Estimativa de Vazão de Rede Baseada em Técnicas de Regressão sobre Dados de Monitoramento

Maria L. Linhares[1], Maria C. M. M. Ferreira[1], Thelmo P. Araújo[1], Roger Immich[2], Rafael L. Gomes[1]

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

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

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

[2]Universidade Federal do Rio Grande do Norte (UFRN)

roger@imd.ufrn.br

Resumo. *Empresas e provedores de internet (ISPs) costumam realizar serviços de monitoramento de rede, fornecendo resultados de testes regulares de desempenho, como vazão, perda, traceroute, atraso, entre outros. As medições de vazão, diferente das demais, são realizadas em intervalos longos, visto que consomem muitos recursos de rede e afetam a Qualidade de Serviço (QoS) e de Experiência (QoE) dos usuários. Todavia, as informações de vazão são fundamentais para um gerenciamento de rede eficaz. Dentro deste contexto, este artigo apresenta uma metodologia inovadora para estimar a vazão de rede (em um determinado momento) a partir do agrupamento de outras medições que não comprometem a QoS e QoE dos usuários, tais como traceroute e atraso. Os experimentos, realizados com dados reais da Rede Nacional de Ensino e Pesquisa (RNP), demonstram que a integração de medições complementares aumenta a precisão das estimativas.*

Abstract. *Companies and internet service providers (ISPs) often perform network monitoring services, providing regular performance test results such as throughput, loss, traceroute, delay, and others. Throughput measurements, unlike the others, are conducted at long intervals due to their high network resource consumption, which impacts the Quality of Service (QoS) and Quality of Experience (QoE) for users. However, throughput information is essential for effective network management. In this context, this article presents an innovative methodology for estimating network throughput (at a given moment) by aggregating other measurements that do not compromise users' Quality of Service (QoS) and Quality of Experience (QoE), such as traceroute and delay. Experiments conducted with real data from the National Education and Research Network (RNP) demonstrate that integrating complementary measurements improves the accuracy of the estimates.*

1. Introdução

A Internet tornou-se pilar fundamental para serviços modernos, desde aplicações comerciais até sistemas críticos de governança. Desafios como lentidão, falhas e interrupções afetam diretamente a Qualidade do Serviço (QoS) e a Experiência do Usuário

(QoE), impactando negativamente os Acordos de Nível de Serviço (SLAs), causando prejuízos financeiros e danos à reputação [Gijon et al. 2024, Souza et al. 2024]. Para enfrentar esses desafios, empresas e provedores de serviços de Internet (ISPs) implementam ferramentas de monitoramento que realizam medições de rede essenciais como vazão, *traceroute*, atraso e perda de pacotes [Ferreira et al. 2024].

As medições de vazão, diferentemente das demais, são realizadas em intervalos longos devido ao seu alto impacto no consumo de recursos da rede [Mok et al. 2021, Silva et al. 2022]. Estas medições envolvem transferência de grandes volumes de dados entre dois pontos, podendo gerar sobrecarga nos equipamentos e impactar negativamente o desempenho geral [Gomes et al. 2014a, Portela et al. 2024a]. A realização frequente destes testes pode interferir diretamente na QoS oferecida aos usuários, causando aumento de latência, redução de velocidade e interrupções momentâneas no tráfego regular [Portela et al. 2024b, Silva et al. 2023]. Por isso, são planejadas estrategicamente em períodos de baixa demanda, minimizando impactos e garantindo precisão nas informações coletadas [Silveira et al. 2023a]. Soluções como o FastBTS [Yang et al. 2021b] têm otimizado este processo, reduzindo tempo de teste e consumo de dados em até 10,7 vezes, mantendo a precisão das medições. No entanto, mesmo com essas otimizações, a predição de métricas de rede com base em indicadores indiretos pode ser uma alternativa ainda mais eficiente, especialmente em cenários onde o custo de medição direta é proibitivo.

A falta de informações frequentes sobre vazão prejudica o gerenciamento da rede ao dificultar a detecção rápida de problemas, comprometer o planejamento de capacidade e o cumprimento de SLAs [Portela et al. 2023, Silveira et al. 2023b]. Esta limitação compromete análises precoces e ajustes dinâmicos, dificultando a identificação de tendências, alocação eficiente de recursos e prevenção de degradações que impactam a QoE, podendo resultar em decisões baseadas em informações incompletas.

Neste contexto, é apresentado um modelo de regressão para estimar a vazão de rede a partir do agrupamento de outras medições que não comprometem QoS e QoE [Gomes et al. 2014b]. Esta abordagem inova ao explorar medições disponíveis com maior frequência e intervalo regular, como atraso, para prever a vazão com horizonte temporal maior. Foram incluídos dados complementares como número de saltos e identificação de links de gargalo, extraídos de arquivos de *traceroute*, enriquecendo o processo de estimativa.

A escolha pela estimativa pontual, em vez de modelos de séries temporais, justifica-se pelas limitações inerentes às medições diretas de vazão, que frequentemente apresentam falhas devido à natureza dinâmica e imprevisível das redes. Problemas como interrupções na coleta de dados, variações bruscas no tráfego ou indisponibilidade temporária de dispositivos dificultam a aplicação de técnicas de séries temporais, que dependem de dados históricos contínuos e consistentes. A abordagem proposta, baseada em medições instantâneas e complementares (como atraso e informações de *traceroute*), mostra-se mais robusta e adaptável, especialmente em cenários onde a coleta contínua de vazão é inviável ou sujeita a imprecisões. Dessa forma, o modelo consegue superar as limitações das medições diretas, oferecendo estimativas confiáveis mesmo em condições de rede desafiadoras.

