

**VIABILIDADE DE UM MODELO DE REGRESSÃO LINEAR MÚLTIPLA PARA
MELHORA DA GESTÃO DE RESÍDUOS DE SERVIÇOS DE SAÚDE NO ESTADO
DE SÃO PAULO**

Universidade Presbiteriana Mackenzie

Coordenadoria de Fomento à Pesquisa - Pró-reitoria de Pesquisa e
Pós-Graduação

Andrew Samuel Batista de Oliveira
Apoio: *PIBIC Mackenzie*

Raquel Cymrot

Escola de Engenharia (EE)

10400785@mackenzista.com.br
raquel.cymrot@mackenzie.br

RESUMO

A gestão de resíduos de serviços de saúde (RSS) representa um desafio significativo, especialmente em regiões subdesenvolvidas, nas quais a infraestrutura inadequada pode gerar graves problemas socioambientais. Visando melhorar a gestão dos RSS, este estudo buscou avaliar a viabilidade do desenvolvimento de um modelo de regressão linear múltipla para a previsão da geração de RSS nos municípios do estado de São Paulo com populações entre 100 mil e 1,5 milhão de habitantes, utilizando como variáveis explicativas indicadores socioeconômicos. O estudo buscou identificar as variáveis que influenciam a quantidade de RSS, avaliar seu impacto por meio de análise de regressão linear múltipla e fornecer um modelo preditivo que apoie a tomada de decisões na gestão de tais resíduos. Os modelos ajustados para os anos de 2021 e 2022 apresentaram alto poder explicativo, com coeficientes de determinação superiores a 96%, além de estabilidade estatística confirmada por validação cruzada com 5 dobras. A etapa de previsão demonstrou que o modelo de 2021 manteve desempenho satisfatório quando aplicado aos dados de 2022, com resultados consistentes em métricas como o erro quadrático médio (MSE), erro absoluto médio (MAE) e o erro percentual médio (MAPE). Esses resultados indicam que o modelo proposto foi estável, podendo ser generalizável e aplicável a cenários futuros, possibilitando atuar como ferramenta de apoio à gestão pública de RSS. Vale ressaltar que futuras pesquisas com dados coletados em períodos sem influência da Covid-19 precisam confirmar que o modelo aqui apresentado continua válido.

Palavras-chave: Resíduos de Serviços de Saúde. Modelo de regressão linear múltipla. Cidades do Estado de São Paulo.

ABSTRACT

Managing healthcare waste (HCW) presents a significant challenge, particularly in underdeveloped regions where inadequate infrastructure can lead to severe socio-environmental problems. This study sought to evaluate the feasibility of developing a multiple linear regression model to predict HCW generation in cities in the state of São Paulo with populations between 100,000 and 1.5 million inhabitants, using socioeconomic indicators as explanatory variables. The objectives were to identify the variables influencing the quantity of HCW, evaluate their impact through multiple linear regression analysis, and provide a predictive model to support decision-making in the management of that waste. The models fitted

for the years 2021 and 2022 showed high explanatory power, with coefficients of determination greater than 96%, in addition to statistical stability confirmed by 5-fold cross-validation. The prediction stage demonstrated that the 2021 model maintained satisfactory performance when applied to 2022 data, with consistent results in metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These results indicate that the proposed model was stable, generalizable, and applicable to future scenarios, potentially serving as a supporting tool for public HCW management. It is important to note that future research with data collected during periods without the influence of Covid-19 needs to confirm that the model presented here remains valid.

Keywords: Healthcare Waste. Multiple linear regression model. Cities in São Paulo State.

1. INTRODUÇÃO

De acordo com a norma brasileira NBR 10004 da Associação Brasileira de Normas Técnicas (ABNT), resíduos sólidos incluem aqueles em estado sólido e semissólido, originários de atividades industriais, domésticas, agrícolas, comerciais, além de serviços de saúde, sendo este último frequentemente classificado como perigoso devido à possível toxicidade e patogenicidade (ABNT, 2004).

A gestão de resíduos de serviços de saúde (RSS) representa um grande desafio para muitas cidades, especialmente onde há falta de infraestrutura adequada para seu manejo e descarte. Esse cenário pode gerar sérios problemas socioambientais, como riscos de transmissão de doenças por contaminação do ambiente, afetando catadores de lixo, trabalhadores de limpeza, profissionais de saúde e a comunidade em geral (UN, 2020).

Os RSS, devido aos seus potenciais riscos biológicos, são regulados por normas específicas, como o Plano de Gerenciamento de Resíduos de Serviços de Saúde (PGRSS), exigido pela Agência Nacional de Vigilância Sanitária (ANVISA), que detalha o manejo adequado de cada etapa da gestão (Brasil, 2018). Além disso, RSS gerados em domicílios, estimados no início da década de 2020 em 1,68 kg/paciente/mês, muitas vezes são descartados como lixo comum por falta de conhecimento dos cuidadores do Serviço de Atenção Domiciliar (SAD) sobre as diretrizes de descarte adequado (Silva, 2020).

Embora os indicadores socioeconômicos sejam temas de muito estudo, não se encontrou na literatura a sua relação com o aumento na geração de RSS, em particular, no estado de São Paulo. Assim, coloca-se a questão: Qual a viabilidade do desenvolvimento de um modelo de regressão linear múltipla que permita melhorar a gestão dos RSS no estado de São Paulo?

Estimativas da quantidade futura de RSS são essenciais ao planejamento de coleta, transporte e tratamento. Indicadores socioeconômicos oferecem síntese informativa e são recorrentes em aplicações de modelagem estatística e previsão (Caldeira; Aquino, 2020).

Portanto, o objetivo da pesquisa foi avaliar a viabilidade do desenvolvimento de um modelo de regressão linear múltipla para a previsão da geração de Resíduos de Serviços de Saúde (RSS) nos municípios do estado de São Paulo com populações entre 100 mil e 1,5 milhão de habitantes, utilizando como variáveis explicativas indicadores socioeconômicos.

Para alcançar esse objetivo, foram identificadas variáveis candidatas, selecionando-se aquelas com maior poder explicativo e avaliou-se o desempenho explicativo e preditivo dos modelos de regressão linear múltipla resultantes.

2. REFERENCIAL TEÓRICO

No Brasil, os resíduos de serviços de saúde são classificados em cinco grupos, A, B, C, D e E, que são compostos por respectivamente “[...] resíduos com possível presença de agentes biológicos” (Brasil, 2005, p. 619), “[...] resíduos contendo substâncias químicas que podem apresentar risco a saúde pública ou ao meio ambiente” (Brasil, 2005, p. 620), “[...] materiais resultantes de atividades humanas que contenham radionuclídeos” sob determinadas condições (Brasil, 2005, p. 621), “[...] resíduos que não apresentem risco biológico, químico ou radiológico à saúde ou ao meio ambiente, podendo ser equiparados aos resíduos domiciliares” (Brasil, 2005, p. 621) e “[...] materiais perfurocortantes e escarificantes” (Brasil, 2005, p. 621).

