ricardo Alves de Olinda

CAMPINA GRANDE - PB
2023

M539e Mendonça, Gabriel Graciano de.

Estudo do índice de desenvolvimento da educação básica do estado da Paraíba através de algoritmos não supervisionados e estatística espacial [manuscrito] / Gabriel Graciano de Mendonca. - 2023.

35 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.

"Orientação : Prof. Dr. Ricardo Alves de Olinda , Departamento de Estatística - CCT. "

1. Educação. 2. Cluster. 3. Estatística espacial. 4. Machine learning. I. Título

21. ed. CDD 519.5

Elaborada por Geovani S. de Oliveira - CRB - 15/1009

Biblioteca Central BC/UEPB

GABRIEL GRACIANO DE MENDONÇA

ESTUDO DO ÍNDICE DE DESENVOLVIMENTO DA EDUCAÇÃO BÁSICA DO ESTADO DA PARAÍBA ATRAVÉS DE ALGORITMOS NÃO SUPERVISIONADOS E ESTATÍSTICA ESPACIAL

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Estatística.

Trabalho aprovado em 27/11/2023.

BANCA EXAMINADORA

Prof.Dr. Ricardo Alves de Olinda (Orientador)
Universidade Estadual da Paraíba (UEPB)

Prof.Dr. Macio Augusto de Albuquerque
Universidade Estadual da Paraíba (UEPB)

Prof.Dr. Luiz Medeiros de Araújo Lima Filho
Universidade Federal da Paraíba (UFPB)

Dedico esse trabalho em especial a minha mãe Katia Suzana Medeiros Graciano, a minha vó Maria das Vitórias Medeiros Graciano e a minha tia Karla Suely Medeiros Graciano.

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus, pela saúde, coragem, oportunidades e todas as bençãos que me proporcionou ao logo da vida e durante os anos de graduação.

Agradeço a toda a minha família, minha mãe Katia Suzana, tia Karla Suely e avó Maria das Vitórias, por todas as palavras de encorajamento, conselhos, paciência e incentivo constante nos momentos difíceis, sem o apoio de vocês essa conquista não seria possível. Assim como a minha namorada, Suzany Silva, pela compreensão, paciência e constante apoio durante minha jornada acadêmica.

Gostaria de agradacer também ao meus colegas de sala Daví Barbosa, Elyda Camyla, Gislânia Cauanny, Hellen Silva e Marcela Araujo, por todas as conversas, momentos de descontração e auxílio nas atividades ao longo do curso. Ao professor e amigo Castor da Paz Filho, por todos conselhos e caronas durante os 5 anos de curso.

Agradeço a todos os professores que tive ao longo do curso, em especial aos professores Tiago Almeida de Oliveira e ao meu orientador e amigo Ricardo Alves de Olinda, por todos os conhecimentos e vivências que me proporcionou ao longo do curso, sendo elas, a oportunidade de participar de um Programa de Iniciação Cientifica, Projetos de Extensão e pesquisas de avaliação de gestões públicas, contribuindo significativamente para o meu desenvolvimento acadêmico, bem como, o aprendizado de técnicas de amostragem, na prática.

“É fácil mentir com estatísticas, mas é difícil dizer a verdade sem elas.”
(Andrejs Dunkels)

RESUMO

Este trabalho teve como objetivo estudar o comportamento do Índice de Desenvolvimento da Educação Básica (Ideb), juntamente com a taxa de aprovação e o nível de aprendizado, selecionando as escolas estaduais de todos os 223 municípios do estado da Paraíba durante o ano de 2019. Para realização das análises foram empregadas técnicas de algoritmos não supervisionados bem como modelagem espacial. Partindo para análise dos dados, foi utilizado técnicas de análise espacial para dados de área, com intuito de observar o comportamento dos indicadores em estudo, tanto em relação a cada município como em relação as mesorregiões do estado. A fim de verificar a presença de autocorrelação espacial em relação aos indicadores taxa de aprovação, aprendizado e o próprio Índice de Desenvolvimento da Educação Básica, foi utilizado o Índice de Moran Global, seguidos pela construção dos mapas Box Map e Lisa Map, visando identificar os municípios que apresentaram alguma dependência espacial. Após avaliar a influência entre os municípios, utilizou-se o método K-Means, a fim de detectar a presença de grupos entre os municípios a partir de suas semelhanças, mesmo não sendo geograficamente adjacentes. Diante dos resultados, observamos que municípios localizados nas mesorregiões do Sertão e na Borborema tendem a apresentar índices superiores comparado com as demais mesorregiões. Além disso, por meio do método K-Means identificamos a presença de 3 grupos, sendo eles: municípios com fluxo alto, aprendizado baixo e seu Ideb regular; municípios com fluxo baixo, aprendizado regular, apresentando seu Ideb muito baixo; municípios com todos os indicadores altos. Tais resultados são fundamentais para o desenvolvimento de estratégias mais eficazes e direcionais, uma vez que identificamos os municípios com deficit e quais indicadores estão influenciando diretamente para isso.

Palavras-chave: educação; estatística espacial; cluster; machine learning.

ABSTRACT

This work aims to study the behavior of the Basic Education Development Index (Ideb), together with the approval rate and the level of learning, selecting only state schools from all 223 municipalities in the state of Paraíba during the year 2019. To carry out the analyses, unsupervised algorithmic techniques were used, as well as spatial modeling. Starting with data analysis, spatial analysis techniques were used for area data, with the aim of observing the behavior of all indicators under study, both in relation to each municipality and in relation to the mesoregions of the state. In order to verify the presence of spatial autocorrelation in relation to the indicators of approval rate, learning and Ideb itself, the Moran Global Index was used, followed by the construction of the Box Map and Lisa Map maps, aiming to identify the municipalities that showed some dependence space. After evaluating the influence between municipalities, the K-Means method was used in order to detect the presence of groups between municipalities based on their similarities, even though they are not geographically adjacent. Given the results, we observed that municipalities located in the Sertão and Borborema mesoregions tend to present higher rates compared to the other mesoregions. Furthermore, through the K-Means method we identified the presence of 3 groups, namely: municipalities with high flow , low learning and regular Ideb; municipalities with low flow, regular learning, presenting very low Ideb; municipalities with all high indicators. Such results are fundamental for the development of more effective and directional strategies, since we identify municipalities with deficiencies and which indicators are directly influencing this.
Keywords: education; spatial statistics; cluster; machine Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

IBGE	Instituto Brasileiro de Geografia e Estatistica
Ideb	Índice de Desenvolvimento da Educação Básica
Saeb	Sistema de Avaliação da Educação Básica
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MEC	Ministério da Educação

LISTA DE SÍMBOLOS

α	Letra grega Alpha
μ	Letra grega Mi
ε	Letra grega Epsilon

SUMÁRIO

1 INTRODUÇÃO

O Índice de Desenvolvimento da Educação Básica (Ideb) foi criado em 2007 pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), de acordo com o Ministério da Educação (MEC), ele foi elaborado com o intuito de medir a qualidade do aprendizado nacional bem como estabelecer metas visando a melhoria do ensino. O cálculo do Ideb é feito a partir da combinação dos dois principais índices de ensino, sendo eles: a taxa de aprovação/ fluxo (rendimento escolar), que apresenta a quantidade de alunos que concluíram o ano escolar dentro do tempo previsto, e o nível de aprendizado, média de desempenho nas avaliações de matemática e português.

Soares e Xavier (2013) afirmam que o Ideb se tornou a forma frequentemente utilizada para analisar a qualidade da educação básica brasileira e devido a isso, tem tido uma influência considerável em relação aos debates sobre educação no pais. Para Fernandes (2007) possuir um indicador referente ao desenvolvimento educacional apresenta algumas vantagens, sendo elas, a detecção de escolas e rede de ensino cujo, estudantes apresentam desempenhos a baixo dos esperados, além disso, ele permite o monitoramento da evolução ao longo do tempo dos alunos. Porém ao optar por um indicador de taxa de aprovação e de desempenho em testes de padronização podem apresentar alguma dificuldade devido à possibilidade de compensação entre eles.

