

1067 - A ADAPTAÇÃO DE BIOMASSA GRANULAR ANAMMOX PARA O PÓS-TRATAMENTO DE EFLUENTE ANAERÓBIO DE ESGOTO SANITÁRIO DOMÉSTICO É FACTÍVEL?

Tobias Ribeiro Barreto⁽¹⁾

Engenheiro Ambiental (UNESP) e mestre em Hidráulica e Saneamento (USP). Foi consultor ambiental em São José do Rio Pardo e professor universitário em Alfenas (UNIFENAS), Guaxupé (UNIFEG) e Ribeirão Preto (Estácio). Doutorando em Hidráulica e Saneamento (USP), com realização de doutorado sanduíche na Universidade de Tecnologia de Delft (TUD), Holanda.

Andre Lermontov⁽²⁾

Engenheiro Químico e mestre em Engenharia Química (UNIFAL). Engenheiro de Segurança do Trabalho (UFSJ). Doutorando em Hidráulica e Saneamento (USP) e professor universitário (UNIFEG).

Endereço⁽¹⁾: Avenida João Dagnone, 1100 Jardim - Santa Angelina - São Carlos - São Paulo - 13563-120 - Brasil - Tel: +55 (16) 99168-8452 - Fax: +55 (16) 3373-9552 e-mail: glegiglio@usp.br

RESUMO

As Cidades Inteligentes (CI) visam aprimorar a prestação de serviços públicos por meio da integração de tecnologias inovadoras, focadas no planejamento, desenvolvimento, operação e governança urbana. Indicadores relacionados ao Saneamento Básico, como o volume de esgoto tratado, o consumo médio per capita de água e as perdas na distribuição, representam uma das principais dimensões de avaliação de uma CI. Nesse contexto, a previsão do volume de água tratada a ser distribuída é crucial para o planejamento estratégico das entidades responsáveis pelo saneamento básico nas cidades, impactando diretamente na qualidade de vida dos cidadãos. Este artigo tem como objetivo realizar uma análise empírica comparando diferentes modelos de séries temporais para a previsão do volume de água distribuída em um município do estado do Rio de Janeiro, Brasil. Os dados utilizados foram extraídos e processados a partir do Sistema Nacional de Informações sobre Saneamento Básico (SNIS).

PALAVRAS-CHAVE: *Previsão, Série Temporal, Saneamento Básico, Água*

INTRODUÇÃO

Com a crescente urbanização, as cidades vêm buscando melhorar a prestação de serviços por meio da implementação de novas tecnologias no planejamento, desenvolvimento, operação e governança das cidades, criando assim as chamadas Cidades Inteligentes [CI], também conhecidas como “Smart Cities” (Cruz et al. 2022). O conceito de Cidades Inteligentes ganhou relevância, em especial na última década (Banni et al. 2022).

Um aspecto essencial das cidades inteligentes é a utilização de dados e tecnologias de comunicação para monitorar e gerenciar diversos aspectos da vida urbana, incluindo transporte, energia, meio ambiente, segurança pública e governança. As cidades inteligentes são caracterizadas pela aplicação de tecnologias de informação e comunicação em todas as áreas da vida urbana, promovendo uma governança mais eficaz, uma economia mais dinâmica e uma maior qualidade de vida para os habitantes (Giffinger et al. 2007). As CI são caracterizadas pela criação de soluções urbanas entre os diversos atores da cidade, incluindo governos, empresas, organizações da sociedade civil e cidadãos. Isso possibilita uma abordagem mais colaborativa e democrática para a gestão urbana, onde os cidadãos têm um papel ativo na definição das políticas e na resolução dos problemas locais (Anthopoulos 2017).

Programas e projetos de CI buscam desenvolver soluções inovadoras para mitigar problemas urbanos e, assim, promover uma melhor qualidade de vida para os cidadãos e a sociedade. Os centros urbanos frequentemente enfrentam desafios estruturais que demandam manutenção contínua (Zhang, 2003). Esses problemas podem ser apontados pela população por meio de canais de comunicação diversos, participação em conselhos e fóruns municipais que recomendam ações para o setor executivo. No contexto do saneamento, as organizações responsáveis pelo tratamento e distribuição de água monitoram o volume disponibilizado para o município, composto pela água fornecida pelas estações de tratamento e reservatórios ao longo do sistema de distribuição. Além de analisar dados históricos, prever o volume de água disponibilizado representa um impacto financeiro e ambiental significativo. Para realizar essa previsão, um modelo de série temporal é utilizado.

