Quaestum

DOI: https://doi.org/10.22167/2675-441X-2025838
Quaestum 2025; 6: e2675838
ISSN: 2675-441X

Artigo

Exploring the correlation between sanitation standards and school performance in Brazil: a clustering analysis

Explorando a correlação entre saneamento básico e performance escolar no Brasil: uma análise de cluster

Karoline Louize Macedo Granja[1]*; Manoel Brod Siqueira[2]

Recebido: ago. 13, 2024
Aceito: abr. 11, 2025

[1]Especialista em Data Science e Analytics. SQS 402, Bloco G, Apt. 201, Asa Sul, 70236-070, Distrito Federal, Brasília, Brasil

[2]Doutorando em Administração de Empresas na Fundação Getúlio Vargas. Quadra 2, Bloco L, Lote 06, 6° andar, Asa Norte, 70040-002, Distrito Federal, Brasília, Brasil

*Autor correspondente: karollouize@gmail.com

Abstract: Sanitation and education are two critical areas requiring immediate attention and improvement in Brazil. Students facing health issues have greater challenges in succeeding academically, as they may experience hospitalizations and difficulties concentrating due to illness. Although various factors can contribute to illness, this paper focuses on diseases resulting from inadequate sanitation conditions. The main objective of this research is to explore the correlation between sanitation standards and academic performance indices, considering sanitation, education, and health indices. In addition, the Brazilian states are clustered based on their indices. Pearson's correlation coefficient was applied to measure the continuous variables' statistical correlation. This research explores both positive and negative correlations. Clustering is an unsupervised machine learning method that groups data into clusters based on their similarities and patterns. The data collected in this study are sourced from prominent government databases, namely the Instituto Brasileiro de Geografia e Estatística (IBGE), the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), and the Sistema Único de Saúde (SUS), covering the period from 2007 to 2020, and all Brazilian states. Results indicate a clear positive correlation between inadequate sanitation standards and higher hospitalization rates, or low sanitation rates and poorer school performance indices. This influences the students' dispersion of age-grade and the non-attendance index, resulting in an adverse whirlpool effect on their academic performance. Thus, it is concluded that the initial hypothesis is confirmed, and that the government should raise its investments in sanitation to support the investments in education.

Keywords: Academic performance; Brazilian education; Brazilian states; correlation analysis; sanitation quality.

Resumo: Saneamento e educação são duas áreas críticas que requerem atenção e melhorias imediatas no Brasil. Estudantes que enfrentam problemas de saúde tem maiores desafios para obter bons resultados acadêmicos, pois podem passar por hospitalizações e dificuldades de concentração devido às doenças. Embora diversos fatores possam contribuir para o adoecimento, este artigo foca especialmente nas doenças resultantes de condições inadequadas de saneamento básico. O principal objetivo foi explorar a correlação entre os padrões de saneamento e os índices de desempenho acadêmico, a partir de indicadores de saneamento, educação e saúde. Para isto, foi aplicado o coeficiente de correlação de Pearson para medir a relação estatística entre as variáveis contínuas. Além disso, os estados brasileiros foram agrupados baseados nos índices mencionados por meio do método de "clustering" - metodologia de "machine learning" não supervisionada que organiza os dados em grupos com base em suas semelhanças e padrões. Os dados utilizados neste estudo foram extraídos de bancos de dados governamentais como o Instituto Brasileiro de Geografia e Estatística (IBGE), o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) e o Sistema Único de Saúde (SUS), referentes ao período de 2007 a 2020, e contemplam todos os estados brasileiros. Os resultados indicam uma clara correlação positiva entre padrões inadequados de saneamento e maiores taxas de hospitalização, bem como entre baixos índices de saneamento e piores indicadores de desempenho escolar. Esses fatores influenciam diretamente a defasagem idade-série e o índice de ausência dos alunos, o que afeita negativamente o desempenho acadêmico. Assim, conclui-se que os resultados confirmam a hipótese inicial de que o governo deve aumentar os investimentos em saneamento básico para dar suporte aos investimentos em educação.

Palavras-chave: Rendimento acadêmico; educação brasileira; estados brasileiros; análise de correlação; qualidade do saneamento.

1. Introduction

Numerous studies highlight the far-reaching economic and social benefits associated with education for both individuals and societies. For individuals, it often results in better health status, lower unemployment, better food habits, and greater engagement in civic and political life. This is one way to offer better opportunities to young people to gain qualifications and secure better jobs[1]. However, education alone is not enough to achieve prosperity. Investment in basic life quality, such as access to drinkable water and sanitation, is equally crucial[2].

The health-sanitation approach reflects the country's social development level[3]. Inadequate access to water supply and sanitation services correlates with poor health indicators and lower economic development. According to the Sistema Nacional de Informações sobre Saneamento (SNIS), 16% of Brazil's population lacks access to a water supply network, 46% are not connected to a sanitation network, and 22% of wastewater remains untreated[4]. Improving integrated sanitation management could expand sanitation network access to 76.5% nationwide[5]. Additionally, providing adequate drinking water facilities would not only enhance human health by reducing the number of hospitalizations for water-related diseases but also positively impact the entire population, bringing Brazil closer to the standards observed in developed countries[6].

The importance of health-sanitation cannot be overstated, particularly in its relation to education. Children exposed to Diseases Associated with Poor Environmental Sanitation (DRSAI) often face health challenges that bring difficulties for their learning. The lack of adequate water and sanitation infrastructure in disadvantaged communities creates vulnerability and fosters social inequality, allowing diseases associated with inadequate sanitation to proliferate, such as cholera, dysentery, typhoid, intestinal worm infections, and polio[7]. The World Health Organization (WHO) also emphasizes that inadequate sanitation can lead to antimicrobial resistance and stunted growth. As a result, affected children exposed to DRSAI miss classes and have difficulty keeping up with their studies[8].

In 2015, the United Nations Member States embraced the 2030 Agenda for Sustainable Development, a comprehensive framework that encompasses 17 Sustainable Development Goals (SDGs) and a call to action for all countries to work together. Recognizing access to water and sanitation as a human right, SDG 6 focuses on ensuring sustainable water and sanitation management for all. Moreover, SDG 6 is closely interlinked with SDG 4.2, which aims to provide quality early childhood development, care, and pre-primary education. By promoting water and sanitation alongside education, the SDGs strive for a more sustainable and inclusive future[9].

The main objective of this study is to explore the positive and negative correlation between sanitation standards and academic performance indices (during primary years – 1st to 5th grades), considering sanitation, education, and health indices. In addition, the Brazilian states are clustered based on their indices.

2. Material and Methods

This study is characterized as quantitative research, which enables the quantification of relationships among variables—specifically, the independent or predictor variable(s) and the dependent or outcome variable. Such an approach also facilitates the use of statistical analyses and the visualization of data summaries within the dataset. Quantitative research designs adopt objective, rigorous, and systematic strategies for generating and refining knowledge[10]. Quantitative research utilizes deductive reasoning and generalization. Deductive reasoning is when the researcher begins their study based on an existing or new theory/framework, and generalization is the process of extending research conclusions drawn from a smaller sample to the larger population based on evidence and results[11].

