Aplicando Decomposição de Valores Singulares na Predição de Vazão de Rede

Maria C. M. M. Ferreira[1], **Maria L. Linhares**[1], **Thelmo P. Araújo**[1], **Rafael L. Gomes**[1]

[1]Universidade Estadual do Ceará (UECE), Fortaleza, Ceará, Brasil.

{clara.mesquita,malu.linhares}@aluno.uece.br

{thelmo.araujo,rafa.lopes}@uece.br

Resumo. *Diversas empresas e provedores de internet (ISPs) realizam serviços de monitoramento de rede para compreender o comportamento da rede e obter dados relevantes para o planejamento estratégico. No entanto, falhas podem ocorrer durante a realização dessas medições, principalmente nas medições de vazão, dificultando a execução de soluções de predição do desempenho da rede. Dentro deste contexto, este artigo propõe uma abordagem que combina imputação de dados com predição para melhorar a qualidade das análises de vazão. A solução baseia-se na técnica de decomposição de valores singulares (SVD) para realizar a imputação de dados, explorando aspectos sazonais das séries temporais que ditam o desempenho de vazão da rede. Após a imputação, as séries são utilizadas em redes neurais recorrentes para predições com alta eficiência. Nos experimentos realizados, usando dados reais do Serviço de Monitoramento da Rede Ipê (Monipê), foi validada a eficácia do modelo proposto, com eficiência superior a soluções existentes.*

Abstract. *Several companies and Internet Service Providers (ISPs) perform network monitoring services to understand network behavior and obtain relevant data for strategic planning. However, failures may occur during these measurements, particularly in throughput measurements, hindering the implementation of network performance prediction solutions. In this context, this article proposes an approach that combines data imputation with prediction to enhance the quality of throughput analysis. The solution is based on the Singular Value Decomposition (SVD) technique for univariate data imputation, leveraging the seasonal aspects of time series that influence network throughput performance. After imputation, the series are used in recurrent neural networks for highly efficient predictions. Experiments conducted with real data from the Ipê Network Monitoring Service (Monipê) validated the proposed model's effectiveness, demonstrating superior efficiency compared to existing solutions.*

1. Introdução

Empresas e Provedores de Serviço de Internet (ISPs) contam com ferramentas avançadas de monitoramento de rede, projetadas para fornecer informações essenciais à gestão e ao planejamento estratégico. Essas ferramentas realizam testes regulares em diversas métricas de desempenho, cuja análise eficaz transforma dados em conhecimento estratégico [Vasileva et al. 2021, Silveira et al. 2023b]. A experiência do usuário na internet está diretamente ligada à qualidade da conexão, sendo a vazão um fator determinante

[Pinheiro et al. 2011, Silva et al. 2022]. A demanda variável por recursos, influenciada por fatores como a hora do dia e a localização dos usuários, exige um ajuste constante na infraestrutura da rede para garantir a qualidade do serviço [Gomes et al. 2014a]. Embora o alto desempenho da rede seja fundamental, a qualidade dos dados coletados para sua análise é igualmente crucial. As ferramentas de monitoramento, apesar de essenciais, apresentam limitações que podem resultar em dados ausentes ou com ruídos, comprometendo a análise e a interpretação dos resultados. Essa situação afeta negativamente a aplicação de técnicas de predição de dados que exploram o comportamento temporal dos dados [Silveira et al. 2023a].

Uma abordagem para mitigar esses problemas é a aplicação de técnicas de pré-processamento e imputação de dados sintéticos [Naf et al. 2022]. Ao preencher as lacunas nos dados e corrigir inconsistências, essas técnicas permitem uma análise mais precisa e completa do desempenho da rede, possibilitando a identificação de padrões, a detecção de anomalias e a previsão de falhas [Gomes et al. 2014b, Silva et al. 2023]. Contudo, técnicas simples de imputação são pouco eficazes para séries temporais porque ignoram características estatísticas fundamentais, como dependência temporal, sazonalidade e tendências. Essas técnicas desconsideram padrões sequenciais e dinâmicas complexas, gerando estimativas inconsistentes, especialmente em dados irregulares ou com *outliers* [Park et al. 2023]. Em trabalhos anteriores dos autores [Portela et al. 2024b, Ferreira et al. 2024, Portela et al. 2024a, Portela et al. 2023], foram analisadas técnicas de imputação em conjunto com modelos de predição e foi realizado um estudo sobre o impacto do comportamento do protocolo TCP sobre o desempenho de vazão. A partir desses trabalhos, percebeu-se a necessidade de aplicação de abordagens mais robustas para habilitar um processo de predição de vazão eficiente.

Dentro deste contexto, este artigo apresenta um modelo de predição de vazão de rede que aplica Decomposição de Valores Singulares (*Singular value decomposition* - SVD) para mitigar falhas de medição (e, consequentemente, lacunas na série temporal de desempenho de vazão) e explorar o aspecto sazonal das séries temporais de desempenho de vazão. Diferentemente de sua aplicação usual [García-Peña et al. 2021], neste trabalho o SVD é aplicado a dados de vazão para predição de séries temporais univariadas [Spiliotis et al. 2020]. Adicionalmente, modelo de predição proposto é adaptativo, pois ajusta a série temporal das medições realizadas por meio de análises estatísticas, como decomposição, identificação de tendência e correção de erros cíclicos. Esse processo gera uma série temporal padronizada, ideal para o treinamento de modelos preditivos. A série tratada serve como insumo para os modelos de Inteligência Artificial (IA), que realizam a predição de uma nova série representando o desempenho esperado da rede, mesmo em casos de falhas de medição, que são superadas pelo uso do SVD.

