# SDG 3 AND BASIC SANITATION: EMPIRICAL EVIDENCE ON THE LINK BETWEEN INFRASTRUCTURE AND PREVENTABLE ILLNESSES

Eduardo Ogawa Cardoso - USP - Universidade de São Paulo

Victória Ribeiro Da Silva - FEA USP

Daielly Melina Nassif Mantovani - FEA USP

**Resumo**

Este estudo teve como objetivo investigar a relação entre fatores socioeconômicos e a prevalência de doenças associadas ao saneamento inadequado. A análise de clusters identificou três agrupamentos distintos, cada um caracterizado por perfis socioeconômicos e de saúde específicos. Notavelmente, o Cluster 2, caracterizado por maiores níveis de desenvolvimento e acesso a recursos, apresentou níveis significativamente mais elevados de doenças associadas ao saneamento inadequado em comparação aos Clusters 0 e 1. Além disso, o estudo baseou-se em dados disponíveis, que podem não refletir toda a gama de fatores relevantes. Os resultados destacam a necessidade de intervenções direcionadas para enfrentar os desafios específicos de cada cluster. O Cluster 2, apesar de seu maior nível de desenvolvimento, requer atenção especial para lidar com os fatores subjacentes que contribuem para a alta prevalência de doenças. Este estudo oferece insights valiosos sobre a heterogeneidade espacial e socioeconômica das doenças associadas ao saneamento inadequado. O uso da análise de clusters representa uma abordagem inovadora para identificar grupos distintos dentro da população e orientar intervenções de forma mais eficaz.

**Palavras-chave:**ODS3. Saneamento básico. Analytics

**Abstract**

This study aimed to investigate the relationship between socioeconomic factors and the prevalence of illness associated with poor sanitation. Cluster analysis identified three distinct clusters, each characterized by unique socioeconomic and health profiles. Notably, Cluster 2, characterized by higher levels of development and access to resources, exhibited significantly higher levels of illness associated with poor sanitation compared to Clusters 0 and 1.The linear regression models were constrained by limitations such as autocorrelation in residuals and potential violations of the normality assumption. Additionally, the study relied on available data, which might not capture the full spectrum of relevant factors. The findings highlight the need for targeted interventions to address the unique challenges faced by each cluster. Cluster 2, despite its higher development level, requires special attention to address the underlying factors contributing to the high prevalence of illness. The use of cluster analysis offers a novel approach to identifying distinct groups within the population and tailoring interventions accordingly.

**Keywords:** SGD3. Sanitation. Analytics

# SDG 3 AND BASIC SANITATION: EMPIRICAL EVIDENCE ON THE LINK BETWEEN INFRASTRUCTURE AND PREVENTABLE ILLNESSES

## INTRODUCTION

Currently, basic sanitation structure is so common that many people are unaware of its importance for the quality of life of communities. In addition to the supply of drinking water and sewage, the term also includes urban cleaning, solid waste management, drainage and management of urban stormwater, cleaning and preventive inspection of the respective urban networks (WHO, 2015). The development of sanitation is slow and gradual throughout history, being driven by the advancement of technologies and public policies highlighting the importance of sanitary health, the protection of drinking water against contamination and the expansion of preventive actions (HELLER et al., 2018).

Even today, around 1 billion people do not have access to sanitation and clean water, causing the death of around 6,000 children daily (SOUZA, 2009). The United Nations (UN) projects that the world population will continue to grow in the coming decades, reaching 8.3 billion in 2030 and 8.9 billion in 2050 (UN, 2017). This scenario makes the current model of economic development unsustainable, whose tendency is to aggravate its consequences over time. The right to health is one of the fundamental rights of every human being, regardless of economic or social conditions (WHO, 1948). In Brazil, basic sanitation policies are under increasing development (GOMES; OLIVEIRA, 2017), where health is a right provided, protected and formalized in the Federal Constitution of 1988, in articles 196 to 200 (BRASIL, 1988).

In 2018, the coverage of the sewage network in the country increased from 42.6% to 53.2%, which means that more than 100 million Brazilians still did not have access to this service. In parallel, with regard to drinking water, around 30 million Brazilians did not benefit from this service (BRASIL, 2018). To mitigate these impacts and build a global strategy, the UN, with the adhesion of 193 countries, launched the 2030 Agenda in 2015. This milestone serves as a strategic guidance for nations to develop public policies aimed at solving the global challenges faced by society (STF, 2020). It is a "global call to action to end poverty, protect the environment and ensure that people everywhere can enjoy peace and prosperity" (UN, 2015). To this end, 17 Sustainable Development Goals (SDGs) were structured, subdivided into indicators and targets, integrating three dimensions: economic, social and environmental.

In 2020 (Law No. 14,026 of July 15, 2020, updating Law No. 11,445/2007), Brazil approved the Legal Framework for Sanitation, aiming to universalize access to basic sanitation services. The framework sets targets to be achieved by 2033: 99% coverage for treated water supply and 90% coverage for sewage collection and treatment. To this end, the Legal Framework introduces important changes compared to the previous legislation. It regionalizes the management of services, placing municipalities in a leading role; allows private companies to operate in a sector that was previously limited to public organizations (CAPOBIANCO et al., 2023); and makes the National Water and Basic Sanitation Agency responsible for regulating the sector (BRASIL, 2020).