Para validar a solução, foram realizados experimentos com dados reais da Rede

Nacional de Ensino e Pesquisa (RNP), extraídos do Serviço de Monitoramento da Rede Ipê (Monipê). Os resultados demonstram que a abordagem integrada, combinando múltiplas medições e treinamento baseado em agrupamentos por origem de comunicação, alcança elevada precisão nas estimativas. A análise baseada no RMSE evidencia a robustez do método, especialmente ao explorar padrões comportamentais semelhantes entre links monitorados. O restante deste artigo está organizado da seguinte forma: A Seção 2 apresenta trabalhos relacionados; a Seção 3 detalha a proposta; a Seção 4 descreve os experimentos realizados; a Seção 4.2 discute os resultados obtidos; e a Seção 5 apresenta conclusões e direções para pesquisas futuras.

2. Trabalhos Relacionados

Esta seção apresenta uma análise crítica dos principais trabalhos recentemente publicados pela comunidade científica sobre medição de desempenho de redes, com foco em técnicas, qualidade de serviço e técnicas de aprendizado de máquina.

Liang *et al.* [Liang et al. 2023] desenvolveram um modelo inovador baseado em deep learning capaz de estimar valores de média e variância da distribuição de probabilidades através da extração de informações de estruturas de grafo. Os autores propuseram um framework de probabilidade de predição para estimar a distribuição do valor de QoS, que pode ser adaptado para diversas aplicações, como tomada de decisão e detecção de anomalias. De forma complementar, Damaskinos *et al.* [Damaskinos et al. 2022] apresentaram um middleware para o sistema operacional Android que estima e controla o impacto computacional de tarefas de aprendizado de máquina em dispositivos móveis, focando especificamente no tempo de processamento e consumo de energia. Embora relevantes, estas abordagens não exploram especificamente o comportamento da rede nem abordam a estimativa de demanda por recursos.

Yang *et al.* [Yang et al. 2021a] propuseram um método híbrido para previsão de tráfego de rede, combinando o modelo ARIMA (AutoRegressive Integrated Moving Average) com uma Rede Neural Backpropagation (BP), otimizada por um algoritmo de Simulated Annealing (SA). O estudo demonstrou que a integração dessas técnicas melhorou a precisão das previsões de tráfego, especialmente em cenários onde os dados apresentam alta variabilidade. Os autores validaram o modelo com dados reais de tráfego de rede, alcançando resultados significativos em termos de precisão e redução de erros de previsão. No entanto, a abordagem de Yang et al. [2021] depende fortemente de dados históricos contínuos e de técnicas de séries temporais, o que pode ser uma limitação em cenários onde os dados de tráfego são escassos ou inconsistentes. Além disso, o foco do trabalho foi principalmente em redes com padrões de tráfego estáveis, o que pode limitar sua aplicabilidade em redes dinâmicas e complexas.

Aldhyani *et al.* [Aldhyani et al. 2020] propuseram um modelo híbrido inteligente para previsão de tráfego de rede, combinando técnicas de Fuzzy C-Means (FCM) e Suavização Exponencial Ponderada (WES) com modelos de séries temporais avançados, como LSTM (Long Short-Term Memory) e ANFIS (Adaptive Neuro-Fuzzy Inference System). O estudo demonstrou que a integração dessas técnicas melhorou a precisão das previsões de tráfego, especialmente em cenários onde os dados apresentam alta variabilidade. No entanto, a abordagem de Aldhyani et al. [2020] depende fortemente de dados históricos contínuos e de técnicas de séries temporais, o que pode ser uma limitação em

cenários onde os dados de tráfego são escassos ou inconsistentes. Além disso, o foco do trabalho foi principalmente em redes celulares, o que pode limitar sua aplicabilidade em redes de backbone ou em cenários com diferentes padrões de tráfego.

A partir desta revisão da literatura, observa-se que, embora existam trabalhos significativos na área, ainda há uma lacuna importante no desenvolvimento de modelos para estimar o desempenho de rede que sejam adaptáveis ao contexto específico da medição e que explorem efetivamente a integração de conjunto de dados de mesma origem como estratégia para superar limitações de dados disponíveis.

3. Proposta de Regressão Multivariada para Estimativa da Vazão

O método proposto estima a vazão de rede por meio de modelos de regressão multivariada, utilizando métricas de desempenho como variáveis explicativas. Entre elas, o atraso, registrado a cada minuto, e as informações de *traceroute*, coletadas a cada dez minutos, possuem frequência significativamente maior que os dados de vazão, medidos em intervalos mais longos devido ao alto custo de coleta.

A abordagem busca mitigar essa limitação integrando métricas de alta frequência para capturar variações detalhadas na rede, ampliando as informações disponíveis e melhorando a precisão das estimativas, mesmo em cenários dinâmicos.

A Figura 1 ilustra a solução proposta. O processo inicia com a coleta de dados de serviços compatíveis com ferramentas como Perfsonar, SolarWinds e Auvik. Em seguida, as informações de *traceroute* são analisadas, extraindo número de saltos e link de gargalo. Cada ponto de comunicação fim a fim é tratado individualmente, unificando *traceroute*, atraso e vazão por concatenação temporal. Por fim, os dados passam por processamento adicional e são utilizados no treinamento dos modelos, aprimorando a estimativa de desempenho da rede.