A ANVISA define como geradores de resíduos de serviços de saúde (RSS):

[...] todos os serviços cujas atividades estejam relacionadas com a atenção à saúde humana ou animal, inclusive os serviços de assistência domiciliar; laboratórios analíticos de produtos para saúde; necrotérios, funerárias e serviços onde se realizem atividades de embalsamamento (tanatopraxia e somatoconservação); serviços de medicina legal; drogarias e farmácias, inclusive as de manipulação; estabelecimentos de ensino e pesquisa na área de saúde; centros de controle de zoonoses; distribuidores de produtos farmacêuticos, importadores, distribuidores de materiais e controles para diagnóstico in vitro; unidades móveis de atendimento à saúde; serviços de acupuntura; serviços de piercing e tatuagem, salões de beleza e estética, dentre outros afins (Brasil, 2018, não paginado).

De acordo com a Organização Mundial de Saúde, cerca de 75% dos resíduos provenientes de atividades relacionadas à saúde são considerados comuns (WHO, 2016). Os 25% restantes são parte dos grupos A, B, C e E de RSS, e possuem normas vigorosas para o seu manuseio. Deve-se efetuar a segregação e identificação adequada, seguindo o Plano de Gerenciamento de Resíduos de Serviços de Saúde da ANVISA, assegurando que o seu descarte seja feito da forma correta, dependendo da classe desse resíduo. Tal plano deve conter as seguintes etapas: manejo, “[...] segregação, acondicionamento, identificação, transporte interno, armazenamento temporário, armazenamento externo, coleta interna, transporte externo, destinação e disposição final ambientalmente adequada dos resíduos de serviços de saúde” (Brasil, 2018, p. 4). Além disso, deve ser compatível e cumprir com todas as normas locais de coleta de resíduos (Brasil, 2018).

O estabelecimento ou serviço gerador é responsável por todas as etapas do gerenciamento de RSS, podendo terceirizar os serviços caso necessário, porém as consequências de sua disposição inadequada continuam sob a responsabilidade do gerador. As informações dos resíduos devem ser declaradas no Manifesto de Transporte de Resíduos (MTR) no Sistema Nacional de Informações Sobre a Gestão de Resíduos Sólidos (SINIR), que é um

“[...] documento numerado, gerado por meio do Sistema Nacional de Informações Sobre a Gestão de Resíduos Sólidos (SINIR), emitido exclusivamente pelo Gerador, que deverá acompanhar o transporte do resíduo até a destinação final ambientalmente adequada” (Brasil, 2020, não paginado).

Uma pesquisa feita por Janik-Karpinska *et al.* (2023) que buscou entender a geração de RSS em países selecionados, mostrou que o Brasil possui uma das maiores taxas de geração de RSS da América Latina, com valores de 2,94–3,3 kg/leito/dia. Isso evidencia a necessidade de se entender quais fatores influenciam na variação dessa taxa, podendo assim facilitar a gestão destes resíduos.

Indicadores socioeconômicos são fundamentais em modelos de regressão linear múltipla para prever a geração de RSS numa localidade, já que capturam variáveis como renda, escolaridade e saúde. O IDH avalia essas três dimensões: expectativa de vida, anos de educação e renda per capita em PPP (UN, 2024). Por falta de disponibilidade de dados do IDH mais atuais, na presente pesquisa, optou-se pela utilização do Índice Paulista de Desenvolvimento Municipal (IPDM), Este indicador, desenvolvido pela Fundação Sistema Estadual de Análise de Dados (SEADE), também sintetiza, para os municípios do Estado de São Paulo, informações sobre riqueza, escolaridade e longevidade. O IPDM é baseado em registros administrativos e atualizado com frequência, sendo uma fonte oficial reconhecida (SEADE, 2024)

Outro indicador é o PIB, que “[...] é a soma de todos os bens e serviços finais produzidos por um país, estado ou cidade, geralmente em um ano” (IBGE, 2025, não paginado). Já o PIB *per capita* é o resultado da divisão do PIB pelo número de habitantes da região (Ipeadata, 2024).

A análise de dados estatísticos relacionando índices socioeconômicos a outros fatores para estudo é amplamente usada, principalmente por meio de regressões lineares e utilização de coeficientes como os de correlação linear de Pearson. Alguns estudos são comentados a seguir.

Pisani Junior, Castro e Costa (2018) elaboraram um modelo de previsão para a geração de resíduos sólidos urbanos tendo como variáveis explicativas o consumo de energia, a população e o PIB *per capita* em 238 municípios no estado de São Paulo.

Consoante a esse estudo, Garlasco *et al.* (2021) buscaram relacionar o aumento da geração de RSS durante a pandemia do COVID-19. Para isso, foram utilizadas análises de tendência por meio de análise de regressão linear e de séries temporais interrompidas, além de teste de correlação linear de Pearson a partir de dados de geração de RSS de janeiro de 2015 a

março de 2021, fornecidos pelo hospital Azienda Ospedaliero-Universitaria Santi Antonio e Biagio e Cesare Arrigo, na cidade de Alexandria, Itália. Os resultados indicaram uma correlação significativa entre a geração de resíduos e o número de admissões e dias de paciente relacionados à COVID-19.

Outra pesquisa, conduzida por Kaposi *et al.* (2024) no centro clínico University of Debrecen Clinical Centre Nagyerdei Campus, Hungria, utilizou análises estatísticas descritivas, correlação linear de Spearman, e regressão robusta para analisar a quantidade de RSS e a variação dessa quantidade ao longo do tempo, e fatores que podem estar associados com estes valores. Foi concluído que existe uma correlação significativa entre, principalmente, a quantidade de RSS e a incidência de infecções relacionadas à assistência à saúde.

3. METODOLOGIA

A presente pesquisa adotou um enfoque quantitativo, utilizando técnicas de regressão linear múltipla para investigar os fatores associados à geração de resíduos de serviços de saúde (RSS) por habitante nos municípios do estado de São Paulo, com base em dados dos anos de 2021 e 2022. O procedimento analítico envolveu as etapas de preparação dos dados, seleção de variáveis, diagnóstico do modelo, validação cruzada e exclusão de observações influentes.

