UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM

INSTITUTO DE COMPUTAÇÃO- ICOMP

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA - PPGI

O USO DE HISTÓRICO DE ACESSO NA DEFINIÇÃO DE TAXA DE BITS EM SESSÕES DE VÍDEO COM TAXA ADAPTÁVEL

Tonny Franck Osaki da Paz

MANAUS-AM

Agosto 2019

Tonny Franck Osaki da Paz

O USO DE HISTÓRICO DE ACESSO NA DEFINIÇÃO DE TAXA DE BITS EM SESSÕES DE VÍDEO COM TAXA ADAPTÁVEL

Dissertação apresentada ao Curso de Pós-Graduação em Informática da Universidade Federal do Amazonas como requisito para a obtenção do grau de Mestre em Informática.

Orientador: Prof. Dr. César Augusto Viana Melo

MANAUS-AM

Agosto 2019

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Paz, Tonny Franck Osaki da
P348u O Uso de Histórico de Acesso na Definição de Taxa de Bits em Sessões de Vídeo com Taxa Adaptável / Tonny Franck Osaki da Paz . 2019
73 f.: il. color; 31 cm.

Orientador: Cesar Augusto Viana Melo
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Medição e Predição de Vazão. 2. Streaming de Vídeo. 3. Vazão Realizável. 4. Aprendizagem de Máquina. I. Melo, Cesar Augusto Viana. II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"O USO DE HISTÓRICO DE ACESSO NA DEFINIÇÃO DE TAXA DE BITS EM SESSÕES DE VÍDEO COM TAXA ADAPTÁVEL "

TONNY FRANCK OSAKI DA PAZ

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. César Augusto Viana Melo - PRESIDENTE

Profa. Eulanda Miranda dos Santos - MEMBRO INTERNO

Prof. Rafael Giusti - MEMBRO EXTERNO

Manaus, 21 de Agosto de 2019

Este trabalho é dedicado a toda minha família,
pelo apoio durante essa fase da minha vida,
especialmente aos meus pais.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pela vida, força, coragem, sabedoria e persistência em alcançar meus objetivos durante esse trabalho. Aos meus familiares e amigos, pelo apoio, incentivo, paciência e compreensão sempre que precisei.

Aos meus Professores, do Instituto de Computação (ICOMP) da Universidade Federal do Amazonas (UFAM), em especial ao professor Dr. César Augusto Viana Melo, por toda atenção, ensinamentos, conselhos e orientação durante o desenvolvimento deste trabalho.

Por fim, a todos os meus amigos e colegas, que me ajudaram e colaboraram de forma direta e indiretamente para o desenvolvimento deste trabalho.

RESUMO

As aplicações de distribuição de vídeo estão entre as que mais geram tráfego na Internet. Estima-se que tais aplicações participarão com 82% de todo o tráfego gerado em 2022. A tecnologia mais usada por essas aplicações fazem a distribuição centrada na ação dos clientes que usam informações da sessão em andamento para tomar decisão sobre a qualidade das imagens e a continuidade da sessão. Essa tomada de decisão tem como um dos principais fatores a vazão fim-a-fim das conexões estabelecidas entre o cliente e os servidores de vídeos. Uma parcela importante das estratégias já implementadas para a tomada de decisão utiliza apenas dados coletados durante a sessão. Nesse cenário, a observação contínua da vazão do canal e suas peculiaridades são essenciais para realização de predições de maior precisão, dada a natureza instável do canal. Entretanto, eventos gerados pela audiência ou pelas estratégias do *TCP* produzem períodos de ausência de observação da vazão, induzindo tais estratégias a comportamentos erráticos. Nesta dissertação, apresenta-se o conceito de taxa realizável, que pode auxiliar as estratégias de adaptação na tomada de decisão em aplicações de vídeo *streaming*. Usam-se dados sobre o canal, coletados em medições ativas, para definição de tal taxa. Métodos de *Aprendizado de Máquina* foram empregados para construção de modelos preditivos. As avaliações dos modelos mostram que as predições são precisas e aplicáveis em sessões de vídeo *streaming*, sendo possível, por exemplo, a redução do período inicial de prospecção, e até mesmo um equilíbrio maior entre continuidade e qualidade das imagens.

Palavras-chaves: Medição e Predição de Vazão; Streaming de Vídeo; Vazão Realizável; Aprendizagem de Máquina.

ABSTRACT

Video related traffic is prevalent on the Internet. By 2022, this traffic will be responsible for 82% of all data transfer throughout the Internet. The HTTP Adaptive Streaming (HAS) is the most popular technology behind modern video applications. Its millstone approaching is to use data collected by clients to build sessions that balance image quality and continuity. The connection between HAS-based clients and servers is a crucial factor in finding that balance. A number of these clients use data from ongoing sessions. In these scenarios, continuous measurements and a clear understanding of essential elements are crucial to estimate the connection throughput due to its natural instability. However, client-side events have the potential to steam slots of silence in those measurements that drive the embedded algorithms for adaptive streaming to erratic behavior. In this work, the concept of achievable rate is introduced based on data collected by proactive monitors. Machine Learning methods are applied to those data to build throughput predictive models, and numerical studies are carried to assess its accuracy. The assessed accuracy shows that the conceived models can improve the decision made by those algorithms that use ongoing sessions data to find the balance between session continuity and image quality.

Key-words: Throughput Measurement; Throughput Prediction; Video streaming; Achievable Throughput; Machine Learning.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

As aplicações de distribuição de vídeo estão entre as que mais geram tráfego na Internet. Um estudo recente mostrou que o consumo de vídeo tem aumentado exponencialmente e estima-se que até 2022 represente mais de 82% de todo o tráfego gerado na Internet, tendo sido responsável por 75% deste tráfego em 2018 [1].

A tecnologia que tem sido empregada pelos grandes distribuidores de conteúdo, i.e., Netflix, YouTube e similares, é o fluxo com taxa de bit adaptável, do inglês *Adaptive Bit Rate (ABR)*, padronizado sob o arcabouço DASH - *Dinamic Adaptive Streaming over HTTP (do inglês, Hypertext Transfer Protocol)*. Com essa tecnologia, a otimização do *playtime*, que expressa a relação entre a duração da sessão e a duração do vídeo, ocorre com a adequação da qualidade da sessão, medida pela combinação de diferentes métricas, às condições do canal de transmissão. Para fim de normalização, assume-se que uma sessão de acesso é o registro da solicitação de um vídeo, do início ao fim. A viabilização dessa adequação deve-se a uma rotina de preparação da mídia que produz k versões do conteúdo original, e essas versões são segmentadas sob a mesma régua temporal, permitindo que se use as diferentes versões sem prejuízo ao tempo de duração da sessão. As versões preparadas são mantidas sob a gerência de um servidor, que também possui um arquivo descritor (MPD) dessas versões, devendo este ser o primeiro objeto acessado na sessão.

As estratégias de adaptação da mídia às condições do canal de transmissão abstraem os detalhes da implementação do serviço de entrega do conteúdo, apreendendo as medidas chaves usadas para decidir quando será feita uma nova requisição e qual a qualidade do próximo segmento a ser acessado. A vazão do canal de transmissão e o tempo de vídeo já transferido não reproduzido são informações chaves para a tomada de decisão acerca da qualidade do próximo segmento.

O desafio da predição da vazão de um canal de transmissão está na identificação das variáveis mais relevantes e na forma como elas se relacionam. A medição passiva caracteriza-se pelo registro dessa vazão durante a troca de mensagens de uma aplicação cujo propósito

é outro que tal medição, enquanto que a medição ativa tem como característica o uso de uma aplicação especificamente projetada para medir tal vazão. No caso de uma medição passiva, o valor medido pode refletir: *i*) a ação dos mecanismos de redução de latência, i.e., a hierarquia de caches; *ii*) a contribuição da alocação dinâmica de recursos, i.e., entrada/saída de servidores do conjunto de recursos disponíveis; *iii*) a carga de processamento que está a espera de atendimento, no servidor para qual uma requisição foi direcionada; e *iv*) o nível de compartilhamento do canal, expresso pela presença de fluxos de natureza relativamente agressiva na demanda por recursos de transmissão. Na medição ativa da vazão, pode-se isolar fatores que se deseja observar para melhor compreensão da sua contribuição na vazão observada. Por exemplo, pode-se: *i*) realizar o agendamento das medições; *ii*) dimensionar os servidores para que estes estejam habilitados a responder às demandas sem inserção de novos atrasos; e *iii*) garantir que a comunicação, durante toda a sessão, envolva somente os sistemas finais, impedindo assim a ação de otimizadores de latência.

Neste trabalho, apresenta-se o conceito da *vazão realizável* de um canal. Para tal, serão utilizadas informações coletadas por medidores ativos de vazão, que incorporam as dinâmicas de uma aplicação de *streaming* de vídeo, portanto, são percebidos pela rede como se fossem uma instância da aplicação mimetizada do cliente. Vislumbra-se o uso da vazão realizável, disponibilizada a partir de modelos preditivos, como fator de ajuste na tomada de decisão das estratégias de adaptação de um fluxo de vídeo.

1.1 Motivação

O emprego da tecnologia ABR nem sempre é sinônimo de engajamento da audiência e qualidade das sessões. Em [2] mostrou-se que 10% das sessões foram abandonadas antes do início da reprodução, e que 8% das sessões iniciadas com sucesso experimentaram pelo menos um evento de interrupção. A predição da largura de banda empregada na grande maioria das estratégias de adaptação de fluxo, mapeadas em [3], usa apenas dados coletados durante a sessão e consequentemente introduz um período de espera na sessão, *startup delay*, que tem a função de prospecção de capacidade de transferência do canal. O *startup delay* está presente tanto na abordagem mono quanto multi servidor, sendo que nesse último caso, ao seu final, um *rank* baseado na vazão das conexões pode ser gerado para os servidores disponíveis [4]. A duração do *startup delay* é dependente da implementação

da estratégia de adaptação do fluxo e seu valor expressa o difícil balanceamento entre a continuidade e a qualidade visual da sessão.

A prospecção inicial mostrou-se necessária, mas ainda é insuficiente para a construção de sessões com alta qualidade. A complexidade gerada pela interação das variáveis presentes no estabelecimento e na manutenção de um canal fim-a-fim é o motivo para tal dificuldade. Por exemplo, em [5], relatou-se como a ação do mecanismo de controle de congestionamento do TCP, que ao reiniciar sua janela de congestionamento (*cong_wind*) após período de silêncio da fonte, induz a estratégia de adaptação ABR a um comportamento errático em espiral. Ao reiniciar a *cong_wind*, o próximo segmento, que teve qualidade definida em função da vazão observada antes do período de silêncio, irá ser transmitido sob a dinâmica da partida lenta (que é o procedimento de inicia-se a reprodução do conteúdo na taxa de bits mais baixa e aos poucos a taxa de bits vai se adaptando a vazão do cliente), gerando uma vazão menor que aquela experimentada anteriormente. Diante de um cenário de severo desalinhamento da taxa de bits dos segmentos e a vazão do canal, impõe-se a redução da qualidade do próximo segmento a ser acessado, retardando a chegada ao estado de prevenção de congestionamento, portanto, reduzindo o período de permanência neste estado. Essa permanência reduzida acarreta em uma percepção incorreta da vazão, conduzindo a estratégia de adaptação a construção de sessões de baixa qualidade, em termo de taxa de bits.