The present study is aligned with SDG 3, which aims to "ensure healthy lives and promote well-being for all at all ages". The objective is to investigate how the coverage of health services and sanitation infrastructure influence the rate of diseases related to inadequate sanitation in Brazilian municipalities and regions. To this end, we raised the following research question:

RQ: Which indicators better contribute to explain illness associated with poor sanitation?

## SANITATION AND HEALTH

The study conducted by Instituto Trata Brasil, covering data from 2008 to 2024, highlights how the lack of essential services impacts public health, especially among the most vulnerable populations. The report identifies diseases such as diarrhea, viral hepatitis, leptospirosis, schistosomiasis, and arboviruses (such as dengue and chikungunya) as directly associated with the absence of basic sanitation. During the analyzed period, there was a trend of reduction in hospitalizations due to Waterborne Diseases Related to Inadequate Environmental Sanitation (DRSAI). However, high rates still persist in regions with lower sanitation coverage, particularly in the North and Northeast regions. Infant mortality also shows a significant correlation with the lack of adequate sanitation infrastructure. The study highlights marked regional disparities in access to sanitation services, which are reflected in differences in disease and mortality rates. The lack of adequate sanitation imposes significant costs on the healthcare system due to increased hospitalizations and treatments for preventable diseases (INSTITUTO TRATA BRASIL, 2025).

Between 2001 and 2009, Brazil recorded an average of 13,449 annual deaths from diseases related to inadequate sanitation, representing approximately 1.31% of all deaths. Furthermore, there was an annual average of 466,351 reported disease cases and 758,750 hospitalizations attributed to these deficiencies, resulting in significant expenses for the healthcare system (TEIXEIRA et al., 2014). In the same vein, the article by Uhr et al. (2016) estimated econometric models aimed at assessing the influence of basic sanitation services on the health of the Brazilian population, considering hospitalizations due to diseases transmitted by contaminated water and inadequate sewage treatment, during the period from 2000 to 2011. The results indicate that the greater the coverage of sanitation services, particularly the sewage collection network, the greater the reduction in morbidity. This suggests that improvements in basic sanitation infrastructure have a positive impact on public health by reducing hospitalizations for waterborne diseases (UHR, 2016).

## METHOD

To initiate the study, a search was conducted in various databases in order to select the indicators related to the objective of this work. The relationship between the indicators and their specificities is presented in Table 1 below, considering data from the year 2020, covering 5,570 Brazilian municipalities, 20 variables, and 9 databases.

Table 1. Variable in the dataset

| Indicator | Code | Source | Meas. Unit |
|---|---|---|---|
| Orçamento municipal para a | budget_city | DataSUS | Reais per capita |

saúde

| | | | |
|---|---|---|---|
| População atendida por equipes de saúde da família | health_coverage | DataSUS | % |
| Unidades Básicas de Saúde | ubs_per_1000 | DataSUS | mil habitantes |
| Perda de água | water_loss | SNIS | IN |
| População atendida com serviço de água | water_coverage | SNIS | % |
| População atendida com esgotamento sanitário | sewer_coverage | SNIS | % |
| Índice de tratamento de esgoto | sewer_treat_index | SNIS | % |
| PIB per capita | pib_per_capita | IBGE PIB dos municipios | R$ per capita |
| Investimento público em infraestrutura por habitante | public_invest_per_capita | SICONFI e IBGE PIB MUNICIPAL | % |
| Resíduos domiciliares per capita | res_per_capita | SNIS | Ton / Hab / Ano |
| Esgoto tratado antes de chegar ao mar, rios e córregos | sewer_treat_perc | Atlas Esgotos - SNIRH/ANA | % |
| Grau de maturidade de proteção ambiental | enviroment_maturity | IBGE (Munic) | % |
| Grau de estruturação de combate à corrupção | corruption_maturity | IBGE (Munic) | % |
| Grau de estruturação das políticas de transparência | transparency_maturity | IBGE (Munic) | % |
| Investimento público | invest_per_capita | SICONFI | R$ per capita |
| Total de receitas arrecadadas | total_revenue_per | SICONFI | % |
| Taxa de Leitos Sus por 100.000 habitantes | bed_index_100k | IEPS Saúde | Taxa |
| Taxa de médicos por 1000 habitantes | physician_index_100k | IEPS Saúde | Taxa |
| Taxa de Enfermeiros por 1000 habitantes | nurse_index_100k | IEPS Saúde | Taxa |
| Doenças relacionadas ao saneamento inadequado | target | DataSUS/SIH | 100 mil habitantes |