Desta forma, um pré-processamento inteligente é realizado através do SVD para destacar os aspectos temporais semanais da série, maximizando a extração de informações relevantes. Isso é fundamental, pois, ao contrário de conjuntos de dados multivariados, em séries unidimensionais não é possível recorrer a informações de outras variáveis [Phan 2020]. Após a imputação, os dados resultantes foram utilizados em redes neurais recorrentes, que exploram as dependências temporais para realizar predições de alta qualidade. Combinando essas estratégias, esta solução oferece uma abordagem robusta e eficiente para melhorar a análise de desempenho de rede, contribuindo para o avanço de

ferramentas de monitoramento e gestão.

Para a avaliação da proposta, foram utilizados os dados reais do Serviço de Monitoramento da Rede Ipê (Monipê)[1], mantido pela Rede Nacional de Ensino e Pesquisa (RNP). Os dados do Monipê, amplamente reconhecidos por sua qualidade e confiabilidade, foram utilizados para validar a solução proposta e analisar seu desempenho no contexto de predição de métricas de rede. Os resultados obtidos demonstraram que o modelo alcançou altos níveis de acurácia na predição, superando modelos preditivos existentes nas principais métricas de avaliação, como RMSE. Adicionalmente, a aplicação de técnicas de imputação de dados mostrou-se eficaz para melhorar o processo de predição, reduzindo significativamente os erros quando comparados aos dados originais.

O restante deste artigo está organizado da seguinte forma: A Seção 2 apresenta os trabalhos relacionados, destacando soluções similares para análise de desempenho de rede. A Seção 3 detalha a proposta deste trabalho, enquanto que a Seção 4 descreve os experimentos realizados e os resultados obtidos. Por fim, a Seção 5 conclui o artigo e apresenta trabalhos futuros.

2. Trabalhos Relacionados

Park et al. [Park et al. 2023] apresentam uma abordagem baseada em modelos de aprendizagem profunda para estimar valores ausentes em dados multivariados de séries temporais. O foco do estudo está no preenchimento de lacunas longas e contínuas, como meses de observações diárias ausentes, em vez de lidar com valores ausentes aleatórios. Apesar de sua relevância, o trabalho não aborda diretamente a integração de técnicas de imputação para melhorar a precisão de tarefas de predição.

A utilização da decomposição de valores singulares (SVD) em imputação de dados é explorada por [García-Peña et al. 2021]. Este trabalho apresenta um estudo inovador sobre o uso de SVD para lidar com *outliers* e melhorar a precisão na geração de dados sintéticos. Embora a metodologia seja promissora, o estudo não considera as características sazonais das séries temporais, nem explora como a imputação pode impactar tarefas de predição.

Ding et al. [Ding et al. 2020] avaliam diversas técnicas de imputação para séries temporais coletadas de dispositivos IoT. O trabalho compara métodos como *Radial Basis Functions*, *Moving Least Squares (MLS)* e *Adaptive Inverse Distance Weighted* com o KNN, destacando o MLS de Lancaster como a técnica com melhor desempenho. Contudo, o estudo não analisa como essas abordagens influenciam a predição das séries temporais, um aspecto essencial para aplicações práticas. Similarmente, na referência [Ahn et al. 2021], os autores realizam uma análise abrangente das principais técnicas de imputação de dados aplicadas a séries temporais. O estudo avalia métodos como interpolação linear, KNN, e redes neurais para prever valores ausentes, destacando o KNN como o método com melhor desempenho em diversos cenários. Apesar disso, o trabalho enfatiza a importância de métodos que capturem padrões temporais complexos, destacando uma lacuna que pode ser abordada por técnicas baseadas em SVD.

Com base nesses trabalhos, nossa proposta busca preencher as lacunas identificadas, combinando a robustez do SVD com a consideração dos aspectos sazonais das séries

[1] monipe-central.rnp.br

temporais para criar uma abordagem eficaz tanto para imputação quanto para predição. Esse diferencial permite lidar com dados de rede de forma mais precisa, melhorando a qualidade das análises e previsões no contexto de desempenho de rede.

3. Proposta

Esta seção apresenta a descrição detalhada do modelo de predição de dados de vazão baseado em SVD e RNN proposto neste artigo. A proposta busca processar dados de medição de vazão entre dois pontos de rede utilizando técnicas avançadas de análise e preparação de séries temporais, com o objetivo de mitigar os efeitos adversos de dados faltantes e melhorar a precisão da predição com redes neurais recorrentes.

A solução é executada em seis etapas principais, conforme ilustrado na Figura 1, descritas a seguir: (i) Coleta de Dados, em que os dados necessários para a predição são coletados diretamente da API do Monipê, (ii) análise de dados, realiza-se uma inspeção detalhada dos dados coletados para identificar falhas, verificar a quantidade disponível de medições e avaliar características gerais do conjunto de dados, como padrões de sazonalidade e lacunas de dados, (iii) tratamento de dados, onde os dados univariados de vazão são transformados em uma matriz de dados normalizados. Cada coluna da matriz representa uma semana diferente, permitindo destacar o comportamento sazonal e facilitando o processo de imputação e análise posterior, (iv) imputação com SVD, em que finalmente a técnica de decomposição por valores singulares (SVD) é aplicada sobre a matriz gerada. Esse processo extrai as informações mais relevantes dos dados, permitindo uma imputação eficiente que reduz os impactos de falhas ou lacunas nos valores originais, (v) após a imputação, os dados processados retornam à forma original como uma série temporal univariada, e, por fim, (vi) predição de dados: as séries temporais imputadas são utilizadas como entrada para redes neurais recorrentes (RNNs), que realizam a predição dos valores futuros de vazão.

3.1. Análise e Tratamento de Dados

Nesta seção serão tratados os aspectos relacionados ao processo de imputação de dados, bem como as possíveis técnicas que poderiam ser aplicadas e o detalhamento da técnica SVD (que é a aplicada na proposta deste trabalho).