Os períodos de silêncio durante uma sessão são provocados por ações em dois espaços: da audiência e da estratégia de adaptação. No espaço da audiência, as pausas na reprodução interrompem, por questões de otimização de recursos, a transferência de novos segmentos, até que a sessão seja retomada. Nesse cenário, o período de silêncio é dependente de uma ação externa à aplicação e, portanto, tem duração aleatória. No espaço da estratégia de adaptação à iminência de sobrecarga do *buffer*, o tempo demandado para a formulação de novas requisições, e a seleção de um novo servidor são ações que também geram períodos de silêncio no canal. A duração do silêncio é dependente da ação, por exemplo, a iminência de sobrecarga do *buffer* pode gerar um silêncio igual à duração de segmento, 5 a 10 segundos.

Outra fonte de ruído no processo de decisão que define a qualidade do próximo

segmento a ser acessado são as tecnologias implementadas para redução da latência. Em [6], mostrou-se que o uso de cache é uma estratégia eficiente para redução de latência no caso do tráfego de vídeo. O reuso do conteúdo desse tráfego em sessões independentes produz uma percepção incorreta da vazão do canal, super dimensionando-a. A distinção das fontes de atendimento de uma requisição exige adaptações do serviço de entrega do conteúdo, e/ou a implementação de elementos de redes que sejam cientes do conteúdo transportado e que possam, portanto, incluir informações úteis para uma melhor interpretação dos recursos que estão à disposição da sessão em curso[7].

As estratégias de adaptação de taxa consideram o estado corrente da sessão para tomada de decisão sobre qual é a taxa de bits mais adequada em um certo instante de tempo[3]. Nesse contexto, os períodos de silêncio têm sido ignorados, ver [5], causando a degradação da sessão, ou têm sido percebidos como fator de imposição de redução da qualidade da sessão, ver [4]. Acredita-se que o tratamento desses eventos demandam informações históricas sobre as sessões de vídeo e como a vazão foi afetada ao longo da sessão.

1.2 Objetivo

Estudar vazão de canais fim-a-fim a partir de medições gerada pela ação de agentes autônomos que simulam o comportamento de uma aplicação de *streaming* de vídeo.

Para alcançar tal objetivo, definiu-se os seguintes objetivos específicos:

1. Analisar e avaliar a base de dados em busca de características e especificações que estejam diretamente relacionadas com estabilidade e vazão dos canais;

2. Avaliar métodos de predição de vazão que sejam representativos das diferentes classes de preditores de séries temporais;

3. Avaliar diferentes abordagens de tratamento dos dados e treinamentos dos métodos de aprendizagem de máquina clássico para predição de vazão que reconheça as limitações da base;

1.3 Contribuições da Dissertação

As principais contribuições deste trabalho são estas:

1. Disponibilizar a base de dados com informações sobre comunicação fim-a-fim em infraestruturas de ISP (do inglês, *Internet Service Provider*) nacionais;
2. Definir e caracterizar o conceito de taxa realizável de uma comunicação fim-a-fim, no contexto de aplicações de fluxo de vídeo com taxa de bits adaptável;

1.4 Estrutura da Dissertação

O Capítulo 2 apresenta conceitos empregados no desenvolvimento desta dissertação, tais como *Streaming* Adaptativo, Qualidade de Experiência, séries temporais, aprendizado de máquina e preditores baseados em médias. O Capítulo 3 apresenta trabalhos relacionados em dois contextos: predição em histórico recente e predição usando histórico de sessões. O Capítulo 4 descreve a base de dados usada, seguido dos tratamentos e análises realizados na mesma. O Capítulo 5 apresenta os resultados de experimentos realizados para avaliar os diferentes modelos de predição de séries temporais. O Capítulo 6 apresenta as considerações finais do trabalho e trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo são apresentados os principais conceitos relacionados à realização desta dissertação. Este capítulo organiza-se da seguinte forma: na Seção 2.1 apresenta-se a tecnologia mais utilizada para transmissão de conteúdo multimídia pela *Internet*. Na Seção 2.2 apresentam-se os conceitos dos métodos de avaliação de QoE (do inglês, *Quality of Experience*). Na Seção 2.3 apresentam-se alguns conceitos fundamentais para entender o processo de adaptação dos vídeos aos reprodutores. Na Seção 2.4 descrevem-se conceitos de séries temporais. A Seção 2.5 apresentam-se os conceitos de Aprendizagem de Máquina e dos demais algoritmos que foram usados durante a pesquisa. Por fim, na Seção 2.6 são apresentadas as considerações finais desta dissertação.

2.1 Streaming de Vídeo Adaptativo

O *streaming* de vídeo com taxa adaptável é a tecnologia mais utilizada para a transmissão de vídeo pela Internet. A sua principal característica é adaptar a taxa de bits do vídeo às condições da rede durante a sessão. A seleção da taxa de bits, i.e., quantidade de bits transmitidos em um determinado intervalo de tempo, é possível através da disponibilidade de versões do mesmo vídeo preparado com taxas de bits distintas, armazenadas no servidor de conteúdo ou servidor de *streaming* [8]. A Figura 1 mostra a ideia geral da transmissão de vídeo com taxa adaptável.

O conteúdo é preparado de forma que possa ser reproduzido um vídeo à medida que é baixado e carregado no *buffer* do reprodutor. A ideia é dividir o conteúdo original de entrada em uma série de fragmentos, com duração entre dois e dez segundos. Esses fragmentos são codificados em taxas de bits diferentes, que são então chamados de segmentos. O acesso ao vídeo a partir desses segmentos considera a vazão do canal fim-a-fim, estabelecida entre aplicação e o servidor. Além disso, as condições atuais da aplicação, especificamente a qualidade de vídeos já transferidos, mas não reproduzida, é considerada.

Normalmente, inicia-se a reprodução do conteúdo na taxa de bits mais baixa, pois,

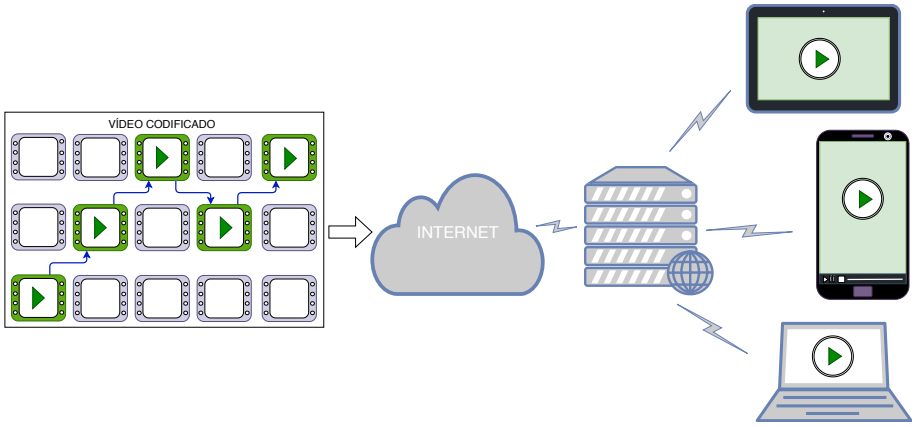


Figura 1 – Tecnologia de *Streaming* Adaptativo de Vídeo

Fonte: Produção própria

não há informação das condições da rede no momento do início da sessão. Se durante a transmissão do vídeo a dinâmica da adaptação de taxa detectar que a taxa de transferência excedeu a taxa de bits do segmento atual, o próximo segmento deve ter taxa de bits mais alta, caso contrário, ele solicita um fragmento com taxa de bits mais baixa. Essa tecnologia foi desenvolvida com finalidade de minimizar as deficiências das abordagens anteriores, permitindo: segmentação do vídeo, a transmissão contínua do *streaming* e o descarte do conteúdo depois de ser apresentado à audiência.

As transmissões adaptativas são baseadas na estrutura de HTTP [9], por oferecer: *i*) facilidade de atravessar *firewall* e NAT (do inglês, *Network Address Translation*); *ii*) toda a lógica de adaptação reside no cliente, reduzindo o requisito de conexões persistentes entre o servidor e a aplicação cliente. A infraestrutura de entrega existente, como caches HTTP e servidores, pode ser adotada de forma transparente[10]. Enfim, o uso do HTTP deve-se à sua ampla difusão entre a comunidade de desenvolvedores de software para a Internet e o conjunto de tecnologias já desenvolvidas para a redução da latência do tráfego.

2.2 Qualidade de Experiência

A Qualidade de Experiência (QoE - *Quality of Experience*) captura a percepção de usuário sobre um serviço que lhe foi prestado. Existem diversos fatores que afetam a QoE e eles estão relacionados ao desempenho do serviço e à experiência do próprio usuário. A aferição da QoE leva em consideração todos os fatores que contribuem para que o usuário

tenha uma boa experiência ao utilizar um sistema ou serviço que incluem[11]:

1. Fatores de influência humana
    a) Processamento de baixo nível (acuidade visual e auditiva, gênero, idade, humor etc)
    b) Processamento de alto nível (processos cognitivos, contexto sociocultural e econômico, expectativas, necessidades e objetivos etc)
2. Fatores de influência do sistema
    a) Relacionado ao gênero do conteúdo (detalhe de vídeo: ação, aventura, romance)
    b) Relacionado à Mídia (codificação, resolução, taxa de amostragem)
    c) Relacionado à rede (largura de banda, atraso, *jitter*)
    d) Relacionado ao dispositivo (resolução de tela, tamanho de exibição)
3. Fatores de influência contextual
    a) Contexto físico (localização e espaço)
    b) Contexto temporal (hora do dia e frequência de uso)
    c) Contexto social (relações inter-pessoais durante a experiência)
    d) Contexto econômico
    e) Contexto da tarefa (multitarefas, interrupções, tipo de tarefa)
    f) Contexto técnico e de informação (relacionamento entre sistemas)

Em resumo, a métrica avalia o nível de desempenho da perspectiva de um serviço prestado. É particularmente relevante para os serviços de distribuição de vídeos pela Internet, devido às altas demandas de tráfego e o desempenho da infraestrutura de acesso. Assim, ao projetar sistemas, a QoE esperada é muitas vezes considerada como uma métrica de saída do sistema.

Em [3], propõe-se a mensuração da QoE das sessões a partir de medidas quantitativas. Seja R_k a taxa de bits de vídeo, B_k a ocupação do buffer, q a qualidade do segmento

e C_k a largura de banda. O modelo considera:

$$\sum_1 = \frac{1}{K}\sum_{k=1}^{K} q(R_k)$$

que mede a qualidade média do vídeo a partir da qualidade média em todos os fragmentos;

$$\sum_2 = \frac{1}{K-1}\sum_{k=1}^{K-1} |q(R_{k+1}) - q(R_k)|$$

a variação média de qualidade, a partir da amplitude das alterações na qualidade de um fragmento para outro;

$$\sum_3 = \sum_{k=1}^{K} \left(\frac{d_k(R_k)}{C_k} - B_k\right)_+$$

o tempo total de rebufferização, quando o tempo de download (d_k)é maior que o tempo total do conteúdo baixado mas não reproduzido.

A métrica de QoE proposta combina as medidas anteriores para os segmentos $1 \ldots K$, como segue:

$$QoE_1^K = \sum_{k=1}^{K} q(R_k) - \lambda \sum_{k=1}^{K-1} |q(R_{k+1}) - q(R_k)| - \mu \sum_{k=1}^{K} \left(\frac{d_k(R_k)}{C_k} - B_k\right)_+ - \mu_s T_s$$

onde T_s define o *startup delay*, λ, μ, μ_s são pesos que regulam a importância dos fatores: variações de qualidade de vídeo, tempo de recarga e *startup delay*.