No ano de 2019, a rede estadual do estado da Paraíba não atingiu nenhumas das metas previstas para a nota do Ideb, cujo, nos anos iniciais do ensino fundamental apresentou nota 4,9 e sua meta era 5, nos anos finais obteve nota 3,8 porem sua meta era 4,2 e em ralação ao ensino médio atingiu notas muito abaixo do esperado no qual sua nota foi 3,6 e sua meta era 4,2 (PARAIBA, 2020). Devido a esse cenário, torna-se de suma importância o estudo dos fatores que compõem o Índice de Desenvolvimento da Educação Básica através da estatística espacial, técnica na qual possibilita a criação de mapas proporcionando deste modo a distribuição geográfica dos índices em estudo, para todos os municípios do estado da Paraíba.

Para Anselin (1995) a análise de estatística espacial é fundamental para o planejamento das cidades, bem como, para tomada de decisões em diversas áreas como a saúde e gestão pública, devido a fornecer informações importantes sobre a distribuição geográfica de fenômenos. Com isso estudos no contexto espacial vêem se tornando cada vez mais comum, dado a disponibilidade de sistema de informação geográfica (SIG) devido ao seu baixo custo e fácil manipulação (CAMARA et al., 2004). Outra vantagem notável da aplicação da estatística espacial é a capacidade de identificar padrões e tendências entre as regiões em estudo, permitindo, deste modo, a detecção de possíveis grupos levando em consideração apenas sua localidade geográfica.

Porém, a partir da aplicação do aprendizado de máquina, ou Machine Learning, conseguimos verificar a presença de grupos ocultos nos dados. Podemos definir Machine Learning como a subárea da inteligência artificial, na qual utiliza dados e algoritmos no intuito de imitar a forma que os seres humanos aprendem. Para Paixao et al. (2022) o principal objetivo de um

modelo de *machine learning* é construir um sistema que aprenda com base em um banco de dados pré-definido e posteriormente gere um modelo de predição, classificação ou detecção.

O *machine learning*, por sua vez, também pode se dividido em diversas áreas. Para Hastie e Tibshirani (2011) os problemas de aprendizagem podem ser categorizados como supervisionado e não supervisionado. No aprendizado supervisionado, o objetivo da análise é prever o resultado de um determinada medida com base em uma série de medidas de entrada, já no aprendizado não supervisionado, o objetivo é descrever as relações e padrões entre os dados de entrada. A diferença entre o aprendizado supervisionado e o não supervisionado se torna fundamental para o direcionamento das abordagens e métodos a partir dos interesses da análise.

Diante disso, o presente trabalho tem como objetivo verifcar a presença de grupos entre os 223 municípios do estado da Paraíba através de algoritmos não supervisionados e de estatística espacial, na qual a partir dessas técnicas é possível verificar a semelhança entres municípios que são geograficamente vizinhos, assim como, a semelhança entre municípios que apresentam uma distância geográfica. Além disso, também possui o objetivo de propor abordagens de classificação na área do desenvolvimento regional do estado da Paraíba, demonstrando como essas técnicas podem solucionar problemas existentes.

2 FUNDAMENTAÇÃO TEÓRICA

Nessa seção buscamos descrever os principais conceitos referentes a estatistica espacial, dando ênfase na análise de dados de área, bem como na análise de agrupamento a partir de algoritmos não supervisionados, via método K-Means.

2.1 Estatistica Espacial

O uso da estatística espacial não é uma técnica atual, de acordo com Camara et al. (2004), no século XIX, John Snow foi um dos pioneiros na incorporação do contexto espacial nas análises. No ano de 1854, a cidade de Londres era acometida por uma epidemia de cólera, na época pouco se sabia sobre os mecanismos da doença, inclusive como era a sua forma de transmissão. Diante disso, John Snow verificou que o maior número de óbitos devido à cólera ocorriam perto de pontos de captação de água, local no qual havia uma concentração de dejetos de pacientes que estavam acometidos da doença. Tal situação demonstra o quanto a relação espacial nos dados contribuem de forma significativa para a compreensão de fenômenos e tomada de decisões.

Estatística Espacial é a área da estatística que busca estudar dados que apresentam um componente geográfico, ou seja, quando o valor de determinada variaval está associado a uma localização espacial, sendo ela através de coordenadas ou polígonos. Silva (2010) define a análise de dados espaciais como um conjunto de técnicas estatísticas, que tem como objetivo descrever padrões presentes em dados geograficos, e estabelecer de forma quantitativa o relacionamento entre variáveis geográficas.

Camara et al. (2004) afirma que existem três tipos de dados para descrever problemas relacionados a análises espaciais, sendo eles:

1. Área com contagem ou Dados de Área;
2. Padrões Pontuais ou Processos Pontuais;
3. Superfícies Contínuas.

Visto que as características dos dados apresentam uma variação, é comum que cada tipo de análise apresente uma metodologia estatística singular para descreve-los (SILVA, 2010). Nesse trabalho, o foco estará voltado para a análise de dados de área, abordagem utilizada para compreender padrões em áreas geograficamente especificadas, como, por exemplo, municípios.

2.1.1 Matriz de Proximidade

De acordo com Camara et al. (2004), a principal ferramenta utilizada para estimar a variabilidade espacial de dados de área é a matiz de proximidade espacial, ou como também é conhecida, matriz de vizinhança. A matriz de proximidade espacial modela a estrutura de

variabilidade espacial, sendo desta forma útil para representar o arranjo espacial dos objetos (SILVA, 2010). Dado um conjunto de **n** áreas podemos construir uma matriz $\mathbf{W}_{(nxn)}$, onde cada elemento W_{ij} representa uma medida de proximidade entre o polígono A_i e A_j. Deste modo a medida de proximidade pode ser calculada de três formas sendo elas:

- $w_{ij} = 1$, se o centróide de A_i está a uma determinada distância de A_j ; caso contrário $w_{ij} = 0$
- $w_{ij} = 1$,se A_i, compartilha um lado comum com A_j, caso contrário $w_{ij} = 0$;
- $w_{ij} = \frac{I_{ij}}{I_i}$, onde I_{ij} é o comprimento da fronteira entre A_i e A_j, I_i é o perímetro de A_i

Nesse trabalho adotaremos a segundo forma citada, forma no qual também é conhecida como movimento da rainha (*queen's movement*).

2.1.2 Média Movél Espacial

A média móvel é um dos indicadores mais utilizados, com o intuito de prever tendências, possibilitando desta forma o cálculo do valor médio de um determinado período (SANTOS; JUNIOR, 2006). Camara et al. (2004) afirma, que a média móvel é uma técnica simples de avaliar a variação da tendência espacial dos dados, assim como, calcular a média dos valores vizinhos. A média móvel é denotada por $\hat{\mu}_i$ associada a um z_i , representando a i-ésima área, calculada a partir dos elementos da matriz de vizinhança w_{ij}, sendo expressa da seguinte forma:

$$\hat{\mu}_i = \sum_{j=1}^{n} w_{ij} z_i$$

podendo ser expresso pelo seguinte modelo estatístico:

$$Y_i = \mu_i + \varepsilon_i^{,} + \varepsilon_{,}$$

na qual Y_i representa o valor da variável na área A_i, μ_i é o valor de Y pertencente a A_i, $\varepsilon_i^{,}$ é o componente espacial de Y_i, com esperança de $\varepsilon_i^{,} = 0$ e ε representa o ruído (variável aleatória independente e identicamente distribuída) (SILVA, 2010).