3.1.1. Técnicas de Imputação de Dados

Tradicionalmente, as técnicas de imputação de dados podem ser classificadas em dois tipos: (i) métodos propostos para dados multidimensionais e (ii) técnicas desenvolvidas para dados unidimensionais. Para dados unidimensionais, muitas técnicas de imputação que são mais simples, mas que consequentemente negligenciam muitos aspectos dos dados, são utilizadas, como média e mediana [Phan 2020]. Além disso, utilizando a interpolação linear assumimos que os valores se conectam por meio de uma função linear [Cho et al. 2020]. Esses métodos nem sempre são interessantes de serem aplicados em séries temporais, uma vez que não consideram as características sazonais dos dados. O método de imputação baseado em k-vizinhos mais próximos (KNN), por outro lado, funciona muito bem e se destaca em inúmeros cenários de imputação de dados, como em [Anil Jadhav and Ramanathan 2019], que compara imputação de dados

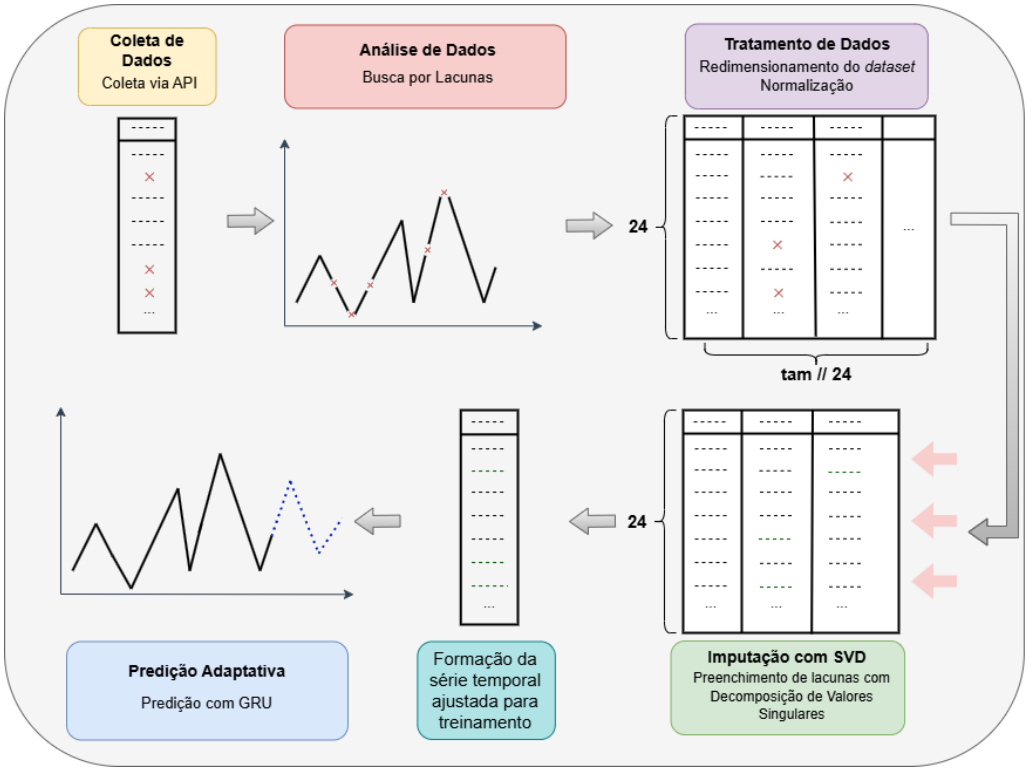

Figura 1. Visão Geral da Solução.

sintéticos em um contexto generalista de conjuntos de dados de valores numéricos; em [Sandra Taylor and Kim 2022], onde essa técnica se mostrou um dos melhores métodos por performar consistentemente em diversas condições, além de ser geralmente o melhor em cenários com mais de 20% de dados faltantes; e em [Ahn et al. 2021], que mostra que o KNN performou de melhor forma entre os métodos apresentados. Trouxemos, com a nossa proposta, a ideia de forçar uma imputação múltipla de conjuntos de dados multivariados em um univariado, forçando uma correlação entre as variáveis para encontrar dados sintéticos que mais se assemelham aos originais, utilizando a decomposição de valores singulares (SVD).

3.1.2. Imputação de Dados Baseada em SVD

O SVD é uma técnica fundamental em álgebra linear que permite fatorar uma matriz A em três outras matrizes, revelando propriedades importantes da matriz original. Ela é amplamente usada em várias áreas, como compressão de dados, aprendizado de máquina, processamento de imagens e análise de sistemas lineares.

É dividido em duas partes: (i) Tratamento e Estruturação dos Dados, antes de aplicar o SVD, uma máscara para os valores faltantes do conjunto de dados original é gerada e os conjuntos de dados e os dados são normalizados, garantindo que todos os valores estejam em uma escala comparável. Além disso, os dados são transformados de univariados para um formato multivariado de forma a apresentar aspectos temporais. No caso deste trabalho, os dados de desempenho de rede geralmente apresentam um padrão sazonal semanal e, pensando nisso, o conjunto de dados univariado foi dividido em várias

colunas com dados de sete dias cada; e, (ii) Processo Iterativo de Imputação, a imputação ocorre em um processo iterativo até atingir o ponto de convergência, definido pelo cálculo do erro quadrático médio (*Root Mean Squared Error*, RMSE) entre as matrizes imputadas em iterações consecutivas. O RMSE para um determinado período de tempo T é definido pela Equação 1, onde $\hat{y}_t$ representa o valor predito e y_t significa o valor real do desempenho no tempo t. Neste trabalho, a convergência é considerada alcançada quando o RMSE é inferior a um limiar predefinido (10^{-3}).