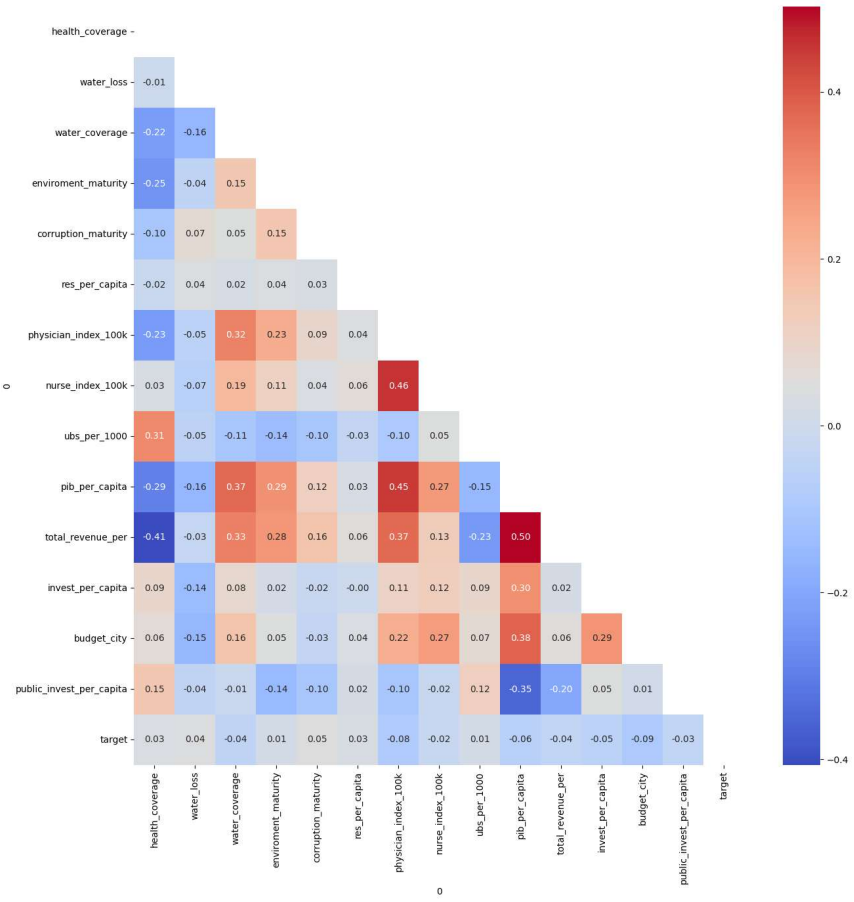
Exploratory Data Analysis (EDA) was conducted to gain initial insights into the data and identify potential patterns and relationships. This process involved a combination of visual and statistical methods. Descriptive statistics, including measures of central tendency (mean, median, mode) and dispersion (standard deviation, range), were calculated for all variables.

Initially missing data were screened within the dataset. The msno library was utilized to visualize the distribution and patterns of missing values, providing insights into the extent of data incompleteness. Subsequently, the K-Nearest Neighbors (KNN) imputation algorithm was employed to replace missing values with estimates derived from the 'k' nearest neighbors based on feature similarity. Hyperparameter tuning was performed to optimize the KNN imputation process. Finally, the msno library was used again to visualize the data distribution after imputation, allowing for an assessment of the imputation's effectiveness and identification of any potential artifacts introduced during the process. The imputed dataset was utilized for subsequent analyses.

A second step involved a log transformation of the data to address potential issues with skewness or non-normality. Log transformation involves applying the logarithm function (typically base 10 or natural log) to each data point. This technique is often effective in reducing skewness and stabilizing the variance, which can improve the performance of subsequent statistical analyses or machine learning models. The choice of log base depends on the specific context and the desired interpretation of the transformed data. In our context it helped gain a better understanding of the variables.