2.3 Largura de Banda e Vazão

A largura de banda de uma faixa de frequência define a quantidade máxima de dados que pode fluir de um ponto a outro em um canal, em um determinado momento [12]. Existem diversos meios para transmissão de dados, cada um com largura de banda específica, definida por suas características construtivas, o que torna a largura de banda dependente do meio de transmissão. Assim, ao afirmar que a largura de banda de um determinado canal é 10,0 Mbps, nem sempre se enviará ou receberá 10,0 Mbps, mas, a quantidade máxima de dados que pode ser transferida por esse canal é de 10,0 Mbps.

A vazão, também conhecida como *taxa de transferência*, é definida como a quantidade total de bits recebida por segundo em um canal [13]. Na definição dessa taxa incluem-se os custos do processamento dos elementos de rede envolvidos na comunicação e as consequências das disputas por uma parte do enlace. Os fatores que afetam a definição da taxa de transferência são:

- Níveis de congestionamento da rede;
- Níveis de sobrecarga do servidor em um horário específico;
- Reduzida largura de banda do canal que conecta os dispositivos à rede;
- A taxa de perda de informação no canal;
- Poder de processamento do hardware (dispositivo, servidor e elementos da rede).

A medição do tráfego que flui por uma rede é um processo importante em relação ao gerenciamento efetivo de largura de banda. Tais medições contêm informações que geram séries temporais que podem ajudar na análise de tráfego de rede, no desenvolvimento de mecanismos de controle, bem como são adequadas para desenvolver modelos que permitam a realização de predições de vazão.

2.4 Série Temporal

Uma série temporal é uma sequência discreta de observações reais obtidas por amostragem de um fenômeno mensurável em intervalos regulares de tempo [14]. Análise de séries temporais permite identificar padrões e consequentemente realizar predição de eventos que mudam com o passar do tempo. O método empregado nesse processo depende do dado envolvido nas predições. Entretanto, as seguintes características devem estar presentes nas séries: registros numéricos do passado e o padrão dos dados, que possivelmente continuará a ocorrer no futuro.

A maioria dos problemas de predição envolve o uso de dados coletados observando-se uma variável de interesse e registrando-se isso em sequências temporais. Neste trabalho, as sequências estudadas registram a vazão observada em sessões de acesso a um serviço de

transmissão de vídeo, ou seja, o número de bits transmitidos durante sessões de vídeos, Figura 2.

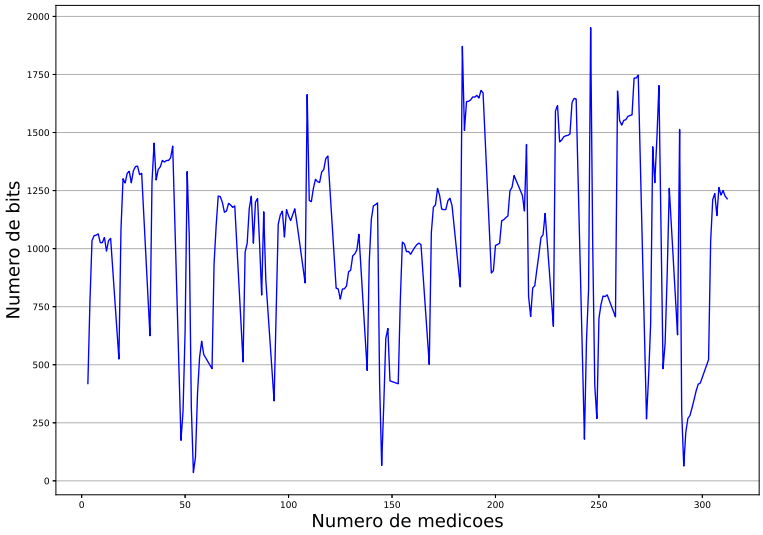


Figura 2 – Histórico de Sessões

Fonte: Produção própria

Predições de séries temporais é uma área importante de aprendizado de máquinas. O emprego das técnicas desenvolvidas sob a égide do aprendizado de máquinas que permite identificar padrões e tendências, possibilitando assim, posteriormente, usar esse conhecimento para prever valores futuros na série. As séries temporais são geralmente decompostas em [15]:

- T_t - a tendência no tempo t, que reflete a progressão a longo prazo da série. Se existe uma tendência quando há uma direção crescente ou decrescente persistente nos dados;

- S_t - padrão sazonal no tempo t, um padrão sazonal existe quando uma série temporal é influenciada por fatores sazonais;

- I_t - o componente irregular (ou 'ruído') no tempo t, que descreve influências aleatórias e irregulares.

Existe um acervo de diferentes modelos para descrever comportamentos de uma série temporal. Para construir esses modelos dependem de vários fatores, tais como o comportamento de fenômenos, natureza dos dados e do objetivo da análise. Os modelos utilizados neste trabalho para identificar o comportamento de tais séries temporais são processos controlados por leis probabilísticas.

2.5 Aprendizado de Máquina - AM

Aprendizado de máquina vista como uma parte da Inteligência Artificial. Algoritmos de AM constroem modelos baseados em dados de amostra, conhecido como 'dados de treinamento', para fazer previsões ou decisões sem ser explicitamente programado para executar a tarefa. [16]. O processo de aprendizagem começa a partir de observações de dados, reais ou sintéticos, como exemplos de experiência direta ou instrução, permitindo que se encontrem padrões ocultos sem que se tenha realizado programação dos algoritmos para tal. Assim, eles podem aprender automaticamente sem intervenção humana ou assistência e se ajustam para tomar ações de acordo com as entradas.

Os métodos de aprendizado estão agrupados conforme as seguintes categorias:

- **Aprendizagem supervisionada:** o computador recebe entradas de exemplo e os resultados desejados para o treinamento com o objetivo de aprender uma regra geral que mapeie entradas para saídas, tudo fornecido pelo responsável pelo treinamento;
- **Aprendizagem não supervisionada:** nenhum rótulo é passado para os métodos e a máquina sozinha deve encontrar uma estrutura ou padrão em sua entrada. A aprendizagem não supervisionada pode ser usada para descobrir padrões ocultos nos dados de entrada ou um meio para um fim (recurso de aprendizagem);
- **Aprendizado por reforço:** é um método que interage com seu ambiente, através de um processo de tentativa e erro. Esse método permite que máquinas determinem, de forma automática, o comportamento ideal dentro de um contexto específico.

Há diferentes algoritmos de aprendizado de máquina. Geralmente eles são agrupados pela categoria do aprendizado ou por similaridade e em forma, ou função (classificação,

regressão, agrupamento, aprendizado profundo etc.). Independentemente da categoria ou função, todas as combinações de algoritmos de aprendizado de máquina consistem no seguinte: representação, avaliação e otimização.

A seguir, apresentam-se os métodos baseados em aprendizado de máquina e média móvel, empregados para realização de predição em séries temporais, gerada pela observação da vazão de conexões estabelecidas no contexto de uma aplicação de distribuição de vídeos.

2.5.1 Regressão Linear - RL

Na estatística, a regressão linear é uma abordagem linear para modelar a relação entre uma variável dependente escalar Y e uma ou mais variáveis independentes (ou variáveis explicativas) denotadas X [17]. Formalmente, essa relação é definida como se segue:

$$Y \approx \beta_0 + \beta_1 X$$

onde β_0 e β_1 são parâmetros do modelo que podem ter seus valores definidos como se segue:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

para $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ e $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

A regressão linear é uma abordagem paramétrica, pois, ela assume uma forma funcional linear para $f(x)$. Os métodos paramétricos possuem vantagem de geralmente serem fáceis de ajustar, devido aos coeficientes terem interpretações simples e os testes de significância estatística serem facilmente realizados. Entretanto, os métodos paramétricos têm a desvantagem de fazer fortes suposições sobre uma forma funcional, i.e., se a tal forma especificada estiver longe da verdade (a forma do problema não corresponde ao que foi assumido) e a precisão da previsão for a meta, o método terá um desempenho ruim e quaisquer resultados retornados serão suspeitos.

A análise de regressão estima a expectativa condicional da variável dependente (y) dada as variáveis independentes ($x_1, x_2, ..., x_k$), isto é, o valor médio da variável dependente

do quanto as variáveis independentes são fixas. Embora possa parecer um pouco simples em comparação com alguns dos mais modernos métodos de AM, a regressão linear ainda é um método de aprendizado estatístico útil e amplamente utilizado como um bom ponto de partida e comparação para métodos mais recentes.

2.5.2 Random Forest - RF

O Random Forest é um meta classificador definido a partir de uma coleção de classificadores estruturados em árvore. Cada árvore é gerada a partir de um conjunto de vetores aleatórios, que pode ser obtido através de seleção aleatória de atributos ("característica") e/ou seleção aleatória de amostra de árvore. Em sua implementação mais simples, os vetores aleatórios são gerados independentemente dos vetores aleatórios anteriores, mas com a mesma quantidade de instâncias e atributos em cada sub-árvore.

O método pode ser usado tanto para classificação quanto para regressão. Sua saída no modo de classificação é construída sob votação de uma multidão de árvores de decisão no tempo de treinamento, que retorna a classe predominante. RF para regressão retorna a média das árvores de decisão da multidão de árvores de decisão para melhorar a precisão preditiva e controlar o viés. Esse processo consiste em dois passos:

- Amostrar k exemplos de treinamento de X (atributos do conjunto de treino), Y (classe); chamam estes X_k, Y_k;
- Treinar uma árvore de regressão f_k em X_k, Y_k.

Após o treino das K árvores, as previsões podem ser calculadas usando a média das previsões de todas as árvores de regressão individuais em x':

$$\hat{f} = \frac{1}{K} \sum_{k=1}^{K} f_k(x')$$

O RF é um método variante do *bagging*, que cria n réplicas da base de treinamento através de seleção aleatória de amostras. A principal diferença entre os métodos é a escolha do tamanho de m "atributos" do subconjunto. Por exemplo, se um RF é construído usando $m = M$, total de "atributos" na base de dados, então isso é simplesmente equivale ao *bagging*.

2.5.3 Naive Bayes - NB

O classificador Naive Bayes é construído com base na hipótese de independência condicional entre os atributos dados à classe, entretanto, essa suposição raramente acontece nos problemas do mundo real. É um algoritmo de aprendizado de máquina muito conhecido, cuja eficiência de classificação é comprovada em aplicações como categorização de documentos e filtragem de *spam* de e-mail [18]. É um modelo probabilístico condicional que representa o possível resultado ou classe de uma nova instância do conjunto, dada à observação de valores atribuídos a outras instâncias do mesmo conjunto.

Usando o teorema de bayes, a probabilidade condicional é definida como segue:

$$p(Y_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid Y_k)\ p(Y_k)}{p(\mathbf{x})}$$

onde $\mathbf{Y}$ é o atributo Classe e $\mathbf{x}$ é um exemplo que contém k atributos $(\mathbf{x}_1, ..., \mathbf{x}_k)$, ele tenta predizer $\mathbf{Y}$ dado $\mathbf{x}$. Como $p(Y|\mathbf{x})$ geralmente não é conhecido, ele deve ser estimado a partir dos dados, utilizando probabilidade a priori $p(Y_k)$, densidade de probabilidade condicional da classe ($p(\mathbf{x} \mid Y_k)$), a probabilidade incondicional do evento $p(x)$ e a probabilidade a posteriori ($p(Y_k \mid \mathbf{x})$) para evidenciar o objeto de classificação.