Reliable data sources are crucial for the integrity and validity of the research. They ensure that the data collected and analyzed are accurate, credible, and replicable, essential for making informed conclusions and recommendations. Therefore, it is crucial to carefully select and use reliable data sources in the materials and methods chapter of your research paper or report. The primary data sources used in this study were the Instituto Brasileiro de Geografia e Estatística (IBGE)[12], the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep)[13], and the Sistema Único de Saúde (SUS); besides articles and data searching platforms such as SciELO and PubMed. Since all data collected is publicly available, there was no need to proceed with interviews.

The data presented covers the period from 2007 to 2017 for IBGE data, and from 2010 to 2020 for Inep and SUS data. In the following section called indices the study clarifies the reason behind choosing different time range in the IBGE data. Brazilian public policies, in general, undergo a comprehensive review every 10 years to assess their implementation and evaluate their impact on the population. For this study, the data analysis was limited to the state level due to the large volume of observations and the lack of computational resources required to process data at the municipal level.

Datasets

The IBGE is a valuable source of information on various socioeconomic indicators, including education and sanitation. IBGE provides data on access to water supply and sanitation in people's houses, how many citizens are living locally, what kind of water supply the city has, and what type of waste treatment they provide. They periodically collect data by conducting interviews with the Brazilian population using a form designed to gather as much relevant information as possible. With this comprehensive approach to data collection, IBGE's data can provide a rich source of insights into the state of education and sanitation in Brazil. All social data, such as each state's extension and its respective population, came from IBGE's open data website. In addition, all data related to wastewater and water treatment came from the Basic Sanitation National Research[14].

The Inep open data website provided all school-related data. This institute has been collecting and releasing annual information on the infrastructure of all basic education institutions in the country through a survey answered by school principals, headteachers, or the designated responsible person since 2014[13]. By using this standardized approach to data collection, Inep's data provides a reliable and comprehensive source of information on school infrastructure in Brazil. From their data, available at the Inep website, it was possible to select some indices such as the academic progression index, non-attendance index, and dispersion age-grade index.

The SUS is Brazil's official health data system, providing information on disease prevalence, hospitalizations, and mortality rates. Researchers can use this database to investigate the relationship between poor sanitation and health outcomes in Brazil. By leveraging the wealth of data available through DataSUS[15], it is possible to identify patterns and trends in disease incidence and prevalence that may be linked to inadequate sanitation infrastructure. Overall, DataSUS serves as a crucial resource for understanding the state of public health in Brazil and the factors that contribute to poor health outcomes in the population.

The variables will be defined based on the previously collected data to start the analysis. Therefore, this paper aims to investigate the correlation between sanitation rates and school performance in Brazil by applying a correlation matrix and clustering.

Data Analysis Methodology

This study applied two methodologies to reach the objectives: correlation using Pearson's correlation coefficient and clustering. Pearson's correlation coefficient is a measurement that quantifies the strength of the association between two variables. Pearson's correlation coefficient r ranges from −1 to +1. Values of −1 or +1 indicate a perfect linear relationship between the two variables, whereas a value of 0 means no linear relationship. The negative values indicate the direction of the association, whereby as one variable increases, the other decreases.

The Pearson's correlation coefficient formula can be represented by the following Equation (1):

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (1)$$

where, r: represents the Pearson's correlation coefficient; x_i: are the individual values of one variable; y_i: are the individual values of the other variable; $\bar{x}$ and $\bar{y}$ are respectively the mean values of the two variables.

Although Pearson's correlation coefficient is a measure of the strength of an association (specifically the linear relationship), it is not a measure of the significance of the association. The significance of an association is a separate analysis of the sample correlation coefficient r using a t-test to measure the difference between the observed r and the expected r under the null hypothesis. Correlation analysis cannot be interpreted as establishing cause-and-effect relationships. It can only indicate how or to what extent variables are associated with each other. The correlation coefficient measures only the degree of linear association between two variables. Any conclusions about a cause-and-effect relationship must be based on the analyst's judgment[16].

Clustering can segment customers based on their behavior or preferences, group similar products together for marketing or inventory purposes, or identify anomalies or outliers in data. Clustering methods can also pre-process data for downstream supervised learning tasks such as classification or regression. The clustering process typically involves selecting an appropriate algorithm, defining a distance metric, and determining the optimal number of clusters to identify. Many clustering algorithms are available, each with its strengths and weaknesses. Standard algorithms include k-means, hierarchical clustering, and density-based clustering. Overall, unsupervised machine learning clustering methods provide a powerful tool for exploring and understanding large datasets without needing labeled data or prior knowledge of the underlying patterns or groupings[17].

The clustering method applied was k-means. The basic idea behind k-means clustering consists of defining clusters so that the total intra-cluster variation (known as total within-cluster variation) is minimized. There are several k-means algorithms available. The standard algorithm is the Hartigan-Wong algorithm (1979), which defines the total within-cluster variation as the sum of squared Euclidean distances between items and the corresponding centroid, as represented by Equation (2):

$$W(C_k)=\sum x_i \epsilon c_k \ (x_i - \mu_k)^2 \quad (2)$$

where, $W(C_k)$ is the Hartigan-Wong algorithm for the cluster C_k, x_i is a data point belonging to the cluster C_k, μ_k is the mean value of the points assigned to the cluster C_k.

Each observation (x_i) is assigned to a given cluster such that the sum of the squared distances of the observation to their assigned cluster centers (μ_k) is minimized[18].

All data was compiled into a spreadsheet, and the average calculation was done using Excel. This Microsoft program enables users to format, organize, and calculate data in a spreadsheet[19].

The compiled data spreadsheet ran in the code as its data source. The software chosen to build the code of this study was R Project version R 4.2.3 for Windows, which is mainly used for statistical computing and graphics. This software provides a wide variety of statistical and graphical techniques. One of R's strengths is how easily well-designed publication-quality plots can be produced; it is an open-source free software with over 10,000 available packages[20]. Some R packages were applied to facilitate correlation and clustering methods analysis, such as cluster, correlation, and corrplot.

Indices

Various indices compose the variables to correlate sanitation rate to school performance and try to capture different sanitation and school performance aspects.

Educational indices:

Non-attendance index: The institute has already calculated and provided this index. It indicates the percentage of students without performance data (approved or not) or movement data (death, dropout, relocation) in the second stage of primary years School Census. Only students who had a high number of absences presented these results. This index is directly related to the academic progression index[21]. This index captures the overall non-attendance rate of students at a school, which could be affected by its sanitation practices. For example, if a city has poor sanitation, students may be more likely to get sick and miss school classes.

Dispersion age-grade index: The institute has already calculated and provided this index. Since the Brazilian educational structure follows a grade system, there is a correspondence between the student's age and the grade the student should be in. Therefore, it calculates the percentage of students older than expected for the current grade. This metric refers to the number of students who are not studying in their appropriate grade due to poor academic performance stemming from high absenteeism and high dropout rates, particularly among students who are frequently sick from DRSAIs.

Academic progression index: The institute has already calculated and provided this index. It is collected at the primary years School Census' second stage in the student status column. At the end of the school year, enrolled students are measured based on their fulfillment of the pre-selected academic performance and attendance requirements. They may be classified as approved, failed, or removed due to abandonment. Abandonment happens when a student stops attending school before the end of the academic year without formally requesting a relocation. This index was selected to validate the hypothesis that students attending schools in areas with poor sanitation conditions do not experience satisfactory academic progress. It measures how many students have a good school performance compared to all students.