$$\mathrm{RMSE}(T) = \frac{1}{\sqrt{T}} \left(\sum_{t=1}^{T} (\hat{y}_t - y_t)^2 \right)^{\frac{1}{2}} . \tag{1}$$

Definindo o SVD de maneira formal, temos que dada uma matriz $A \in \mathbb{R}^{m \times n}$, sua decomposição em valores singulares é definida como apresentado na Equação 2, onde $U \in \mathbb{R}^{m \times m}$ é a matriz ortogonal cujas colunas são os autovetores de AA^T; $\Sigma \in \mathbb{R}^{m \times n}$ é a matriz diagonal contendo os valores singulares $\sigma_1, \sigma_2, \ldots, \sigma_r$ de A, com $r = \min(m, n)$; e $V^T \in \mathbb{R}^{n \times n}$ é a transposta da matriz ortogonal V, cujas colunas são os autovetores de $A^T A$. Os valores singulares σ_i são definidos como as raízes quadradas dos autovalores de $A^T A$ ou AA^T, de acordo com a Equação 3, sendo que λ_i são os autovalores de $A^T A$ ou AA^T e satisfazem $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$.

$$A = U \Sigma V^T \tag{2}$$

$$\sigma_i = \sqrt{\lambda_i} \tag{3}$$

De maneira geral, as Matrizes U, Σ e V possuem as seguintes propriedades: $U^T U = I_m$: U é uma matriz ortogonal; $V^T V = I_n$: V é uma matriz ortogonal; e, A matriz Σ é diagonal e tem a forma definida na Equação 4, onde $\sigma_1, \sigma_2, \ldots, \sigma_r$ sendo os valores singulares.

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \tag{4}$$

Posteriormente, a etapa de Reconstrução da Matriz no SVD consiste em usar os valores singulares e os vetores singulares para reconstituir ou aproximar a matriz original A de acordo com a Equação 5, onde σ_i é o i-ésimo valor singular, u_i é o i-ésimo vetor singular à esquerda (coluna de U) e v_i^T é o i-ésimo vetor singular à direita (linhas de V^T). Assim, cada termo $\sigma_i u_i v_i^T$ representa uma "contribuição"da "direção"u_i e v_i pelo valor σ_i. Por fim, podemos aproximar A utilizando apenas os k maiores valores singulares (sendo $k < r$), através da Equação 6, onde $U_k \in \mathbb{R}^{m \times k}$ contém as k primeiras colunas da matriz U, $\Sigma_k \in \mathbb{R}^{k \times k}$ contém os k maiores valores singulares de Σ em sua diagonal, e $V_k \in \mathbb{R}^{n \times k}$ contém as k primeiras colunas da matriz V. Isso reduz a complexidade computacional e captura as informações mais importantes da matriz.

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T, \tag{5}$$

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T, \tag{6}$$

Como dito anteriormente, este processo iterativo é executado, incrementando k, até que se alcance um ponto de convergência, resultando em uma matriz de medições de rede mais aproximada do padrão encontrado. Em seguida, essa matriz é revertida para a escala original e convertida novamente em série temporal. Utilizando a máscara de dados faltantes gerada anteriormente, apenas as lacunas são preenchidas, preservando-se todos os valores originais. Esse procedimento conclui a imputação, fornecendo valores de medição coerentes com o comportamento característico do conjunto de dados.

3.1.3. Análise comparativa entre as Abordagens

Os métodos de Média e Mediana Móvel são soluções simples e computacionalmente eficientes, adequadas para séries temporais com padrões básicos. Enquanto a Média Móvel é sensível a *outliers*, a Mediana Móvel oferece maior robustez e é mais indicada para dados com flutuações esporádicas. Ambos os métodos, no entanto, são limitados na captura de padrões complexos, como sazonalidade ou não linearidade. De forma semelhante, a Interpolação Linear se destaca pela simplicidade e baixo custo, sendo eficaz para séries com transições suaves e consistentes, mas ineficaz em cenários que demandam a identificação de padrões avançados. Por outro lado, o KNN e o SVD são métodos mais robustos e adequados para séries temporais complexas. O KNN é flexível e eficiente na captura de relações intrincadas, embora seu alto custo computacional seja uma limitação em grandes conjuntos de dados. Já o SVD se sobressai por explorar padrões latentes em dados multivariados, oferecendo alta precisão em séries sazonais após iterações avaliadas por RMSE. Além disso, sua robustez a *outliers* e capacidade de reduzir ruídos tornam o SVD uma escolha ideal para cenários com alta variabilidade e padrões temporais estruturados.

3.2. Modelo de Predição

Nesta seção são descritos os aspectos relacionados ao processo de predição aplicado na proposta, o qual aplica uma abordagem adaptativa e resiliente perante a falhas de medição (e, consequentemente, "quebras"ou "buracos"na série temporal). A organização de séries temporais regulares, essencial para análises preditivas confiáveis, torna-se particularmente difícil diante desse problema. Estes pontos são detalhados nas Seções 3.2.1 e 3.2.2.

3.2.1. Ajuste da Série Temporal para Predição

A análise de séries temporais sazonais frequentemente utiliza o método *Seasonal and Trend Decomposition using Loess* (STL), que permite decompor a série em seus componentes de tendência e sazonalidade. Esse método emprega o ajuste Loess para estimar relações não lineares nos dados, proporcionando uma compreensão mais detalhada da série temporal. O processo inicia-se com a suavização dos valores da série por meio de um filtro *low-pass*, que preserva ciclos de baixa frequência e atenua oscilações de alta frequência, reduzindo o ruído e obtendo um componente de tendência mais suave. Em seguida, a tendência extraída é subtraída da série original, resultando em uma versão sem tendência. Sobre essa nova série, aplica-se um filtro sazonal para estimar a componente de sazonalidade, que posteriormente é removida para obter a série residual.