Foi construído um banco de dados baseado em pesquisas secundárias, contendo dados oficiais obtidos no Sistema Nacional de Informações sobre Saneamento (SNIS) (SNIS, 2024), Instituto Brasileiro de Geografia Estatística (IBGE) (IBGE, 2024), Fundação Sistema Estadual de Análise de Dados (SEADE) (SEADE, 2024). No banco de dados final, após a exclusão de municípios com ausência de dados cruciais (como a quantidade de RSS gerada) e a aplicação dos filtros relacionados à faixa populacional de interesse, restaram 52 municípios em 2021 e 64 municípios em 2022.

Os anos de 2021 e 2022 foram impactados pela pandemia de COVID-19, principalmente 2021, que gerou efeitos relevantes na estrutura e no funcionamento dos serviços de saúde em todo o país. Esse contexto pode ter impactado os resultados, com aumento na geração de resíduos de serviços de saúde, especialmente em municípios que concentram unidades hospitalares regionais ou que receberam maior demanda assistencial. Ainda assim, esses dois anos foram escolhidos para esta análise por representarem o período mais recente com dados disponíveis e consistentes tanto sobre indicadores de RSS quanto sobre variáveis socioeconômicas relevantes, possibilitando uma avaliação com dados mais relevantes.

A abordagem inicial consistiu em realizar uma regressão com todas as variáveis para identificar possíveis fatores como multicolinearidade, existência de pontos influentes, e avaliar a capacidade preditiva inicial do modelo. Para refinar o conjunto de variáveis explicativas, aplicou-se um procedimento de seleção *stepwise*, com critérios de entrada e saída baseados em nível de significância de 15% ($\alpha = 0,15$), complementado pelo critério do coeficiente C_p de Mallows. Esse coeficiente atua como uma medida do erro quadrático médio total padronizado para o modelo de regressão. O C_p avalia a qualidade do ajuste de um modelo com p parâmetros em relação ao modelo completo (aquele com todas as variáveis candidatas), comparando seus erros médios. O princípio básico é que, se um modelo possui tendenciosidade negligenciável, o valor de C_p tende a ser aproximadamente igual ao número de parâmetros do modelo (p). Por outro lado, valores de C_p significativamente maiores que p indicam a presença de viés relevante, sugerindo que o modelo não captura adequadamente a variabilidade dos dados. Assim, modelos com C_p próximos, mas não superiores a p , são preferidos, pois representam um equilíbrio entre subajuste e sobreajuste. Essa abordagem favorece a seleção de modelos parcimoniosos, com bom desempenho preditivo, sem incluir variáveis irrelevantes (Montgomery; Runger, 2021).

Com o modelo ajustado, foram realizados testes estatísticos para verificar a adequação aos pressupostos da regressão linear múltipla. A normalidade dos resíduos foi avaliada por meio do teste de Shapiro-Wilk, escolhido por ser reconhecido como um teste relevante para detecção de desvios da normalidade, tanto em distribuições simétricas quanto assimétricas (Razali; Wah, 2011). Essa capacidade é ainda mais significativa em amostras pequenas, cenário em que outros testes como Kolmogorov-Smirnov e Anderson-Darling tendem a apresentar baixo poder estatístico. Além disso, Souza *et al.* (2023) demonstraram que o desempenho do teste Shapiro-Wilk melhora consideravelmente à medida que o tamanho da amostra aumenta, mas que já atinge níveis de precisão confiáveis a partir de 17 unidades amostrais, mesmo em contextos de dados reais com variabilidade acentuada, como a produtividade de grãos de soja. Essa evidência reforça a robustez e a sensibilidade do teste para validar a suposição de normalidade dos resíduos no modelo de regressão linear múltipla.

A aproximada homoscedasticidade (variâncias dos resíduos constante para todos os níveis da variável resposta) e média zero dos resíduos foi verificada por meio da análise gráfica dos resíduos versus valores ajustados e a independência entre os resíduos foi verificada analisando-se o gráfico dos resíduos em função da ordem das observações. Além disso, a existência de multicolinearidade entre as variáveis explicativas foi verificada pelo Fator de

Inflação da Variância (FIV), que indica quanto da variância estimada dos coeficientes de regressão está inflada devido à correlação entre as variáveis independentes, sendo considerado que valores superiores a 4 já podem indicar multicolinearidade problemática (Montgomery; Runger, 2021).

Observações com potencial influência maior nos resultados foram identificadas utilizando a Distância de Cook, uma medida que avalia o quanto cada observação individual afeta as estimativas dos coeficientes da regressão. Essa métrica considera simultaneamente o tamanho do resíduo da observação e sua posição em relação ao conjunto de dados, ou seja, combina o erro do ponto com o quanto ele se distancia do centro das demais observações. Assim, a Distância de Cook permite identificar observações que, por serem atípicas podem exercer alta influência e distorcerem significativamente os resultados do modelo. Valores grandes da distância de Cook indicam observações altamente influentes (Montgomery; Runger, 2021). Esses pontos devem ser avaliados e, se necessário, removidos das análises para garantir a estabilidade e confiabilidade do modelo ajustado.

Para avaliar a capacidade preditiva dos modelos desenvolvidos, foi utilizada a técnica de validação cruzada k -fold com 5 dobras (*5-fold cross-validation*). Esse procedimento consiste em particionar o conjunto de dados em k subconjuntos de tamanho aproximadamente igual. Em cada iteração, um desses subconjuntos é separado como conjunto de teste, enquanto os $k-1$ restantes são utilizados para treinar o modelo. O processo se repete k vezes, garantindo que todas as observações sejam usadas tanto para treinamento quanto para teste. Ao final, as métricas de desempenho, como o coeficiente de determinação médio (R^2) e o erro quadrático médio (MSE), são calculadas a partir da média dos resultados obtidos nas iterações, fornecendo uma estimativa mais realista da *performance* preditiva do modelo em dados não vistos (James *et al.*, 2023).

A escolha de cinco dobras foi fundamentada no tamanho da amostra, que variou entre 50 e 60 observações. Segundo Kohavi (1995), o uso de 5 a 10 dobras é especialmente recomendado para amostras pequenas ou moderadas, uma vez que proporciona equilíbrio entre viés e variância na estimativa do erro. Além disso, a validação cruzada é preferível em relação à divisão simples dos dados (*holdout*) por oferecer maior estabilidade nas estimativas e reduzir a sensibilidade à forma como os dados são particionados. Dessa forma, a aplicação da validação cruzada 5-fold neste estudo contribuiu para uma avaliação menos suscetível à variabilidade na separação entre treino e teste dos modelos de regressão linear múltipla (James *et al.*, 2023).