All data for these indices are available on the Inep open data website. You can access it by clicking on the Educational Indices section. The tables contain more detailed information, such as whether the school is located in a rural or urban area and whether it is a public or private school. To ensure the data's accuracy, certain filters were applied. The total filter was used to categorize data by localization and school types. Additionally, only the index numbers representing the initial years (1st grade to 5th grade) were selected, as is highlighted in Figure 1. This choice aligns with the government's responsibility to support students during these mandatory schooling years and is consistent with the goals of SDG 4. An average of the index numbers was calculated between 2010 and 2020 to get a number that represents the indexes for this time period.

Ministério da Educação
Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

Taxa de Distorção Idade-Série - Brasil, Regiões Geográficas e Unidades da Federação - 20

Taxa de Distorção Idade-Série por Localização e Dependência Administrativa, nos Níveis de Ensino Fundamental e Médio - Brasil, Regiões Geográficas e Unidades da Federação

Ano	Unidade Geográfica	Localização	Dependência Administrativa	Ensino Fundamental de 8 e 9 anos										
				Total	Anos Iniciais	Anos Finais	1º Ano	2º Ano	3º Ano	4º Ano	5º Ano	6º Ano	7º Ano	8º Ano	9º Ano
2019	Norte	Total	Total	24.2	17.6	33.0	4.4	6.7	19.1	25.2	28.7	34.4	34.5	31.6	30.7

Figure 1. Inep data source sample screenshot
Source: Inep[13]

Health indices:

Hospitalization index: This index represents the number of hospitalizations caused by several infectious and parasitic diseases primarily transmitted through contaminated water. To get this data, this study searched in the Internation Classification of Diseases (ICD) published by WHO which aims to classify all diseases by groups[22]. This transmission factor is directly correlated to sanitation and water supply. Data was collected from the SUS open data website, DataSUS, to obtain the numerator number. This was done by navigating to the Tabnet option, then selecting Epidemiological and Morbidity, and finally choosing the first option, General, by hospitalization location – since 2008. The chosen time frame was from 2010 to 2020, focusing on the ICD First Chapter from A00 to A09, which represents infectious diseases primarily transmitted through contaminated water.

The final equation for this index in each state can be represented by the following Equation (3):

$$HI_{total} = \frac{\sum_{i=2010}^{2020} hospitalizations_i}{10 \times AVG\ (population)} \quad (3)$$

where, $HI_{total:}$ is the Hospitalization index (HI); $\sum_{i=2010}^{2020}$: is the sum of the total number of hospitalizations for each year within the period from 2010 to 2020, $AVG_{population}$: is the state's average population during the same period. The Hospitalization index (HI) was calculated by summing the total number of hospitalizations for each year from 2010 to 2020 and dividing the result by ten to obtain the average annual hospitalizations. This average was then divided by the state's average population during the same period, as further detailed in the social indices section.

The chosen date range aligns with the date range of the education indices data source. This alignment was necessary because neither the health nor the education data sources cover the 2007–2017 period specified by the IBGE's National Research on Basic Sanitation.

Sanitation indices

Water treatment index: This metric quantifies the per capita consumption of treated water. It is important to note that not all cities solely rely on treated water for consumption. This measurement will be correlated with health and education indices to assess their interrelation.

Wastewater treatment index: Like the water treatment index, this metric measures the per capita volume of treated wastewater. While there may not be a direct correlation between this metric and school performance, it is essential to highlight that wastewater treatment plays a vital role in disease prevention and protecting water sources from contamination.

All sanitation information was sourced from IBGE's open data platform, SIDRA, through research under the acronym PNSB, which stands for National Research on Basic Sanitation. By clicking on the Water Supply category, it is possible to extract data from tables numbered 1773, 1361, and 7479. These tables provide insights into metrics such as the volume of treated water, the volume of treated wastewater, and the number of cities equipped with wastewater treatment infrastructure. Since this data comes from government research on sanitation, the PNSB data does not offer the flexibility to select specific time intervals due to its ten-year data collection cycle, with the latest release being in 2017. The volumes of water treatment and wastewater treatment were normalized by the population of each respective state to derive the corresponding percentages.

Social indices

Number of habitants: Discrete numeric variable representing the state's population. The average population between 2010 and 2020 was used to calculate the number. It was possible to get the values on IBGE's system called SIDRA, where it is simple to select how detailed the data will be (Brazil, Regions, States, Cities) and the range of years (2010-2020). The average population for each state was calculated by summing each year's population and dividing by ten. Since this data does not come from the PNSB research, selecting the same date range as the health and education indices was possible.

Territory size: Continuous numeric variable representing the size of the state's territory. The term territorial areas was searched on IBGE's website to access this data, and the Excel file was downloaded. No data wrangling was needed here.

It is essential to consider the data limitations in this research. Firstly, in the educational context, index numbers corresponding to the initial years were selected, ranging from 1st grade to 5th grade. This selection aligns with the government's responsibility to support students during these grades. An average of the index numbers for that time frame was calculated for those indices where no number represented the interval from 2010 to 2020.

Lastly, the data provided by the IBGE does not allow choosing specific time intervals when related to the PNSB data. This limitation arises from their ten-year data collection cycle, with the most recent data release in 2017, containing information from the last 10 years (2007-2017).

Using differing date ranges is a common challenge when working with Brazilian public data sources. The article Educação Infantil no Estado de São Paulo: Condições de Atendimento e Perfil das Crianças[23] exemplifies this issue, as it had to reconcile data from Inep and IBGE with varying temporal coverage. Despite these discrepancies, meaningful analysis is still possible, as this study aims to explore the correlation between sanitation standards and school performance in Brazil.

Notice that data from a time range was collected to calculate the average numbers for index calculation purposes. However, this study is not about a time series since all data was estimated to get the average of those years. The objective of this study is to understand the correlation between indices, not to understand their behavior through time.

3. Results and Discussion

The dataset variables are prepared to generate the correlation matrix and serve as inputs for data analysis. The study summary (Table 1) with each input and output allows for individual analysis of the clustering model distribution and its variable indices.

Table 1. Study table summary containing each Brazilian state and its respective indices, inputs, and cluster classifications

Name	Region	Non- attendance index (%)	Academic progression index (%)	Dispersion age-grade index (%)	Water treatment index (volume/ inhabitants)	Wastewater treatment index (volume/ inhabitants)	Hospitalization index (number/ inhabitants)	Cluster
Rondônia	North	4.18	92.25	14.99	0.08	0.005	0.0079	1
Acre	North	5.94	90.45	24.24	0.19	0.009	0.0056	2
Amazonas	North	4.07	89.70	21.52	0.18	0.008	0.0037	2
Roraima	North	3.98	93.78	15.46	0.33	0.064	0.0038	3
Pará	North	5.84	87.34	26.54	0.09	0.002	0.0087	1
Amapá	North	5.35	90.10	21.82	0.18	0.008	0.0029	2
Tocantins	North	3.61	94.05	11.83	0.20	0.026	0.0048	4
Maranhão	Northeast	4.41	92.30	17.21	0.11	0.005	0.0101	1
Piauí	Northeast	5.60	89.22	22.69	0.21	0.009	0.0092	1
Ceará	Northeast	3.46	95.30	12.38	0.12	0.023	0.0052	4
Rio Grande do Norte	Northeast	5.00	89.22	17.30	0.16	0.024	0.0050	2
Paraiba	Northeast	6.13	89.83	20.43	0.14	0.031	0.0057	1
Pernambuco	Northeast	4.53	91.25	18.20	0.15	0.022	0.0051	2