A decomposição STL permite a reconstrução da série original somando novamente os componentes de tendência, sazonalidade e resíduos. Esse método é particularmente útil para a análise e modelagem de dados com padrões sazonais bem definidos. Além disso, é altamente flexível, podendo lidar com diferentes períodos sazonais e ser aplicado tanto a séries temporais aditivas quanto multiplicativas.

3.2.2. IA para Predição

A Gated Recurrent Unit (GRU) é uma evolução das redes neurais recorrentes (RNNs) e possui uma estrutura menos complexa em comparação às Long Short-Term Memory (LSTMs). Enquanto as LSTMs utilizam dois estados diferentes, o estado da célula e o estado oculto, para manter memórias de longo e curto prazo, respectivamente, as GRUs utilizam apenas um estado oculto. Esse estado é responsável por manter simultaneamente as dependências de curto e longo prazo, sendo atualizado através de dois mecanismos principais: a porta de redefinição (*reset gate*) e a porta de atualização (*update gate*)

As portas da GRU são vetores que decidem quais informações devem ser passadas para a saída. A porta de redefinição define como combinar a nova entrada com a memória anterior, enquanto a porta de atualização controla a quantidade de memória anterior que será mantida. Essa arquitetura permite que a GRU lide eficientemente com o problema de dissipação do gradiente comum em RNNs tradicionais e seja treinada para capturar dependências de longo prazo em séries temporais, sem perder informações relevantes. Além disso, estudos adicionais, como o trabalho de análise preditiva em séries temporais financeiras, também indicam que a GRU é frequentemente superior a modelos como LSTM e bi-LSTM. [Yamak et al. 2020]. A simplicidade e eficiência da GRU a tornam uma excelente escolha para tarefas de predição em séries temporais. Desta forma, GRU foi a técnica de IA escolhida para integrar a solução proposta e fazer parte do modelo de predição que irá receber os dados de vazão com as imputações através de SVD.

4. Experimentos

Esta seção irá descrever a forma como os experimentos realizados foram configurados, bem como apresentar o resultado desses experimentos realizados. É válido ressaltar que o código desenvolvido, bem como os dados utilizados nos experimentos, estão disponíveis no repositório do projeto[2] e com as instruções necessárias para reprodutibilidade.

4.1. Configuração dos Experimentos

Nos experimentos utilizamos os dados da RNP, através do Serviço de Monitoramento da Rede Ipê (MonIPÊ)[3], a fim de trazer uma análise totalmente focada em cenários reais. O MonIPÊ utiliza o padrão de monitoramento do antes citado perfSONAR, de forma que as medições de Vazão ocorrem a cada 4 horas. Contudo, ocorrem falhas de medição de vazão (por diversos fatores), resultando em perda de valores de medição e consequentemente dados de baixa qualidade para análise. Essas falhas, nos dados estudados, podem chegar a mais de 50% em alguns casos.

[2]https://github.com/LarcesUece/SVD-IMPUTATION-SBRC-2025/

[3]https://redeipe.rnp.br/

Para uma maior consistência no nosso trabalho, nomeamos 8 enlaces para o estudo: AP-BA, BA-PA, CE-RO, ES-PR, GO-SE, MA-RJ, PB-ES e RO-SE. Esses *links* conectam Pontos de Presença (PoPs) localizados em diferentes regiões do Brasil, abrangendo o Norte, Nordeste, Sudeste, Centro-Oeste e Sul do país, o que reflete também características de tráfego distintas ao longo do dia e da semana. Essa distribuição é importante, pois os padrões de tráfego em redes de longa distância são influenciados pelas características regionais, como demanda de usuários, densidade populacional e atividades econômicas.

Após esse procedimento, diferentes técnicas de imputação foram usadas para preencher as lacunas dos dados referentes às medições de vazão. Nesse processo, vimos uma escassez de métodos estatísticos que realmente se adequassem a esse cenário, de forma que trouxemos nossa solução de imputação baseada em SVD adaptada às séries temporais. Esse procedimento permitiu uma avaliação controlada e consistente da eficácia das técnicas, simulando condições reais de dados faltantes enquanto mantinha a estrutura temporal original dos dados.

Na etapa da predição, a ausência de dados (ou a presença de lacunas) poderia comprometer o desempenho do modelo de predição, dificultando a captura de padrões temporais relevantes e a geração de previsões precisas, então utilizamos esses conjuntos de dados tratados, agora representando três meses completos de medição de dados de vazão. Para a predição, os dados foram separados em 80% para treinamento e 20% para testes e validação dos modelos. A imputação não apenas restaurou as séries temporais tratadas, mas também assegurou que os modelos fossem treinados em dados que preservassem padrões sazonais, tendências e características originais da série. Em seguida, foram realizadas quatro rodadas de treinamento com GRU, visando avaliar o desempenho do preditor e identificar a melhor configuração para previsões de vazão de rede (ou seja, a melhor técnica de imputação).

Para avaliar com melhor precisão as técnicas de imputação e o desempenho dos modelos de predição frente a valores reais, foi utilizado o RMSE (descrito anteriormente). Adicionalmente, foi realizada uma análise baseada na medida de Dynamic Time Warping (DTW), que é amplamente utilizada para avaliar a similaridade entre séries temporais [Li 2021], permitindo comparar padrões mesmo quando os dados possuem deslocamentos ou deformações no tempo. A DTW calcula a distância acumulada entre dois conjuntos de dados, identificando o alinhamento ótimo entre os pontos temporais. Essa métrica é especialmente relevante em séries temporais onde é fundamental preservar características sazonais, tendências e variações, como em dados de desempenho de rede. Com isso, é possível validar os valores gerados pelas técnicas, além de avaliar a predição pelos modelos utilizando as mesmas técnicas de preenchimento.