Uma série temporal consiste em uma sequência de observações da mesma natureza, coletadas em intervalos de tempo regulares. Esse processo de previsão requer um valor de entrada (variável original), que é processado por um modelo preditivo para gerar um valor de saída (variável predita). A precisão da previsão é então avaliada por meio de métricas específicas, que analisam os erros e a confiabilidade do modelo. No contexto de previsão de demanda de água, técnicas de aprendizado de máquina têm sido cada vez mais aplicadas. Por exemplo, o trabalho de Sajjanshetty et al. (2023) explora o uso de ensembles, que combinam modelos de aprendizado de máquina e modelos estatísticos para prever o consumo de água de uma comunidade. Os autores empregaram algoritmos como SARIMAX, Prophet e XGBoost, sendo os dois primeiros de natureza estatística e o último, um algoritmo de aprendizado de máquina.

Há também pesquisas que utilizam o algoritmo Prophet em combinação com redes neurais convolucionais profundas e mecanismos de atenção para prever a demanda de água em áreas urbanas (Liu et al., 2024). Esses modelos avançados, como o “Long Short-Term Memory” [LSTM], contribuem para aprimorar a capacidade preditiva em contextos em que a complexidade e a variabilidade dos dados são elevadas. Assim, um passo inicial importante é avaliar a capacidade do Prophet para prever a demanda de água em cidades brasileiras, estabelecendo uma base de comparação para o desenvolvimento futuro de métodos preditivos mais sofisticados. O objetivo deste artigo é, portanto, apresentar uma análise experimental que explore a previsão de demanda de água potável.

O objetivo deste artigo é apresentar uma análise experimental com o objetivo de realizar uma previsão de demanda de água potável, para isso 4 modelos de série temporal – Prophet, SARIMAX, Theta e STLwES serão aplicados e seus resultados serão comparados.

OBJETIVOS

Este artigo tem como objetivo realizar uma análise empírica comparando diferentes modelos de séries temporais para a previsão do volume de água distribuída em um município do estado do Rio de Janeiro, Brasil. Os dados utilizados foram extraídos e processados a partir do Sistema Nacional de Informações sobre Saneamento Básico (SNIS).

METODOLOGIA UTILIZADA

Neste trabalho, o dataset é composto por uma variável independente (tempo) e uma variável dependente (volume de água distribuída). Cada linha do conjunto de dados representa o volume de água que passou em um determinado ponto em uma janela de tempo de 1 dia. O Volume de água se encontra representado em metros cúbicos (m³). O conjunto de dados cobre o período de 09/07/2023 às 06:30:00 até 15/12/2023 às 01:20:00, totalizando 27.401 registros. Dessa forma, foi aplicada uma metodologia de análise definida em 3 etapas:

Na Etapa 01, etapa de obtenção dos dados, o conjunto foi gerado Sistema Nacional de Informações do Saneamento Básico .

Na Etapa 02 foi aplicado o script desenvolvido em Python 3.11. Esse script realiza todo o processo dessa etapa em diante. A seguir é apresentado as técnicas de tratamento dos dados utilizadas nesse estudo. O Algoritmo é responsável por realizar o pré-processamento dos dados tratando os dados da seguinte forma: A série foi agregada para a frequência diária. Valores inferiores e superiores a 3 desvios padrões foram substituídos pela média aritmética da série. Valores acima de 87.000 m³ e inferiores a 83.000 m³, ambas as regras foram elaboradas após análise de um especialista em distribuição de água. Posteriormente a série será dividida em série de treino (contendo 2/3 dos dados) e a série de teste (contendo os restantes 1/3 dos dados). Em seguida o treinamento utilizando diferentes modelos é iniciado e o resultado é discutido na etapa seguinte.

Na Etapa 3 foi realizada uma análise comparativa entre os modelos de séries temporais testadas, utilizando a métrica Erro Percentual Médio Absoluto Simétrico Negativo [NSMAPE].

O script também é responsável por importar os dados coletados, transformar e agregar os dados em intervalos de tempo adequados para análise, separar a base de treino e de teste, treinar o modelo e avaliar seus resultados. Nesse estudo, com objetivo de realizar a análise, o mesmo dataset foi submetido aos seguintes modelos Prophet, SARIMAX, Theta e STLwES. Um breve resumo das principais características de cada modelo é apresentado a seguir:

O Prophet é um modelo de série temporal desenvolvido com foco em previsão de dados com sazonalidade e tendência, e é especialmente eficaz em séries com padrões complexos e irregulares. Sua abordagem baseia-se na decomposição aditiva da série em três componentes principais: tendência, sazonalidade e feriados (ou eventos especiais). Cada componente é modelado separadamente, permitindo ao Prophet capturar padrões sazonais e tendências de longo prazo com flexibilidade.