Name	Region	Non- attendance index (%)	Academic progression index (%)	Dispersion age-grade index (%)	Water treatment index (volume/ inhabitants)	Wastewater treatment index (volume/ inhabitants)	Hospitalization index (number/ inhabitants)	Cluster
Alagoas	Northeast	5.78	89.77	20.60	0.16	0.012	0.0048	2
Sergipe	Northeast	5.44	87.01	23.90	0.19	0.024	0.0026	2
Bahia	Northeast	7.18	87.90	23.96	0.15	0.046	0.0063	1
Minas Gerais	Southeast	2.79	97.45	6.56	0.20	0.052	0.0039	4
Espirito Santo	Southeast	3.43	93.87	13.23	0.21	0.052	0.0048	4
Rio de Janeiro	Southeast	6.41	90.87	19.15	0.33	0.046	0.0030	3
São Paulo	Southeast	1.55	97.56	4.45	0.26	0.097	0.0027	5
Paraná	South	1.63	95.19	6.84	0.19	0.086	0.0041	5
Santa Catarina	South	2.39	96.55	7.96	0.21	0.036	0.0035	4
Rio Grande do Sul	South	2.41	93.20	13.55	0.24	0.031	0.0045	4
Mato Grosso do Sul	Central-West	3.80	90.11	17.30	0.22	0.053	0.0044	4
Mato Grosso	Central-West	4.75	97.69	6.62	0.28	0.050	0.0046	4
Goiás	Central-West	4.54	95.35	10.92	0.15	0.053	0.0044	4
Distrito Federal	Central-West	3.12	94.58	10.20	0.19	0.109	0.0026	5

Source: Original results from the research

This study begins with Figure 2, which illustrates the interrelation between the indices and their respective scales.

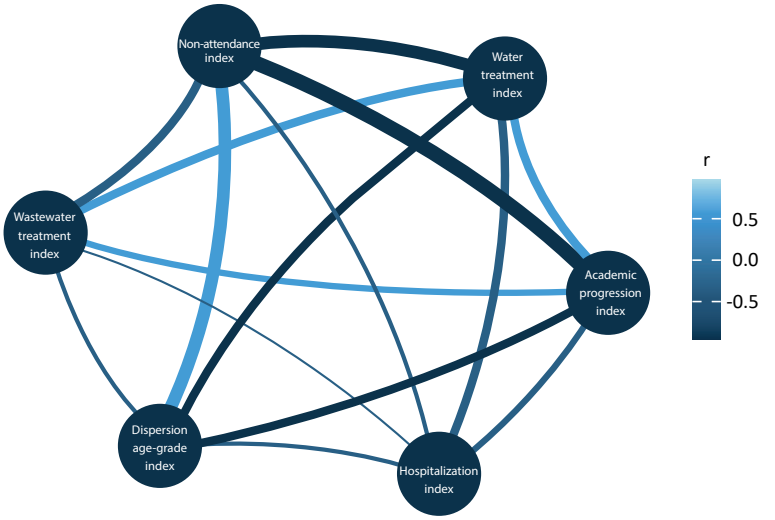

Figure 2. Diagram illustrating the interrelation between the indices and their scales
Source: Original results from the research
Note: The darker the blue color, the stronger the negative correlation between the indices. The thickness of the line connecting the two indices indicates the magnitude of the correlation, which can range from -1 to 1

An analysis of Figure 2 reveals that better academic performance is negatively correlated with a lower non-attendance rate, which is negatively correlated with the water treatment index. This last index has a positive correlation with the academic progression index. Therefore, states with higher water treatment rates will present better school performance, fewer school dropouts, and fewer hospitalizations caused by DRSAIs.

Plotting the correlation matrix using circles (Figure 3) allows a better understanding of the relationship between indices.

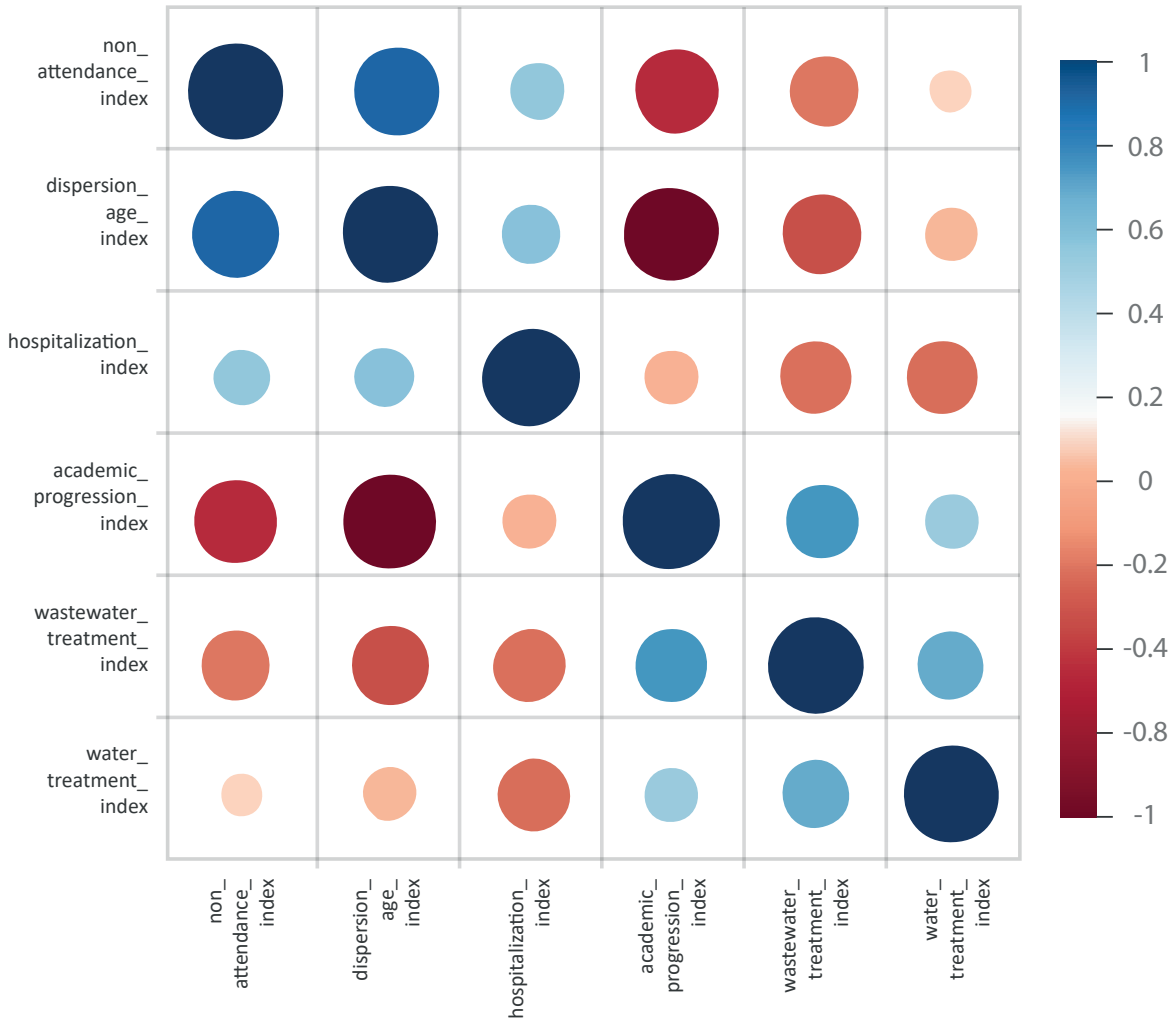

Figure 3. Correlation matrix using circles
Note: The blue circle represents a positive correlation. The red circle represents a negative correlation. The darker the color, the closer the correlation rate is to the extremes. The circle size represents Pearson's correlation magnitude, ranging from -1 to 1
Source: Original results from the research

Several positive correlations within the dataset can be identified. For example, there are positive correlations between the non-attendance index and the dispersion age-grade index, the non-attendance index and the hospitalization index, and the dispersion age-grade index and the hospitalization index. These findings suggest that when there is a higher rate of student hospitalizations, students are likely to miss classes and not progress according to their grade level.