O classificador Naive Bayes é caracterizado por [18]:

- boa eficiência computacional;
- baixa variância;
- aprendizado incremental;
- robusto ao ruído;
- robusto em valores ausentes.

A eficiência computacional na modelagem e previsão de um problema é uma característica de destaque do classificador NB. Sua eficiência vem da possibilidade de fácil paralelização e capacidade de manuseio de grande número de atributos sem necessidade

de realizar seleção desses atributos. Apesar de ser um classificador probabilístico simples e considerado "ingênuo" por ser baseado no pressuposto de forte independência entre os atributos, o modelo consegue excelente desempenho em vários problemas.

2.5.4 K-Vizinhos Mais Próximos - KNN

O KNN (do inglês, *K Nearest Neighbors*) é um método baseado em observações de instâncias vizinhas e referido como um método de classificação clássico. O método assume que as instâncias podem ser representadas em um espaço, e.g. espaço euclidiano, em seguida, para rotular uma nova instância desconhecida x, as K instâncias mais próximas similarmente da nova instância x são identificadas para atribuição da classe à instância x. Por exemplo, usando k_i para representar o número de elementos na vizinhança de x que são da classe Y_i, temos:

$$P(Y_i|x) \approx \frac{k_i}{K}$$

Essa rotulação, após identificar as K instâncias mais próximas da nova instância x, acontece de forma diferente para ambos os tipos de modelo de KNN: para classificação usa-se a probabilidade condicional para a classe, ou seja, a saída é apoiada na classe mais comum entre os K vizinhos mais próximos, definida como segue [19]:

$$P(Y|X) = \frac{1}{K} \sum_{i \epsilon N_0} I(y_i)$$

onde N_0 são os K pontos nos dados de treinamento que estão mais próximos; para regressão, a saída é o valor médio do alvo entre os k vizinhos mais próximos, definida como:

$$\hat{f}(x) = \frac{1}{K} \sum_{x_i \epsilon N_0} y_i.$$

Ambos os tipos podem ser usados pondo pesos às contribuições dos vizinhos, de modo que os vizinhos mais próximos contribuam mais do que os mais distantes.

O método é sensível à localização dos dados no espaço. Um bom K pode ser selecionado usando várias técnicas heurísticas, por exemplo, escolher via método *bootstrap* [20]. Por esse motivo, a melhor escolha de K depende totalmente do dado, valores de K muito altos geralmente reduzem o efeito do ruído durante a rotulação, entretanto isso faz com que o método considere dados menos similares.

2.5.5 Média Móvel - MA

A média móvel (do inglês, *Moving Average* - MA) é um método usado no processo de suavização de dados, especialmente para dados de séries temporais que visam estimar a tendência dos dados [21]. Pode ser vista como uma janela de tempo que se move ao longo do sinal de entrada, executando uma média local dos quadros vizinhos relacionados à continuidade da sequência do alvo. Seu valor é continuamente recalculado à medida que novos dados são disponibilizados, ela segue descartando o valor mais antigo e adicionando o mais recente. O limiar da janela pode ser entre curto e longo prazo dependendo da aplicação, e seus parâmetros são definidos em concordância.

Um modelo de média móvel é conceitualmente uma regressão linear do valor atual da série em relação à janela corrente e em termos do erro observado anteriormente. O método visa reduzir o efeito das variações temporárias nos dados, melhorar o ajuste dos dados com um processo chamado "suavização" para mostrar a tendência dos dados mais claramente e destacar qualquer valor acima ou abaixo da tendência. Uma média móvel simples (do inglês, *simple moving average* - *SMA*) é definida como segue:

$$SMA = \frac{M_1 + M_2 + \cdots + M_{n-1}}{n} = \frac{1}{n}\sum_{i=1}^{n-1} M_i$$

onde os valores de M representam valores em uma janela de tempo e n o tamanho da janela.

O uso da taxa de transferência média do passado para prever futuros é comumente usado em muitas aplicações. O modelo auto-regressivo ARIMA (do inglês, *Autoregressive Integrating Moving Average*), por exemplo, consiste em três componentes, os modelos $AR(p)$, $MA(q)$ e $ARMA(p, q)$, onde p é o número de termos autorregressivos e q é o número de previsão defasada na equação de previsão [22]. Embora os modelos como ARIMA e Média Harmônica, (HM) possam resolver muito bem o problema de previsão em séries temporais, a precisão da predição ainda não é muito alta para a série com grandes flutuações, principalmente quando há ausência de dados.

2.5.6 EWMA

A Média Móvel Exponencialmente Ponderada (do inglês, *Exponentially Weighted Moving Average* - EWMA) é uma média ponderada dos últimos n valores, em que a

ponderação diminui exponencialmente a cada dado/período, isto é, dar mais importância aos dados mais recentes e à medida que os dados ficam antigos a significância da informação diminui exponencialmente. Essa média reage de maneira mais significativa às pequenas mudanças dos dados do que uma SMA, que aplica peso igual à todas as n observações.

Apesar do método responder mais rapidamente às flutuações dos valores do que uma SMA, o método é mais sensível aos ruídos nos dados. A EWMA é definida como segue:

$$\hat{S}_t = B\left[S_t + AS_{t-1} + A^2S_{t-2} + \cdots + A^nS_{t-n}\right] + A^{n+1}S_{t-(n+1)}$$

onde B é um fator de suavização constante entre 0 e 1, A é $(1-B)$, os S's são observações de períodos e o t indica a ordenação temporal das observações.

2.5.7 Média Harmônica - HM

A média harmônica (do inglês, *Harmonic Mean* - HM) é uma média usada para fornecer uma agregação entre os operadores max e min, e é amplamente usada como uma ferramenta para agregar dados de tendência central [23]. Geralmente, é considerada uma técnica de fusão de informações numéricas de dados. É calculado dividindo o número de observações pelo inverso multiplicativo, definido como $\frac{1}{x}$ ou x^{-1}, de cada valor x na série, como segue:

$$\hat{H}_i = K \,/\, \left(\sum_{k=1}^{K} \frac{1}{H_{i,-k}}\right) z$$

onde $\hat{H}_i$ é a taxa de transferência futura prevista e $H_{i,-k}$ denota a vazão média do $k^{th}(k = 1, 2, ...K)\Delta - \textbf{segundos}$ antes da sessão i começar.

A HM é muito usada no contexto de problemas que envolvam grandezas inversamente proporcionais, tal como a vazão de redes. O método tende a mitigar o impacto de valores altos (como os *outliers*), e agravar o impacto dos valores baixos, fazendo com que cada um dos dados tenha peso igual. Diante disso, é utilizada para recuperar informações e estimar dados de séries temporais.

2.6 Considerações Finais do Capítulo

Neste capítulo, apresentou-se sobre os principais conceitos que envolvem o contexto em que a pesquisa está baseada. Discutiu-se sobre a estruturação da transmissão de conteúdo de *streaming* adaptativos, a percepção dos usuários ao desfrutar desse tipo de serviço, fatores de influência na transmissão de conteúdo e sobre as séries temporais. Os métodos preditores usados neste trabalho foram descritos, bem como os conceitos associados aos métodos. Especificamente, o conceito de aprendizagem de máquina e suas variações, i.e. os algoritmos de AM.

3 TRABALHOS RELACIONADOS

Existem dois tipos de abordagens para fazer predições de vazão: i) baseada em sessões em curso e ii) baseada em histórico de acesso. A primeira prevê a vazão do canal através de expressões matemáticas e levando em consideração o *buffer* e a vazão alcançada dos n últimos segmentos baixados. A segunda prevê o valor contínuo de séries temporais do tráfego da rede, medidas anteriores no mesmo caminho, analisando e identificando os padrões de acessos para criar modelos de predição através de técnicas de aprendizagem de máquinas e modelagens matemáticas. Nas próximas seções serão discutidos trabalhos que realizam a predição da vazão de um canal estabelecido para transferir dados de uma sessão de vídeos.

3.1 Predição Baseado em Sessões em Curso

O uso da taxa de transferência média do passado para prever futuros é comumente usado em muitas aplicações. Essa abordagem analisa a vazão, buffer do player de vídeo e a informação do(s) último(s) segmento(s) baixado(s) para predizer a vazão e, em consequência, a qualidade do próximo segmento a ser baixado, por intermédio de expressões matemáticas e análise dessas informações.

Em [24], realizam um estudo sistemático sobre o desempenho comparativo dos algoritmos de predição existentes no contexto das redes móveis. Para tal, utilizam quatro algoritmos de previsão baseados em média de taxa de transferência amplamente utilizado: média aritmética, HM, média geométrica e EWMA. Mostra-se que a predição de vazão é um problema desafiador e que várias abordagens funcionam bem para diferentes cenários de rede. Além disso, mostram que os dados de passados recentes podem fornecer informações mais precisas do que os dados de um passado mais distante e que estão menos correlacionado com a vazão futura.

Em [25], propõe-se um modelo estocástico para prever a taxa de transferência em redes móveis. O modelo considera os ruídos que conduzem a predição imprecisa, tais como

fenômenos aleatórios e informações imprecisas sobre localização do usuário. Mostra-se que não existe um único preditor capaz de considerar os diversos cenários que influenciam a vazão de um canal de redes móveis, i.e., as flutuações rápidas do canal, o deslocamento e as informações estatísticas gerais de um usuário.

O modelo de Markov tem sido utilizado no reconhecimento de padrões temporais. Em [26], utiliza-se um Modelo de Markov Oculto (do inglês, *Hidden Markov Model*) para lidar com o histórico de séries temporais e o Modelo Gaussiano Misto (do inglês, *Gaussian Mixture Model*) para avaliar o fator de flutuação com variância total ao prever a vazão. Compara-se a precisão do método proposto com outros três métodos de previsão convencional: regressão linear, regressão linear ponderada localmente e um modelo estocástico. Mostra-se que o método proposto pode identificar a flutuação dos dados de forma eficaz e prever as saídas dos próximos 100 segundos, com alta precisão em várias situações.

3.2 Predição Usando Histórico de Sessões de Vídeos

O uso das medições de vazão do passado para prever vazões de redes no futuro vem sendo bastante pesquisado, como uma nova alternativa para melhorar o QoE e QoS (do inglês, *Quality of Service*) da transmissão de conteúdos. Essa abordagem analisa e busca encontrar melhores conjuntos de características, fortemente correlacionadas, para criar modelos de predição de vazão robustos, inclusive no início das sessões de vídeos.

Em [24], mostra-se resultados de um estudo comparativo de algoritmos de AM para predição de vazão em redes móveis, além dos algoritmos baseado em média mencionados na seção anterior. Considerou-se três algoritmos de predições que usam dados históricos para realizar as predições, i.e. *Multiple Linear Regression - MLR*, *Neural Network Regression*, e *Support Vector Regression*. As avaliações usaram dados reais e consideraram métricas de desempenho, valor da informação da transferência passada, performance da predição da taxa de transferência, horizonte de previsão e o impacto das localizações geográficas. Em suas análises, o modelo MLR destacou-se pela baixa complexidade, tendo apresentado desempenho comparável aos demais.