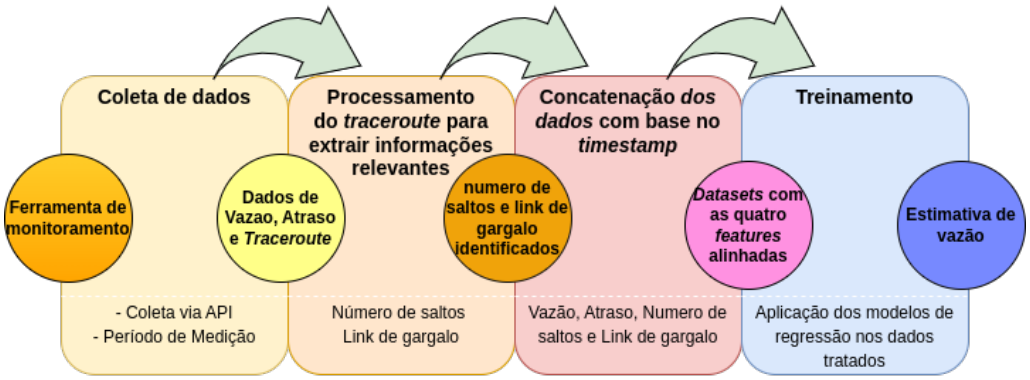

Figura 1. Visão Geral da Solução.

Cada conjunto de dados está associado a pares de pontos de comunicação específicos para um período de tempo selecionado. Com granularidades e características distintas, esses dados exigem um processo robusto de análise e tratamento para viabilizar sua utilização nos modelos de regressão.

Nesse contexto, os dados deste estudo foram obtidos da Rede Nacional de Ensino e Pesquisa (RNP), cuja ampla cobertura nacional e diversidade de cenários de rede tornam-na uma fonte ideal para a modelagem. Essa heterogeneidade foi essencial para a construção de estimativas mais robustas e alinhadas às condições reais das redes brasileiras.

3.1. Coleta e Processamento dos Dados

Conforme mencionado anteriormente, este trabalho utilizou três tipos de medições de rede: *traceroute* (coletado a cada dez minutos, extraindo número de saltos e identificação do link de gargalo), atraso (medido a cada minuto) e vazão (registrada a cada quatro horas, com lacunas significativas nos registros). O número de saltos foi determinado a partir dos dados de *traceroute*, baseando-se na contagem de links ao longo da rota de comunicação, representando o total de saltos intermediários entre a origem e o destino. Por outro lado, o link de gargalo foi identificado considerando as capacidades de rede informadas pela RNP, o que permitiu localizar o ponto de maior limitação de tráfego na rota. A identificação desses gargalos foi visualizada na Figura 2, que ilustra a capacidade de cada link da rede.

A etapa de junção dos dados é realizada com base nos *timestamps* próximos para sincronizar as medições, onde os *dataset* de Vazão, Atraso e *traceroute* são unificados em um único conjunto de dados. O processo consiste em ordenar todos os *dataset* pela coluna de *timestamp*, seguido pela junção sequencial dos dados com uma tolerância máxima de 10 minutos entre as medições correspondentes. Primeiramente, o dataset de Vazão foi utilizado como base temporal para a sincronização dos dados. Os *datasets* de Atraso e *traceroute* foram então unidos ao de Vazão, tomando como referência o *timestamp* das medições de Vazão e selecionando, para cada registro, os dados de Atraso e *traceroute* com *timestamps* mais próximos. Este método garantiu que todas as informações fossem alinhadas temporalmente de forma precisa, resultando em um *dataset* final coeso onde cada entrada representa um momento específico com todos os parâmetros relevantes (Vazão, Atraso e rotas de *traceroute*) devidamente sincronizados.

A escolha do intervalo de 10 minutos para unir as medições de vazão, atraso e *traceroute* foi feita para equilibrar granularidade e disponibilidade dos dados. Com o atraso registrado a cada minuto, o *traceroute* a cada 10 minutos e a vazão a cada 4 horas, esse intervalo garante correspondência entre as medições, evitando lacunas e perda de informações relevantes. Além disso, reduz o impacto de flutuações momentâneas, minimizando ruídos e preservando padrões significativos. Intervalos menores poderiam causar inconsistências, enquanto maiores comprometeriam a correlação. Assim, o intervalo de 10 minutos assegura uma sincronização eficiente e representativa.

Para garantir a robustez do modelo a ruídos e medições incorretas, os dados passaram por um processo de pré-processamento que incluiu a remoção de outliers e a normalização das métricas.

3.2. Modelos de Regressão

A escolha por modelos de regressão multivariada, em vez de abordagens baseadas em séries temporais como Redes Neurais Recorrentes (RNNs), foi motivada pela natureza das medições de vazão, que ocorrem em intervalos longos e apresentam lacunas frequentes. Modelos como LSTM exigem dados históricos contínuos para realizar previsões, o

que pode ser problemático em redes com medições irregulares e falhas na coleta. Em contraste, a regressão permite estimar a vazão com base em métricas complementares de maior frequência, como atraso e *traceroute*, sem exigir continuidade temporal rigorosa. Modelos como Random Forest, XGBoost e Gradient Boosting foram escolhidos por sua capacidade de lidar com não linearidade, alta dimensionalidade e variabilidade nos dados, sem o alto custo computacional do deep learning. O SVR foi incluído por sua robustez a outliers, e o Elastic Net, por combinar regularizações L1 e L2, sendo ideal para variáveis correlacionadas. Redes neurais convencionais foram descartadas devido à alta demanda computacional, priorizando-se alternativas mais eficientes e interpretáveis, como o Gradient Boosting, que equilibra precisão e custo computacional.