2.1.3 Índice de Moran Global

Um aspecto fundamental da análise exploratória, consiste na caracterização da dependência espacial, apresentando como os valores estão correlacionados no espaço (CAMARA et al., 2004). Diante disso, para verificar se os valores observados de um determinado polígono apresenta alguma dependência, isto é, autocorrelação espacial, relacionada com seus adjacentes, utiliza-se o índice de Moran global, expressa pela seguinte equação:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \overline{z})(z_j - \overline{z})}{\sum_{i=1}^{n}(z_i - \overline{z})^2}$$

na qual temos:

n: Número de áreas em estudo;

w_{ij}: Elementos que compoem a matriz de proximidade espacial ou vizinhança espacial;

z_i: O valor do atributo considerado na área i;

$\overline{z}$: O valor médio do atribributo na região em estudo.

Camara et al. (2004) ainda destaca que, a partir do índice de Moran, formula-se as seguintes hipóteses:

$$\begin{cases} H_0: & \text{Há independencia espacial;} \\ H_1: & \text{Há autocorrelação espacial} \end{cases}$$

Diante disso, temos a rejeição da hipótese nula, ou seja, evidências que há autocorrelação espacial, quando o p-valor resultante do teste for inferior ao nível de significância, sendo utilizado na maioria das vezes α=0,05 (5% de significância). Segundo Marques et al. (2010), a interpretação da estatística do índice de Moran, é semelhante à da interpretação da correlação entre duas variáveis aleatórias, devido ao índice variar de -1 a 1, indicando ausência de dependência espacial quando for igual a 0. Por outo lado, quando os valores forem próximos de 0, representam uma autocorrelação espacial não significativa.

2.1.4 Indicadores Locais de Associação Espacial (LISA)

Os indicadores locais são conhecidos por estabelecer um índice de associação, levando em consideração cada área (MARQUES et al., 2010). Para Camara et al. (2004) tais indicadores produzem um valor específico em cada área, possibilitando, desta forma, a identificação de grupos (*clusters*). Portanto, o índice local de Moran pode ser calculado para cada área de i, a partir dos valores normalizados de Z_i (valor do atributo), expresso da seguinte forma:

$$I_i = \frac{z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} z_j^2}$$

Corfome Anselin (1995), o indicador local de autocorrelação espacial, deve atender dois objetivos especificos, sendos eles, possibilitar a identificação de padrões de associação

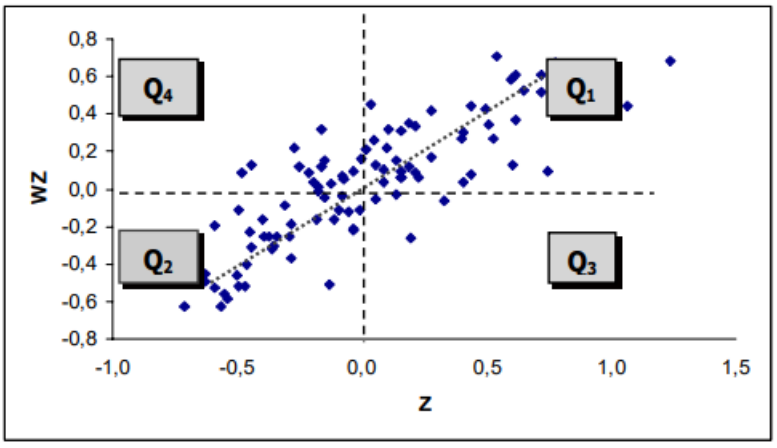
espacial, e garantir que a soma dos indicadores do LISA, devem ser proporcional ao índice global. Para construção do LISA map, mapa responsável pela avaliação do nível de significância das comparações, Anselin (1995) classifica os valores de significância em 5 grupos: não significativos, 0,05 de significância (95% de confiança), 0,01 de significância (99% de confiança), 0,001 de significância (99,9% de confiança) e 0,0001 de significância (99,99% de confiança) (MARQUES et al., 2010). Nesse trabalho iremos usar apenas os quatro primeiros grupos de significância citados.

2.1.5 Diagrama de Espalhamento de Moran

Conforme Anselin (1996), o gráfico de espalhamento de Moran, indica a forma do grau de associação linear entre um conjunto de valores observados (Z), e a média ponderado dos seus vizinhos (WZ). Essa representação linear entre Z e WZ serve para fundamentar as definições acerca dos processos autorregressivos espaciais, que são frequentemente usados para descrever o método gerado por trás da dependência espacial.

Nesse contexto, é gerado um gráfico bidimensional de Z (abscisa) por WZ (ordenada), com base nos seus valores normalizados, permitindo assim a análise da variabilidade espacial (CAMARA et al., 2004). O diagrama é dividido em quatro quadrantes, apresentados na Figura 1. Nunes (2015) explica os quadrantes da seguinte forma:

Figura 1 – Diagrama de Espalhamento de Moran



Fonte: Camara et al. (2004)

- Q1 (Alto-Alto) Valores positivos com médias dos vizinhos positivas e Q2 (Baixo-Baixo) Valores negativos com média dos vizinhos negativas. Ambos os quadrantes apresentam um associação possitiva, ou seja, vizinhos que apresentam valores semelhantes.
- Q3 (Alto-Baixo) Valores positivos com média dos vizinhos negativas e Q4 (Baixo-Alto) Valores negativos com médias dos vizinhos positivas. Ambos os quadrantes apresentam

associação negativa, ou seja, vizinhos com valores distindos.

2.2 Aprendizado de Máquina

O aprendizado de máquina está presente no nosso dia a dia, solucionando problemas, às vezes até sem nossa percepção, ele está presente desde assistentes virtuais (como alexa e siri) até nas sugestões das plataforma de streaming, na qual realiza sugestões de acordo com o perfil de cada usuário. Sabe-se que nós seres humanos estamos em constante evolução, visto que a cada dia adquirimos conhecimentos novos. Segundo Souza (2020), no aprendizado de máquina, o processo de aprendizagem é obtido de forma similar a nossa, uma vez que o programa precisar de algumas informações iniciais, e a partir dessas informações geram seus próprios conhecimentos, conforme os resultados obtidos.

O aprendizado de máquina pode ser dividido em diversas formas, sendo às duas principais o aprendizado supervisionado, que pode ser definido como regressão ou classificação, e o aprendizado não supervisionado, definido como clusterização/agrupamento. De acordo com Maia (2023), no modelo supervisionado os dados de entrada e saída são definidos e estão relacionados entre si, para isso, é necessario uma fase de treino, em que, uma parte dos dados são postos no sistema, com a intenção de ajustar os parâmetros do modelo, a fim de prever o resultado no conjunto de amostras de treinamento. Em contra partida, no aprendizado não supervisionado, apenas os dados de entrada são definidos, com o objetivo de verificar padrões ou estruturas existente no conjunto de dados.

2.2.1 Análise de Cluster

A análise de cluster, ou análise de agrupamento, é definido por Hair et al. (2009) como um conjunto de técnicas multivariadas, empregadas no intuito de agrupar objetos de acordo com suas características. Favero e Belfiore (2017) complementam mencionando que os agrupamentos devem ser homogêneos internamente e heterogêneos entre si, isto é, observações de um grupo devem apresentar semelhança em relação as variáveis em estudo, e consequentemente devem diferir das observações dos demais grupos.

No aprendizado não supervisionado, a máquina realiza o agrupamento dos dados que possuem alguma similaridade, sintetizando, desta forma, os dados em *clusters* com base em suas semelhanças. Portanto, a clusterização pode ser definida como a análise do nível de similaridade entre os *clusters* e seus elementos, que se baseia em uma função de distância, em que, quanto menor for sua distância, mais semelhantes são os dados (SOUZA, 2020). Podemos considerar a análise de agrupamento como uma análise exploratórias, ou de interdepêndencia, devido as suas aplicações não apresentarem características preditivas, para observações não contidas na análise inicial (FAVERO; BELFIORE, 2017).