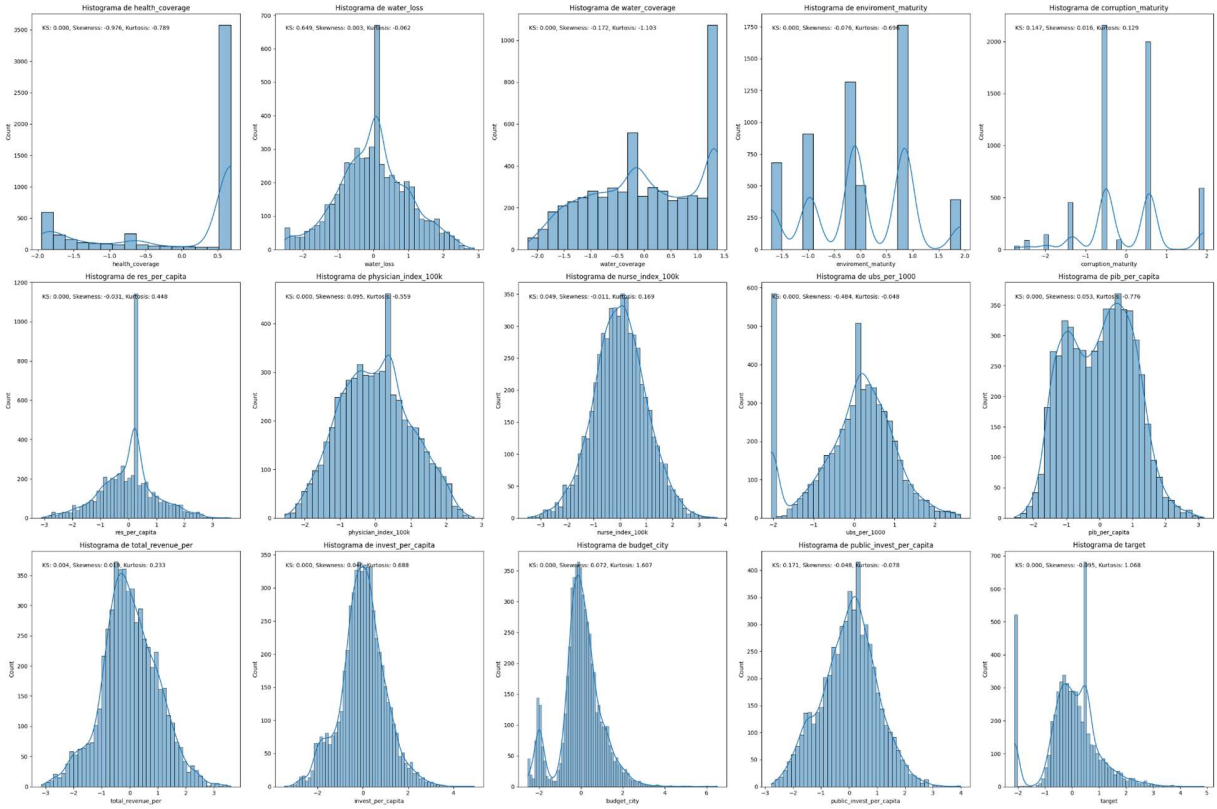
Furthermore, a correlation plot was generated to assess the presence of multicollinearity among the features. By focusing on the lower triangular part of the correlation matrix (Figure 1), redundant information and potential issues arising from highly correlated features were effectively identified and addressed. This step is crucial for building robust and interpretable models, as multicollinearity can hinder model performance and make it difficult to understand the individual contributions of features. In this particular case, none of the variables presents any particularly significant high correlation, allowing us to exclude the possibility of multicollinearity.

Figure 1. Correlation matrix of the predictors




Normality assessment employed histogram plots and Q-Q plots to assess the distribution of the data. Histogram plots (Figure 2) provide a visual representation of the frequency distribution of the data, allowing for identification of potential skewness, kurtosis, and outliers. Q-Q plots, on the other hand, compare the quantiles of the data to the quantiles of a theoretical distribution, typically a normal distribution. By examining the deviations from the expected diagonal line in the Q-Q plot, one can assess the normality of the data and identify potential departures from the assumed distribution. These visual assessments are crucial for selecting appropriate statistical models and ensuring the validity of assumptions underlying subsequent analyses.

Figure 2. Histograms for normality assessment




To further assess the distribution of each variable, the histograms were augmented with key statistical measures. The Kolmogorov-Smirnov test was included to quantify the goodness-of-fit between the observed data distribution and a theoretical normal distribution. Kurtosis and skewness values were also incorporated to provide insights into the shape of the distribution. Kurtosis measures the "tailedness" of the distribution, while skewness indicates the degree of asymmetry. These additional metrics, alongside the visual representation of the histograms, provided a more comprehensive understanding of the distributional characteristics of each variable, aiding in the identification of potential transformations or model selection strategies.

Although not all the variables are normally distributed, we chose to include them in the next analysis because we ultimately aimed to understand the general influence. Data analysis included a Linear Regression Model to find which features related to sanitation policies (Table 1) significantly influence illnesses caused by poor sanitation (dependent variable). In addition, cluster analysis was conducted to segment the municipalities in the database and find a strategic positioning in each group.

## RESULTS

The initial regression model (H1) model had an adjusted R-squared of 0.242, indicating that the predictors explain about 24.2% of the variance in "illness associated with poor sanitation", considering a linear model (Figure 4). This suggests that the included predictors have some

explanatory power, but a significant portion of the variation remains unexplained. Looking at the coefficients, several predictors have statistically significant relationships with the target variable, the most important is Hospital Bed Rate (bed_index_100k) which has the largest standardized coefficient, suggesting it has the strongest association with "illness associated with poor sanitation". Other significant predictors include Population covered by the Family Health Strategy (FHS) (health_coverage), Water Loss (water_loss), Number of physicians per 100.000 inhabitants (physician_index_100k), and Number of Nurses available per 100.000 inhabitants (nurse_index_100k). This suggests that the health infrastructure has a relevant relation to the incidence of the target diseases. Curiously variables related to sanitation infrastructure and government policies did not show a significant relation to the target variable.

The H1 linear regression model explains a moderate portion of the variance in "illness associated with poor sanitation". Further investigation and potential model adjustments, such as addressing the autocorrelation issue and considering alternative model specifications, are necessary to improve the model's performance and gain a more comprehensive understanding of the factors influencing illness associated with poor sanitation.