A seguir, são descritos os modelos utilizados neste estudo.

- O *Random Forest Regressor* é um modelo de aprendizado de máquina que utiliza um conjunto (*ensemble*) de árvores de decisão. Cada árvore é treinada com uma amostra aleatória dos dados (*bootstrap sampling*) e uma seleção aleatória de variáveis em cada divisão. A estimativa final é a média das estimativas das árvores, reduzindo o risco de *overfitting* e proporcionando robustez a ruídos nos dados. Este modelo é descrito por Breiman [Breiman 2001] em seu trabalho seminal sobre *Random Forests*, que destaca sua eficiência em problemas de alta dimensionalidade e sua robustez a ruídos nos dados.
- O *Gradient Boosting Regressor* é um método sequencial que ajusta modelos simples (*weak learners*) para corrigir erros residuais do modelo anterior, utilizando o gradiente da função de perda. Friedman [Friedman 2001] apresenta este modelo como uma abordagem inovadora e altamente eficiente para problemas de regressão, destacando seu desempenho em dados complexos.
- O *XGBoost Regressor* é uma implementação otimizada do *Gradient Boosting*, que utiliza regularização para evitar *overfitting* e algoritmos eficientes para computação paralela. Chen e Guestrin [Chen and Guestrin 2016] apresentam sua implementação em escala, destacando sua popularidade em competições de aprendizado de máquina.
- O *K-Nearest Neighbors Regressor* faz estimativas com base na média das observações mais próximas no espaço de características. Este modelo, introduzido por Cover e Hart [Cover and Hart 1967], é simples e intuitivo, mas pode ser sensível à escolha de k e à escala das variáveis.
- O *CatBoost Regressor* é otimizado para lidar com variáveis categóricas de forma eficiente, utilizando um algoritmo baseado em árvores. Prokhorenkova *et al*. [Prokhorenkova et al. 2018] detalham sua abordagem inovadora para lidar com viés e variância em variáveis categóricas.

Também foram testados outros modelos, incluindo LightGBM, SVR, Elastic Net, AdaBoost Regressor, regressão linear e polinomial. No entanto, esses métodos mostraram desempenho inferior no contexto desta pesquisa, reforçando a escolha dos modelos destacados anteriormente.

Para mitigar o risco de overfitting, especialmente em modelos mais complexos como Random Forest, Gradient Boosting e XGBoost, foi utilizada a técnica de validação cruzada (cross-validation) durante o treinamento. A validação cruzada permite avaliar a generalização do modelo, garantindo que ele não se ajuste excessivamente aos dados de treinamento. Além disso, hiperparâmetros como a profundidade máxima das árvores e o número de estimadores foram ajustados para evitar modelos excessivamente complexos, que poderiam capturar ruídos em vez de padrões reais nos dados.

A seleção de diversos modelos de regressão é crucial para avaliar a robustez e eficácia das abordagens em diferentes cenários de variabilidade dos dados. Métodos como Random Forest, XGBoost e Gradient Boosting combinam técnicas avançadas de aprendizado de máquina e ensemble, proporcionando maior capacidade de generalização, mesmo em dados de alta dimensionalidade. Já modelos como regressão linear e polinomial são analisados por sua simplicidade e eficiência computacional em cenários com relações menos complexas entre variáveis. Técnicas mais recentes, como CatBoost e LightGBM, foram incluídas por sua eficiência no tratamento de grandes volumes de dados e variáveis categóricas, comuns em medições de redes. Os hiperparâmetros de cada modelo foram ajustados utilizando *Random Search*, explorando combinações que maximizam o desempenho.

4. Avaliação e Resultados Experimentais

Esta seção apresenta a configuração dos experimentos realizados (Subseção 4.1) e os resultados obtidos (Subseção 4.2). Além disso, o código desenvolvido e os dados utilizados nos experimentos estão disponíveis no repositório do projeto[1], juntamente com instruções detalhadas para garantir sua reprodutibilidade.

4.1. Configuração de Experimentos

Para os experimentos com dados reais, foram utilizados conjuntos fornecidos pela RNP via Serviço MonIPÊ, que adota o padrão perfSONAR. As medições ocorrem em diferentes frequências: vazão a cada quatro horas, *traceroute* a cada dez minutos e atraso a cada minuto. Foram analisados pares de comunicação entre Pontos de Presença (PoPs) em diferentes estados do Brasil: PR-PI, RJ-GO, RO-CE, SC-PB, MA-SP, BA-ES, PA-SE, AP-TO, MT-MS, AC-RJ e RS-AM. A seleção abrange todas as regiões do país e a maioria dos estados, considerando diversidade geográfica, infraestrutura, distâncias e padrões de tráfego. A inclusão de pares extremos, como RS-AM (Sul-Norte), e intra-regionais, como MT-MS (Centro-Oeste), garante uma avaliação ampla e representativa das condições de rede no Brasil. Alguns estados do Brasil não foram incluídos nos pares de PoPs analisados devido à insuficiência de dados, o que comprometeria a confiabilidade da análise. A Figura 2 ilustra a rede Ipê da RNP e seus links de comunicação.