Em [27], propõe-se uma estratégia baseada em *buffer* que se ancora na previsão

de *throughput* de longo prazo. A estratégia utiliza um método de predição de taxa de transferência, não mencionado no trabalho, para usuários móveis, com base em informações históricas da rede de acesso, e programa o tempo de download do vídeo. No cenário em que o método foi avaliado, o usuário se desloca no sistema de transporte público, especificamente no metrô, e são considerados ainda o horário e o padrão de deslocamento dos usuários, i.e., número de viagens e as suas durações.

Em [28], avaliam-se os desafios da predição da largura de banda em transmissões de vídeos. Um método de predição é proposto baseado na agregação de dados coletados de um histórico de sessões. A escolha dos dados a serem considerados representa um desafio para a construção do modelo, assim como a sua incompletude. O método proposto é comparado com preditores simples, e.g. último amostra, e alguns algoritmos convencionais de aprendizado de máquina.

Em [2], mostra-se que a seleção imprópria das taxas de bits iniciais, devido à indisponibilidade da predição de vazão, leva a uma alta probabilidade de interrupção na sessão. Ademais, propõe-se uma abordagem híbrida para predição da vazão que combina várias heurísticas sobre a vazão, são elas: LSU - *Last Sample from the Same User*, HSU - *Historical Samples from the Same User*, HSS - *Historical Samples from the Same Subnet* e LSS - *Last Samples from the Same Subnet*. Avaliou-se a heurística utilizando dados de sessões longas, geradas pelos clientes, e os resultados apresentados mostram que a predição tem impacto positivo na criação das novas sessões de acesso de um serviço de *streaming* de vídeo, em especial no início da reprodução.

3.3 Considerações Finais do Capítulo

Neste capítulo discutiu-se sobre os trabalhos relacionados, abordando-se seus experimentos realizados e os resultados obtidos. O uso de preditores de vazão baseados em características únicas, como a taxa do downlink ou horário, apresentam bons resultados, no entanto, tendem a desconsiderar outros aspectos importantes que contribuem para a definição da vazão de um canal, e.g., a relação entre o ISP, servidor e horário. Como será mostrado mais a frente, as combinações de várias características podem ter um impacto mais significativo na predição de vazão, quando comparado como os preditores que usam

uma única característica.

4 A BASE

Neste capítulo apresenta-se a base de dados utilizada para a realização deste trabalho, que contém informações de diversas infraestruturas de comunicação. Essas informações foram coletadas no contexto do projeto Neubot, que tem em sua origem a verificação de conformidade das infraestruturas monitoradas em relação à neutralidade no tratamento de fluxos oriundos de diferentes aplicações.

A base utilizada foi coletada ao longo de quatro anos, de 2015 a 2018, e contém mais de 1100 horas de medições feitas em três grandes regiões metropolitanas do Brasil - Rio de Janeiro, Belo Horizonte e São Paulo. Nessas cidades foram monitoradas as infraestruturas das operadoras de acesso com maior capilaridade.

O capítulo está organizado da seguinte forma: Na Seção 4.1 apresenta-se uma descrição detalhada e na Seção 4.2 apresentam-se algumas estatísticas descritivas da base de dados. Na Seção 4.3 descreve as técnicas usadas no tratamento da base de dados. E por fim, na Seção 4.4 são apresentadas as considerações finais deste capítulo.

4.1 Projeto Neubot: medindo a vazão a partir da borda

O projeto Neubot monitora os provedores de acesso à Internet para testar a neutralidade do serviço prestado [29]. Usam-se os bots instalados em máquinas de participantes voluntários, que simulam o comportamento de aplicações de Vídeo *Streaming*, P2P (do inglês, *peer-to-peer*) e *Web Browser*. Esses bots enviam requisições para servidores instalados no M-Lab[30] e registram os dados e as dinâmicas observadas em cada sessão de acesso.

A Figura 3 mostra como os clientes do projeto estão distribuídos pelo planeta, tendo como maior destaque na América do Sul a presença de voluntários brasileiros. O Neubot realiza aproximadamente 10.000 testes por dia envolvendo mais de 1.000 endereços IP em cerca de 100 países e 1.000 sistemas autônomos (grupo de de redes IP, sob o controle de uma gerência técnica e uma mesma politica de roteamento). Os bots vídeo *streaming* estão programados para realizar dois testes, diariamente, como um serviço executando em

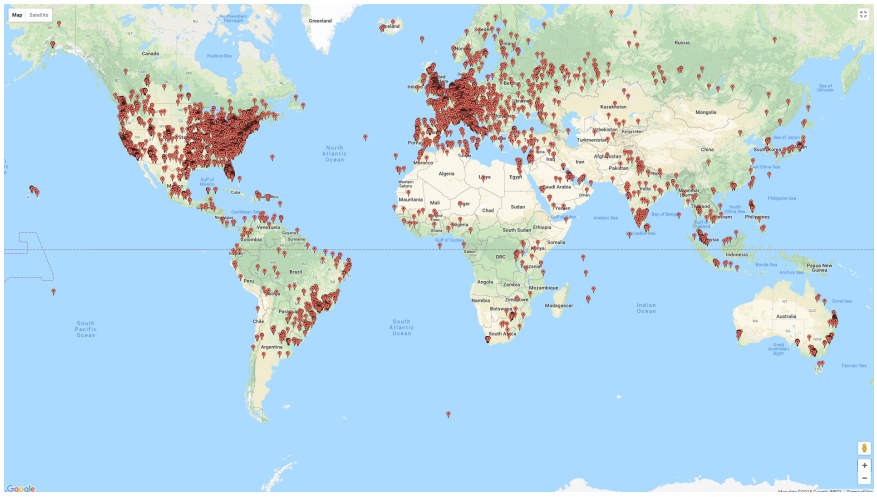


Figura 3 – Distribuição Geográfica dos Clientes Neubot

Fonte: Neubot Data Analysis [31]

segundo plano e a cada 30 minutos realiza uma análise, mas tal agendamento depende diretamente do padrão de conexão dos voluntários. Os voluntários que permanecem mais tempo conectado permitem um monitoramento maior das suas infraestruturas. As informações associadas a cada sessão são enviadas para um repositório público.

As instâncias do bot vídeo *streaming*, em cada sessão de acesso, geram 15 requisições para o servidor. Cada segmento requisitado contém 2 segundos de vídeo e é disponibilizado em uma das seguintes taxas de bits [29]: 100, 150, 200, 250, 300, 400, 500, 700, 900, 1.200, 1.500, 2.000, 2.500, 3.000, 4.000, 5.000, 6.000, 7.000, 10.000, 20.000 kbps. O algoritmo de adaptação da taxa de bit, implementado pelo bot vídeo *streaming* usa a vazão medida do último segmento requisitado, para definir a taxa de bits do próximo segmento. O primeiro segmento é acessado sempre na menor taxa de bits (100 kbps). O algoritmo é definido como segue [29]:

```
if EDT > PLAY_TIME:
  REL_ERR = 1 - EDT / PLAY_TIME
  EAB = EAB + REL_ERR * EAB
  EAB = max(min_rep_bitrate, EAB)
else:
  EAB = size_of_segment / EDT
  EAB = max(all_rep_bitrate < EAB)
```

onde EAB é a vazão estimada, o EDT é o tempo de download, o PLAY_TIME é a duração de reprodução do segmento, o size_of_segment é o tamanho do segmento em kbit e os min_rep_bitrate e all_rep_bitrate representam a taxa de bits mínima e todas as taxas de bits disponíveis no arquivo de manifesto (arquivo que descreve a mídia), sucessivamente.

As requisições são realizadas usando uma conexão HTTP persistente, com o servidor configurado para receber um número ilimitado de requisições, por um tempo indeterminado. Cada requisição de teste é descrita com propriedades que caracterizam o cliente Neubot, o servidor, a conexão e cada um dos segmentos acessados. Após cada teste, o cliente envia os resultados, em formato JSON, ao servidor que os armazena e disponibiliza ao público. Aos dados coletados em cada requisição foram acrescidas as seguintes informações: localização geográfica e propriedade de rede, especificamente o nome País, Cidade e Sistema Autônomo (do inglês, *Autonomous System*). Essas informações foram obtidas usando os serviços disponíveis em maxmind.com [32].

4.2 Análise da Base

Nesta seção é apresentado o estudo realizado na base de dados Neubot. Esse estudo é motivado pelo objetivo específico 1: analisar e avaliar a base de dados em busca de características e especificações que estejam diretamente relacionadas com estabilidade e vazão dos canais. Primeiro, mostra-se o padrão de popularidade dos bots nas regiões metropolitanas brasileira e os ISP mais monitorados pelo projeto. Em seguida, apresentam-se os padrões de frequência de acesso dessas medições e a duração das sessões de vídeo. Por fim, apresenta-se um estudo da esparsidade dos dados na base.

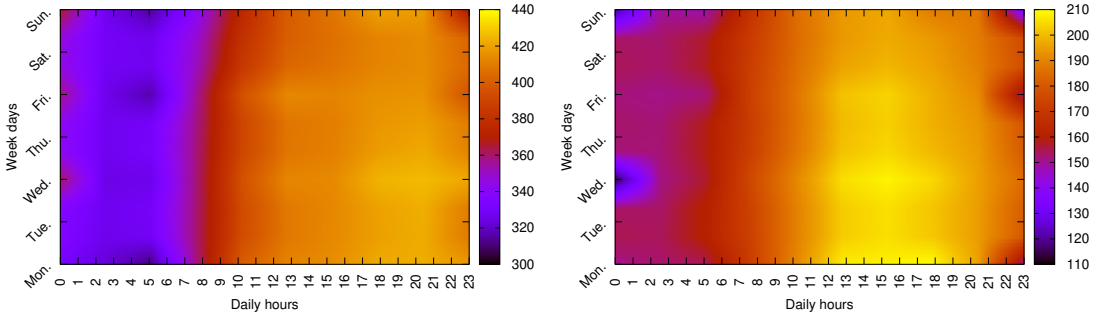
A Tabela 1 mostra o resumo de dados dos ISP da Claro Telecom S.A., da BR (Claro), a Telefônica Brasil S.A. (Telefônica) e a Telemar Norte Leste S.A., da BR (Telemar), que foram os provedores mais monitorados nas três maiores regiões metropolitanas do Brasil, em termos de número de sessões e duração.

Supõe-se que a combinação de um grande número de voluntários e a frequência da coleta fornecerá um retrato mais preciso dos serviços de rede fornecidos por um ISP. Para verificar tal suposição, os mapas de calor foram construídos com base na frequência das

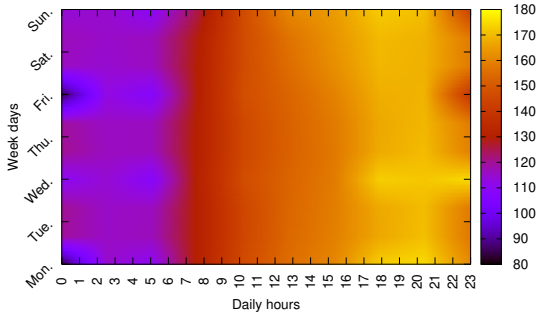
Cidade	ISP	2015-2016			2017-2018		
		Sessões (k)	Dur (hrs)	Bots (n)	Sessões (K)	Dur (hrs)	Bots (n)
São Paulo	Telefônica	18.9-15.2	114-92	66-63	13.3-15.4	74-89	67-267
	Claro	4.5-5.2	18-29	18-20	6.7-12.0	37-71	69-109
Rio de Janeiro	Claro	3.9-6.6	23-39	12-9	10.0-9.3	60-56	27-60
	Telemar	1.2-1.9	08-12	12-11	3.6-2.9	26-19	19-30
Belo Horizonte	Claro	4.2-5.2	25-28	9-11	6.7-7.3	40-40	19-79
	Telefônica	1.2-1.7	08-11	10-12	3.2-6.1	20-39	17-25

Tabela 1 – Resumo de dados Coletados por Região e Provedor

sessões coletadas em períodos da semana, conforme mostra a Figura 4. Delimitado por esse período de tempo, as medições foram agrupadas usando dias da semana e a hora em que as primeiras solicitações ocorreram e foram enviadas para análise. O calor expressa o número de sessões para cada uma hora. Os dados foram coletados dos provedores da Telefônica Brasil e Claro Telecom que atendem cidades de São Paulo, Rio de Janeiro e Belo Horizonte, respectivamente.