Figure 3. Regression Model

**Linear Regression**

Model Summary - target

| | | | | | Durbin-Watson | | |
|---|---|---|---|---|---|---|---|
| Model | R | $R^2$ | Adjusted $R^2$ | RMSE | Autocorrelation | Statistic | p |
| $H_0$ | 0.000 | 0.000 | 0.000 | 0.538 | 0.223 | 1.554 | < .001 |
| $H_1$ | 0.505 | 0.255 | 0.242 | 0.469 | 0.175 | 1.646 | < .001 |

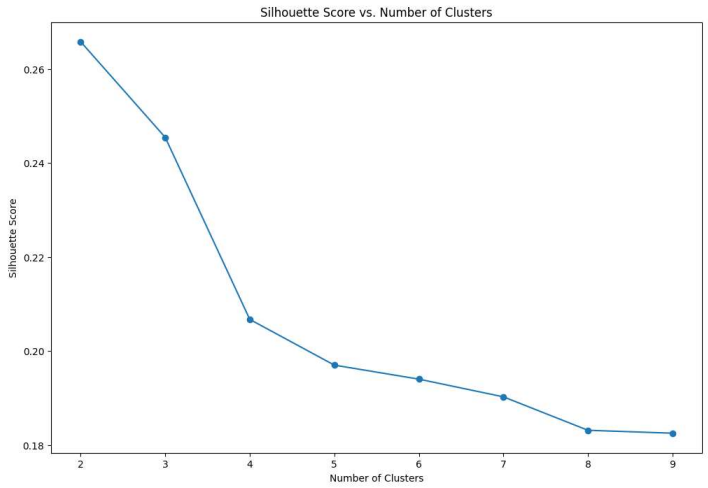
Coefficients

| Model | | Unstandardized | Standard Error | Standardized | t | p |
|---|---|---|---|---|---|---|
| $H_0$ | (Intercept) | 4.812 | 0.017 | | 286.112 | < .001 |
| $H_1$ | (Intercept) | 3.052 | 0.471 | | 6.481 | < .001 |
| | health_coverage | 0.382 | 0.080 | 0.154 | 4.759 | < .001 |
| | water_loss | -0.164 | 0.053 | -0.086 | -3.070 | 0.002 |
| | water_coverage | 0.214 | 0.127 | 0.058 | 1.679 | 0.093 |
| | sewer_coverage | -0.014 | 0.039 | -0.012 | -0.358 | 0.720 |
| | sewer_treat_index | 0.076 | 0.044 | 0.048 | 1.724 | 0.085 |
| | enviroment_maturity | 0.059 | 0.101 | 0.017 | 0.586 | 0.558 |
| | corruption_maturity | -0.105 | 0.181 | -0.016 | -0.580 | 0.562 |
| | res_per_capita | 0.006 | 0.064 | 0.002 | 0.088 | 0.930 |
| | physician_index_100k | -0.321 | 0.069 | -0.190 | -4.675 | < .001 |
| | nurse_index_100k | -0.246 | 0.109 | -0.084 | -2.249 | 0.025 |
| | bed_index_100k | 0.733 | 0.060 | 0.378 | 12.236 | < .001 |
| | ubs_per_1000 | 0.083 | 0.042 | 0.057 | 1.953 | 0.051 |
| | pib_per_capita | 0.020 | 0.071 | 0.010 | 0.277 | 0.782 |
| | total_revenue_per | -0.020 | 0.035 | -0.017 | -0.557 | 0.578 |
| | invest_per_capita | 0.030 | 0.032 | 0.026 | 0.923 | 0.356 |
| | budget_city | -0.025 | 0.039 | -0.019 | -0.652 | 0.515 |
| | public_invest_per_capita | -0.011 | 0.034 | -0.009 | -0.312 | 0.755 |

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the dataset while preserving the most important information. This technique transforms a set of correlated variables into a new set of orthogonal (uncorrelated) variables called principal components. Each principal component is a linear combination of the original variables, ordered by the amount of variance they explain in the data. Then, we selected a 3 components solution as it was responsible for more than 50% of the variance. The components were employed as variables in the k-means cluster analysis.

To determine the optimal number of clusters for the k-means algorithm, we conducted a silhouette analysis. This method evaluates the quality of clustering by measuring how similar an object is to its own cluster compared to other clusters. The silhouette coefficient ranges from -1 to 1, where higher values indicate better-defined clusters. By plotting the average silhouette width against the number of clusters, we identified the number of clusters that resulted in the highest average silhouette score, providing an objective criterion for selecting the optimal cluster solution. As can be seen below (Figure 4), the optimal number is 2, however, we chose 3 clusters since 2 could overfit our model.