A decomposição aditiva torna o modelo intuitivo, pois a série temporal é considerada como uma soma desses componentes, o que simplifica o ajuste de dados não estacionários e facilita a detecção de padrões repetitivos e variações sazonais (Taylor e Letham, 2018). Um dos aspectos mais notáveis do Prophet é sua capacidade de lidar com dados incompletos e com outliers de maneira robusta. A metodologia do Prophet incorpora técnicas de suavização e modelagem estatística que permitem que o modelo seja menos sensível a valores ausentes e a variações extremas na série, adaptando-se automaticamente ao comportamento dos dados históricos.

Essa robustez é possível porque o Prophet assume uma distribuição sazonal de Fourier e uma função linear ou logística para a tendência. Com isso, o modelo detecta variações de longo prazo e sazonais sem a necessidade de ajustes constantes nos parâmetros, tornando-o altamente eficiente para séries temporais com alta variabilidade (Zeng et al., 2020). O Prophet também permite ajustes finos para o controle da sazonalidade e da tendência, facilitando o ajuste a séries temporais com ciclos sazonais irregulares ou variáveis.

Além disso, ele integra o uso de feriados ou eventos customizados que podem impactar os dados, como datas comemorativas ou mudanças nas políticas de uma organização. O modelo permite ao usuário adicionar intervalos de confiança, essencial para lidar com incertezas na previsão. Essa flexibilidade torna o Prophet uma ferramenta poderosa e adequada para séries temporais com diferentes níveis de complexidade, principalmente em contextos em que as séries apresentam padrões de sazonalidade não tradicionais (Candanedo et al., 2021).

O modelo “Seasonal Auto-Regressive Integrated Moving Average with Exogenous Regressors” [SARIMAX] é uma extensão avançada do modelo ARIMA, que é amplamente utilizado para modelar séries temporais com sazonalidade e variáveis exógenas. A estrutura do SARIMAX combina componentes autorregressivos (AR), de média móvel (MA) e integração (I) para capturar padrões de dependência temporal, e adiciona uma camada sazonal

que permite o ajuste de ciclos repetitivos em intervalos regulares.

A presença de variáveis exógenas permite que o modelo SARIMAX integre variáveis externas ao processo de previsão, o que é útil em casos em que o comportamento da série temporal é influenciado por fatores adicionais, como condições econômicas ou eventos externos (Box et al., 2015). O funcionamento do SARIMAX baseia-se na aplicação de parâmetros sazonais e de dependência temporal para capturar tendências e padrões em séries não estacionárias. Ele utiliza uma decomposição multiplicativa ou aditiva da série, dependendo da natureza dos dados, para modelar a sazonalidade.

A tendência é capturada pelo termo de integração (I), que transforma séries temporais não estacionárias em estacionárias por meio de diferenciação. Ao incorporar o termo sazonal, o SARIMAX consegue ajustar sazonalidades de diferentes frequências, como padrões anuais ou mensais, tornando-o robusto para séries com sazonalidades complexas (Hyndman e Athanasopoulos, 2018). Além disso, o SARIMAX permite prever o impacto de variáveis exógenas, que podem ter influência significativa sobre o comportamento da série temporal.

Variáveis exógenas são fatores externos que podem ser incluídos no modelo, permitindo que o SARIMAX capture relações causais que influenciam diretamente a variável de interesse. Este aspecto é útil em setores como economia, finanças e energia, onde a previsão precisa levar em conta variações nas condições de mercado ou fatores climáticos. Com essa capacidade, o SARIMAX se torna uma ferramenta poderosa para séries temporais que exigem ajuste sazonal e consideração de fatores externos, oferecendo previsões mais acuradas em contextos em que múltiplas variáveis interagem com a série temporal principal (Lütkepohl, 2005).

O modelo Theta é uma técnica de previsão de séries temporais que combina suavização exponencial e decomposição para capturar tendências e padrões sazonais de maneira robusta. Originalmente introduzido por Assimakopoulos e Nikolopoulos (2000), o modelo Theta transforma a série temporal em várias "linhas Theta", que representam diferentes versões suavizadas da série original.

A abordagem mais comum utiliza duas linhas Theta: uma linha de tendência linear e uma linha de suavização, combinando-as para obter uma previsão final. Esse processo permite que o modelo capture tanto a tendência de longo prazo quanto as variações sazonais, tornando-o especialmente eficaz para séries com sazonalidade menos rígida e tendências sutis (Assimakopoulos e Nikolopoulos, 2000). A essência do modelo Theta está na decomposição da série em componentes de tendência e sazonalidade, ajustando as linhas Theta com base em parâmetros de suavização específicos. O componente de tendência é frequentemente modelado com uma linha reta, enquanto o componente sazonal é ajustado utilizando técnicas de suavização exponencial.