Segundo Favero e Belfiore (2017), para realizar tal análise, não é necessário ter conhecimentos de álgebra nem estatística, ao contrário de outras técnicas multivariadas, o pesquisador

necessita apenas de um objetivo principal e a partir disso escolher a medida de distância mais apropriada de acordo com sua base de dados. Outras características da análise de cluster é a capacidade de lidar com: conjunto de dados de alta dimensão, sua aplicabilidade em diferentes tipos de dados e habilidade de realizar o agrupamento de diferentes tamanhos (DONI, 2004).

2.2.2 Medidas de Semelhança e Distância

Maia (2023) descreve que, a etapa inicial da aplicação da análise de cluster é a definição de uma medida de distância ou semelhança. Tais medidas são utilizadas como referência para a atribuição de cada observação em um grupo. Hair et al. (2009) destaca que, as medidas de distância representam uma métrica de dissimilaridade, na qual, quanto maiores seus valores, menor será sua similaridade, em outras palavaras, para utilizar a distância como uma medida de similaridade, é aplicado a relação inversa.

No contexto da análise de agrupamento, Neto e Moita (1998) discutem que, a semelhança entre duas amostras podem ser expressas, por uma função de distância entre dois pontos representativos destas amostras em um espaço n-dimensional. Para Hair et al. (2009), mesmo que as proximidades aparentem ser conceitos simples, há diversas medidas de distâncias, tendo cada uma dela características específicas.

Em seguida, são apresentadas algumas distâncias que se enquadram na definição de medidas de dissimilaridade.

Distância Euclidiana

Hair et al. (2009) afirma que é a mediada mais conhecida e utilizada, na qual também é chamada de distância em linha reta. Ela é definida pela seguinte expressão:

$$d_{pq} = \sqrt{(X_{1p} - X_{1q})^2 + (X_{2p} - X_{2q})^2 + ... + (X_{kp} - X_{kq})^2} = \sqrt{\sum_{j=1}^{k} (X_{jp} - X_{jq})^2}$$

em que X_{jp} e X_{jq} representam os valores dos indivíduos p e q, para as variáveis j.

Distância Quadrática Euclidiana

Representa a soma dos quadrados das diferenças sem o cálculo da raiz quadrada, acelerando o tempo computacional (HAIR et al., 2009). Sendo expressa por:

$$d_{qe} = (X_{1p} - X_{1q})^2 + (X_{2p} - X_{2q})^2 + ... + (X_{kp} - X_{kq})^2 = \sum_{j=1}^{k} (X_{jp} - X_{jq})^2$$

Distância Manhattan

A distância de Manhattan na maioria dos casos, exibem resultados semelhantes à distância Euclidiana, entretanto, nela o efeito de uma grande disparidade em uma dimensões de um elemento é reduzida, uma vez que a medida não esta elevada ao quadrado (DONI, 2004). Sendo definida por:

$$d_{xy} = |X_1 - Y_1| + |X_2 - Y_2| + ... + |X_p - Y_p| = \sum_{i=1}^{p} |X_i - Y_i|$$

Distância Chebychev

Hair et al. (2009) apresenta a distância de Chebychev, como a distância que é representada pela maior diferença entre todas as variáveis do agrupamento, sendo possível o uso de variáveis com escalas diferentes. Sua definição é dado por:

$$d_{ch} = mx(|X_1 - Y_1|, |X_2 - Y_2|, |X_p - Y_p|)$$

2.2.3 Métodos Hierárquicos

Os métodos hierárquicos, correspondem a uma cadeia de divisões de agrupamento $n-1$, cujo, n representa o número de observações, tendo como resultado uma estrutura de hierarquia (HAIR et al., 2009). Podemos dividir esse método em dois subgrupos, sendo eles os métodos aglomerativos ou divisivos. Para Favero e Belfiore (2017) se todas as observações forem consideradas independentes, com base na sua distância, são formados sucessivos grupos até que no estágio final, reste apenas um agrupamento. Em contrapartida, se todas as observações forem tidas como agrupadas, e após cada etapa, cada observação formar um grupo menor, até que posteriormente haja grupos individuais, é um processo divisivo.

Em outras palavras Doni (2004) descreve que o método aglomerativo, como cada elemento representando um grupo, inicialmente, e a cada estagio os elementos são ligados de acordo com sua similaridade, finalizando com apenas um grupo contendo todos os elementos. Ele também define os métodos divisivos, como uma forma inversa ao método aglomerativo, ou seja, há apenas um grupo inicial contendo todos os elementos e é divido em dois subgrupos, de tal modo que os elementos de cada subgrupo esteja distante dos outros.

2.2.4 Métodos Não Hierárquicos

Os métodos não-hierárquicos, são técnicas de agrupamento desenvolvidos com o intuíto de agrupar e organizar os elementos em numero pré definido de grupos. Doni (2004) descreve que os métodos não hierarquicos são mais rápidos, comparado com os hierárquicos, uma vez

que não é necessário calcular nem armazenar a matriz de similaridade. Dentre as abordagens de agrupamentos, temos que a via *K-Means* é o mais utilizada e popularmente conhecida, a qual, é empregada para segmentar um conjunto de dados em grupos, a partir das caratéristicas em comum de cada indivíduo.

Método *K-Means*

Podemos definir o método *K-Means* como uma algorítimo de aprendizado não supervisionado, utilizado para agrupar um conjunto de dados em k grupos (MAIA, 2023). Favero e Belfiore (2017) apresenta as seguintes etapas no procedimento *K-Means*:

1. Definir o número inicial de grupos e seus respectivos centroides;
2. Deve-se selecionar a observação que se encontra mais próxima de um centroide e realoca-la para esse grupo. Devido a esse movimento, outro grupo perde essa observação, portanto, é necessário recalcular os centroides dos grupos;
3. Deve-se repetir o passo anterior, até que não seja mais possível deslocar nenhuma observação por maior proximidade a um centroide de outro grupo.

Para complementar, Doni (2004) menciona algumas características desse método, sendo elas a sensibilidade a ruido, ou seja, elementos com valores discrepantes podem distorcer os dados, a tendência em criar determinados grupos e a inadequação do método para identificar grupos que não possuem formas convexas.

2.3 Materiais

Como mencionado ao longo do trabalho, os dados em análise são referentes ao Índice de Desenvolvimento da Educação Básica (Ideb). Especificadamente, foram selecionadas as notas do ensino médio, apenas das escolas estaduais do estado da Paraíba, referentes ao ano de 2019, retirados do site QEdu (2023). Ao observar os dados, constatou-se que 21 municípios, cerca de 9,41% do total de municípios na Paraíba, expostos na Tabela 1, apresentavam informação faltantes referentes a variável aprendizado, essa lacuna influenciava diretamente para que o Ideb fosse 0 também.

Tabela 1 – Municípios com dados faltantes referente ao índice de aprendizado.

Municípios com dados faltantes	
Alagoa Grande	Mulungu
Alagoinha	Nova Floresta
Arara	Pedro Régis
Assunção	Riachão
Cacimbas	Santa Inês
Conde	São José do Sabugi
Cuitegi	São José dos Ramos
Diamante	Serraria
Joca Claudino	Sertãozinho
Logradouro	Sumé
Mataraca	

Fonte: Elaborado pelo autor, 2023.