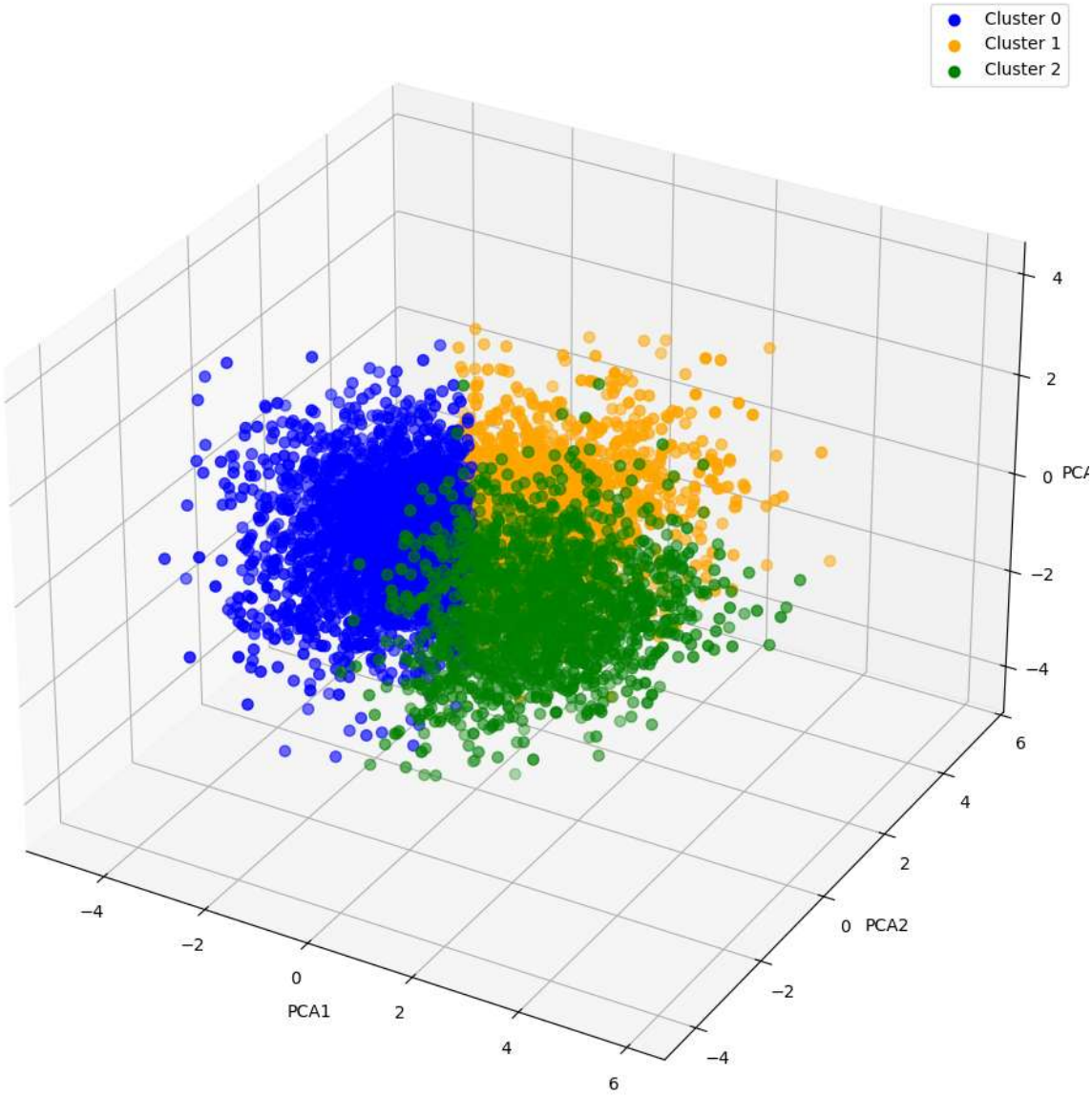
Figure 4. Silhouette analysis




Based on the results of the silhouette analysis, we proceeded with k-means clustering. This iterative algorithm partitions the data into k clusters by minimizing the within-cluster sum of squares. In each iteration, the algorithm assigns each data point to the nearest cluster centroid and then recalculates the centroid based on the new cluster assignments. This process continues until the cluster assignments stabilize. To visualize the clustering results, we generated a 3D plot using the first three principal components. Each data point is represented as a point in this 3D space, and points belonging to the same cluster are colored differently. This visualization provides a visual representation of the clustering structure, allowing for an intuitive understanding of how the data points are grouped based on their similarity in the principal component space. By examining this

plot (Figure 5), we can gain insights into the underlying patterns and relationships within the data and assess the effectiveness of the clustering algorithm.

Figure 5. Cluster distribution




To further investigate the impact of clustering on the target variable ("illness associated with poor sanitation"), the cluster assignments were added back to the original dataset. Subsequently, a exploratory data analysis (EDA) was conducted. These analyses provided insights into the variability and potential differences in the prevalence of illness across the clusters. Building upon the EDA, hypothesis testing was performed to formally assess whether statistically significant differences existed in the target variable between the identified clusters. The Kruskal-Wallis test, a non-parametric test suitable for comparing groups when the assumption of normality is violated, was employed. The results of the Kruskal-Wallis test indicated significant differences

in the prevalence of "illness associated with poor sanitation" across the clusters ($p < 0.001$). Subsequent post-hoc Tukey tests were conducted to identify specific pairwise differences between the clusters, revealing significant differences in the target variable between Cluster 2 and both Cluster 0 and Cluster 1.