In addition, the correlation matrix found another set of positive correlations among the water treatment index, wastewater treatment index, and academic progression index. These correlations indicate that students' academic performance increases as sanitation indices improve. With better sanitation facilities in place, students are less likely to miss classes or repeat grade levels, leading to improved academic outcomes.

Lastly, Figure 4 shows the correlation chart containing Pearson's correlation coefficients and their corresponding magnitude levels (*. **. or ***). Correlations with higher magnitude were observed between the hospitalization index and wastewater treatment index, hospitalization index and water treatment index, wastewater treatment index and water treatment index, wastewater treatment index and dispersion age-grade index, wastewater treatment index and academic progression index, wastewater treatment index and non-attendance index, dispersion age-grade index and academic progression index, dispersion age-grade index and non-attendance index, academic progression index and non-attendance index, academic progression index and water treatment index, and finally, the hospitalization index and dispersion age-grade index.

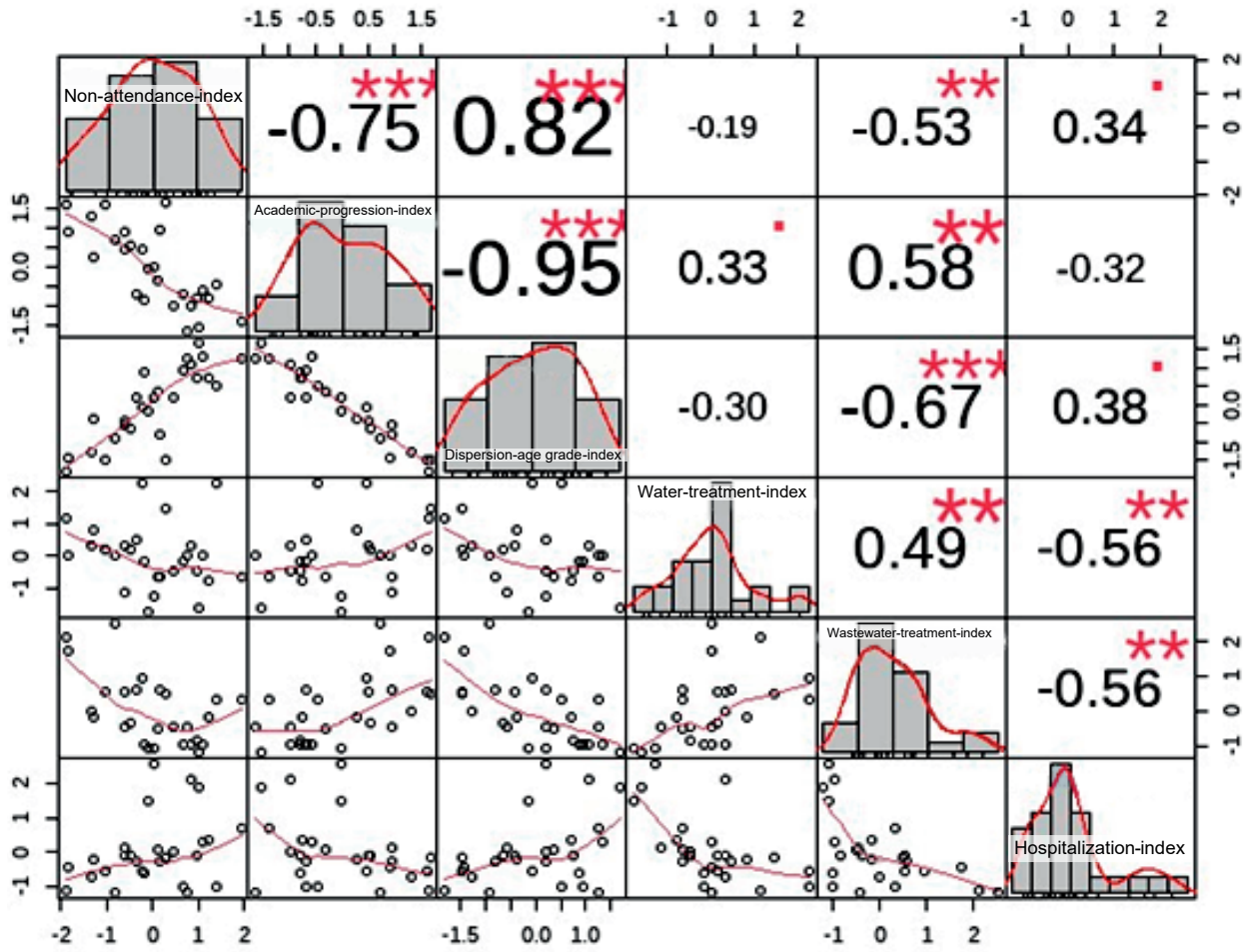

Figure 4. The correlation chart provides information on variable distributions, scatter plots, correlation values, and magnitudes

Note: The star-shaped mark (*) represents the corresponding p-value magnitude levels. *p < 0.05, **p < 0.01 and ***p < 0.001
Source: Original results from the research

Another objective of this paper is to compare different states of Brazil based on their sanitation rates and their impact on various education-related indicators, such as school absenteeism, dropout rates, academic progression, and sanitation standards. After analyzing the correlations, it is possible to go a step further and start clustering the Brazilian states into clusters considering their indices' similarities. Once the tree is constructed, data can be partitioned into any number of clusters by cutting the tree at the appropriate level. Three standard options for hierarchical clustering are single linkage, average linkage, and complete linkage. These options differ in their definition of the distance between two clusters. Single linkage defines the minimum distance over all pairs. Average linkage takes the average distance over all pairs, and complete linkage uses the maximum distance over all pairs[24]. After comparing the three hierarchical clustering options, notable differences emerge. In the case of a single linkage, the resulting clusters tend to exhibit an elongated shape, corresponding to a higher number of clusters and necessitating a larger dendrogram height (Figure 5A). Similarly, the average linkage method (Figure 5B) also yields elongated clusters. Consequently, the decision was made to employ the complete linkage approach (Figure 5C), as it generates clusters with a more compact shape, leading to a reduced number of clusters and smaller inter-cluster distances.