Para solucionar esse problema, optou-se por realizar a média dos municípios vizinhos, para cada município que apresentava dados faltantes em relação ao índice de aprendizado. Posteriormente, utilizando a taxa de aprovação, bem como, o nível de aprendizado aplicamos a fórmula do Ideb, conforme definido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, INEP (2018) , sendo apresentada da seguinte forma:

$$Ideb_{ji} = N_{ji}P_{ji}$$

$$0 \leq N_j \leq 10; 0 \leq P_j \leq 1; 0 \leq IDEB \leq 10$$

Sendo:

- i = Ano do exame (Saeb e Prova Brasil) e do Censo Escolar;
- N_{ji} = Média da proficiência em Língua Portuguesa e Matemática, normalizadas para um indicador variando entre 0 e 10, dos alunos da unidade j, alcançada em determinada ano do exame;
- P_{ji} = Indicador da taxa de aprovação da etapa de ensino dos alunos da unidade j;

Também incorporamos o shapefile dos municípios e das mesorregiões do estado da Paraíba, obtidos do site do IBGE, sendo essencial para construção dos mapas ao longo do trabalho. Todas as análises foram realizadas por meio do software R, versão 4.2.0 (RCran), por meio da IDE RStudio (TEAM, 2023).

3 RESULTADOS E DISCUSSÃO

Nessa sessão aprensentamos os resultados das análises estatísticas citados anteriomente. Primeiramete foi realizado uma breve análise exploratória, em seguida foram aplicadas as técnicas de análise de dados de área e por fim a aplicação do aprendizado não supervisionado via K-Means .

3.1 Análise exploratória

Na Tabela 2 apresentamos a análise descritivas dos dados, análise na qual nos permite compreender de melhor formas o comportamento dos índices em estudo, denotando as principais medidas de tendência central, assim como seus pontos máximo e mínimo.

Tabela 2 – Análise descritiva do Índice de Desenvolvimento da Educação Básica, Nível de aprendizado e Taxa de aprovação.

	Taxa de aprovação	Aprendizado	Ideb
Mínimo	0,54	3,05	2,13
1º Quartil	0,78	4,06	3,30
Mediana	0,84	4,35	3,60
Media	0,83	4,36	3,65
3º Quartil	0,90	4,68	4,00
Máximo	1,00	5,44	5,20

Fonte: Elaborado pelo autor, 2023.

Interpretando a análise descritiva, observou-se que o Ideb das escolas estaduais do estado da Paraíba apresentou uma variação de 2,13 á 5,2 tendo como média 3,65. No que se refere ao nível de aprendizado, tal índice variou de 3,05 a 5,44 tendo sua media de 4,36. Também foi possivel identificar que a menor taxa de aprovação encontrada foi de 54%, enquanto sua média foi de 83%.

3.2 Análise espacial

Com a finalidede de verificar se as variáveis em estudo apresentam dependência espacial, foi realizado a análise de Moran global, na qual, podemos observar na Tabela 3.

Tabela 3 – Análise do Índice de Moran, visando verificar a presença de dependêndia espacial.

	Taxa de aprovação	Aprendizado	Ideb
Estatística de Moran	0,108	0,311	0,300
Variância	0,001	0,001	0,001
P-Valor	0,003	<0,001	<0,001

Fonte: Elaborado pelo autor, 2023.

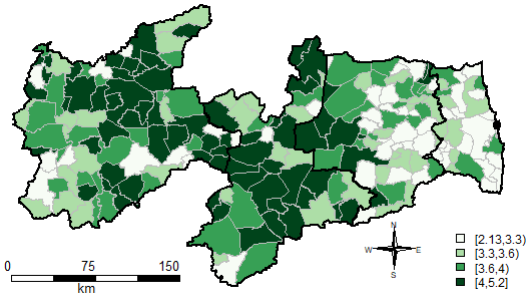
Logo ao nível de 1% de significância rejeitamos a hipótese nula, ou seja, temos evidencias que os índices apresentam autocorrelação espacial. Tais resultados são similares ao de Oliveira (2023), em que realizou uma análise espacial da qualidade da educação pública do Brasil, abrangendo os anos iniciais e finais. Em seus resultado Oliveira (2023) observou a rejeição da hipótese nula ao nível de 0,001% de significância, concluindo que, ao investigar a qualidade da educação em âmbito nacional, também é evidente a presença autocorrelação espacial.

Após identificar a presença da autocorrelação espacial, é de interesse compreender a natureza da dependência espacial e desta forma, identificar possiveis grupos de municípios que exercem algum tipo de influência sobre seus vizinhos, para isso prosseguimos a análise de forma mais detalhada. Com o intuito de alcançar esses objetivos, realizamos a elaboração dos gráficos de quartis, *Box Map* e o *LISA Map* .

3.2.1 Mapa de quartis

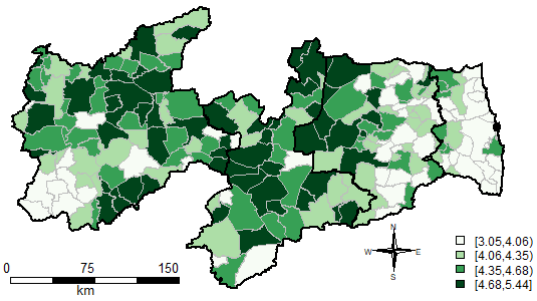
A partir do mapa de quartis, conforme as Figura 2, 3 e 4, conseguimos observar a distribuição espacial dos índices em estudo, sendo fundamental para identificação de padrões e tendências espaciais. Diante disso foi realizado o respectivo mapa para todos os índices:

Figura 2 – Distribuição espacial do Índice de Desenvolvimento da Educação Básica do estado da Paraíba no ano de 2019



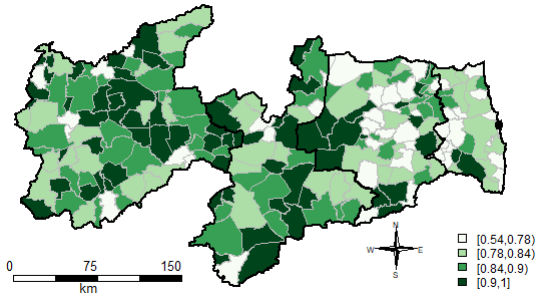
Fonte: Elaborado pelo autor, 2023.

Figura 3 – Distribuição espacial do nível de aprendizado do estado da Paraíba no ano de 2019



Fonte: Elaborado pelo autor, 2023.

Figura 4 – Distribuição espacial da taxa de aprovação do estado da Paraíba no ano de 2019



Fonte: Elaborado pelo autor, 2023.

Na Figura 2, a qual se refere ao Ideb, podemos observar que os municípios situados nas mesorregiões do sertão paraibano e na borborema apresentam notas superiores quando comparado com as demais, além disso, podemos perceber uma tendência na diminuição do índice à medida que se aproxima da mesorregião da mata paraibana (litoral), de forma que nenhum município localizado à direita da cidade de Campina Grande apresenta nota superior a 4.

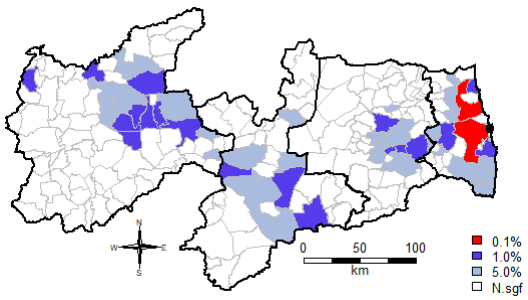
Em relação ao nível de aprendizado, apresentado na Figura 3, podemos interpretar de forma similar a variável Ideb, no entanto, é pertinente ressaltar que alguns municípios localizados no sudoeste da mesorregião do sertão (sendo eles Santa Inês, Conceição, Diamante, Ibira, Santana de Mangueira, Cural Velho), apresentam pontuações inferiores a 4,06, diferença na qual se destaca em relação aos demais municípios que integram essa mesorregião.

Por outro lado, no tocante a taxa de aprovação, Figura 4, nota-se um equilíbrio relativo, entretanto os municípios localizados na mesorregião do agreste apresentam suas taxas de aprovação ligeiramente inferiores.