(a) Telefônica Brasil S.A em São Paulo (b) Claro Telecom S.A. no Rio de Janeiro



(c) Claro Telecom S.A. em Belo Horizonte

Figura 4 – Sessões de Medições no Intervalo de uma Semana

Fonte: Produção própria

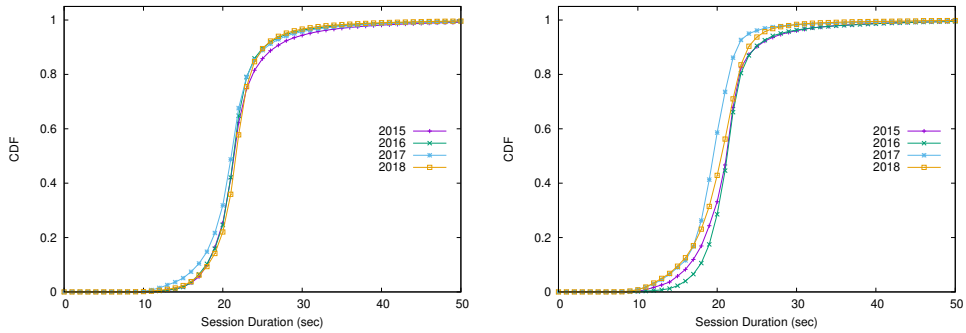
Existem padrões distintos em dois períodos do dia: das 1:00 às 6:00 e das 10 às 22 horas. O segundo tem, para a duração e frequência das sessões exploradas, os maiores valores registrados sobre os dados coletados. A ação de exploração dos bots está fortemente relacionada ao tempo de conexão do dispositivo. De fato, as sessões de exploração são previamente agendadas [29], e os dados sugerem uma tendência de investigar infraestruturas durante períodos de sobrecarga, ou seja, quando os usuários estão envolvidos em atividades profissionais, à tarde e ao anoitecer, e/ou estão envolvidos com entretenimento no avançar da noite.

À medida que o número de bots ativos aumenta, o padrão observado tende a ser mais correlacionado ao número de dispositivos conectados, como sugerido pelos dados coletados em São Paulo. No intervalo de 2:00 às 5:00 o número de sessões coletadas diminuiu. Isso levanta uma incerteza em relação à duração da sessão, uma vez que uma infraestrutura de rede super dimensionada terá sessões com baixa latência e alta vazão.

Para verificar a duração média das sessões coletadas, calculou-se a função de distribuição acumulada (do inglês, *cumulative distribution function - CDF*) dessas sessões no intervalo de 2015 a 2018 (Figura 5a). Nos dados coletados em 2015, a grande maioria das sessões, i.e. 93%, dura entre 15 e 30 segundos, e 98% de todas as sessões duram até 40 segundos. Um padrão semelhante foi identificado para os outros três anos, com forte tendência em torno dos 22 segundos de duração. De fato, o número de sessões que duram 22 segundos supera as demais sessões.

O ISP mais observado, a Telefônica Brasil de São Paulo, teve a duração das sessões de coleta avaliada ao longo daqueles quatro anos (Figura 5b). O padrão de duração das sessões coletadas no referido ISP é muito semelhante ao padrão apresentado anteriormente, 95% de todas as sessões duram entre 15 e 30 segundos, e 98% duram até 40 segundos. A mudança em torno da duração de 22 segundos é menos proeminente durante os outros três anos, especialmente em 2017.

Apesar de sua importância, em tamanho, o padrão geral de duração da sessão não é determinado pelo padrão identificado nos dados coletados do ISP da Telefônica Brasil de São Paulo. Retirou-se esses dados da base, e a base resultante teve a CDF recalculada para mostrar esse deslocamento de 22 segundos. Especificamente, pelo menos 91% de todas as



(a) Todos os ISPs (b) Telefonica Brasil S.A em São Paulo

Figura 5 – CDF da Duração das Sessões

Fonte: Produção própria

sessões duraram entre 15 e 30 segundos e 98% duraram até 40 segundos.

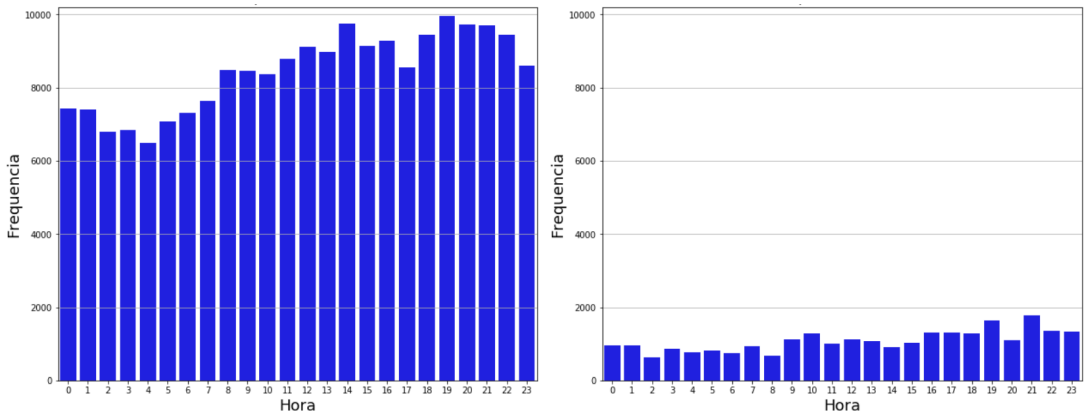
Do ponto de vista do aplicativo vídeo *streaming*, esse padrão de 22 segundos significa que as infraestruturas testadas foram capazes de fornecer sessões de vídeo de 30 segundos antes do tempo de reprodução programado. Esse resultado está apoiado em algum nível de estabilidade da vazão, durante as sessões, ou, devido à natureza do aplicativo imitado, vídeo *streaming* com taxa de bit adaptável, a qualidade das sessões entregues é questionável.

A esparsidade do conjunto de dados foi calculada usando as informações das medições do par dia-hora. Para realizar esse estudo, os dados foram filtrados usando sua marca de tempo e a taxa de download. As sessões de medição foram executadas em planos de 2,0 Mbps, realizadas durante os dias úteis e finais de semana. Além disso, a esparsidade foi calculada considerando os seguintes períodos: 24 horas e horário nobre, das 10h às 22h, ver Tabela 2.

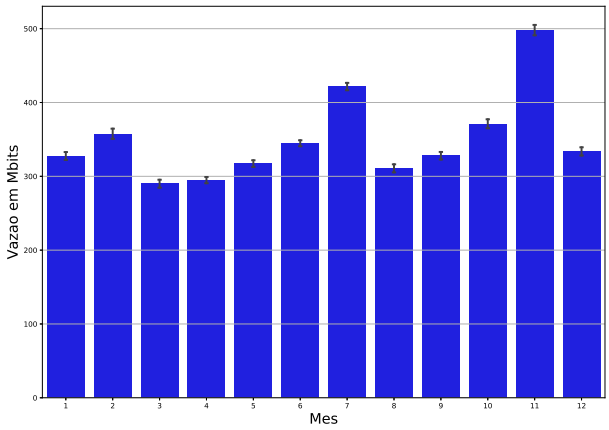
Cidade	ISP	Esparsidade			
		24 horas		Horário nobre	
		Sab-Dom	Seg-Sex	Sab-Dom	Seg-Sex
São Paulo	Telefônica	0.713478	0.702738	0.689947	0.678363
Belo Horizonte	Claro	0.882066	0.885063	0.862254	0.866709
Rio de Janeiro	Claro	0.88941	0.884299	0.873707	0.884299

Tabela 2 – Avaliação de Esparsidade nos Dados de Planos de 2,0 Mbps

A esparsidade tem os valores mais baixos durante o horário nobre, das 10h às 22h. O modelo oportunista de bots vincula sua ativação ao horário em que a máquina hospedeira está conectada. Embora o número de medições seja limitado ao longo do dia, os dados sugerem que os bots realizam coletas de forma independente. Essa independência de ações força o número de bots *online* a ser elevado para que se possa ter um retrato do serviço oferecido pelos ISP. A partir dos dados coletados, pode-se ver que o dilema do dia versus hora diminui ao longo do tempo, o que sugere que a hora da coleta é mais importante do que o dia.



(a) Planos de dados até 2,0 Mbps (b) Planos de dados acima de 2,0 Mbps



(c) Vazão média dos planos até 2,0 Mbps durante os meses

Figura 6 – Frequência e Flutuação Média das Medições

Fonte: Produção própria

Analisou-se também a base de dados no início, meio e fim do ano. Tal análise considerou a frequência das sessões de acesso, vazão máxima de cada sessão de acesso e a frequência dos diferentes planos (1,0 Mbps, 2,0 Mbps, 4,0 Mbps, 8,0 Mbps e acima de 8,0

Mbps). Assim, foi possível verificar um padrão de flutuação das vazões entre os meses do ano. A vazão observada nas medições dos meses iniciais do ano, de março até julho, tende a se repetir e apresenta uma tendência de aumento da vazão média nos meses seguintes, i.e., de agosto até novembro, ver Figura 6c.

Verificou-se que a maioria das medições concentra-se nos planos de até 2,0 Mbps, ver Figura 6a. Somando-se os demais planos identificados não se tem nem 50% das medições dos planos de até 2,0 Mbps, ver a Figura 6b. Assim, decidiu-se realizar experimentos utilizando somente os dados de planos até 2,0 Mbps, visto que é o plano com maior número de medições na base, representando mais de 82% dessas medições.

4.3 Tratamentos Realizados: filtragem e engenharia de atributos

Os métodos de AM requerem dados representativos para construção de modelos efetivos. A coleta de dados é um passo importante, uma vez que os conjuntos de dados representativos variam não apenas de um problema para outro, mas também de um período para o outro.

Característica	Exemplo	Descrição
uuid	87c750ff-e5b2-4b40	ID do usuário.
real_address	216.158.92.121	IP do cliente Neubot, visto pelo servidor.
platform	linux2	O sistema operacional, e.g. “linux2”, “win32”.
internal_address	192.168.55.55	Endereço IP do cliente.
remote_address	184.105.23.83	Endereço IP do servidor.
version	0.004016008	Número da versão do Neubot.
timestamp	1496390549	Hora em que o download do segmento foi iniciado, em Unix.
connect_time	0.02710292226402089	Latência ou RTT (do inglês, *Round Trip Time*).
iteration	1	Número de sequencia do segmento solicitado.
rate	1500	Taxa de bits do segmento solicitado, em kbits.
elapsed	0.24498810980003327	Duração do download, em segundos.
elapsed_target	2	Duração esperada do download, em segundos.
received	375130	Quantidade de bytes recebidos.