Finalmente, foram realizados testes de previsão *out-of-time* utilizando o modelo ajustado com dados do ano de 2021, para prever a quantidade de RSS produzida em 2022. Os resultados dessas previsões foram então comparados com os dados reais de 2022, permitindo avaliar a robustez preditiva e a aplicabilidade prática do modelo desenvolvido.

Todas as análises estatísticas foram conduzidas utilizando Microsoft Excel®, Minitab® Statistical Software v. 22, e a linguagem de programação Python.

4. RESULTADOS E DISCUSSÃO

A análise de regressão linear múltipla foi conduzida separadamente para os anos de 2021 e 2022, com o objetivo de identificar os principais determinantes da geração de resíduos de serviços de saúde (RSS) por habitante nos municípios paulistas com uma população entre cem mil e um milhão e meio de habitantes. As variáveis utilizadas nos modelos são descritas no Quadro 1.

Quadro 1 – Dicionário de variáveis utilizadas no modelo de regressão

Variável	Descrição
despesa_tot_rss	Valor total gasto pelo município com o manejo de resíduos de serviços de saúde (RSS).
envia_rss	Variável binária que indica se a cidade envia seus RSS para tratamento em outro município.
taxa_rss/rt	Proporção dos RSS em relação ao total de resíduos coletados na cidade.
ipdm	Índice Paulista de Desenvolvimento Municipal, que avalia o desenvolvimento dos municípios com base em indicadores sociais, econômicos e ambientais.
PIB per capita	Produto Interno Bruto per capita do município, representando a riqueza gerada por habitante.
mx_total	Taxa de mortalidade geral do município.
estab_por_hab	Número de estabelecimentos de saúde por habitante, indicando a disponibilidade de infraestrutura de saúde.
mx_infantil	Taxa de mortalidade infantil, representando o número de óbitos de crianças menores de um ano por mil nascidos vivos.
Pop_tot	População total do município.

Fonte: Os autores (2025)

Após a aplicação do procedimento de seleção de variáveis por meio do método *stepwise*, complementado pela análise de melhores subconjuntos, o modelo final foi composto pelas variáveis: taxa_rss/rt, despesa_rss_hab, envia_rss, pop_tot e mx_total. Essas variáveis foram selecionadas de forma consistente para os anos de 2021 e 2022, resultando em modelos estruturalmente idênticos em termos de preditores. As análises foram precedidas por uma avaliação dos pressupostos do modelo, incluindo normalidade dos resíduos, independência dos resíduos, média zero e homoscedasticidade dos resíduos, coeficiente de Mallows adequado, ausência de multicolinearidade e identificação de valores influentes.

Para 2021, as suposições de uso do modelo foram satisfeitas, não havendo nenhum ponto influente, sendo realizada uma regressão linear múltipla pelo método *stepwise*, e análise de melhores subconjuntos, conforme Tabela 1. O resultado da regressão linear múltipla usando-se o *stepwise* foi um modelo com 5 variáveis, significativo, uma vez que seu p-valor foi 8,74E-33. Este modelo apresentou os coeficientes de determinação R^2 igual a 96,67% e R^2 ajustado igual a 96,31%, evidenciando a alta capacidade explicativa. O Cp de Mallows igual a 5,5141 foi inferior e próximo a 6 que é igual ao número coeficientes beta do modelo (número de variáveis acrescido de um, referente à constante).

Tabela 1 – Melhores Subconjuntos (Modelo 2021)

Variáveis	Cp de	R ²	Nº de
	Mallows	ajustado	Variáveis
Modelo Completo	10	0,9627	9
('pop_tot', 'despesa_rss_hab', 'envia_rss', 'taxa_rss/rt', 'estab_hab', 'mx_total', 'PIB_per_capita')	7,7601	0,9629	7
('pop_tot', 'despesa_rss_hab', 'taxa_rss/rt', 'mx_infantil', 'mx_total', 'PIB_per_capita')	6,725	0,9629	6
('pop_tot', 'despesa_rss_hab', 'envia_rss', 'taxa_rss/rt', 'ipdm', 'mx_total', 'PIB_per_capita')	7,5474	0,9631	7
('pop_tot', 'despesa_rss_hab', 'envia_rss', 'taxa_rss/rt', 'mx_total')	5,5141	0,9631	5
('pop_tot', 'despesa_rss_hab', 'envia_rss', 'taxa_rss/rt', 'mx_infantil', 'mx_total')	6,4891	0,9631	6

Fonte: Dados da pesquisa (2025)

A Tabela 2 apresenta as variáveis do modelo, sendo que, dentre as variáveis escolhidas pelo método de *stepwise*, apenas mx_total não foi isoladamente significativa ao nível de significância de 5%. A Tabela 2 também apresenta a estimativa pontual e por intervalo dos coeficientes estimados, os níveis descritivos do teste de significância destes coeficientes, bem como os respectivos Fatores de Inflação da Variância (FIV), os quais não ultrapassaram 4, indicando que não houve evidência de multicolinearidade no modelo.

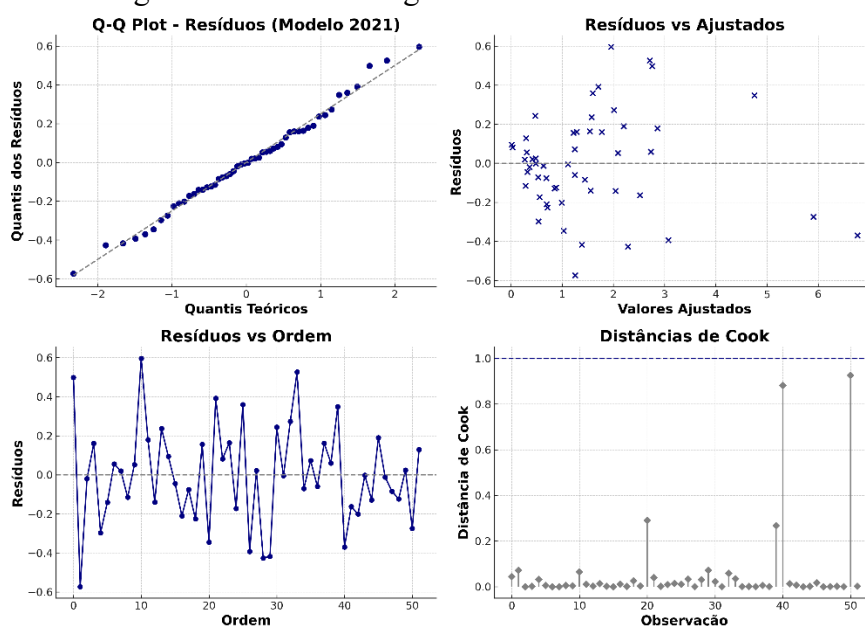
Tabela 2 – Coeficientes estimados, p-valor e FIV (Modelo 2021)