3.2.2 LISA Map

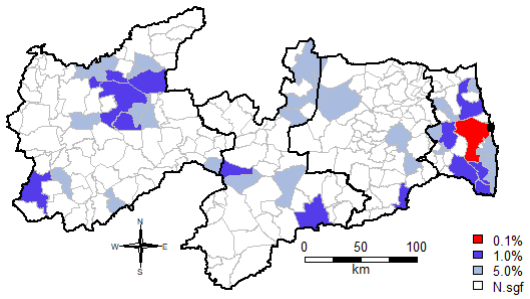
De acordo com Silva (2010), o *LISA Map* indica as regiões que apresentam correlação local diferente das outras. Através de tais mapas, Figura 5, 6 e 7, é possível identificar os municípios que apresentam uma correlação espacial mais significativa em comparação com os demais.

Figura 5 – LISA Map Ideb



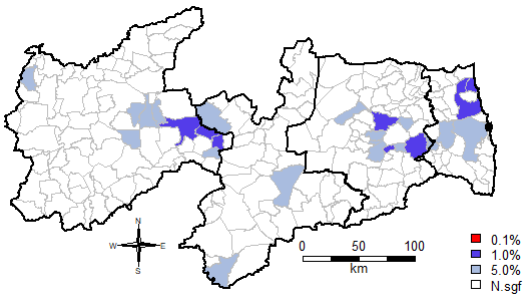
Fonte: Elaborado pelo autor, 2023.

Figura 6 – LISA Map Aprendizado



Fonte: Elaborado pelo autor, 2023.

Figura 7 – LISA Map Taxa de Aprovação



Fonte: Elaborado pelo autor, 2023.

De maneira equivocada (SANTOS; JUNIOR, 2006), apresenta o *LISA Map* e realiza sua interpretação, considerando os níveis de significância a 95%, 99% e 99,9%. O termo correto a ser empregado para essa porcentagem seria "nível de confiança", uma vez que os níveis de significância são de 5%, 1%, 0,1% e 0,001%. Como mencionado anteriormente utilizaremos apenas os três primeiros níveis citados.

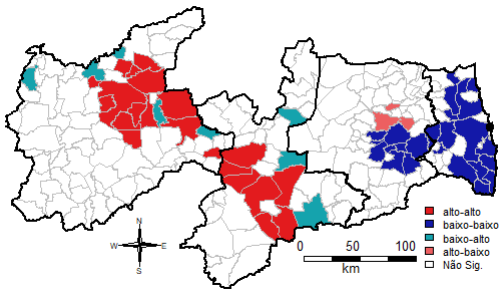
Logo, ao analisarmos o LISA Map referente ao Ideb, Figura 5, observamos que 2 municípios se destacam, na qual apresentam significância de 0,1% , sendo eles os municípios de Santa Rita e Rio Tinto. Também foi possível identificar a presença de 19 municípios significativos ao nivel de 1%, bem como e 35 municípios ao nível de 5% de significância. Já no que diz respeito ao aprendizado, Figura 6, verificamos que apenas um município foi significativo ao nível de 0,1%, sendo ele Santa Rita, além disso, identificamos 14 municípios com 1% de significância e 31 com 5%. Por fim, ao analisar a taxa de aprovação, apresentado na Figura 7, não identificamos a presença de municipios ao nível de significância de 0,1%, esse cenário pode ser consequência da homogeneidade dos dados. Entretanto, vale se destacar a presença de 8 municípios com

significância de 1% e outros 19 ao nível de 5%.

3.2.3 Box Map

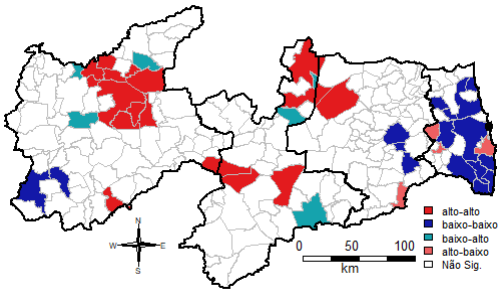
O *Box Map*, corforme as Figura 8, 9 e 10, é uma extensão do gráfico de espalhamento de Moran, na qual os elementos de cada quadrante são representados com uma cor específica (SILVA, 2010). A fim de indentificar os que apresentam relação entre si foi contuidos os graficos a seguir :

Figura 8 – *BoxMap* do Índice de Desenvolvimento da Educação Básica do estado da Paraíba no ano de 2019



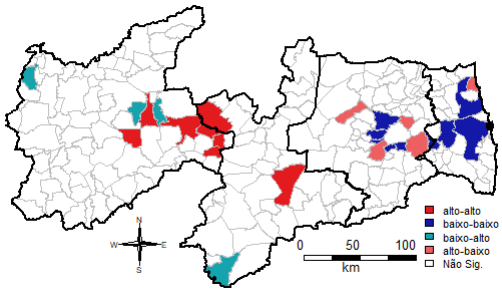
Fonte: Elaborado pelo autor, 2023.

Figura 9 – *BoxMap* do nível de aprendizado do estado da Paraíba no ano de 2019



Fonte: Elaborado pelo autor, 2023.

Figura 10 – *BoxMap* da taxa de Aprovação do estado da Paraíba no ano de 2019



Fonte: Elaborado pelo autor, 2023.

De maneira análoga a (OLIVEIRA, 2023), destacamos os municípios considerados alto-alto em vermelho e baixo-baixo em azul. Todavia, além de considerar que os municípios exibem valores semelhantes entre vizinhos, é possível também, definir que eles tendem a influenciar seus vizinhos a alcançarem valores semelhantes (ou inverso) aos seus.

Por meio da Figura 8, reverente às notas do Ideb, constata-se que as cidades situadas no sertão e na borborema tentem a influenciar seus vizinhos de maneira positiva, em contrapartida, a maioria dos municípios presentes na mata paraibana apresentam seus índices baixos e, adicionalmente, tendem a influenciar seus vizinhos de forma negativa.

Para nível de aprendizado, plotado na Figura 9, observa-se que as cidades de Conceição e Diamante, localizadas no sudoeste da mesorregião do sertão, são aquelas que além de apre-

sentarem notas baixas exercem uma influência negativa sobre seus vizinhos, seguindo para a mesorregião da mata paraibana, percebe-se que os municípios em sua maioria, impactam de maneira negativa seus vizinhos, no entanto, mesmo com esse detalhe, cidades como João Pessoa, Mari e Sobrado apresentam suas notas "altas"mesmo sob influência negativa dos seus adjacentes.

Referente ao índice da taxa de aprovação, apresentado na Figura 10, podemos observar uma diminuição a respeito da quantidade de municípios que apresentam influência. Como visto no mapa de quartis, esse índice evidenciou um comportamento mais homogêneo comparado com as demais. Entretanto, vale destacar o município de São Sebastião de Umbuzeiro, localizado no sul da mesorregião da borborema, e o município de Triunfo, localizado no noroeste da mesorregião do sertão, que mesmo apresentando uma taxa de aprovação baixa, influência positivamente seus vizinhos. Já na mesorregião da mata paraibana destaca-se a cidade de Baia da Traição, sendo o único município que apresenta taxa de aprovação alta, porém influencia negativamente seus vizinhos.