4.2. Resultados

Esta seção apresenta os resultados obtidos a partir da análise dos experimentos realizados com um conjunto de dados reais, onde as Figuras 2, 3 e 4 apresentam os dados referentes ao RMSE, a análise de DTW e uma análise sobre o percentual de evolução da proposta diante das técnicas existentes, respectivamente.

Os resultados da análise de RMSE (Figura 2) fornecem uma avaliação quantitativa da precisão das imputações e predições realizadas com os diferentes métodos testados. Na

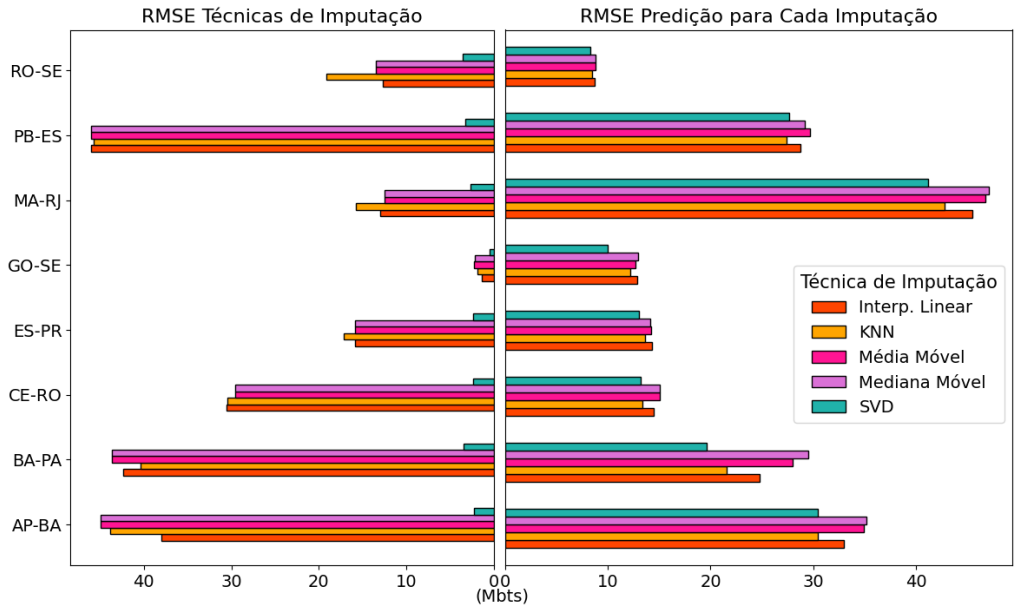

Figura 2. Resultados do RMSE.

análise de imputação de dados, o método baseado em SVD apresentou consistentemente os menores valores de RMSE em todos os links avaliados. Isso indica que o SVD foi mais eficaz em minimizar os desvios médios entre os valores imputados e os dados reais ausentes. Em contraste, os métodos de Interpolação Linear e KNN tiveram RMSE significativamente maiores, demonstrando uma maior incerteza e menos precisão na imputação. Nos resultados de predição da Figura 2, o SVD novamente se destacou com os menores valores de RMSE para todos os links analisados. Esses valores indicam que as predições realizadas a partir de dados imputados com SVD estavam mais próximas dos valores reais observados, sugerindo uma preservação superior das características temporais subjacentes e dos padrões de rede relevantes. A análise revela que o SVD apresentou desempenho superior tanto na imputação quanto na predição de dados, resultando em valores de RMSE mais baixos em comparação com os outros métodos avaliados. Matematicamente, isso reflete a capacidade do SVD de minimizar a soma dos desvios quadráticos médios em relação aos dados reais, evidenciando maior precisão em capturar a estrutura e os padrões subjacentes dos dados de rede.

No que se refere a DTW, os resultados da Figura 3 indicam que a proposta apresentou os menores valores DTW nas três métricas analisadas: sazonalidade (0.054), tendência (0.021) e observação (0.060), sendo estes valores substancialmente menores que os dos demais métodos. Essa consistência destaca a capacidade do SVD em preservar as características temporais originais dos dados. Padrões sazonais foram melhor preservados (DTW de sazonalidade mais baixo), tendências de longo prazo mostraram uma fidelidade significativamente maior (DTW de tendência reduzido) e as características gerais das observações foram menos distorcidas (DTW de observação reduzido). Em contextos de dados de rede, onde padrões temporais são essenciais para a detecção de anomalias e para a análise de desempenho, essa preservação é crítica.

A robustez demonstrada pelo SVD em preservar a estrutura temporal dos dados é particularmente importante ao alimentar modelos de predição como a GRU. Mode-

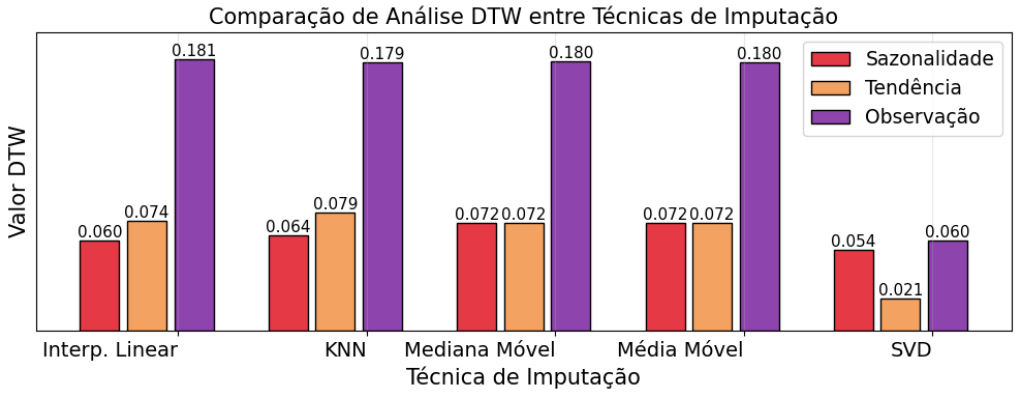

Figura 3. Comparação de Análise DTW entre diferentes técnicas de Imputação

los como a GRU dependem de entradas que mantêm tanto as tendências de longo prazo quanto as variações de curto prazo para capturar as dependências temporais de forma eficaz. Dados distorcidos ou que perdem características sazonais podem prejudicar severamente o desempenho preditivo da GRU. Nesse sentido, a capacidade do SVD de preservar padrões em diferentes escalas temporais — como sazonalidade, tendências e variações — maximiza o potencial da GRU em gerar previsões confiáveis e precisas.