Variável	Coeficiente	p-valor	Intervalo com 95% de confiança	FIV
const.	-0,3589	0,215	[-0,9338; 0,2159]	
taxa_rss/rt	2,3541	1,23E-18	[21668; 254144]	1,3260
despesa_rss_hab	0,0481	1,02E-06	[0,0373; 0,0588]	1,5611
envia_rss	-0,2100	0,031	[-0,3994; -0,0206]	1,0115
pop_tot	3,05E-07	0,039	[1,65E-8; 5,93E-7]	1,0084
mx_total	0,0496	0,089	[-0,0078; 0,1069]	1,1964

Fonte: Dados da pesquisa (2025)

Conforme Figura 1, foi verificada a aderência dos resíduos à distribuição Normal ($p = 0,303$), bem como que os resíduos tiveram média aproximadamente zero, com variância aproximadamente constante e independentes. Já o gráfico das distâncias de Cook mostra que nenhuma cidade foi considerada como ponto influente no modelo, uma vez que nenhum valor ultrapassou o limite de corte de 1.

Figura 1 – Gráficos diagnósticos do modelo de 2021

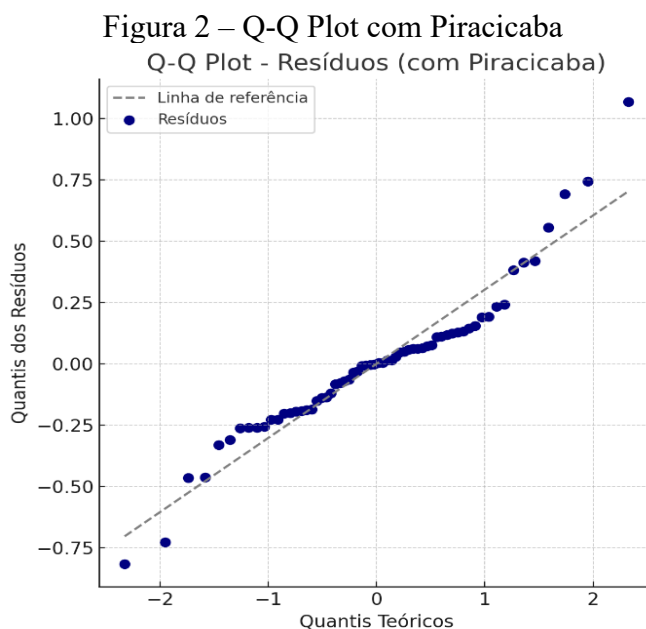


Fonte: Dados da pesquisa (2025)

Em 2022, as cidades de Assis, Caraguatatuba, Piracicaba e Santos foram removidas do modelo completo de regressão por comprometerem a aderência dos resíduos à distribuição Normal e/ou apresentarem distâncias de Cook superiores a um, indicando influência excessiva sobre os parâmetros do modelo. A exclusão dessas observações foi essencial para garantir a robustez estatística e a validade dos pressupostos do modelo ajustado.

Embora o motivo exato para o comportamento discrepante dessas cidades não possa ser determinado com precisão a partir dos dados disponíveis, a principal hipótese está associada à presença de hospitais regionais de referência em suas respectivas regiões de saúde. O Hospital

Regional de Assis (DRS IX), o Hospital Regional do Litoral Norte em Caraguatatuba (DRS XVII), o Hospital Regional de Piracicaba (DRS X) e o polo hospitalar de Santos (DRS IV, Baixada Santista) atuam como centros de atendimento para diversos municípios vizinhos. Esse fator, aliado ao contexto da pandemia de COVID-19, pode ter resultado em uma geração de RSS desproporcional à população residente dessas cidades, afetando diretamente a relação entre os preditores do modelo e a variável dependente. Para Piracicaba, o p-valor do teste de Shapiro-Wilk foi de 0,042, o que mostrou a necessidade de uma análise específica para essa cidade. Analisando o QQ-plot, na Figura 2, é evidente que há um ponto descolado da reta teórica, o qual é justamente a cidade de Piracicaba. Optou-se então por remover Piracicaba do modelo.



Fonte: Dados da pesquisa (2025)

Com as cidades filtradas, foi feita uma análise de melhores subconjuntos. Dentre os 5 melhores modelos, apresentados na Tabela 3, o melhor R^2 ajustado foi de 0,9593, com as variáveis `pop_tot`, `despesa_rss_hab`, `envia_rss`, `taxa_rss/rt` e `mx_total`. O Cp de Mallows para esse modelo é menor do que o número de variáveis mais um (referente à constante), o que é desejável.

Tabela 3 – Melhores Subconjuntos (Modelo 2022)

Variáveis	Cp de Mallows	R ² ajustado	Nº de Variáveis
Modelo Completo	10	0,9583	9
('pop_tot', 'despesa_rss_hab', 'envia_rss', 'taxa_rss/rt', 'ipdm', 'mx_total')	6,7321	0,9585	6
('pop_tot', 'despesa_rss_hab', 'envia_rss', 'taxa_rss/rt', 'mx_infantil', 'mx_total')	6,665	0,9586	6
('pop_tot', 'despesa_rss_hab', 'envia_rss', 'taxa_rss/rt')	4,5539	0,9586	4
('pop_tot', 'despesa_rss_hab', 'envia_rss', 'taxa_rss/rt', 'estab_hab', 'mx_infantil', 'ipdm', 'mx_total')	8,5232	0,9587	8
('pop_tot', 'despesa_rss_hab', 'envia_rss', 'taxa_rss/rt', 'mx_total')	4,7337	0,9593	5

Fonte: Dados da pesquisa (2025)

Também foi realizada uma regressão linear múltipla para 2022 utilizando o método *stepwise*, que resultou no mesmo conjunto de variáveis selecionado pelo critério dos melhores subconjuntos. Optou-se por manter esse modelo não apenas pelo alto R² ajustado (95,93%), mas principalmente por ele replicar as variáveis explicativas do modelo de 2021, o que reforça a estabilidade das relações identificadas entre os anos. O modelo final de 2022 apresentou um coeficiente de determinação R² de 96,27% e p-valor de 2,85E-37, confirmando sua significância estatística.

A Tabela 4 também apresenta a estimativa pontual e por intervalo dos coeficientes estimados, os níveis descritivos do teste de significância destes coeficientes, bem como os respectivos Fatores de Inflação da Variância (FIV), os quais não ultrapassaram 4, indicando que não houve evidência de multicolinearidade no modelo.