Tabela 3 – Informações coletadas durante uma sessão de acesso dos bots de vídeo streaming

Em geral, coleta de dados com registro de eventos do mundo real apresentam: incompletudes, imprecisão e inconsistências. Por esse motivo os dados da base Neubot foram tratados, visando a redução de ruídos que levem os algoritmos a predições equivocadas. As

técnicas usadas nesta etapa vão desde à remoção de medições e características irrelevantes, passando pela atribuição de valores padrões e engenharia de atributos. A Tabela 3 apresenta exemplos das medições que formam a base de dados estudada.

Na remoção de medições, tiraram-se sessões que apresentavam poucas medições de acesso, i.e., sessões que apresentavam quantidade inferior às 15 medições padrão das sessões. Assim como, clientes com poucas sessões de medição, i.e., clientes com poucas medições durante os quatro anos da base coletada. Consideraram-se irrelevantes as que não possuíam informações explicativas das condições da conexão de sessões cliente-servidor no contexto de *streaming* de vídeo, e.g. platform e version, por esse motivo removeram-se tais informações da base de dados.

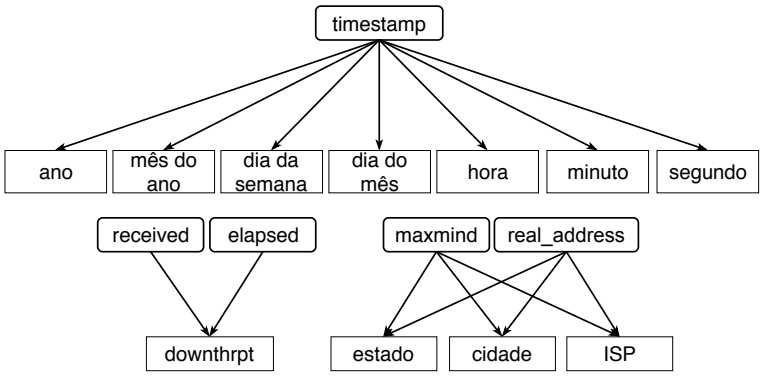


Figura 7 – Engenharia de atributos aplicada na base Neubot

Fonte: Autoria própria

Na atribuição de valores padrões, alterou-se valores do tipo alfabético e alfanumérico para um código apropriado de categoria, e.g., uuid, remote_address, city, ISP etc, pois dados em formato numérico são preferidos por muitos dos métodos de AM. Na engenharia de atributos, utilizou-se de conhecimento específico sobre o problema para criar novos atributos que possuem correlação com a medida alvo, ver Figura 7. Os dados usados para gerar novas características estão em destaque, parte superior da figura. Um exemplo que destaca a engenharia de atributos empregada foi o tratamento realizado no atributo *timestamp*, que produziu os atributos *dia, mes, dia_semana, hora, minuto, segundo* e o *ano*.

No contexto de AM, afirma-se que as características são propriedades individuais mensuráveis de algo que se tem interesse. Na base de dados usada neste trabalho cada

Característica	Exemplo	Descrição
uuid	75	ID do usuário.
city	212	Código da cidade.
ISP	10321123.0	Código do provedor de serviço.
remote_address	1841052383	Código do servidor.
connect_time	0.02710292226402089	Latência ou RTT (do inglês, *Round Trip Time*).
iteration	1	Número de sequencia do segmento solicitado.
rate	1500	Taxa de bits do segmento solicitado, em kbits.
elapsed	0.24498810980003327	Duração do download, em segundos.
dia	21	Dia do mês (em inteiro).
mes	9	O mês em que ocorreu a medição (em inteiro).
ano	2018	Ano em ocorreu a medição (em inteiro).
dia_semana	2	Dia da semana em ocorreu a medição (em inteiro).
hora	3	Hora em ocorreu a medição (em inteiro).
minuto	45	Minuto em ocorreu a medição (em inteiro).
segundo	32	Segundo em ocorreu a medição (em inteiro).
received	375130	Quantidade de bytes do segmento transferido.
downthrpt	325.23464	Vazão (em kbps).

Tabela 4 – Informações da base de dados após tratamento

entrada representa um segmento de vídeo solicitado por um cliente Neubot vídeo *streaming* e cada coluna expressa uma característica dessa solicitação, Na Tabela 4 apresentam-se os atributos da base de dados após realizados os tratamentos, em que cada linha da tabela representa uma característica da base de dados utilizada. Para cada um dos atributos da tabela há um exemplo do seu valor e uma pequena descrição sobre.

4.4 Considerações Finais do Capítulo

Neste capítulo apresentou-se a base de dados que foi utilizada nesta pesquisa e os tratamentos que foram realizados. O projeto *Neubot* foi presentado, especificamente, tratou-se de como as são monitoradas as sessões e como acontece a coleta de informações. Analisou-se a base com a finalidade de adquirir conhecimento sobre o padrão de acesso dos usuários e da vazão das sessões. Ademais, utilizou tais conhecimentos para oportunizar melhorias nos treinamentos dos modelos, i.e., o estudo serviu como guia para o tratamento, a manipulação dos dados e estratégias utilizadas durante os treino e validação dos modelos de AM.

5 RESULTADOS NUMÉRICOS

Neste capítulo apresentam-se os resultados obtidos com o emprego de técnicas de Aprendizado de Máquina (AM) na previsão da taxa de vazão de uma conexão estabelecida para transportar segmentos de sessões de vídeo viabilizadas usando a tecnologia de *streaming* adaptativo. Essas sessões possuem características temporais, espaciais e próprias de uma comunicação realizada usando a Internet.

Para a realização dos estudos deste capítulo usaram-se os dados coletados na infraestrutura da operadora Telefônica Brasil S.A., que opera na região metropolitana da cidade de São Paulo. Entre os diversos planos oferecidos pela operadora, retirou-se somente medições realizadas em planos de até 2,0 Mbps. Essa escolha foi motivada pela representatividade desses planos na base, que respondem por $82,8\%$ das medições. Ademais, para o estudo selecionou-se oito métodos previsores de vazão procedente do objetivo específico 2, sendo: quatro baseados em AM, três baseados em média móvel e um baseado na última requisição da sessão.

A base com as medições consideradas, chamada de base telefônica, foi dividida usando a técnica de validação cruzada (do inglês, *Cross Validation - CV*) *holdout*. O método consiste em dividir a base de dados em dois subconjuntos: treino e teste. O primeiro é utilizado para identificar os padrões e possui a maior quantidade de dados , e.g., 2/3 da base. O subconjunto teste é usado para validar o modelo e possui o restante dos dados da base, e.g., 1/3 da base.

Nas seções seguintes mostram-se: i) as métricas usadas para a avaliação dos modelos, Seção 5.1; ii) as curvas de aprendizagem de quatro algoritmos na identificação de padrões da base Telefônica, Seção 5.2, e por fim iii) os resultados obtidos com os diversos métodos de predição quando dependência temporal e espacial da base são consideradas, seções 5.3 e 5.4, e quando a estabilidade da sessão de medição é o critério principal para filtragem da base, Seção 5.5.

5.1 Métricas de Avaliação

Os algoritmos Random Forest (RF), Naive Bayes, Regressão Linear e KNN foram usados para treinar modelos que identificam as correlações entre as características de uma requisição e a vazão (taxa de bits) do canal por onde a requisição foi transferida. A base coletada entre os anos de 2015 e 2018 foi usada para treinar os modelos e, ao final da fase de treino, os modelos gerados foram validados usando a base de teste. Os algoritmos baseados em médias móveis e de última amostra não precisam ser treinados, somente testados, uma vez que suas predições ocorrem durante o curso da transmissão.

Existem algumas métricas de avaliação para determinar a precisão das previsões, a mais comum é calcular o erro dos algoritmos de predição, dividindo o valor predito pelo valor real. Avaliaram-se os resultados dos experimentos realizados usando as seguintes métricas: Erro Absoluto Médio (do ingês, *Mean Absolute Error - MAE*) e a Raiz do Erro Quadrático Médio (do inglês, *Root Mean Squared Error - RMSE*), que são calculadas como segue:

$$\text{MAE} = \frac{\sum_{t=0}^{T-1} |\hat{y}_t - y_t|}{T}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{t=0}^{T-1} |\hat{y}_t - y_t|^2}{T}}$$

onde $\hat{y}_t$ é o valor da predição e y_t é o valor de fato da instância t. Nessas métricas, quanto menor é o erro, melhor será a predição do algoritmo. Formalmente, o valor do RMSE é maior que o do MAE, e quanto maior a diferença entre o MAE e o RMSE mais o erro de predição está disperso, i.e., o erro da predição varia mais em relação à média absoluta.

Utilizou-se também a métrica R^2 (coeficiente de determinação), que é comumente usada em métodos de predições. Essa métrica representa a proporção da variância na vazão previsível do preditor. Seu melhor resultado possível é 1.0, que indica que uma grande proporção da variabilidade na resposta foi explicada pelo modelo. Um número próximo de zero indica que o modelo de regressão não explicou grande parte da variabilidade na resposta. Além disso, pode ser negativa, indicando que o modelo pode ser arbitrariamente pior. A métrica é calculada como segue:

$$R^2 = 1 - \frac{\sum_{t=0}^{T-1} (\hat{y}_t - y_t)^2}{\sum_{t=0}^{T-1} (\bar{y} - y_t)^2}$$

onde $\bar{y}$ é o valor médio das observações, $\hat{y}$ é o valor previsto e T é o número de observações.

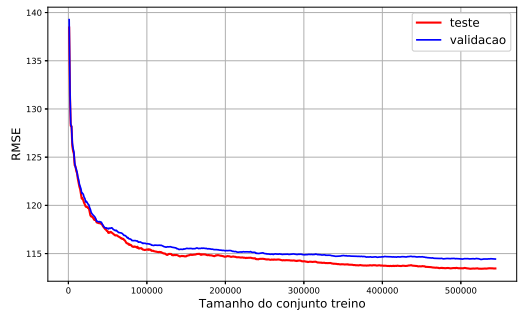
5.2 Curvas de Aprendizagem

O processo de definição de modelo de predição baseado em AM está ancorado na minimização do erro das predições. Os fatores ajustáveis desse erro são: o viés e a variância. A solução dada ao impasse criado pelo ajuste desses dois fatores é o que define a precisão do modelo. Em geral, os modelos com poucos parâmetros apresentam alta variância enquanto modelos com muitos parâmetros apresentam maior viés. Para avaliar o impacto desses fatores no erro do modelo as curvas de aprendizagem têm sido empregadas com frequência.