Cluster 0: This cluster appears to represent areas with lower levels of development and access to essential services. It shows lower values in variables like health_coverage, water_coverage, environment_maturity, and corruption_maturity. Additionally, indicators like physician_index_100k, nurse_index_100k, and ubs_per_1000 suggest lower healthcare infrastructure. Economic indicators such as pib_per_capita, total_revenue_per_capita, and invest_per_capita are also significantly lower in this cluster.

Cluster 1: This cluster likely represents areas with intermediate levels of development. It shows higher values in most variables compared to Cluster 0, indicating better access to healthcare, improved infrastructure, and higher economic activity. However, it still falls behind Cluster 2 in terms of overall development indicators.

Cluster 2: This cluster appears to represent the most developed areas. It exhibits the highest values in almost all variables, including health_coverage, water_coverage, environment_maturity, corruption_maturity, and economic indicators. This cluster likely has the best access to healthcare services, infrastructure, and economic resources.

In summary, the analysis of the descriptive statistics reveals distinct characteristics for each cluster. Cluster 0 represents areas with lower development levels, characterized by limited access to healthcare, infrastructure, and economic resources. Cluster 1 represents areas with intermediate development, showing improvements in various indicators compared to Cluster 0. Cluster 2 represents the most developed areas, with high values across most variables, indicating better access to services, infrastructure, and economic opportunities. These distinctions can be valuable for policymakers and researchers to understand the varying needs and challenges faced by different regions or communities.

Furthermore, these characteristics help demonstrate statistically significant differences in the prevalence of illness associated with poor sanitation across the three clusters. The Kruskal-Wallis test, which is a non-parametric test used to compare groups when the assumption of normality is violated, revealed a significant difference among the clusters ($p < 0.001$). Post-hoc Tukey tests were conducted to identify specific pairwise differences. The results indicate that Cluster 0 has significantly lower levels of illness compared to Cluster 2 (Mean Difference = -0.157, $p < 0.001$). Similarly, Cluster 1 also shows significantly lower levels of illness compared to Cluster 2 (Mean Difference = -0.205, $p < 0.001$). However, the difference between Cluster 0 and Cluster 1 was not statistically significant. This suggests that areas belonging to Cluster 2, which likely represent the most developed areas as indicated by the previous descriptive statistics analysis, experience significantly higher levels of illness associated with poor sanitation compared to the other two clusters.

**CONCLUSION**

Our study revealed a complex relationship between various socioeconomic and health-related factors and the prevalence of illness associated with poor sanitation. The linear regression model, while showing some predictive power, exhibited limitations due to low R-squared values and the presence of autocorrelation in the residuals. Most of the predictors were not significant, only variables related to health infrastructure showed significant relations to the target variable. Cluster analysis, on the other hand, provided valuable insights into the heterogeneity within the data. The identified clusters, characterized by distinct patterns in socioeconomic and health indicators, demonstrated significant differences in the prevalence of illness associated with poor sanitation. Cluster 2, characterized by higher levels of development and access to resources, exhibited significantly higher levels of illness compared to Clusters 0 and 1. These findings highlight the importance of considering contextual factors and adopting a multi-faceted approach to address the challenges of poor sanitation and related health outcomes. Actions regarding the Legal Framework for Sanitation may consider these tangent features (health infrastructure and heterogeneity among the cities) to design policies appropriate to different contexts. Further research, incorporating more robust modeling techniques and exploring potential interactions between variables, is warranted to deepen our understanding of the underlying mechanisms driving these observed patterns.

## REFERENCES


BRASIL. Câmara dos Deputados. Constituição da República Federativa do Brasil de 1988. Disponível em: https://www2.camara.leg.br/legin/fed/consti/1988/constituicao-1988-5-outubro-1988-322142-publicacaooriginal-1-pl.html . Acesso em: 10 dez. 2024.

BRASIL. Ministério do Desenvolvimento Regional. Diagnósticos anteriores do SNIS: água e esgoto (2016). Disponível em: https://www.gov.br/mdr/pt-br/assuntos/saneamento/snis/diagnosticos-anteriores-do-snis/agua-e-esgotos-1/2016. Acesso em: 10 dez. 2024.

BRASIL. Senado Federal. Conteúdo normativo do direito à saúde. Disponível em: https://www2.senado.leg.br/bdsf/bitstream/handle/id/598844/001220122_Conteudo_normativo_direito_saude.pdf. Acesso em: 10 dez. 2024.

BRASIL. Supremo Tribunal Federal. Resolução 710, de 2020. Disponível em: https://portal.stf.jus.br/hotsites/agenda-2030/assets/img/RESOLUCAO710-2020.PDF. Acesso em: 10 dez. 2024.

BRASIL. LEI Nº 14.026, DE 15 DE JULHO DE 2020. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/l14026.htm

CAPOBIANCO, João Paulo Ribeiro; SANTOS, Gesmar Rosa dos; CHECCO, Guilherme Barbosa; MENDES, Alesi Teixeira. Saneamento básico no Brasil: perfil do investimento público para a universalização e promoção do direito humano à água. *Boletim Regional, Urbano e Ambiental*, n. 29, p. 1–22, 2023. Disponível em: https://repositorio.ipea.gov.br/handle/11058/12163. Acesso em: 28 maio 2025.