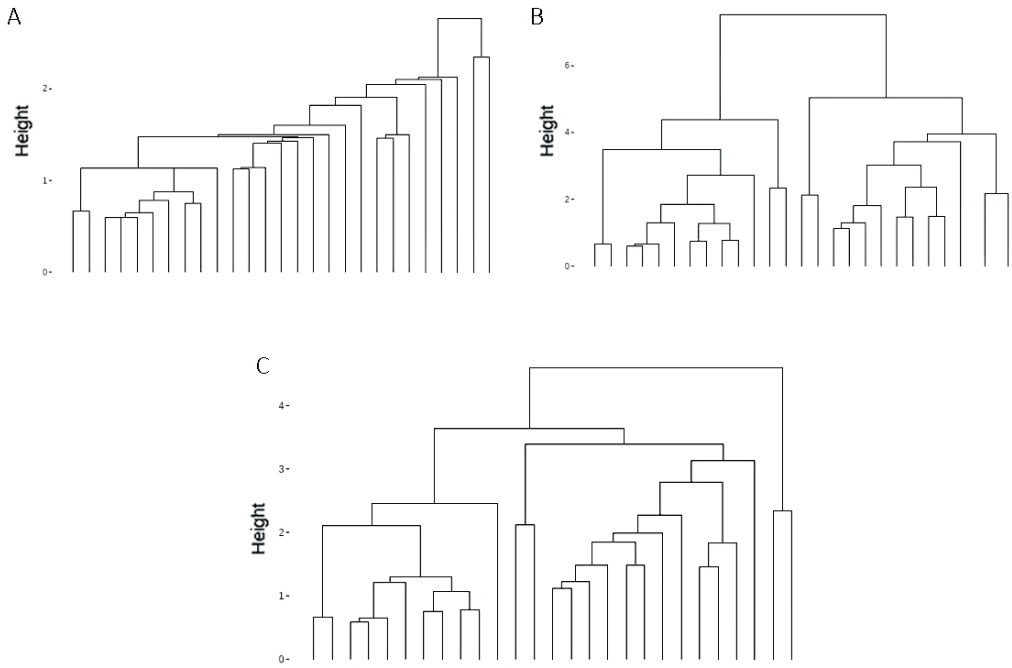

Figure 5. Cluster dendrogram applying A) single linkage, B) average linkage, C) complete linkage
Source: Original results from the research

Using the Elbow Method to determine the optimal number of clusters, as shown in Figure 6, the analysis indicated that selecting five clusters produced the most suitable outcome. This was based on examining the line's inclination and the extent of its variation along the y-axis. Notably, Figure 6 provides additional insights into the cluster selection process. It shows a significant variation in the total within the sum of squares before reaching the threshold of five clusters. However, when the number of clusters was six or more, the variation along the y-axis became smaller, indicating less noticeable differences. Therefore, the five clusters option was chosen as it provided a good balance between capturing variation and avoiding an excessive number of clusters.

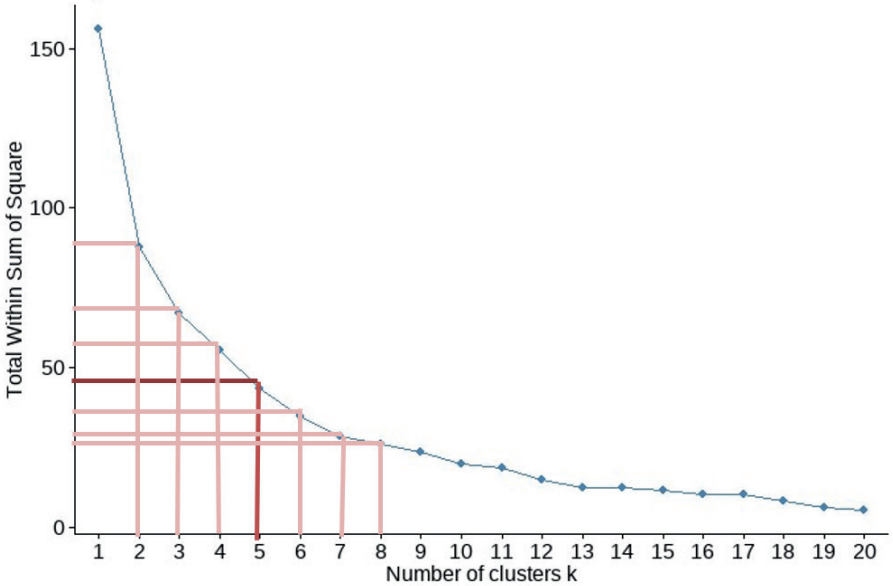

Figure 6. Plot of the elbow method applied to choose the optimal number of clusters
Source: Original results from the research

Plotting a new cluster dendrogram that combines the information from hierarchical clustering and the number of clusters makes it easier to identify the clusters and determine the number of states within each cluster. This visualization provides a more comprehensive understanding of the clustering results and facilitates the interpretation of the data.

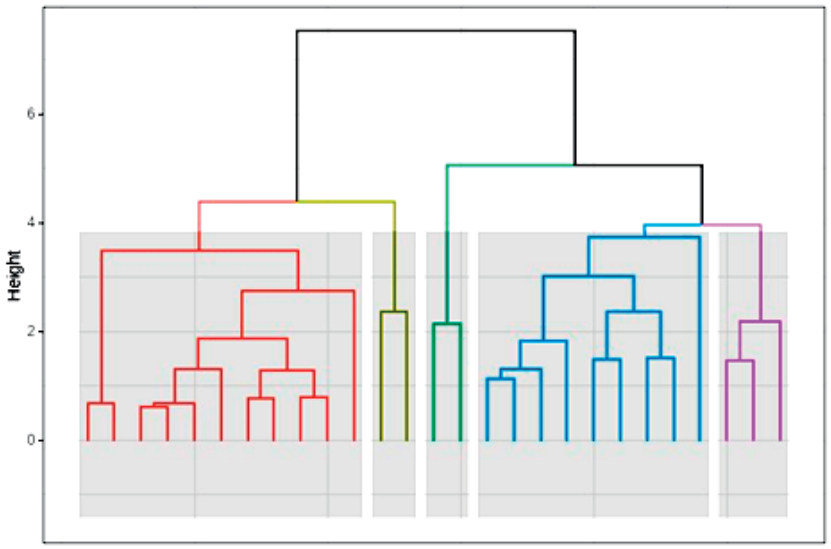

Figure 7. Cluster dendrogram applying complete linkage and five clusters
Source: Original results from the research

An Analysis of Variance (ANOVA) was conducted upon further investigation of the clusters. ANOVA is a statistical method used to compare the means of multiple groups and determine if significant differences exist between them. Table 2 illustrates the small Mean Sq values for Hierarchical cluster, indicating minimal distances between distinct clusters. Additionally, even smaller Mean Sq values for residuals confirm that the distances between observations within each cluster are also minimal. The distances between elements within the clusters are smaller compared to the distances between the clusters themselves. This indicates that the clustering division is meaningful, and all elements are positioned near others exhibiting high similarity. Furthermore, the obtained p-values validate that all the selected indices significantly contribute to forming at least one cluster.