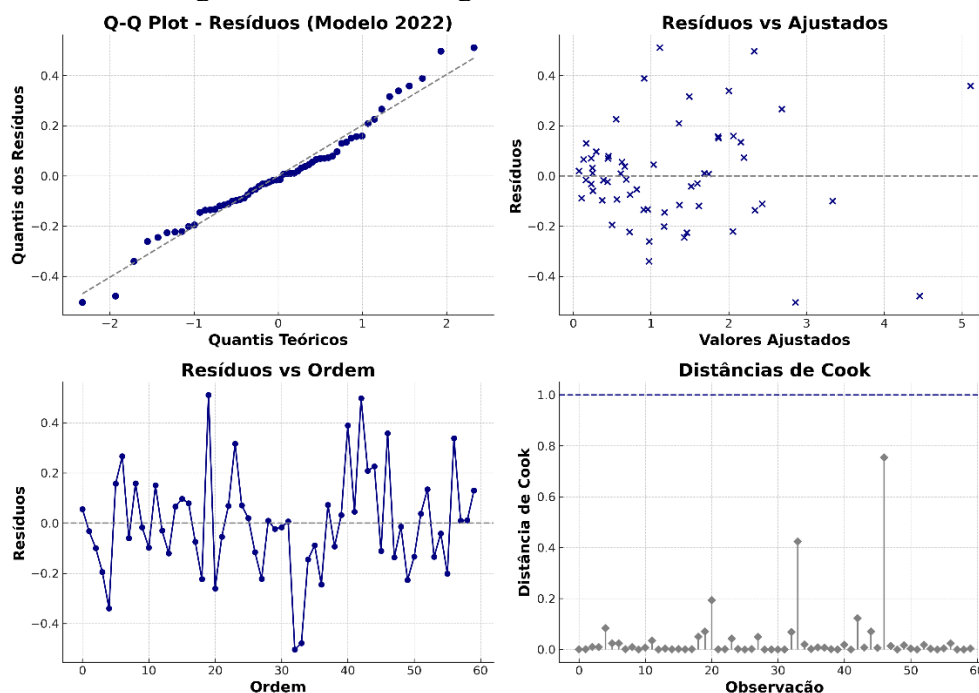
Tabela 4 – Coeficientes estimados pontualmente e por intervalos, p-valor e FIV (Modelo 2022)

Variável	Coeficiente	p-valor	Intervalo com 95% de confiança	FIV
const	-0,2109	0,3210	[-0,6329; 0,2112]	
taxa_rss/rt	2,6250	5,18E-26	[240772; 284223]	1,5348
despesa_rss_hab	0,0275	4,6478	[0,0150; 0,0400]	1,4840
envia_rss	-0,1242	0,1654	[-0,3013; 0,0529]	1,1707
pop_tot	2,92E-7	0,0337	[1,34E-8; 5,60E-7]	1,0639
mx_total	0,0339	0,2152	[-0,0203; 0,08815]	1,0564

Fonte: Dados da pesquisa (2025)

A aderência da distribuição dos resíduos à distribuição Normal não foi rejeitada ($p = 0,275$), conforme Figura 3. No gráfico Q-Q plot dos resíduos, é visível a diferença causada pela remoção da cidade de Piracicaba. No gráfico de resíduos vs. ajustados pode-se verificar as suposições de média zero e variância constante dos resíduos e pelo gráfico de resíduos vs. ordem, a ausência de padrões atesta a independência dos resíduos. Já por meio do gráfico que apresenta as distâncias de Cook calculadas para cada uma das cidades, verificou-se estar todas abaixo de um, que é limite de corte.

Figura 3– Gráficos diagnósticos do modelo de 2022



Fonte: Dados da pesquisa (2025)

Os resultados da validação cruzada com 5 dobras para os modelos dos anos de 2021 e 2022 são apresentados na Tabela 5. Ambos os modelos R^2 médio superiores a 0,96, com erro quadrático médio (MSE) inferior a 0,12 e erro absoluto médio (MAE) inferior a 0,25. Embora os valores de R^2 e R^2 ajustado sejam ligeiramente superiores para 2021, o modelo de 2022 apresentou menores erros médios, indicando maior precisão pontual nas previsões.

Tabela 5 – Resultados da Validação Cruzada

Modelo	R^2 Médio	R^2 Ajustado Médio	MSE Médio	RMSE Médio	MAE Médio	MAPE Médio (%)
2021	0,9684	0,9639	0,1182	0,3438	0,2400	23,04%
2022	0,9644	0,9601	0,0688	0,2622	0,1894	25,20%

Fonte: Dados da pesquisa (2025)

O erro absoluto percentual médio (MAPE) foi de 23,04% para 2021 e 25,20% para 2022, indicando que, embora haja uma leve diferença na acurácia percentual, ambos os modelos apresentam desempenhos próximos. A semelhança entre os indicadores de erro e os altos valores de R^2 em ambos os anos reforçam a estabilidade do modelo e sua capacidade de generalização. Isso sugere que as variáveis selecionadas mantêm sua relevância preditiva ao longo do tempo, mesmo diante de possíveis variações operacionais ou contextuais entre os anos.

Em termos práticos, esses resultados apontam para a viabilidade do uso das variáveis obtidas no modelo proposto como ferramenta auxiliar na previsão da geração de resíduos de

serviços de saúde, com margem de erro controlada e desempenho preditivo em diferentes contextos amostrais.

Para testar a capacidade preditiva do modelo ajustado com dados de 2021, foi realizada uma previsão *out-of-time*, aplicando-se os coeficientes estimados em 2021 diretamente sobre os dados explicativos do ano de 2022. Foram excluídas da base de 2022 as cidades de Assis, Caraguatatuba, Piracicaba e Santos, que apresentaram comportamento atípico ou violaram pressupostos estatísticos na etapa anterior de ajuste do modelo.

Os resultados dessa previsão estão apresentados na Tabela 6. O modelo de regressão ajustado com base nos dados de 2021 manteve desempenho satisfatório ao ser aplicado sobre os dados de 2022. O coeficiente de determinação ($R^2 = 0,9447$) indica que aproximadamente 94,5% da variação na geração de RSS em 2022 foi explicada pelas variáveis selecionadas no modelo, mesmo sem reestimativa dos coeficientes, o que sugere forte capacidade de generalização.