Os experimentos utilizaram dados do primeiro semestre de 2024 (janeiro a julho), abrangendo PoPs em quase todos os estados. As capacidades dos enlaces variam entre 300 Gbps e 1 Gbps, influenciando os padrões de comunicação e refletindo diferentes condições de rede no país. Essa seleção permite uma análise ampla e representativa da infraestrutura nacional. Os critérios de seleção dos dados foram: (i) Heterogeneidade geográfica, já que a localização dos PoPs impacta o número de enlaces no caminho fim a fim e a carga na infraestrutura; (ii) Capacidade da rede, pois a variação na carga dos enlaces altera o comportamento da comunicação entre diferentes PoPs. Esses fatores garantem uma amostra representativa da Rede Ipê.

Foi empregada a técnica de divisão de dados conhecida como *train-test split*, na qual 80% dos dados foram destinados ao treinamento do modelo e 20% para testes e validação. Essa divisão assegura uma análise sólida do desempenho dos modelos de regressão, considerando a diversidade e as características dos pontos de comunicação. Além

[1]http://github.com/LarcesUece/ESTIMATIVA-REGRESSAO-WGRS-2025

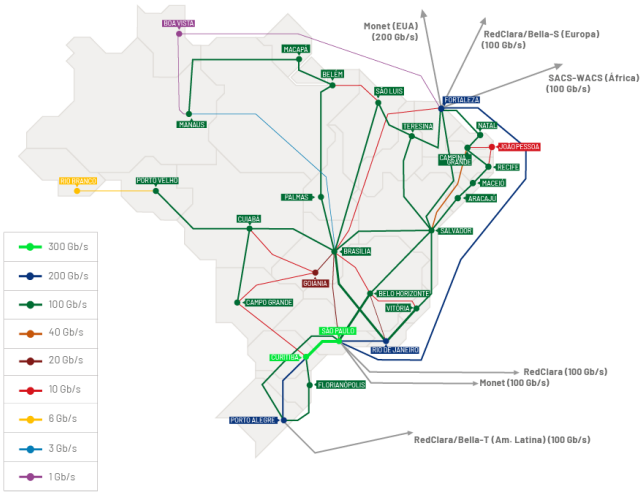

Figura 2. Rede Ipê da RNP.

disso, para garantir uma avaliação mais robusta e confiável do desempenho do modelo, foi utilizada a validação cruzada (*cross-validation*) com 3 folds. Essa abordagem permite que o modelo seja treinado e validado em diferentes subconjuntos dos dados de treinamento, reduzindo o risco de overfitting e proporcionando uma estimativa mais precisa da generalização do modelo para dados não vistos.

Após o treinamento, o *Root Mean Square Error* (RMSE) foi calculado para comparar os valores estimados com os reais. Essa métrica, normalizada com base na média dos valores reais, permite validar a precisão das estimativas realizadas pelos modelos. Em seguida, a acurácia dos resultados foi avaliada com base em uma classificação fornecida pela RNP, o que possibilitou uma análise mais detalhada e consistente do desempenho dos modelos.

Com base nos resultados preliminares, foram selecionados cinco modelos que apresentaram melhor desempenho para uma análise mais aprofundada. Dada a extensão do estudo, que envolve onze datasets distintos e onze modelos de regressão diferentes, a inclusão de todas as tabelas com hiperparâmetros específicos para cada combinação modelo-dataset tornaria o artigo excessivamente extenso e potencialmente confuso para o leitor. Por esta razão, optou-se por não incluir estas informações detalhadas no corpo do artigo. Todos os hiperparâmetros utilizados em cada modelo e para cada dataset, bem como as configurações experimentais completas, estão disponibilizados de forma integral e organizada no repositório público do trabalho, garantindo assim a transparência e reprodutibilidade dos resultados sem comprometer a concisão do documento.

4.2. Resultados

Foram analisados os valores de RMSE (Root Mean Squared Error) e o desempenho dos modelos para cada par de pontos de comunicação. Os modelos com menor RMSE médio, baixa variabilidade de erros e maior consistência nas estimativas foram selecionados para uma análise mais detalhada. Nesse processo, os modelos KNeighborsRegressor, GradientBoostingRegressor, XGBRegressor, RandomForestRegressor e CatBoostRegressor se destacaram, conforme mostrado na Tabela 1.

A Figura 3 apresenta a análise do Root Mean Square Error (RMSE) em megabits/s

Modelo	Média	Desvio Padrão	Mínimo	Máximo	Mediana
KNeighborsRegressor	**6.98**	5.56	2.09	20.98	5.00
GradientBoostingRegressor	**8.66**	7.31	3.43	27.74	5.08
XGBRegressor	**9.25**	8.85	3.09	33.59	4.72
RandomForestRegressor	**9.61**	8.25	3.42	31.67	5.23
CatBoostRegressor	**10.45**	10.21	3.70	37.60	4.71
AdaBoostRegressor	11.42	10.11	4.21	37.69	6.38
LGBMRegressor	12.06	9.65	4.80	37.66	9.37
PolynomialRegression	11.60	10.42	4.13	39.44	6.52
ElasticNet	12.60	9.70	4.38	37.41	13.43
LinearRegression	12.65	9.82	4.42	37.90	13.22
SVR	13.20	11.08	4.07	42.09	13.65

Tabela 1. Avaliação Comparativa dos Modelos de Regressão.

entre os diferentes pares de pontos de comunicação e os cinco modelos de regressão selecionados.