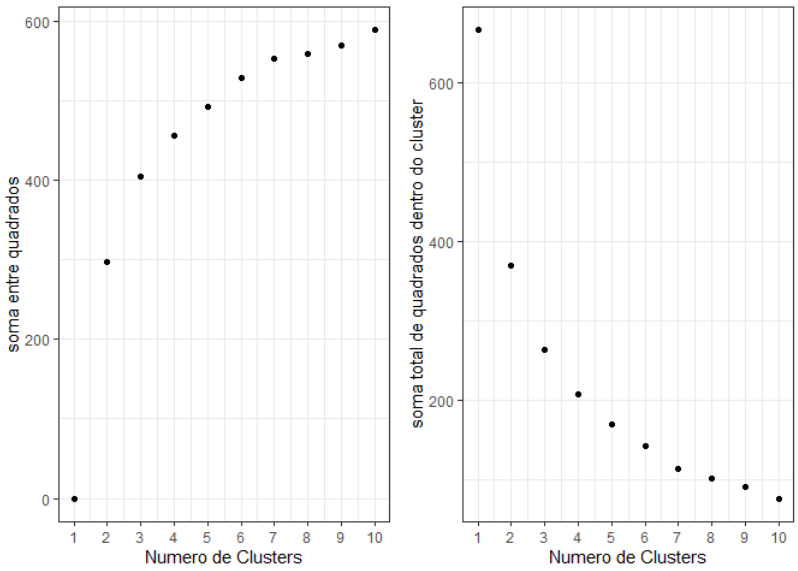
3.3 Aprendizado não supervisionado

Para a aplicação do aprendizado não supervisionado, a primeira técnica a ser realizada foi a padronização dos dados, com intuito de uniformizar a escala de todos os índices, tal passo se torna primordial para a sequência da análise, assegurando que nenhum íncide apresente influência sobre as outras, visto que a taxa de aprovação apresenta valores entre 0 e 1 e o nível de aprendizado entre 0 e 10. Nesse trabalho a forma de padronização escolhida foi a normalização dos dados, a qual consiste nos valores do individuo (nesse caso os municípios) subtraídos de sua média, divididos pelo desvio padrão.

Em seguida realizamos o método elbow (ou regra do cotovelo), a partir de duas métricas de qualidade de cluster, sendo elas a soma das distâncias ao quadrado entre os centroide dos grupos e o centroide geral (betweenss), bem como a soma das distâncias ao quadrado entre cada ponto e o centroide do seu grupo (tot.withinss). Tais medidas descrevem a separação entre os grupos indicando a distância média entre seus centroides, e a outra representa a variação interna dos grupos, ou seja, o quanto os pontos estão bem agrupados dentro de cada grupo.

Na Figura 11, podemos observar as medidas betweesnss e tot.withinss, possibilitando uma compreensão mais clara da variação entre os clusters, bem como, auxiliando na identificação do número ideal de grupos.

Figura 11 – Método Elbow para verificar o número ideal de grupos



Fonte: Elaborado pelo autor, 2023.

Diante dos gráficos acima podemos concluir que o número ideal de grupos é entre três e quatro, uma vez que não observamos decaimento nem crecimento significativo dos valores. Para melhor explicabilidade e interpretação assumimos que o número ideal de grupos é três. Diante disso, realizamos a divisão do banco de dados via método K-Means e posteriormente realizamos a média de todos os índices para cada grupo, afim de verificar o comportamento dos municípios dentro de cada grupo bem como a discrepância entre eles. Tais resultados estão apresentados na Tabela 4:

Tabela 4 – Média dos índices em cada grupo após o agrupamento .

	Taxa de aprovação	Aprendizado	Ideb
Grupo 1	0,86	4,04	3,48
Grupo 2	0,71	4,27	3,04
Grupo 3	0,89	4,78	4,26

Fonte: Elaborado pelo autor, 2023.

- Grupo 1: Os municípios que apresentam taxa de aprovação alta porém apresentam um nível de aprendizado baixo, fazendo com que seu Ideb seja “regular”.
- Grupos 2: Os municípios que apresentam um nível de aprendizado regular porém apresentam uma taxa de aprovação muito baixa, influenciando seu Ideb a ser baixo também.
- Grupo 3: Os municípios que apresentam tanto a taxa de aprovação como o nível de aprendizado alto, conseguintemente apresentando os melhores valores do Ideb.

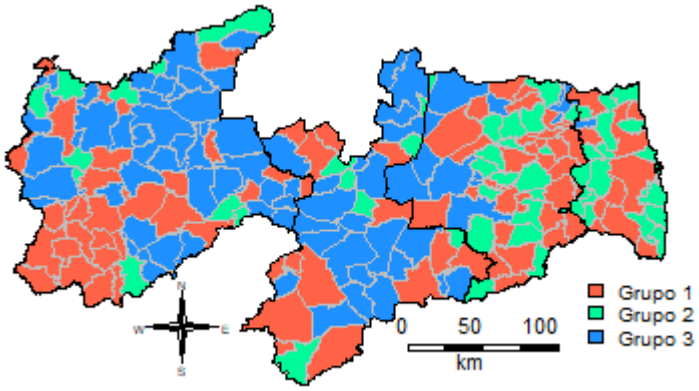
Com base nessas informações é possível observar o comportamento dos grupos de forma gráfica, na qual através do pacote *factoextra* geramos o *cluster plot*, apresentado na Figura 12, gráfico no qual nos permite visualizar a distribuição dos municípios nos grupos gerados. Logo podemos observar 90 municípios pertencentes ao grupo 1 (representados de vermelho), 55 municípios no grupo 2 (representados de verde) e 78 municípios no grupo 3 (representados de azul), facilitando assim a compreensão das características dos municípios nos seus respectivos grupos.

Figura 12 – *Cluster plot*, identificando o comportamento dos grupos gerados

Fonte: Elaborado pelo autor, 2023.

Outra forma de observar a divisão dos grupos é através do contexto espacial, apresentado na Figura 13, na qual podemos combinar a estatística espacial com o método *K-Means*:

Figura 13 – Identificação dos grupos gerados no contexto espacial



Fonte: Elaborado pelo autor, 2023.

Logo, ao analisar a Figura 13, é possível identificar que os municípios pertencentes ao grupo 3 estão localizados predominantemente nas mesorregiões do Sertão e da Borborema, de forma similar aos mapas do Ideb e do nível do aprendizado. Na mesorregião do Agreste observamos um comportamento mais homogêneo em comparação com as demais mesorregiões, na qual verificamos uma presença considerável de municípios pertencentes aos três grupos. Em contra partida, na mesorregião da Mata Paraibana, prevalece a presença de municípios pertencentes aos grupos 1 e 2.

4 CONCLUSÃO

Com os resultados obtidos podemos concluir que o trabalho atingiu seus objetivos predefinidos, na qual através de técnicas de algoritmos não supervisionados e modelagem espacial, conseguimos verificar de forma eficaz a presença de grupos (clusters) entre os 223 municípios do estado da Paraíba, em relação aos dados do Índice de Desenvolvimento da Educação Básica referentes o ano de 2019. Diante da combinação dessas duas técnicas, foi possível verificar a similaridade entre os municípios, levando em consideração o contexto espacial, ou seja, municípios que são geograficamente adjacentes, bem como a semelhança entre municípios que apresentam distancias geográficas.

A análise espacial proporcionou a identificação de forma precisa dos municípios que necessitam de um incentivo adicional, a fim de que consigam evoluir ao mesmo nível dos demais municípios do estado. Além disso, observou-se que todos os índices em estudos apresentaram dependência espacial, na qual foi possível destacar, que municípios situados nas mesorregiões do sertão paraibano e na borborema apresentam índices educacionais superiores comparados com as demais. De forma paralela, à medida que se aproximamos da mesorregião da Mata Paraibana, observa-se uma tendência decrescente desses índices, na qual ela apresenta o maior número de municípios com índices educacionais muitos baixos.

Através da aplicação do método K-Means foi possível identificar a presença de três grupos distintos dentre os 223 municípios da Paraíba, na qual dentro de cada grupo conseguimos identificar com precisão, qual fator desempenha a maior influência sobre Índice de Desenvolvimento da Educação Básica (Ideb).