HELLER, Léo et al. Saneamento: um direito de todos. Disponível em: https://cee.fiocruz.br/sites/default/files/2_Leo%20Heller%20et%20al_saneamento.pdf . Acesso em: 10 dez. 2024.

INSTITUTO TRATA BRASIL. Saneamento é saúde: Como a falta de acesso à infraestrutura Básica afeta a incidência de doenças relativas ao saneamento ambiental inadequado no Brasil?. 2025. Disponível em: https://tratabrasil.org.br/wp-content/uploads/2025/03/ESTUDO-COMPLETO-Saneamento-e-saude-Como-a-falta-de-acesso-a-infraestrutura-basica-afeta-as-incidencias-de-doencas-relacionadas-ao-saneamento-ambiental-inadequado-no-Brasil-TRATA-BRASIL.pdf?utm_source=chatgpt.com. Acesso em 28/05/2025

NAÇÕES UNIDAS. Objetivos de Desenvolvimento Sustentável (ODS). Disponível em: https://brasil.un.org/pt-br/sdgs#:~:text=Os%20Objetivos%20de%20Desenvolvimento%20Sustent%C3%A1vel%20s%C3%A3o%20um%20apelo%20global%20%C3%A0,de%20paz%20e%20de%20prosperidade. Acesso em: 10 dez. 2024.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. Sanitation. Disponível em: https://www.who.int/topics/sanitation/es/. Acesso em: 10 dez. 2024.

PEREIRA, Rafael. A efetivação do compliance ambiental diante da motivação das certificações brasileiras. Disponível em: https://www.researchgate.net/publication/323352329_A_EFETIVACAO_DO_COMPLIANCE_AMBIENTAL_DIANTE_DA_MOTIVACAO_DAS_CERTIFICACOES_BRASILEIRAS . Acesso em: 10 dez. 2024.

SOUZA, Francisco de Assis Salviano de. A política nacional de saneamento básico no Brasil. Disponível em: http://www.senado.leg.br/comissoes/ci/ap/AP20091130_FranciscodeAssisSalvianodeSousa.pdf . Acesso em: 10 dez. 2024.

TEIXEIRA, Júlio César; OLIVEIRA, Guilherme Soares de; VIALI, Amanda de Mello; MUNIZ, Samuel Soares. Estudo do impacto das deficiências de saneamento básico sobre a saúde pública

no Brasil no período de 2001 a 2009. *Engenharia Sanitária e Ambiental*, Rio de Janeiro, v. 19, n. 1, p. 87–96, jan./mar. 2014. Disponível em: https://www.scielo.br/j/esa/a/phssQJJDhpFtNjB7dLtwW4b/?format=html. Acesso em: 28 maio 2025.

UHR, Júlia Gallego Ziero; SCHMECHEL, Mariana; UHR, Daniel de Abreu Pereira. Relação entre saneamento básico no Brasil e saúde da população sob a ótica das internações hospitalares por doenças de veiculação hídrica. *Revista de Administração, Contabilidade e Economia da Fundace*, v. 7, n. 2, p. 1–15, 2016. Disponível em: https://racef.fundace.org.br/index.php/racef/article/view/104/0. Acesso em: 28 maio 2025.

UNFPA. Population momentum. Disponível em: https://population.un.org/wpp/publications/Files/PopFacts_2017-4_Population-Momentum.pdf . Acesso em: 10 dez. 2024.

**APPENDIX**

```
#Algorithm for KNN Inputation
from sklearn.impute import KNNImputer

# Create a copy of the DataFrame to store the results
df_processed = df.copy()

# Iterate through the columns
for col in df.columns:
    # Calculate the percentage of missing values in the current column
    missing_percentage = df[col].isnull().sum() / len(df) * 100

    # Drop the column if the percentage of missing values is greater than 25%
    if missing_percentage > 25:
        df_processed = df_processed.drop(col, axis=1)
        print(f"Column '{col}' dropped due to more than 25% missing values.")
    # Use KNN imputation if the percentage of missing values is less than or equal to 25%
    elif missing_percentage <= 25 and missing_percentage > 0:
        # Apply KNN imputation only to numerical columns
        if pd.api.types.is_numeric_dtype(df[col]):
          imputer = KNNImputer(n_neighbors=5) # You can adjust the number of neighbors
          df_processed[col] = imputer.fit_transform(df_processed[[col]])
          print(f"KNN imputation applied to column '{col}'.")
        else:
```

```
        #Handle other data types (e.g. categorical) if needed.
        #Example using mode for categorical:
        df_processed[col] = df_processed[col].fillna(df_processed[col].mode()[0])
        print(f"Mode imputation applied to column '{col}'.")
    else:
        print(f"Column '{col}' has no missing values.")

new_df = df_processed.iloc[:, 10:]
```