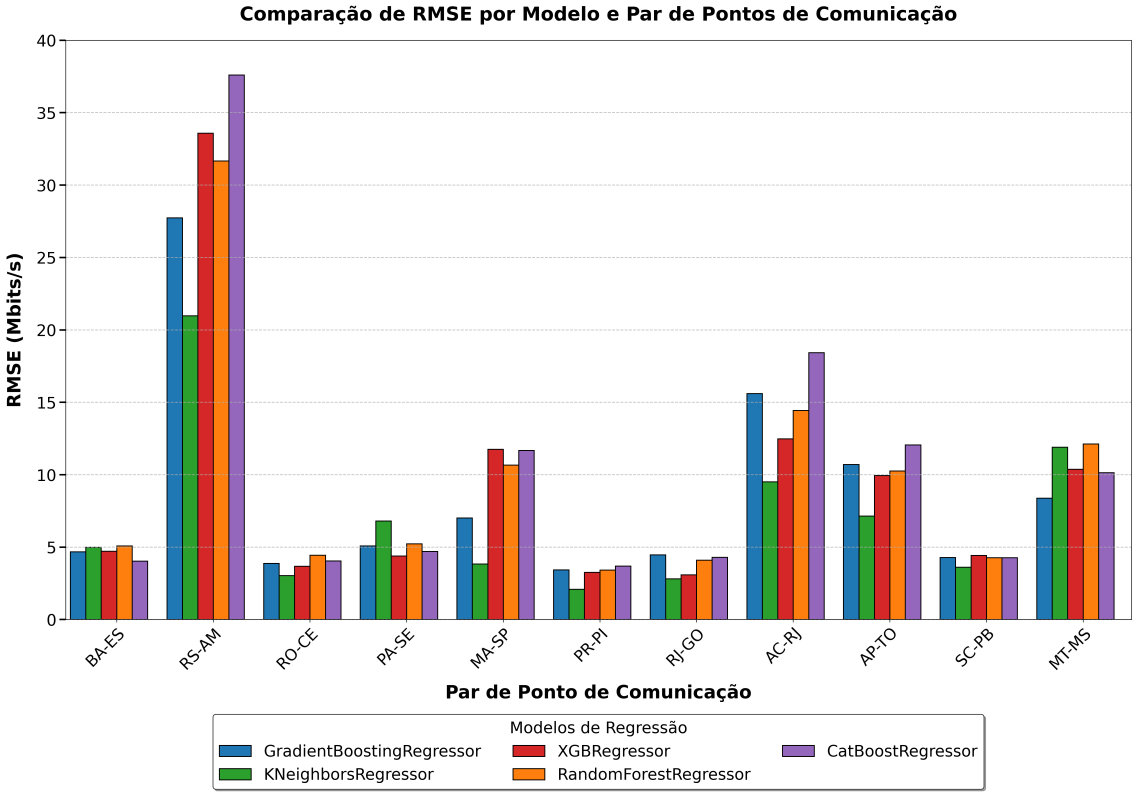

Figura 3. Resultados do RMSE das Estimativas de Vazão Para Cada Par de Ponto de Comunicação.

A avaliação do Erro Quadrático Médio (RMSE) entre diferentes pares de pontos de comunicação revela um panorama complexo do desempenho de modelagem preditiva. A análise desvenda insights críticos sobre os desafios e pontos fortes de diversas técnicas de regressão na estimativa de fluxo de tráfego de rede.

Os valores de RMSE, variando de 3 a 38 Mbits/s, destacam a natureza intrincada

da previsão de tráfego de rede. A conexão RS-AM (Rio Grande do Sul à Amazônia) se destaca como o cenário mais desafiador, com valores de RMSE alcançando até 38 Mbits/s. Essa variação pode ser atribuída a fatores fundamentais como a complexidade geográfica e a heterogeneidade da infraestrutura de rede entre diferentes regiões do país.

Em contrapartida, conexões como PR-PI (Paraná a Piauí), RO-CE (Rondônia ao Ceará) e SC-PB (Santa Catarina a Paraíba) exibem precisão notável, com valores de RMSE entre 3 e 8 Mbits/s. Esse resultado sugere que as características locais da rede podem ser mais influentes do que a distância geográfica na determinação da precisão da estimativa.

Um aspecto crítico da análise reside na proporcionalidade do RMSE em relação às taxas de fluxo médio. Mesmo no cenário mais desafiador — a conexão RS-AM — o RMSE de aproximadamente 35 Mbits/s representa apenas 4% da taxa de fluxo média de 860 Mbits/s. Isso contextualiza o erro, demonstrando a robustez dos modelos utilizados.

A variabilidade de desempenho entre diferentes modelos de regressão oferece insights preciosos. O KNeighborsRegressor se destaca, apresentando performance excepcional em padrões de tráfego complexos, especialmente na conexão RS-AM. Já a conexão MA-SP (Maranhão a São Paulo) revela a sensibilidade dos modelos, com RMSE variando de 4 (KNeighborsRegressor) a 11 (XGBRegressor e CatBoostRegressor).