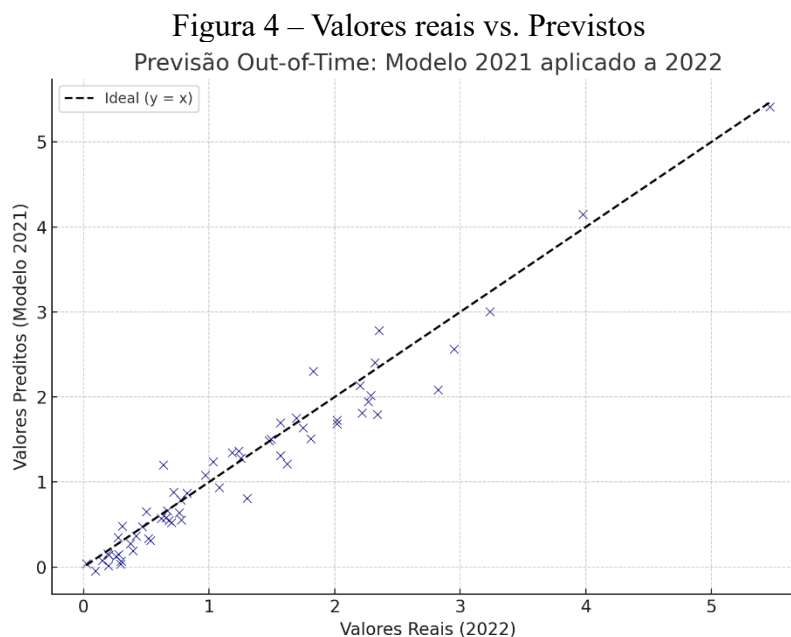
Tabela 6 – Resultados da previsão

Métrica	Valor
R^2	0,9447
MSE	0,0605
RMSE	0,2459
MAE	0,1891
MAPE	26,51%

Fonte: Dados da pesquisa (2025)

Além disso, os valores de erro obtidos: $MSE = 0,0605$; $RMSE = 0,2459$ e $MAE = 0,1891$, são todos consistentes com os resultados obtidos na validação cruzada, evidenciando que o modelo não apenas ajusta bem os dados conhecidos, como também se comporta de maneira estável frente a dados de um novo período. A proximidade entre os erros médios reforça a confiabilidade do modelo em aplicações práticas.

A Figura 4 apresenta o gráfico de dispersão entre os valores reais observados em 2022 e os valores preditos pelo modelo de 2021. A reta tracejada representa o ideal de predição perfeita. A proximidade dos pontos com essa linha indica visualmente que o modelo conseguiu prever a maior parte dos casos com boa precisão, mesmo em um cenário com possíveis variações contextuais entre os anos.



Fonte: Dados da pesquisa (2025)

Os resultados obtidos sugerem que o modelo proposto, além de apresentar bom ajuste interno, também possui capacidade de generalização temporal, podendo ser utilizado como ferramenta preditiva em planejamentos futuros voltados à gestão de resíduos de serviços de saúde.

5. CONCLUSÃO

A presente pesquisa teve como objetivo avaliar a viabilidade do desenvolvimento de um modelo de regressão linear múltipla para previsão da geração de Resíduos de Serviços de Saúde (RSS) nos municípios do estado de São Paulo, com populações entre 100 mil e 1,5 milhão de habitantes, utilizando como variáveis explicativas indicadores socioeconômicos.

A partir da incorporação de variáveis socioeconômicas relevantes, buscou-se construir uma ferramenta de apoio à gestão pública, capaz de auxiliar no planejamento e na destinação adequada dos RSS.

Os modelos ajustados para os anos de 2021 e 2022 apresentaram alto poder explicativo, com coeficientes de determinação (R^2) superiores a 96%, além de significância estatística relevante. As variáveis selecionadas (taxa de RSS sobre resíduos totais, despesa per capita com RSS, população total, taxa de mortalidade geral e o envio de resíduos para outros municípios) mostraram-se consistentes na explicação da variabilidade na geração de RSS, reforçando a estabilidade estrutural do modelo ao longo do tempo.

Os testes de validação cruzada com 5 dobras indicaram que os modelos mantêm desempenho estável e previsões confiáveis, mesmo em contextos amostrais distintos. A proximidade entre os valores de erro médio absoluto (MAE), erro quadrático médio (MSE) e o erro percentual médio (MAPE) entre os anos avaliados confirma a capacidade de generalização do modelo, reduzindo a dependência de um cenário específico para manutenção da acurácia preditiva.

A etapa de previsão *out-of-time*, utilizando os coeficientes de 2021 para prever os dados de 2022, confirmou a estabilidade do modelo proposto. A exclusão de quatro cidades com comportamento atípico, justificada por análises estatística, resultou em melhorias na aderência aos pressupostos do modelo. O desempenho do modelo preditivo manteve-se satisfatório, com R^2 acima de 94% e erros próximos aos obtidos nas validações cruzadas, reforçando sua aplicabilidade em análises preditivas.

A principal limitação do estudo diz respeito à ausência de dados específicos sobre a relevância regional de cada município no sistema de saúde, o que poderia explicar parte das discrepâncias observadas. Além disso, a análise foi limitada aos anos de 2021 e 2022, período ainda impactado pela pandemia de COVID-19, por falta da divulgação de dados mais recentes. Estudos futuros poderiam expandir a análise para outros períodos e incorporar variáveis relacionadas ao fluxo regional de pacientes ou à complexidade dos serviços prestados nos municípios.

Em síntese, os resultados obtidos indicam que o modelo desenvolvido possui potencial como ferramenta de apoio à gestão de RSS. Sua implementação pode auxiliar gestores públicos na alocação de recursos e na formulação de políticas voltadas à sustentabilidade e segurança sanitária. A continuidade de estudos nessa linha, considerando cidades com populações de outros tamanhos ou com a ampliação do intervalo temporal e a inclusão de novos preditores, poderá consolidar ainda mais a utilidade prática da modelagem estatística no contexto da gestão de resíduos em saúde pública.

6. REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). **NBR 10004**: Resíduos sólidos – Classificação. Rio de Janeiro, 2004. Disponível em: <https://analiticaqmresiduos.paginas.ufsc.br/files/2014/07/Nbr-10004-2004-Classificacao-De-Residuos-Solidos.pdf>. Acesso em: 01 mar. 2024.

BRASIL. Ministério da Saúde. Agência Nacional de Vigilância Sanitária. **Resolução RDC nº 222, de 28 de março de 2018**. Regulamenta as Boas Práticas de Gerenciamento dos Resíduos

de Serviços de Saúde, e dá outras providências. Brasília, 2018. Disponível em: <https://www.in.gov.br/web/dou/-/resolucao-rdc-n-222-de-28-de-marco-de-2018-8436194>. Acesso em 05 mar. 2024

BRASIL. Ministério do Meio Ambiente. Conselho Nacional do Meio Ambiente. **Resolução Conama nº 358**, de 29 de abril de 2005. Dispõe sobre o tratamento e a disposição final dos resíduos dos serviços de saúde e dá outras providências. Brasília, 2005. Disponível em: https://as.org.br/docs/Resolucao_CONAMA_358.pdf. Acesso em: 16 mar. 2024.