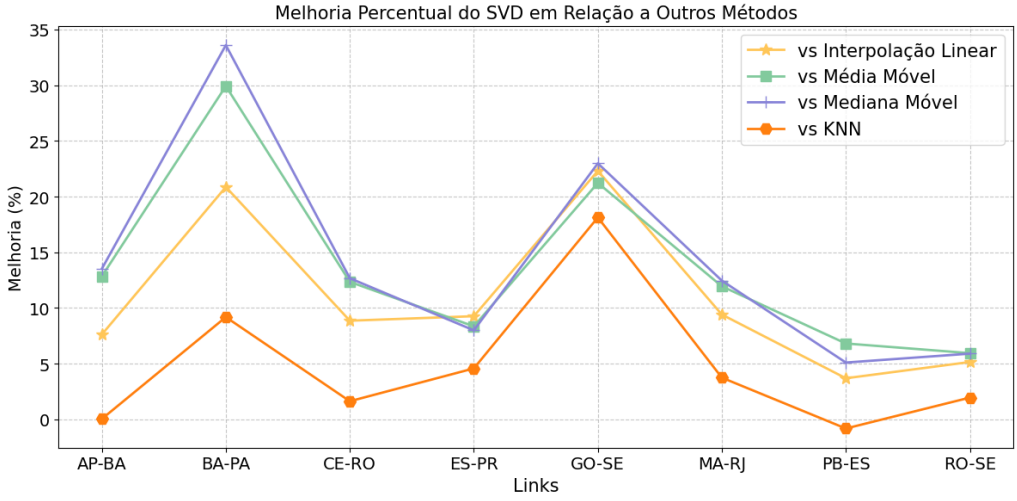

Figura 4. Melhoria Percentual da Proposta em Relação a Outras Técnicas.

A Figura 4 indica que o método de imputação baseado em SVD, considerando os aspectos temporais da série, apresenta consistentemente melhorias significativas no RMSE das predições de vazão. Além disso, é possível observar variações importantes no desempenho dependendo do ponto de comunicação analisado e do método comparativo. Especificamente, os links BA-PA e GO-SE destacaram-se, com o link BA-PA apresentando o melhor desempenho geral. Essa variabilidade de desempenho entre os links pode ser atribuída a características específicas do tráfego em cada conexão, como padrões de comportamento da rede e a quantidade e qualidade dos dados disponíveis para cada link.

Apesar dessas diferenças, o método proposto demonstrou flexibilidade e eficácia em todos os cenários analisados. A melhoria média ao longo de todos os links foi mais

acentuada em comparação com os métodos de mediana móvel e média móvel. Quando comparado ao KNN, o SVD ainda apresentou melhores resultados, embora as diferenças tenham sido menos pronunciadas, sugerindo que o KNN é potencialmente o segundo método mais eficiente entre os avaliados. Por outro lado, a interpolação linear apresentou um desempenho intermediário, sendo consistentemente superada pelo SVD.

4.3. Discussão Final

A análise dos resultados demonstrou que a proposta baseada em SVD superou os métodos tradicionais de imputação (como Mediana Móvel, Média Móvel, Interpolação Linear e KNN), tanto na imputação quanto na predição de dados, conforme evidenciado pelos valores de RMSE. Esses resultados destacam a capacidade do SVD de preservar padrões sazonais, tendências de longo prazo e características gerais das séries temporais, características essenciais para garantir a confiabilidade em cenários de rede onde os padrões temporais desempenham um papel crítico.

A preservação mais precisa das características temporais pelo SVD também foi refletida na superioridade do modelo de predição GRU ao utilizar os dados imputados, reforçando a importância de fornecer entradas de alta qualidade para modelos preditivos. Essa integração permitiu reduzir significativamente os erros de previsão e capturar as dinâmicas complexas dos dados de rede, mesmo em cenários com períodos prolongados de dados ausentes e padrões de tráfego variáveis.

Além disso, a metodologia utilizada, que remove dados proporcionalmente às falhas reais do conjunto de dados, aumentou a relevância prática dos resultados, demonstrando que o método proposto é robusto e adaptável a diferentes contextos. Isso amplia sua aplicabilidade em monitoramento de redes de computadores e outras áreas que dependem de séries temporais confiáveis para análise e decisão. Portanto, este trabalho não apenas valida a eficácia da proposta com SVD, mas também contribui para o avanço de soluções voltadas à predição de dados de rede. O método proposto oferece uma base sólida para futuros estudos e aplicações práticas, promovendo análises mais precisas e uma gestão mais eficiente de redes de computadores.

5. Conclusão

Este trabalho apresentou uma abordagem para imputação e predição de séries temporais no contexto de dados de vazão, integrando o processo de predição com a técnica SVD e permitindo ao modelo de predição se tornar resiliente perante a falhas de medição. Combinando a técnica de decomposição de valores singulares com modelos de redes neurais recorrentes, como a GRU, conseguimos alcançar resultados superiores em termos de preservação de características temporais e eficácia da predição.