Table 2. ANOVA results applied to each index

Item	Hierarchical cluster		Residuals
	Mean Sq[1]	p-value[2]	Mean Sq[1]
Non-attendance index	4.264	6.56E-05	0.407
Academic progression index	4.471	2.34E-05	0.369
Dispersion age-grade index	5.006	8.95E-07	0.272
Water treatment index	3.972	2.38E-04	0.460
Wastewater treatment index	5.462	1.76E-08	0.189
Hospitalization index	4.842	2.73E-06	0.301

Source: Original results from the research
Note: [1]Mean Sq: Mean square values are variance estimates. These values are used in ANOVA and Regression analyses to determine whether model terms are significant; [2]p-value: The p-value is the probability of observing the given value of the test statistic, or a more extreme one, under the null hypothesis

Additionally, Table 3 informs each cluster's average indices, providing some more inputs to observe. Cluster 1 comprises states with the highest non-attendance index, the second smallest academic progression index, the second highest dispersion age-grade index, the smallest water treatment index, the second smallest wastewater treatment index, and the highest hospitalization index. On the other hand, cluster 5 represents the states with the smallest non-attendance index, the highest academic progression index, the smallest dispersion age-grade index, the highest water treatment index, the highest wastewater treatment (almost twice the average of the second-best cluster 3 with 0.055), and the lowest hospitalization index.

Table 3. Cluster x Indices average summary

Item	Cluster				
	1	2	3	4	5
Non-attendance index (%)	5.413	4.955	5.195	3.464	2.100
Academic progression index (%)	89.936	88.520	92.325	94.841	95.777
Dispersion age-grade index (%)	20.485	24.030	17.305	11.150	7.163
Water treatment index	0.156	0.135	0.330	0.203	0.213
Wastewater treatment index	0.018	0.005	0.055	0.042	0.097
Hospitalization index	1.777	1.403	1.337	0.337	0.151

Source: Original results from the research

Lastly, Table 4 draws a comparison between the conclusive clustering and Brazil's factual regions. Clusters 1 and 2 are composed only of states from the North and Northeast, pointing to the urgent need for investment in sanitation. Despite their geographical separation, cluster 3 unites Rio de Janeiro and Roraima, revealing parallel sanitation and school performance patterns. Among the clusters, the larger and more diverse cluster 4 encompasses nine states from all corners of Brazil. Notably, cluster 5, characterized by the most favorable index values, comprises Brasilia, Paraná, and São Paulo, none located in the North or Northeast regions.

Table 4. Clustering x Brazil regions

State	Region	Cluster
Rondônia	North	1
Pará	North	1
Maranhão	Northeast	1
Piauí	Northeast	1
Paraiba	Northeast	1
Bahia	Northeast	1
Acre	North	2
Amazonas	North	2
Amapá	North	2
Rio Grande do Norte	Northeast	2
Pernambuco	Northeast	2
Alagoas	Northeast	2
Sergipe	Northeast	2
Roraima	North	3
Rio de Janeiro	Southeast	3
Tocantins	North	4
Ceará	Northeast	4
Minas Gerais	Southeast	4
Espirito Santo	Southeast	4
Santa Catarina	South	4
Rio Grande do Sul	South	4
Mato Grosso do Sul	Central-West	4
Mato Grosso	Central-West	4
Goiás	Central-West	4
Sao Paulo	Southeast	5
Paraná	South	5
Distrito Federal	Central-West	5

Source: Original results from the research

There is relatively limited research combining sanitation and education, especially in Brazil. It is possible to mention Evaluation of the Efficiency of Basic Sanitation Integrated Management in Brazilian Municipalities[5], Investment in drinking water and sanitation infrastructure and its impact on waterborne diseases dissemination: The Brazilian case[6], Water and sanitation in schools: a systematic review of the health and educational outcomes[25], Water, sanitation and hygiene interventions for acute childhood diarrhea: a systematic review to provide estimates for the Lives Saved Tool[26], Water, sanitation and hygiene (WASH) in schools in Brazil pre-and peri-Covid-19 pandemic: Are schools making any progress?[27], Methodology for a Comprehensive Health Impact Assessment in Water Supply and Sanitation Programmes for Brazil[28], and Unequal geographic distribution of water and sanitation at the household and school level in Sudan[29]. The articles affirm the urgency and need to invest in an efficient wastewater treatment system and recognize the non-health sector's responsibilities in improving children's health. The articles not only corroborated the hypothesis of this study but also quantified how investing in sanitation could lead to improvements in school performance indices.

To reaffirm the significance of this hypothesis and explore the correlation between sanitation and education, a clustering method was applied based on six variables (three related to sanitation and three related to education) collected from government open data sources. Analyzing the results over a ten-year period (2010-2020), including the characteristics of each cluster and the average indices for each cluster, made it possible to get results that overwhelmingly supported the theories presented in the literature.

A clear negative correlation emerged between high hospitalization rates or poor sanitation standards and school performance indices. Educational indicators such as the non-attendance index or age-grade dispersion positively correlate with health indicators, adversely affecting school performance. Furthermore, states with lower school performance indices tended to cluster together and exhibited higher hospitalization rates or inferior water/wastewater treatment. Conversely, states with better sanitation conditions were grouped with similar states that also demonstrated better school performance indices.

The clustering results match with data from IBGE, which indicates that 99.6% of Brazilian cities, the equivalent of 84% of its population, receive water treatment. However, of the 22 cities that don't receive a clean water supply, 20 are located in Brazil's North and Northeast regions–primarily in the clusters with the worst indices (Clusters 1 and 2). Regarding wastewater treatment, only 60.3% of the Brazilian cities, or 64% of the population, have this service. While the Southeast region covers 96.5% of its cities, the North region only covers 16.2%[30].

These results serve not only to confirm the hypotheses of the correlation between sanitation standards and academic performance, but also the clustering results corroborate how sanitation influences education by grouping together similar indices of states. All indices chosen are statistically relevant for the model, and it reaffirms the literature hypothesis and results. Future studies should explore these correlations using Principal Component Analysis (PCA) and Linear Regression tools. These methods can provide deeper insights into how the data's different parts are connected.

4. Conclusion

This study's findings confirm a positive correlation between sanitization standards and school performance indices (during primary years – 1st to 5th grades) in Brazil. Conversely, a positive correlation is evident between lower sanitization standards and higher levels of hospitalization, increased non-attendance rates, and greater disparity in age-grade alignment.

The clustering analysis reveals compelling insights. States with superior academic rates and commendable sanitization standards tend to cluster together, suggesting a positive relationship between these factors. On the other hand, states characterized by poor sanitization standards, elevated hospitalization levels, and inferior school performance form a distinct cluster, highlighting the detrimental impact of inadequate sanitization on educational outcomes. When comparing the cluster results with the actual regional distribution, it becomes evident that the North and Northeast regions would benefit from increased attention and investment in improving their sanitation conditions. None of the states from these areas are categorized within cluster 5, the most favorable group, while a significant majority of them fall under clusters 1 and 2. Future studies should focus on applying Linear Regression to understand the interaction between variables and try to predict their behavior based on the dataset.

Author contributions: All authors contributed to Conceptualization; Data Acquisition; Methodology; Data Analysis; Investigation; Writing and Editing.

How to cite: Granja, K.L.M.; Siqueira, M.B. 2025. Exploring the correlation between sanitation standards and school performance in Brazil: a clustering analysis. Quaestum 6: e2675838.