Ao combinar essas duas componentes, o modelo Theta consegue capturar variações de curto prazo e tendências de longo prazo, sendo muito eficaz para dados que apresentam mudanças graduais ou intermitentes. A simplicidade e a eficiência do modelo o tornam uma escolha popular em aplicações onde a série temporal apresenta padrões de mudança não lineares e oscilatórios (Petropoulos e Kourentzes, 2015).

Além disso, o modelo Theta é frequentemente considerado uma extensão do método de suavização exponencial, porém com um componente adicional que permite ajustar a série para diferentes ângulos (linhas Theta), o que é particularmente útil para capturar padrões sazonais e variações que não seguem uma estrutura rígida. Esta flexibilidade faz do modelo Theta uma opção viável para séries temporais complexas, onde a sazonalidade e a tendência variam de forma menos previsível.

A eficácia do modelo foi demonstrada em várias competições de previsão, como a M3-Competition, onde o Theta foi um dos modelos mais precisos para séries temporais de frequência anual, destacando-se por sua simplicidade e precisão (Makridakis et al., 2018).

O modelo STLwES (Seasonal-Trend decomposition using LOESS with Exponential Smoothing) combina a decomposição de séries temporais com suavização exponencial para capturar padrões de tendência e sazonalidade. A decomposição STL separa a série temporal em três componentes principais: tendência, sazonalidade e ruído. A suavização exponencial (ES) é então aplicada às componentes de tendência e sazonalidade, permitindo que o modelo reaja rapidamente a mudanças nos padrões de dados.

Esse processo é útil para séries temporais com sazonalidade complexa e flutuações não lineares (Cleveland et al., 1990). A decomposição STL é especialmente eficaz porque utiliza um método de suavização não paramétrico (LOESS), que pode capturar padrões sazonais não lineares. Ao aplicar a suavização exponencial aos componentes extraídos, o modelo STLwES consegue prever séries temporais com tendências dinâmicas e sazonalidade flexível. A suavização exponencial, por sua vez, reage rapidamente a mudanças recentes nos dados, o que é particularmente útil para séries com sazonalidade variável ao longo do tempo, como séries de vendas sazonais (Hyndman e Athanasopoulos, 2018). A combinação de STL e suavização exponencial permite que o STLwES produza previsões confiáveis em séries temporais complexas, onde a estrutura de sazonalidade e tendência varia de forma imprevisível. Essa abordagem torna o STLwES robusto para diferentes tipos de séries, especialmente aquelas com sazonalidades irregulares ou que não seguem um padrão fixo. Este modelo é amplamente usado em setores como economia e clima, onde as séries temporais apresentam variações sazonais dinâmicas (Macaulay et al., 2019).

Para avaliar qual é o modelo que melhor se aplica nesse estudo será usado o Erro Percentual Médio Absoluto

Simétrico Negativo. É uma métrica de avaliação usada para medir a precisão das previsões em séries temporais e outras análises de dados. Essa métrica é uma variação do SMAPE (Symmetric Mean Absolute Percentage Error), que foi desenvolvida para solucionar problemas do MAPE (Mean Absolute Percentage Error), especialmente em séries temporais. O SMAPE calcula o erro percentual médio absoluto de forma simétrica, evitando a supervalorização de grandes erros em valores pequenos e promovendo uma avaliação balanceada entre as previsões e os valores reais (Makridakis, Spiliotis, e Assimakopoulos, 2018).

A versão negativa dessa métrica, facilita a comparação de desempenho, onde valores maiores indicam melhores previsões, uma vez que o valor negativo foi incluído para uniformizar o uso em modelos onde o maior valor indica o melhor desempenho. Essa métrica reduz o impacto de erros absolutos extremos que podem ocorrer quando os valores reais estão próximos de zero, problema comumente visto em setores como finanças, onde valores baixos ou instáveis podem gerar valores de erro desproporcionalmente altos.

Ao adotar o SMAPE, evitam-se os vieses das métricas tradicionais, proporcionando uma avaliação mais justa e informativa das previsões (Hyndman e Athanasopoulos, 2018). A escolha dessa métrica para avaliar séries temporais complexas é justificada pela sua capacidade de fornecer uma avaliação equilibrada entre diferentes magnitudes de dados, favorecendo previsões onde a precisão percentual é crucial para a tomada de decisões. Essa métrica tem sido útil em contextos industriais e comerciais, pois evita o viés gerado por grandes erros relativos e promove uma avaliação mais estável e adaptada a diferentes tipos de dados (Petroopoulos et al., 2014).