Os resultados enfatizam a importância da seleção criteriosa de modelos para segmentos de rede específicos. Ficou evidente que a estimativa de tráfego de rede é mais influenciada pela infraestrutura local e padrões de tráfego do que pela distância geográfica. Apesar das variações, os modelos mantêm percentuais de erro baixos, garantindo estimativas de tendência confiáveis, essenciais para a gestão de redes.

As descobertas sublinham três pontos fundamentais para modelagem de desempenho de redes: a necessidade de seleção de modelo sob medida para segmentos de rede específicos, a consideração de características locais na modelagem e o reconhecimento de que pequenos percentuais de RMSE podem fornecer insights valiosos para otimização de redes.

A pesquisa demonstra que, embora nenhum modelo prove-se universalmente superior, uma abordagem detalhada e cautelosa de seleção de modelo pode aprimorar significativamente as estimativas de fluxo de tráfego de rede.

A Tabela 2 apresenta os resultados de acurácia por faixas de vazão, de acordo com os critérios de classificação estabelecidos pela RNP. Para essa análise, os valores de vazão foram divididos em cinco faixas distintas: abaixo de 200 Mbps, entre 200-500 Mbps, entre 500-800 Mbps, entre 800-1000 Mbps e acima de 1000 Mbps. A acurácia foi calculada com base na comparação entre os valores preditos e os valores reais dentro de cada faixa de vazão. Ou seja, foi verificado se os valores preditos caíam dentro da faixa correspondente ao valor real, considerando a precisão das estimativas em cada intervalo de vazão. Isso permitiu uma avaliação mais precisa da capacidade do modelo em estimar a vazão dentro dos limites definidos para cada faixa.

Os resultados demonstram uma excelente performance de classificação por faixas para a maioria dos modelos de ensemble testados, com acurácia de 100% em dez dos onze pares de pontos de comunicação analisados. Apenas a conexão RS-AM apresen-

Pontos de Comunicação	RFReg	GBReg	XGBReg	CBReg	KNNReg
PA-SE	100.00%	100.00%	100.00%	100.00%	100.00%
PR-PI	100.00%	100.00%	100.00%	100.00%	100.00%
RS-AM	92.54%	94.03%	93.28%	93.28%	96.27%
AP-TO	100.00%	100.00%	100.00%	100.00%	100.00%
SC-PB	100.00%	100.00%	100.00%	100.00%	100.00%
AC-RJ	100.00%	100.00%	100.00%	100.00%	100.00%
MT-MS	100.00%	100.00%	100.00%	100.00%	100.00%
BA-ES	100.00%	100.00%	100.00%	100.00%	100.00%
RJ-GO	100.00%	100.00%	100.00%	100.00%	100.00%
RO-CE	100.00%	100.00%	100.00%	100.00%	100.00%
MA-SP	100.00%	100.00%	100.00%	100.00%	100.00%

Tabela 2. Tabela de Acurácia por Modelo e Pares de Ponto de Comunicação.

tou acurácia ligeiramente inferior, variando entre 92,54% e 96,27% dependendo do modelo utilizado, com o KNeighborsRegressor (KNNReg) alcançando o melhor desempenho (96,27%). Para fins de concisão, os modelos são referidos na tabela com as seguintes abreviações: RandomForestRegressor (RFReg), GradientBoostingRegressor (GBReg), XGBRegressor (XGBReg), CatBoostRegressor (CBReg) e KNeighborsRegressor (KNNReg).

Estes resultados são particularmente relevantes do ponto de vista prático, pois mesmo nos casos onde o RMSE é mais elevado, os modelos conseguem classificar corretamente a faixa de vazão, o que é suficiente para muitas aplicações de gerenciamento de rede e planejamento de capacidade.

4.3. Análise Comparativa e Considerações Finais

A análise comparativa dos modelos revela que KNeighborsRegressor, GradientBoostingRegressor e XGBRegressor destacam-se como as alternativas mais eficazes para estimativa de vazão em redes de comunicação. Como mostrado na Tabela 1, estes modelos apresentaram consistentemente métricas superiores de desempenho.

O KNeighborsRegressor demonstrou desempenho excepcional, alcançando o menor RMSE médio (6.98) entre todos os modelos avaliados. Este resultado, combinado com sua acurácia perfeita (100%) em 10 dos 11 cenários de fonte-destino analisados, reforça a hipótese de que os padrões de tráfego de rede frequentemente seguem comportamentos locais bem definidos. No par RS-AM, único cenário onde não atingiu 100% de acurácia, o modelo ainda obteve o melhor resultado comparativo (96.27%), superando todos os outros algoritmos. Isso sugere que sua capacidade de identificar e explorar similaridades em padrões de tráfego é particularmente valiosa em cenários com maior variabilidade ou complexidade topológica.

O GradientBoostingRegressor demonstrou excelente equilíbrio entre precisão e robustez, com RMSE médio de 8.66 e mediana de 5.08. Sua abordagem iterativa de aprendizado, que aprimora modelos em sequência corrigindo erros de estimativas anteriores, mostra-se especialmente adequada para capturar as relações não-lineares presentes nos dados de tráfego de rede. Com acurácia de 100% em 10 dos 11 cenários e 94.03% no

par RS-AM, o modelo apresenta confiabilidade consistente mesmo em condições desafiadoras, sugerindo capacidade de generalização em diferentes contextos de rede.