Tais resultados podem servir como base sólida para construção de políticas públicas no contexto espacial, visando a homogeneização da qualidade da educação em todo o estado da Paraíba. No âmbito do aprendizado não supervisionado, a verificação das áreas que necessitam de uma atenção mais específica servem como guia para construção de projetos mais preciso e direcionais, dado a presença de cenários diversos, entres os municípios. Deste modo, os presentes resultados não são apenas uma análise do cenário educacional, mas indo além se torna uma ferramenta fundamental para a tomada de decisão e implementação de projetos públicos, tendo como objetivo a evolução da educação do estado da Paraíba de forma mais rápida e econômica.

Espera-se também que com esse trabalho o estudo da estatística espacial e de técnicas de classificação seja mais divulgada e que com isso chame mais a atenção do corpo discente para estudo e aplicação do mesmo. Uma vez que ao compartilhar tais métodos enriquecemos a pesquisa acadêmica e aplicada, bem como contribuindo para o desenvolvimento de soluções mais eficazes em diversas áreas de estudo.

REFERÊNCIAS

ANSELIN, L. Local indicators of spatial association—lisa. *Geographical Analysis*, v. 27, p. 93–115, 1995. Disponível em: <https://dces.webhosting.cals.wisc.edu/wp-content/uploads/sites/128/2013/08/W4_Anselin1995.pdf>. Citado 3 vezes nas páginas 13, 17 e 18.

ANSELIN, L. The moran scatterplot as an esda tool to assess local instability in spatial association. In: ____. *Spatial Analytical Perspectives on GLS*. Londres: Taylor & Francis, 1996. p. 111–125. Disponível em: <https://dces.qa.webhosting.cals.wisc.edu/wp-content/uploads/sites/128/2013/08/W4_Anselin1996.pdf>. Acesso em: 27 out. 2023. Citado na página 18.

CAMARA, G. et al. Análise espacial de área. In: ____. *Análise Espacial de Dados Geográficos*. Brasília: EMBRAPA, 2004. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/cap5-areas.pdf>. Acesso em: 17 jul. 2023. Citado 5 vezes nas páginas 13, 15, 16, 17 e 18.

DONI, M. V. *Análise de cluster: Métodos hieraquicos e de particionamento*. 2004. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) - Faculdade de Computação e Informática, Universidade Presbiteriana Mackenzie. Disponível em: <http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF>. Acesso em: 10 out. 2023. Citado 3 vezes nas páginas 20, 21 e 22.

FAVERO, L. P.; BELFIORE, P. *Análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata*. [S.l.]: ELSEVIER, 2017. Citado 3 vezes nas páginas 19, 21 e 22.

FERNANDES, R. *Índice de Desenvolvimento da Educação Básica (Ideb)*. 2007. Disponível em: <https://www2.unifap.br/gpcem/files/2011/09/IDEB-_Texto_para_discuss%C3%A3o26.pdf>. Acesso em: 24 out. 2023. Citado na página 13.

HAIR, J. F. et al. *Análise Multivariada de Dados*. [S.l.]: Bookman, 2009. Citado 3 vezes nas páginas 19, 20 e 21.

HASTIE, T. J.; TIBSHIRANI, R. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2011. Citado na página 14.

INEP. *Índice de Desenvolvimento da Educação Básica – Ideb*. 2018. Nota técnica no INEP. Disponível em: <https://download.inep.gov.br/educacao_basica/portal_ideb/o_que_e_o_ideb/Nota_Tecnica_n1_concepcaoIDEB.pdf>. Acesso em: 06 jun. 2023. Citado na página 23.

MAIA, D. E. G. *Utilização do aprendizado não supervisionado via K-Means para a tomada de decisão no mercado financeiro*. 2023. Trabalho de Conclusão de Curso (Bacharelado em Estatística) - Universidade Estadual da Paraiba. Disponível em: <https://dspace.bc.uepb.edu.br/jspui/bitstream/123456789/29144/4/TCC%20-%20Dami%c3%a3o%20Everton%20Gomes%20Maia.pdf>. Acesso em: 9 ago. 2023. Citado 3 vezes nas páginas 19, 20 e 22.

MARQUES, A. P. da S. et al. *Análise exploratória de dados de área para índices de furto na mesorregião de Presidente Prudente- SP*. 2010. Anais do III Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da geoinformação. Disponível em: <https://docplayer.com.br/89467863-Analise-exploratoria-de-dados-de-area-para-indices-de-furto-na-mesorregiao-de-\presidente-prudente-sp.html>. Acesso em: 24 ago. 2023. Citado 2 vezes nas páginas 17 e 18.

NETO, J. M. M.; MOITA, G. C. Uma introdução à análise exploratória de dados multivariados. *Scielo*, v. 21, p. 467–469, 1998. Disponível em: <https://www.scielo.br/j/qn/a/b64d96fbT5jMHmnc48SdXnr/#>. Citado na página 20.

NUNES, F. G. Análise exploratória espacial de indicadores de desenvolvimento socioambiental das regiões de planejamento do norte e nordeste goiano. *Ateliê Geografico*, v. 7, 2015. Disponível em: <https://revistas.ufg.br/atelie/article/view/19809>. Citado na página 18.

OLIVEIRA, J. A. de. Análise espacial da qualidade da educação publica no brasil em 2011 e 2021. *Doity*, 2023. Disponível em: <https://doity.com.br/anais/xvieec/trabalho/278076>. Citado 2 vezes nas páginas 25 e 28.

PAIXAO, G. M. de M. et al. Machine learning na medicina: Revisão e aplicabilidade. *Scientific Electronic Library Online*, v. 118, p. 95–102, 2022. Disponível em: <https://www.scielo.br/j/abc/a/WMgVngCLbYfJrkmC65VFCkp/#>. Citado na página 13.

PARAIBA, G. *Ideb: Paraíba bate meta nas séries iniciais do ensino fundamental, mas fica abaixo nos anos finais e no ensino médio*. 2020. G1. Disponível em: <https://g1.globo.com/pb/paraiba/noticia/2020/09/15/ideb-paraiba-bate-meta-nas-series-iniciais-do-ensino-fundamental-mas-fica-abaixo-nos\-anos-finais-e-no-ensino-medio.ghtml>. Acesso em: 10 jul. 2023. Citado na página 13.

QEDU. *Use dados. Transforme a educação*. 2023. Disponível em: <https://qedu.org.br/>. Acesso em: 11 ago. 2023. Citado na página 22.

SANTOS, L. dos; JUNIOR, A. A. R. Análise espacial de dados geográficos: A utilização da exploratory spatial data analysis - esda para identificação de áreas críticas de acidentes de trânsito no município de são carlos (sp). *Sociedade e Natureza*, v. 18, p. 97–107, 2006. Disponível em: <https://seer.ufu.br/index.php/sociedadenatureza/article/view/9251/5695>. Citado 2 vezes nas páginas 16 e 27.

SILVA, N. C. N. da. *Análise de dados de área aplicada a dois indicadores econômicos de mesorregiões do estado de Minas Gerais*. Dissertação (Mestrado) — Universidade Federal de Lavras, Lavras, mar. 2010. Citado 4 vezes nas páginas 15, 16, 27 e 28.

SOARES, J. F.; XAVIER, F. P. Pressupostos educacionais e estatísticos do ideb. *Scientific Electronic Library Online*, v. 34, p. 903–923, 2013. Disponível em: <https://www.scielo.br/j/es/a/JLzr4qdx89rjrNXnydNcvcy/>. Citado na página 13.

SOUZA, M. C. C. *Uma análise do algoritmo k-means como introdução ao aprendizado devmáquinas.* 2020. Monografia (Licenciatura em Matemática) - Universidade Federal do Tocantins. Disponível em: <https://repositorio.uft.edu.br/handle/11612/1764>. Acesso em: 13 out. 2023. Citado na página 19.

TEAM, R. C. *R: The Project for Statical Computing*. 2023. Disponível em: <https://www.r-project.org/>. Acesso em: 5 set. 2023. Citado na página 23.