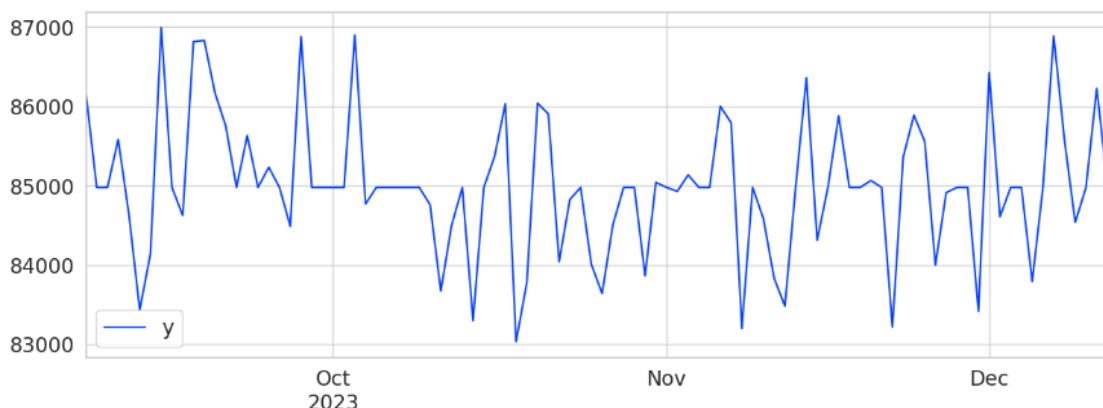
Com o objetivo de analisar o comportamento da série temporal de volume disponibilizado de água tratada no município estudado, utilizou-se os dados diários do período de 08/09/2023 a 14/12/2023, formando uma série de aproximadamente 4 meses, com 27.401 observações correspondentes a um período de 5 minutos cada, essas informações foram agregadas a nível de dia, ficando com 99 observações, conforme a Tabela 1.

Tabela 1. Amostra do dataset utilizado

Indice	ds	Y (m³)
1	2023-09-08	86.160
2	2023-09-09	84.977
3	2023-09-10	84.977
...
96	2023-12-11	86.224
97	2023-12-12	85.032
98	2023-12-13	84.776
99	2023-12-14	83.448

Fonte: Elaborado pelo autor

Para a coluna 'ds', temos o carimbo de data, no formato ano, mês e dia. Na coluna 'y' temos o valor do volume de água disponibilizada, em metros cúbicos. Para uma análise prévia gerou-se, inicialmente, um gráfico da série original. Essa primeira observação é importante para identificar possíveis discrepâncias ou erros que possam existir na extração dos dados, bem como entender, superficialmente, como é o comportamento da curva em relação ao tempo estudado. Após aplicação dos tratamentos de pré-processamento de dados, descritos na seção de método desse trabalho, o novo gráfico da série foi gerado, permitindo analisar se ainda existe possíveis erros. A série assume um comportamento estacionário, pois não apresenta uma tendência bem definida e picos muito elevados. Os dados da série foram observados:



Fonte: Elaborado pelo autor

Seguindo as etapas da metodologia, o conjunto de dados com periodicidade de 1 dia foi dividido em 2 subconjuntos. O subconjunto de treino contendo 2/3 das observações, e o restante (1/3) será o utilizado para realização dos testes e validações.

Tabela 2. Amostra do dataset de Treino

Indice	ds	Y (m³)
1	2023-09-08	86.160
2	2023-09-09	84.977
3	2023-09-10	84.977
...
62	2023-11-10	83.200
63	2023-11-09	84.977
64	2023-11-10	84.584
65	2023-11-11	83.832

Fonte: Elaborado pelo autor

A série de treino apresenta uma média de 84.988,43 e um desvio padrão de 891,02, indicando uma variação moderada em torno da média, com a maioria dos valores concentrados em uma faixa relativamente estreita. O valor mínimo registrado foi 83.032, enquanto o máximo foi 86.992, sugerindo um intervalo de aproximadamente 3.960 unidades entre os extremos. Os quartis apontam para uma distribuição centralizada: 25% dos valores estão abaixo de 84.624 e 75% estão abaixo de 85.232. O valor mediano (50%) é de 84.977,02, bastante próximo da média, o que sugere uma distribuição simétrica e indica que a série não apresenta desvios acentuados ou outliers significativos.

Tabela 3. Amostra do dataset de Teste

Indice	ds	Y (m³)
66	2023-11-12	83.480
67	2023-12-06	84.977
68	2023-12-07	86.360
...