References

[1] Murray, J. 2009. The Wider Social Benefits of Higher Education: What do We Know about Them? Australian Journal of Education 53(3): 230-244. https://doi.org/10.1177/0004944109053003.
[2] Lonescu, D.D.; Ionescu, A.M.; Jaba. E. 2013. The investments in education and quality of life. Journal of Knowledge Management. Economics and Information Technology 3(6): 12. Available at: https://econpapers.repec.org/article/sppjkmeit/spi13-12.htm. Accessed: Mar. 1, 2023.
[3] Castro, J.E.; Heller, L. 2009. Interfaces and inter-sector approaches: water, sanitation and public health. In: Castro, J.E.; Heller, L. Water and Sanitation Services. Routledge. Oxfordshire. UK. https://doi.org/10.4324/9781849773751.
[4] Pereira, M.A.; Marques, R.C. 2022. Technical and Scale Efficiency of the Brazilian Municipalities' Water and Sanitation Services: A Two-Stage Data Envelopment Analysis. Sustainability 14(1): 199. https://doi.org/10.3390/su14010199.
[5] Cavalcanti, A.; Teixeira, A.; Pontes, K. 2020. Evaluation of the Efficiency of Basic Sanitation Integrated Management in Brazilian Municipalities. International Journal of Environmental Research and Public Health 17(24): 9244. https://doi.org/10.3390/ijerph17249244.
[6] Ferreira, D.C.; Graziele, I.; Marques, R.C.; Gonçalves, J. 2021. Investment in drinking water and sanitation infrastructure and its impact on waterborne diseases dissemination: The Brazilian case. Science of the Total Environment 779: 146279. https://doi.org/10.1016/j.scitotenv.2021.146279.
[7] Charlesworth, S.M.; Kligerman, D.C.; Blackett, M.; Warwick, F. 2022. The Potential to Address Disease Vectors in Favelas in Brazil Using Sustainable Drainage Systems: Zika, Drainage and Greywater Management. International Journal of Environmental Research and Public Health 19(5): 2860. https://doi.org/10.3390/ijerph19052860.
[8] UNICEF; WHO. State of the World's Sanitation: An Urgent Call to Transform Sanitation for Better Health. Environments. Economies and Societies. New York: United Nations Children's Fund [UNICEF] and the World Health Organization. 2020. Available at: https://www.who.int/publications/i/item/9789240014473. Accessed: Mar. 5, 2023.
[9] United Nations. Sustainable Development Goals. 2015. Available at: https://sdgs.un.org/goals. Accessed: Feb. 23, 2023.
[10] Sousa, V.D.; Driessnack, M.; Mendes, I.A.C. 2007. An overview of research designs relevant to nursing: Part 1: quantitative research designs. Revista Latino-Americana de Enfermagem 15(3): 502-507. https://doi.org/10.1590/S0104-11692007000300022.
[11] Lakatos, E.M.; Marconi, M.D.A. 2003. Fundamentos de metodologia científica. 5ed. Atlas, São Paulo, SP, Brasil. Available at: https://ria.ufrn.br/jspui/handle/123456789/3097. Accessed: Jun. 2, 2025.
[12] Instituto Brasileiro de Geografia e Estatística [IBGE]. Dados Abertos. Available at: https://www.ibge.gov.br/acesso-informacao/dados-abertos.html. Accessed: Mar. 1, 2023.
[13] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [Inep]. Taxa de distorção idade-série. Available at: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/taxas-de-distorcao-idade-serie. Accessed: Mar. 1, 2023.
[14] Pesquisa Nacional de Saneamento Básico. Sistema IBGE de Recuperação Automática [SIDRA]. Available at: https://sidra.ibge.gov.br/pesquisa/pnsb/pnsb-2017. Accessed: Mar. 1, 2023.
[15] Ministério da Saúde DataSUS. Dados Abertos. Available at: https://datasus.saude.gov.br/acesso-a-informacao/. Accessed: Mar. 1, 2023.
[16] Benesty, J.; Chen. J.; Huang. Y.; Cohen. I. 2009. Pearson correlation coefficient. In: Noise reduction in speech processing. Springer Topics in Signal Processing, vol 2. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5.
[17] Montgomery, D.C.; Peck, E.A.; Vining, G.G. 2021. Introduction to linear regression analysis. John Wiley & Sons, USA.
[18] Bock, H.H. 2007. Clustering Methods: a history of k-means algorithms. In: Brito, P.; Cucumel, G.; Bertrand, P.; Carvalho, F. Selected contributions in data analysis and classification. Studies in classification, data analysis, and knowledge organization. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73560-1_15.
[19] Microsoft Support. Available at: https://support.microsoft.com/en-us/office/what-is-excel-94b00f50-5896-479c-b0c5-ff74603b35a3. Accessed: Jan. 12, 2025.
[20] R Project "about R". Available at: https://www.r-project.org/about.html. Accessed: Sept. 5, 2023.
[21] Governo Federal. Portal de Dados Abertos. Available at: https://dados.gov.br/dados/conjuntos-dados/. Accessed: Jan. 12, 2025.
[22] International Classification of Diseases [ICD] 11th Revision available at: https://icd.who.int/pt/. Published in May 2019.
[23] Fernandes, F.S.; Domingues, J.D.R. 2017. Educação infantil no estado de São Paulo: condições de atendimento e perfil das crianças. Educação e Pesquisa 43(1): 145-160. https://doi.org/10.1590/S1517-9702201701155227.
[24] Heyer, L.J.; Kruglyak, S.; Yooseph. S. 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Research 9(11): 1106-1115. https://doi.org/10.1101/gr.9.11.1106.
[25] Jasper, C.; Le, T.T.; Bartram, J. 2012. Water and Sanitation in Schools: A Systematic Review of the Health and Educational Outcomes. International Journal of Environmental Research and Public Health 9(8): 2772-2787. https://doi.org/10.3390/ijerph9082772.
[26] Darvesh, N.; Das, J.K.; Vaivada, T.; Gaffey, M.F.; Rasanathan; K.; Bhutta, Z.A.; Social Determinants of Health Study. 2017. Water, sanitation and hygiene interventions for acute childhood diarrhea: a systematic review to provide estimates for the Lives Saved Tool. BMC Public Health 17(4): 776. https://doi.org/10.1186/s12889-017-4746-1.
[27] Poague, K.I.; Blanford, J.I.; Martínez, J.A.; Anthonj, C. 2023. Water, sanitation and hygiene (WASH) in schools in Brazil pre-and peri-COVID-19 pandemic: Are schools making any progress?. International Journal of Hygiene and Environmental Health 247: 114069. https://doi.org/10.1016/j.ijheh.2022.114069.
[28] Klingerman, D.C.; Cardoso, T.A.O.; Cohen, S.C.; Azevedo, D.C.B.; Toledo, G.d.A.; Azevedo, A.P.C.B.; Charlesworth, S.M. 2022. Methodology for a Comprehensive Health Impact Assessment in Water Supply and Sanitation Programmes for Brazil. International Journal of Environmental Research and Public Health (19): 12776. https://doi.org/10.3390/ijerph191912776.
[29] Cha, S.; Jin, Y.; Elhag, M.S.; Kim, Y.; Ismail, H.A.H.A. 2021. Unequal geographic distribution of water and sanitation at the household and school level in Sudan. PLoS One 16(10): e0258418. https://doi.org/10.1371/journal.pone.0258418.
[30] Agência de Notícias IBGE. Available at: https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/28324-pnsb-2017-abastecimento-de-agua-atinge-99-6-dos-municipios-mas-esgoto-chega-a-apenas-60-3. Accessed: May 14, 2025.