As próximas etapas deste trabalho consistem em aprimorar a solução para torná-la ainda mais completa e eficaz. Pretendemos adicionar a capacidade de analisar séries temporais multivariadas, o que permitirá considerar outras métricas importantes para o desempenho da rede, como o atraso e a perda de pacotes. Além disso, exploraremos o uso de modelos combinados de aprendizado de máquina, buscando aprimorar a precisão das previsões e a capacidade de adaptação do modelo a diferentes cenários.

Agradecimentos

Pesquisa parcialmente financiada pelo CNPq (Processos 405940/2022-0 e 303877/2021-9) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 88887.954253/2024-00

Referências

Ahn, H., Sun, K., and Kim, K. (2021). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials and Continua*, 70:767–779.

Anil Jadhav, D. P. and Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933.

Cho, B., Dayrit, T., Gao, Y., Wang, Z., Hong, T., Sim, A., and Wu, K. (2020). Effective missing value imputation methods for building monitoring data. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2866–2875.

Ding, Z., Mei, G., Cuomo, S., Li, Y., and Xu, N. (2020). Comparison of estimating missing values in iot time series data using different interpolation algorithms. *International Journal of Parallel Programming*, 48(3):534–548.

Ferreira, M. C., Ribeiro, S. E., Nobre, F. V., Linhares, M. L., Araújo, T. P., and Gomes, R. L. (2024). Mitigating measurement failures in throughput performance forecasting. In *2024 20th International Conference on Network and Service Management (CNSM)*, pages 1–7.

García-Peña, M., Arciniegas Alarcón, S., Krzanowski, W., and Duarte Vogel, D. (2021). Missing-value imputation using the robust singular-value decomposition: Proposals and numerical evaluation. *Crop Science*, 61.

Gomes, R. L., Bittencourt, L. F., and Madeira, E. R. M. (2014a). A similarity model for virtual networks negotiation. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, page 489–494, New York, NY, USA. Association for Computing Machinery.

Gomes, R. L., Bittencourt, L. F., Madeira, E. R. M., Cerqueira, E., and Gerla, M. (2014b). An architecture for dynamic resource adjustment in vsdns based on traffic demand. In *2014 IEEE Global Communications Conference*, pages 2005–2010.

Li, H. (2021). Time works well: Dynamic time warping based on time weighting for time series data mining. *Information Sciences*, 547:592–608.

Naf, J., Spohn, M.-L., Michel, L., and Meinshausen, N. (2022). Imputation scores.

Park, J., Müller, J., Arora, B., Faybishenko, B., Pastorello, G., Varadharajan, C., Sahu, R., and Agarwal, D. (2023). Long-term missing value imputation for time series data using deep neural networks. *Neural Computing and Applications*, 35(12):9071–9091.

Phan, T.-T.-H. (2020). Machine learning for univariate time series imputation. In *2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6.

Pinheiro, B., Nascimento, V., Gomes, R., Cerqueira, E., and Abelem, A. (2011). A multimedia-based fuzzy queue-aware routing approach for wireless mesh networks.

In *2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN)*, pages 1–7.

Portela, A., Linhares, M. M., Nobre, F. V. J., Menezes, R., Mesquita, M., and Gomes, R. L. (2024a). The role of tcp congestion control in the throughput forecasting. In *Proceedings of the 13th Latin-American Symposium on Dependable and Secure Computing*, LADC '24, page 196–199, New York, NY, USA. Association for Computing Machinery.

Portela, A. L., Menezes, R. A., Costa, W. L., Silveira, M. M., Bittecnourt, L. F., and Gomes, R. L. (2023). Detection of iot devices and network anomalies based on anonymized network traffic. In *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, pages 1–6.

Portela, A. L. C., Ribeiro, S. E. S. B., Menezes, R. A., de Araujo, T., and Gomes, R. L. (2024b). T-for: An adaptable forecasting model for throughput performance. *IEEE Transactions on Network and Service Management*, pages 1–1.

Sandra Taylor, Matthew Ponzini, M. W. and Kim, K. (2022). Comparison of imputation and imputation-free methods for statistical analysis of mass spectrometry data with missing data. *Briefings in bioinformatics*, 23(1):913–933.

Silva, M., Ribeiro, S., Carvalho, V., Cardoso, F., and Gomes, R. L. (2023). Scalable detection of sql injection in cyber physical systems. In *Proceedings of the 12th Latin-American Symposium on Dependable and Secure Computing*, LADC '23, page 220–225, New York, NY, USA. Association for Computing Machinery.

Silva, M. V., Mosca, E. E., and Gomes, R. L. (2022). Green industrial internet of things through data compression. *International Journal of Embedded Systems*, 15(6):457–466.

Silveira, M. M., Portela, A. L., Menezes, R. A., Souza, M. S., Silva, D. S., Mesquita, M. C., and Gomes, R. L. (2023a). Data protection based on searchable encryption and anonymization techniques. In *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, pages 1–5.

Silveira, M. M., Silva, D. S., Rodriguez, S. J. R., and Gomes, R. L. (2023b). Searchable symmetric encryption for private data protection in cloud environments. In *Proceedings of the 11th Latin-American Symposium on Dependable Computing*, LADC '22, page 95–98, New York, NY, USA. Association for Computing Machinery.

Spiliotis, E., Assimakopoulos, V., and Makridakis, S. (2020). Generalizing the theta method for automatic forecasting. *European Journal of Operational Research*, 284(2):550–558.

Vasileva, P., McKee, S., Penev, A., and Vukotic, I. (2021). Ps-dash –analysis, monitoring and visualization of network measurements. In *2021 International Conference Automatics and Informatics (ICAI)*, pages 93–96.

Yamak, P. T., Yujian, L., and Gadosey, P. K. (2020). A comparison between arima, lstm, and gru for time series forecasting. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '19, page 49–55, New York, NY, USA. Association for Computing Machinery.