O XGBRegressor destacou-se por apresentar a menor mediana de RMSE, indicando desempenho superior em pelo menos 50% dos casos analisados. Este comportamento, combinado com sua acurácia de 100% na maioria dos cenários (93.28% no par RS-AM), sugere que o modelo é particularmente eficaz em condições típicas de operação de rede. Suas características intrínsecas de regularização e mecanismos para controle de overfitting contribuem para maior robustez em cenários com alta dimensionalidade ou variabilidade, tornando-o uma escolha confiável para implementação em ambientes operacionais.

A análise integrada dos resultados de RMSE e acurácia por faixa demonstra que os três modelos destacados (KNeighborsRegressor, GradientBoostingRegressor e XGBRegressor) oferecem um excelente equilíbrio entre precisão, robustez e consistência para estimativa de vazão em redes.

Além da precisão, a escolha do modelo deve considerar os custos computacionais, especialmente em cenários de larga escala. Modelos mais simples, como KNeighborsRegressor, oferecem alternativas computacionalmente leves para contextos com restrições de processamento. Já modelos mais complexos, como XGBRegressor e GradientBoostingRegressor, demandam maior capacidade computacional durante o treinamento, mas são eficientes na inferência, tornando-os viáveis para aplicações em tempo real. Além disso, a configuração dos hiperparâmetros, como o número de estimadores e a profundidade das árvores, foi otimizada para reduzir o custo computacional sem comprometer a precisão das estimativas. Além disso, a necessidade de agregar e sincronizar medições de diferentes granularidades também impõe desafios no pré-processamento, impactando a performance geral. Dessa forma, a seleção do modelo ideal deve buscar um compromisso entre precisão, eficiência e escalabilidade para viabilizar soluções robustas em ambientes operacionais.

5. Conclusão

A estimativa de desempenho de redes é crucial para o planejamento estratégico de provedores de serviços de Internet e a qualidade da experiência do usuário. As lacunas em séries temporais, causadas por falhas nas medições, representam um desafio, mas a abordagem de estimativa pontual proposta contorna essa limitação, permitindo previsões robustas mesmo com dados incompletos.

Embora validado com dados da RNP, o modelo apresentado é genérico e aplicável a outras redes, como corporativas ou de provedores de Internet, adaptando-se a diferentes topologias e padrões de tráfego por meio de métricas como atraso e número de saltos. Além disso, apesar do uso de *traceroute* para identificar links de gargalo e saltos, essa dependência pode ser mitigada quando o *traceroute* não está disponível. Nesses casos, métricas como atraso e perda de pacotes podem inferir a topologia da rede.

As principais contribuições deste trabalho incluem: (1) uma metodologia eficaz para estimativa pontual de vazão, (2) validação com dados reais de uma rede de grande porte, (3) comparação de desempenho entre modelos de regressão, e (4) diretrizes práticas para seleção e implementação de modelos em diferentes cenários operacionais, considerando o equilíbrio entre precisão, custo computacional e escalabilidade.

Futuros trabalhos explorarão técnicas de aprendizado profundo, abordagens híbridas e a validação do modelo em outros cenários, como redes de data centers e móveis.

6. Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (N^o 303877/2021-9 e N^o 405976/2022-4) pelo apoio financeiro e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

Aldhyani, T. H. H., Alrasheed, M., Alqarni, A. A., Alzahrani, M. Y., and Bamhdi, A. M. (2020). Intelligent hybrid model to enhance time series models for predicting network traffic. *IEEE Access*, 8:130431–130451.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Damaskinos, G., Guerraoui, R., Kermarrec, A.-M., Nitu, V., Patra, R., and Taiani, F. (2022). Fleet: Online federated learning via staleness awareness and performance prediction. *ACM Trans. Intell. Syst. Technol.*, 13(5).

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.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Gijon, C., Mahmoodi, T., Toril, M., Luna-Ramírez, S., and Bejarano-Luque, J. (2024). Sla-driven traffic steering in b5g systems with network slicing. *IEEE Transactions on Vehicular Technology*.

Gomes, R. L., Bittencourt, L. F., and Madeira, E. R. (2014a). A bandwidth-feasibility algorithm for reliable virtual network allocation. In *2014 IEEE 28th International Conference on Advanced Information Networking and Applications*, pages 504–511.

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.

Liang, W., Li, Y., Xu, J., Qin, Z., Zhang, D., and Li, K.-C. (2023). Qos prediction and adversarial attack protection for distributed services under dlaas. *IEEE Transactions on Computers*, pages 1–1.

Mok, R. K. P., Zou, H., Yang, R., Koch, T., Katz-Bassett, E., and Claffy, K. C. (2021). Measuring the network performance of google cloud platform. In *Proceedings of the*

21st ACM Internet Measurement Conference, IMC '21, page 54–61, New York, NY, USA. Association for Computing Machinery.

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.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, pages 6638–6648.

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.

Souza, M. S., Ribeiro, S. E. S. B., Lima, V. C., Cardoso, F. J., and Gomes, R. L. (2024). Combining regular expressions and machine learning for sql injection detection in urban computing. *Journal of Internet Services and Applications*, 15(1):103–111.

Yang, H., Li, X., Qiang, W., Zhao, Y., Zhang, W., and Tang, C. (2021a). A network traffic forecasting method based on sa optimized arima–bp neural network. *Computer Networks*, 193:108102.

Yang, X., Wang, X., Li, Z., Liu, Y., Qian, F., Gong, L., Miao, R., and Xu, T. (2021b). Fast and light bandwidth testing for internet users. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*, pages 1011–1026, Berkeley, CA. USENIX Association.