96	2023-12-11	86.224
97	2023-12-12	85.032
98	2023-12-13	84.776
99	2023-12-14	83.448

Fonte: Elaborado pelo autor

A série de teste possui uma média de 84.954,52 e um desvio padrão de 883,72, indicando uma variação relativamente baixa em torno da média. O valor mínimo é 83.216 e o máximo é 86.888, resultando em uma amplitude de 3.672 unidades entre os extremos. Os quartis mostram uma distribuição bastante centralizada: 25% dos valores estão abaixo de 84.608 e 75% estão abaixo de 85.360. A mediana, de 84.977,02, está muito próxima da média, sugerindo uma distribuição simétrica e indicando que a série é homogênea, sem outliers significativos. Esses dados refletem uma série com variações limitadas e centralizadas ao redor da média, apontando para uma consistência nos valores.

A aplicação do algoritmo foi realizada com o objetivo de identificar o modelo mais adequado para a previsão da série temporal em análise. Para avaliar a precisão dos modelos, utilizou-se a métrica de Erro Percentual Médio Absoluto Simétrico Negativo. Essa métrica permite uma avaliação equilibrada dos modelos, onde valores menores indicam previsões mais precisas, pois refletem um menor desvio entre os valores previstos e os reais. Na tabela a seguir é apresentado o resultado da métrica de avaliação para cada modelo:

Tabela 4. Avaliação entre os modelos

Modelo	Erro Percentual Médio Absoluto
	Simétrico Negativo
Prophet	-0,2711
SARIMAX	-0,2534
Theta	-0,2978
STLwES	-0,3000

Fonte: Elaborado pelo autor

Ao comparar os resultados dos modelos utilizando a métrica de Erro Percentual Médio Absoluto Simétrico Negativo, observa-se que o modelo STLwES apresentou o pior desempenho, com um valor de erro de -0,300, seguido pelo modelo Theta com -0,2978. Já o Prophet e o SARIMAX, obtiveram valores de -0,2711 e -0,2534, respectivamente. Esses resultados indicam que o modelo SARIMAX foi o mais eficaz em capturar os padrões da série temporal analisada, resultando em previsões mais precisas e com menor erro percentual médio. A vantagem do STLwES e do Theta em relação aos demais modelos podem estar relacionada à sua abordagem de decomposição e suavização, que lida de forma robusta com componentes de sazonalidade e tendência. O Prophet e o SARIMAX, mostraram-se mais adequados para essa série específica, possivelmente devido às suas características em capturar algumas variações sazonais e de tendência. Sendo o SARIMAX o modelo com o melhor resultado nessa avaliação. A análise estatística dos dados revelou uma média de 82.609,54 m³ e um desvio padrão de 1.339,76 indicando uma dispersão moderada dos valores em torno da média. Para verificar a normalidade da distribuição dos resíduos, foi aplicado o teste de Shapiro-Wilk, que apresentou uma estatística de 0,9235 e um p-valor de 0,0512. Como o p-valor é inferior ao nível de significância adotado ($\alpha = 0,05$), rejeita-se a hipótese alternativa, indicando que os dados seguem uma distribuição normal. Esse resultado sugere que a distribuição apresenta aproximação com a normalidade.

Fazendo a previsão para os mesmos períodos do da série de teste, utilizando o modelo SARIMAX temos:

Tabela 5. Amostra da previsão modelo SARIMAX

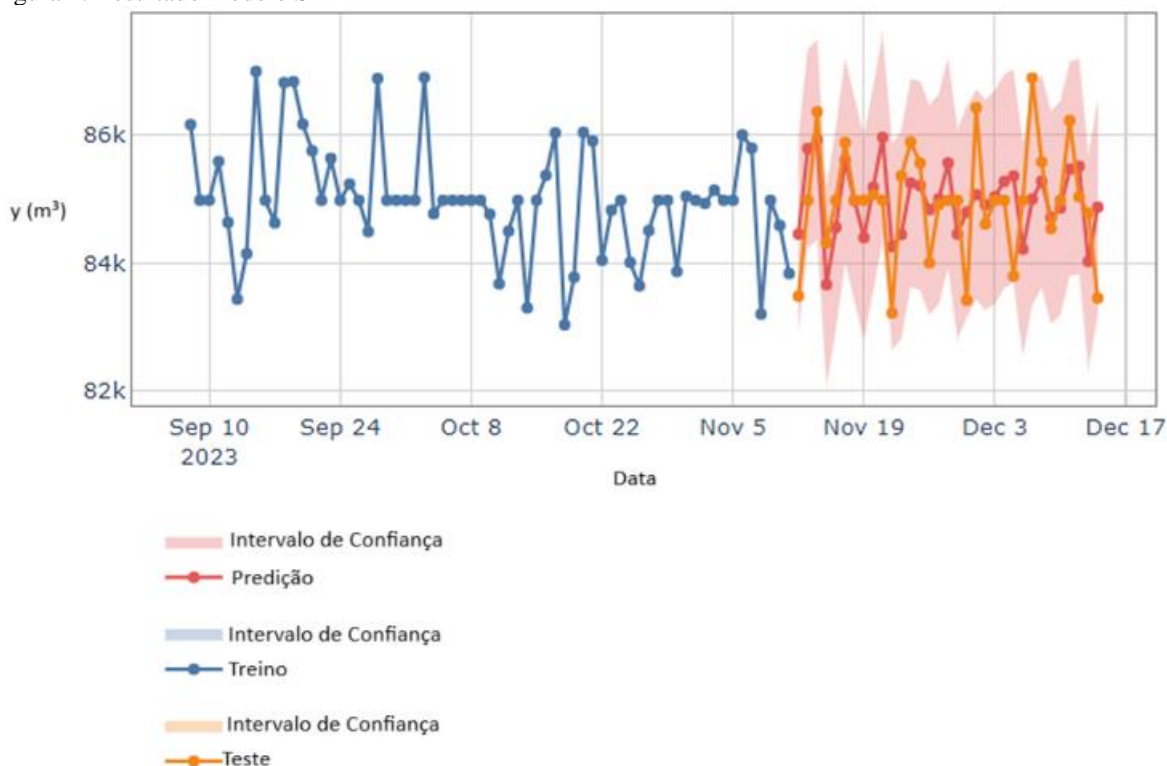
Índice	ds	Y (previsto)	Limite Inferior	Limite Superior
66	2023-11-12	84.448,18	82.913,22	85.983,13

67	2023-11-13	85.786,312	84.228,63	87.343,99
68	2023-11-14	85.923,26	84.364,16	87.482,37
...
96	2023-12-11	85.467,68	83.788,42	87.146,93
97	2023-12-12	85.508,50	83.828,75	87.188,25
98	2023-12-13	84.022,90	82.342,06	85.703,75
99	2023-12-14	84.869,92	83.188,57	86.551,27

Fonte: Elaborado pelo autor

Para melhor visualização, é apresentado um gráfico de linha contendo o resultado obtido pelo modelo SARIMAX, comparando a série de teste e o resultado da previsão:

Figura 2. Resultado modelo SARIMAX



Fonte: Elaborado pelo autor

A análise da coluna y revela um valor médio de 84.964,37, com um desvio padrão de 545,85, indicando uma variação relativamente baixa em torno da média e sugerindo uma consistência nos valores observados. Os valores de y variam de um mínimo de 83.659,91 a um máximo de 85.961,05, com os quartis 25% e 75% posicionados em 84.552,55 e 85.275,69, respectivamente, o que demonstra uma distribuição concentrada e relativamente simétrica em torno da média apresenta intervalos de confiança que fornecem estimativas de limite inferior e superior para cada valor de y, com médias de 83.331,86 e 86.596,88, respectivamente. Esses intervalos indicam que, com alta probabilidade, os valores de y devem estar dentro dessa faixa, garantindo maior confiabilidade nas previsões. A diferença entre os limites inferior e superior de cada intervalo também está em torno de 2.265 unidades, o que reforça a precisão dos dados observados e sua estabilidade, tornando-os adequados para uso em análises preditivas ou modelos de séries temporais.

Estes resultados põem em evidência a aplicação de modelos preditivos com menor risco de erros significativos para a área de negócio. A consistência dos valores dentro da faixa dos intervalos de confiança sugere que os dados

podem ser utilizados de forma confiável para projeções, aumentando a robustez das análises e suportando decisões estratégicas baseadas em previsões

CONCLUSÕES E RECOMENDAÇÕES

Nesse trabalho, o objetivo foi apresentar uma análise experimental realizada para previsão de demanda de água potável. Para isso foi utilizada uma comparação de 4 modelos de série temporal, conforme descrito no método da pesquisa, e, posteriormente foi realizada análise da métrica de avaliação das predições. Embora as diferenças entre os modelos sejam sutis, o SARIMAX demonstra consistentemente um desempenho ligeiramente superior. No contexto de CI, as previsões de demanda de água podem beneficiar significativamente a gestão urbana ao apoiar a implementação de sistemas inteligentes de monitoramento e análise de dados.