BRASIL. Ministério do Meio Ambiente. **Portaria Nº 280**, de 29 e junho de 2020. Brasília, 2020. Disponível em: <https://www.in.gov.br/en/web/dou/-/portaria-n-280-de-29-de-junho-de-2020-264244199>. Acesso em: 23 mar. 2024.

CALDEIRA, J. F.; AQUINO, D. S. Influência de indicadores socioeconômicos no consumo per capita de água para os municípios do estado da Bahia. **Colloquium Exactarum**, Presidente Prudente, v. 12, n. 1, p. 111-124, 15 set. 2020. Disponível em: <https://journal.unoeste.br/index.php/ce/article/view/3590/3046>. Acesso em: 10 mar. 2024.

FUNDAÇÃO SISTEMA ESTADUAL DE ANÁLISE DE DADOS (SEADE). **Seade Municípios**. 2024. São Paulo, 2024. Disponível em: <https://municipios.seade.gov.br/>. Acesso em: 02 abr. 2024.

GARLASCO, J.; CANEPARI, A.; GIACOBONE, G.; FUNICELLI, G.; COTRONEO, A.; ZOTTI, C. M. Impact of COVID-19 on healthcare waste generation: correlations and trends from a tertiary hospital. **European Journal of Public Health**, v. 31, out. 2021. Disponível em: <https://doi.org/10.1093/eurpub/ckab164.135>. Acesso em: 25 mar. 2024.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Censo Demográfico 2022**. Rio de Janeiro, 2024. Disponível em: <https://censo2022.ibge.gov.br/panorama/>. Acesso em 23 mar. 2024.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **O que é o PIB**. Rio de Janeiro, 2025. Disponível em: <https://www.ibge.gov.br/explica/pib.php>. Acesso em 04 abr. 2024.

IPEADATA. **Produto interno bruto (PIB) per capita**. Brasília, 2024. Disponível em: ipeadata.gov.br/ExibeSerie.aspx?serid=38375#:~:text=O%20PIB%20per%20capita%20é,nominal%20e%20a%20população%20residente. Acesso em: 04 abr. 2024.

JANIK-KARPINSKA, E.; BRANCALEONI, R.; NIEMCEWICZ, M.; WOJTAS, W.; FOCO, M.; PODOGROCKI, M.; BIJAK, A. Healthcare Waste—A Serious Problem for Global Health. **Healthcare**, v.11, n. 2, p. 242, Basel, 2023. Disponível em: <https://doi.org/10.3390/healthcare11020242>. Acesso em: 04 abr. 2024.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R.; TAYLOR, J. An Introduction to Statistical Learning: with Applications in Python. **Springer Texts in Statistics**, Cham, 2023. Disponível em: https://doi.org/10.1007/978-3-031-38747-0_1. Acesso em: 15 abr. 2025.

KAPOSI, A.; NAGY, A.; GOMORI, G.; KOCSIS, D. Analysis of healthcare waste and factors affecting the amount of hazardous healthcare waste in a university hospital. **Journal of Material Cycles and Waste Management**, v. 26, p. 1169–1180, 2024. Disponível em: <https://doi.org/10.1007/s10163-024-01890-1>. Acesso em: 25 mar. 2024.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: 14 th INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, Montreal, Quebec, Canada, August 20-25, 1995. **Proceedings of the 14th international joint conference on Artificial intelligence**, v. 2, p. 1137 -1143. Disponível em: <https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf>. Acesso em: 24 mar. 2025.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. 7 ed. Rio de Janeiro: LTC, 2021.

PISANI JUNIOR, R.; CASTRO, M. C. A. A. de.; COSTA, A. Á. da. Desenvolvimento de correlação para estimativa da taxa de geração *per capita* de resíduos sólidos urbanos no estado de São Paulo: influências da população, renda *per capita* e consumo de energia elétrica. **Engenharia Sanitaria e Ambiental**, v. 23, n. 2, p. 415–424, mar. 2018. Disponível em: <https://doi.org/10.1590/S1413-41522018167380>. Acesso em: 17 mar. 2024.

RAZALI, N. M.; WAH, Y. B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. **Journal of Statistical Modeling and Analytics**, v. 2, n. 1, p. 21–33, 2011. Disponível em: <https://www.researchgate.net/publication/267205556>. Acesso em: 14 maio 2025.

SILVA, T. T. S. da. **Resíduos de saúde provenientes de serviços de atenção domiciliar e suas implicações no ambiente e na saúde pública**. 2020. Dissertação (Mestrado em Engenharia Civil e Ambiental) – Universidade Federal de Pernambuco, Caruaru, 2020. Disponível em: <https://repositorio.ufpe.br/handle/123456789/38728>. Acesso em: 11 mar. 2024.

SISTEMA NACIONAL DE INFORMAÇÕES SOBRE SANREAMENTO (SNIS). Ministério das Cidades. **Série Histórica**. Brasília, 2024. Disponível em: <http://app4.mdr.gov.br/serieHistorica/residuosSolidos/index#>. Acesso em: 20 mar. 2024.

SOUZA, R. R.; TOEBE, M.; CHUQUEL MELLO, A. C.; BITTENCOURT, K. C. Sample size and Shapiro-Wilk test: an analysis for soybean grain yield. **European Journal of Agronomy**, v. 142, p. 126666, 2023. Disponível em: <https://doi.org/10.1016/j.eja.2022.126666>. Acesso em: 22 maio 2025.

UNITED NATIONS (UN). UN Development Programme. **Human Development Index (HDI)**, New York, 2024. Disponível em: <https://hdr.undp.org/data-center/human-development-index#/indicies/HDI>. Acesso em: 23 mar. 2024.

UNITED NATIONS (UN). UN Environment Programme. **Healthcare waste: what to do with it?** New York, 09 abr. 2020. Disponível em: <https://www.unep.org/news-and-stories/story/healthcare-waste-what-do-it>. Acesso em: 29 fev. 2024.

WORLD HEALTH ORGANIZATION (WHO). **Protecting Health Through Healthcare Waste Management**. Fact Sheet, Geneva, mar. 2016. Disponível em: [https://www.who.int/docs/default-source/wpro---documents/hae---regional-forum-\(2016\)/hcwmanagement-factsheet-rfhe.pdf?sfvrsn=6d59d7e_2%20%E2%80%93](https://www.who.int/docs/default-source/wpro---documents/hae---regional-forum-(2016)/hcwmanagement-factsheet-rfhe.pdf?sfvrsn=6d59d7e_2%20%E2%80%93). Acesso em: 16 mar. 2024.