Por exemplo, sensores instalados em redes de distribuição de água podem fornecer dados contínuos sobre o consumo e a pressão da água, permitindo ajustes rápidos e precisos na operação das estações de tratamento. Com esses dados, é possível otimizar progressivamente os modelos preditivos de água disponibilizada. Para uma organização, as análises de previsão de demanda de água funcionam como ferramentas de suporte e auxílio na tomada de decisões durante a operação das estações de tratamento de água. Além disso, esse tipo de previsão possibilita uma alocação mais eficiente dos produtos químicos utilizados no tratamento da água e um melhor uso de energia elétrica nessas instalações, uma vez que a organização poderá prever, com antecedência, o volume de água que deverá ser tratado. Para o cidadão, ter acesso a esse tipo de previsão aumenta a transparência no acompanhamento dos serviços de abastecimento de água.

Para trabalhos futuros, recomenda-se a comparação com outros modelos de séries temporais e algoritmos de aprendizado de máquina, como XGBoost. Outro ponto importante é buscar uma volumetria de dados maior do que a aplicada neste estudo, uma vez que conjuntos de dados mais amplos aumentam a probabilidade de capturar tendências necessárias para criar previsões de maior qualidade. Além disso, a inclusão de uma tabela de feriados e datas importantes pode resultar em alguma melhoria no modelo Prophet, a elaboração desse tipo de tabela e a inclusão dela no modelo poderia gerar resultados melhores. Dessa forma, outras agregações de frequência também podem ser exploradas, como frequências diárias e semanais. Tais frequências não foram abordadas neste trabalho devido ao baixo volume de dados disponível para essas periodicidades.

REFERÊNCIAS BIBLIOGRÁFICAS

- Anthopoulos, L. 2017. Smart utopia vs smart reality: Learning by experience from 10 smart city cases. *Cities*, 63:128–148.
- Assimakopoulos, V.; Nikolopoulos, K. 2000. The Theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, 16: 521-530.
- Banni, M.; Rosseti, I.; de Oliveira, D. 2022. Hurricane: um serviço para gerência de dados de aplicações de cidades inteligentes. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*: 151–163.
- Box, G. E.; Jenkins, G. M.; Reinsel, G. C.; Ljung, G. M. 2015. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.
- Candanedo, L. M.; Beltrán, C. A.; Rosado, E. 2021. Time series forecasting with Prophet: An application to energy consumption data. *IEEE Access*, 9:26302-26312.
- Chen, J.; Li, Y.; Chen, H. 2019. Forecasting electricity consumption using STL decomposition and ARIMA model. *Energy Reports*, 5:961-969.
- Cleveland, R. B.; Cleveland, W. S.; McRae, J. E.; Terpenning, I. 1990. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 63-73.
- Cruz, M. M.; Bernardini, F. C.; Nunes, V. T.; Bastos, C. A. 2022. Uma análise empírica dos planejamentos estratégicos das capitais brasileiras com foco na construção de cidades inteligentes. In *Anais do X Workshop de Computação Aplicada em Governo Eletrônico*, 157–167.
- Giffinger, R.; Fertner, C.; Kramar, H.; Meijers, E., et al. 2007. City-ranking of european medium-sized cities. *Cent. Reg. Sci. Vienna UT*. 1:1–12.
- Liu, C.; Liu, Z.; Yuan, J.; Wang, D.; Liu, X. 2024. Urban water demand prediction based on attention mechanism graph convolutional network-long short-term memory. *Water*, 16:831.
- Lütkepohl, H. 2005. *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media.

- Macaulay, R.; Al-Hussaini, L.; Meader, N. 2019. Time series forecasting in health care using STL decomposition and exponential smoothing. *Health Informatics Journal*, 25(3):584-597.
- Makridakis, S.; Spiliotis, E.; & Assimakopoulos, V. 2018. The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34:802-808.
- Petropoulos, F.; Kourentzes, N. 2015. Forecast combinations for intermittent demand. *Journal of the Operational Research Society*, 66:914-924.
- Petropoulos, F.; Nikolopoulos, K.; Assimakopoulos, V. 2014. An evaluation of the theta model in the M3-Competition. *International Journal of Forecasting*, 30:494-502.
- Sajjanshetty, A. S.; Jayanth, V.; Mohan, R.; Pahari, S.; Deepti, C. 2023. Estimation of community water consumption using multivariate ensemble approach. In *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, 1: 1–5.
- Taylor, S. J.; Letham, B. 2018. Forecasting at scale. *The American Statistician*, 72:37-45.
- Zeng, Q.; Liu, J.; Ma, Z. 2020. A deep learning-based forecasting model for time series data using Prophet and LSTM. *Applied Sciences*, 10:1-12.