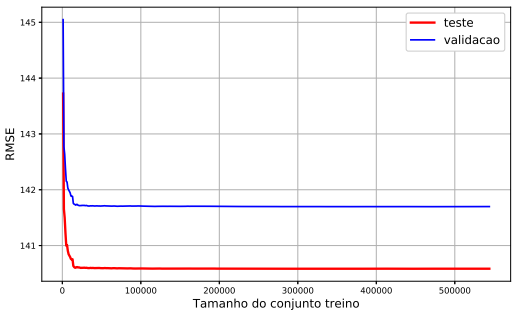
Para fim de definição, aqui descrevem-se os parâmetros dos algoritmos de aprendizagem de máquina utilizados nesta pesquisa durante o processo de treino. Para o Random Forest utilizou-se 400 estimadores, o MSE como função para medir a qualidade de uma divisão, profundidade máxima das árvores foi de 400, número mínimo de amostras para dividir um nó foi 10 e a semente usada pelo gerador de números aleatórios foi fixa em 42, assim como todos os algoritmos aqui descrito que possuíam o tal parâmetro. Para o KNN utilizou-se 10 vizinhos, a métrica da distância foi Minkowski, a função de peso usada na previsão foi a "distance" e o algoritmo usado para calcular os vizinhos mais próximos foi o "BallTree". Para o Naive Bayes e Regressão Linear, por não haver muitos parâmetros de regularização para esses algoritmos, utilizou-se os seus parâmetros padrão. Os algoritmos baseados em Média Móvel, todos eles se utilizara janela de tamanho 3 e suas respectivas definições padrão.

Neste trabalho, implementou-se o protocolo de geração das curvas de aprendizagem que considera três subconjuntos a partir da base avaliada. O primeiro, chamado de subconjunto treino, é formado por 70% da base, o segundo, chamado de teste, é formado por 15% da base, e o terceiro, subconjunto de validação, é formado por 15% da base. O modelo avaliado é treinado variando-se o tamanho do subconjunto de treino até o seu tamanho máximo, ao final de cada fase de treino, o erro do modelo é calculado para os dois outros subconjuntos. Após agregado as bases coletadas durante os anos de 2015 a 2018, as medições em cada subconjunto foram escolhidas de forma aleatória.

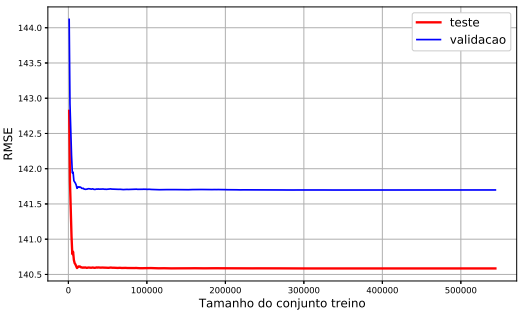
O comportamento da curva de aprendizagem permite que se avalie o impacto do viés e da variância nas predições feitas. Além disso, pode-se verificar se a precisão dos resultados é impactada pelo tamanho da base. Especificamente, se o modelo avaliado continua aprendendo a medida que é treinado com uma base cada vez maior, portanto, indicando se novos dados precisariam ou não ser coletados. Na Figura 8, mostra-se a evolução das curvas de aprendizado para os modelos definidos usando os seguintes algoritmos: Random Forest, Naive Bayes, Regressão Linear e KNN.



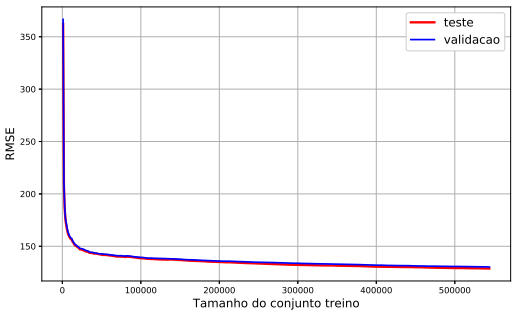
(a) Random Forest



(b) Naive Bayes



(c) Regressão Linear



(d) KNN

Figura 8 – Curva de Aprendizado dos Diferentes Algoritmos

Fonte: Produção própria

Os modelos baseados no Naive Bayes e Regressão Linear têm a sua precisão melhorada com o aumento do tamanho da base de treinamento, mas ocorre uma estabilização da precisão a partir de um certo ponto. Esse evento mostra que o modelo deixa de identificar padrões que ainda estão presentes na base, ou as hipóteses dos modelos não se verificam totalmente, ou o conjunto de características consideradas não permite a generalização dos modelos produzidos na etapa de treinamento. O distanciamento entre às duas curvas sugere que os modelos apresentam uma alta variância que pode ser reduzida com a inclusão de novas características, se elas estiverem disponíveis.

Os modelos baseados no RF e KNN têm curvas de aprendizagem bem distintas das verificadas anteriormente, isto é, ocorre redução do erro na medida que se usa mais dados do subconjunto treinamento. Além disso, as curvas geradas com as bases validação e teste permanecem próximas o que sugere que as características aprendidas pelos modelos são gerais, e que a chegada de novos dados irá ajudar no aumento da precisão do modelo. O modelo definido a partir do KNN é o que apresenta maior redução no erro, e que melhor captura os padrões das medições tomadas da base.

5.3 Explorando Dependência Temporal

A correlação temporal do tráfego de redes é um fenômeno bem documentado na literatura ([33], [28], [34]). Em geral, os trabalhos mostram que as flutuações do tráfego apresentam padrões distintos quando a escala de tempo é variada. Mais especificamente, o tráfego de certos dias, tendo-se a semana como escala de agregação, possui padrão que distingue o final de semana dos demais dias. Quando se considera a escala das horas dos dias, padrões distintos emergem para diferentes períodos do dia.

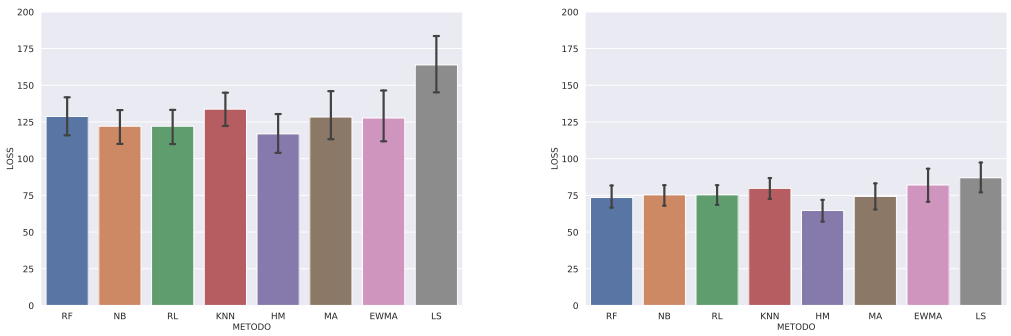
A seguir, mostram-se resultados obtidos na predição da vazão quando a base de dados foi tratada para explorar padrões temporais. A primeira avaliação considerou a dependência temporal em uma escala de 12 meses, tendo-se gerado a série analisada a partir da agregação dos dados coletados. A motivação dessa abordagem foi a construção de um modelo de predição para períodos específicos de um ano. A segunda avaliação considerou a esparsidade dos dados coletados, tendo-se estabelecido três períodos ao longo do dia, associados ao descanso, trabalho e lazer. A motivação dessa abordagem foi verificar possível impacto da esparsidade dos dados, que varia bastante com o período do dia, na construção de modelos de predição.

5.3.1 A Sazonalidade das Medições

Realizou-se um experimento prévio com as bases dos anos de 2015 e 2016. No experimento, a formação da base de teste e treino ocorreu em ordem cronológica, onde se agregaram os dados dos dois anos em ordem, como se fossem de apenas um ano, tendo-se retirado 30% das medições em três períodos diferentes, no início, no meio e no final do ano.

Tal experimento resultou em erro de validação maior ao utilizar a base de teste do início do ano e menor erro ao utilizar a base de teste do final do ano, causado pela tendência de aumento da vazão no padrão identificado na base, ver Figura 6c.

Assim, para verificar os efeitos da sazonalidade nas medições, i.e., padrões que emergem em certos períodos, realizou-se o experimento definitivo usando os dados agregados dos anos de 2015 até 2018. Nesse experimento, usou-se somente a separação da base em que o teste é construído a partir dos 30% das medições do final do ano e o restante usado para treino, pois foi a manipulação de menor erro de predição durante o experimento prévio. A Figura 9 mostra os resultados do experimento definitivo.



(a) RMSE (b) MAE

Figura 9 – A Sazonalidade das Medições e seus Efeitos na Predição da Vazão: RMSE e MAE

Fonte: Produção própria

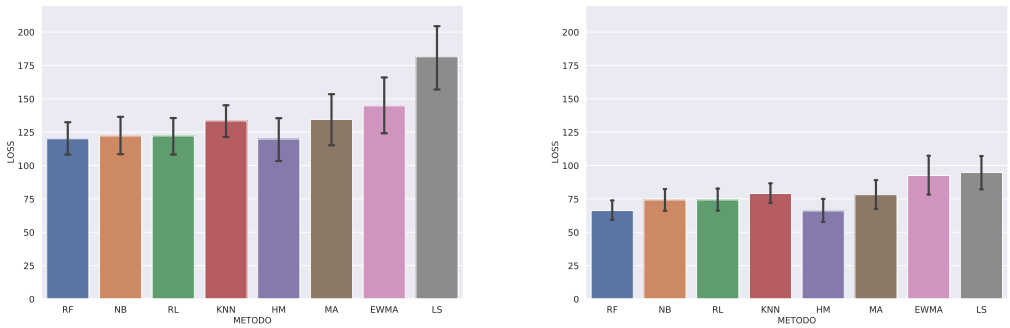
Tal abordagem objetivou melhorar o desempenho das predições dos métodos realizando uma agregação temporal das bases, tratando-as como se fossem todas do mesmo ano. Tal tratamento procurou suprir a deficiência da pouca quantidade de dados e a grande esparsidade, devido a pouca quantidade de medições por intervalo de tempo.

O uso das bases produz resultados com erros próximo para os diferentes tipos de métodos: os baseados em sessões históricas e os baseados em sessões em curso. O método que apresentou melhor resultado foi o HM. Esse método consegue bons resultados devido à natureza dos dados, média conservativa que tenta mitigar valores altos e acentuar os baixos para capturar a tendência central dos dados.

5.3.2 A Esparsidade da Base

A princípio os resultados anteriores de AM não se demostram serem tão bons, quando se compara os modelos de média móvel, 5.3.1. Portanto, realizou-se uma análise de frequência das medições, onde se verificou que há padrões de horários que mudam no decorrer do dia e da semana, Figura 4. Além disso, a base apresenta esparsidade distinta quando considera que ao longo do dia há variação de demanda.

Dessa forma, considerou-se o dia dividido em três períodos - descanso(00:00∼10:00h), trabalho(10:00∼18:00h) e lazer(18:00∼24:00h). Para a construção das bases de treino e teste dos modelos, considerou a esparsidade dos dados naqueles períodos. Desta forma, garantiu-se que durante o treino, dos modelos de AM, a base utilizada seja representativa na esparsidade dos dados pelos períodos estabelecidos.



(a) RMSE das bases 2015 a 2018 (b) MAE das bases 2015 a 2018

Figura 10 – A Esparsidade da Base e seus Efeitos na Predição de Vazão: RMSE e MAE

Fonte: Produção própria

A Figura 10 mostra nos resultados obtidos com o experimento e a conclusão geral é que houve redução do erro, em relação ao tratamento que desconsiderou a esparsidade da base, medido pelo RMSE e MAE. O modelo de média harmônica foi o que apresentou menor erro entre os algoritmos que fazem predição na sessão em curso, já os que fazem predição baseada em sessões históricas, o Random Florest foi o que apresentou o menor erro. O erro dos modelos de AM deram uma reduzida com relação ao tratamento anterior, e a diferença entre os dois tipos de abordagem de predição é insignificante. No entanto, o erro medido é contínua sendo alto para que considere o uso de algum dos seus modelos em detrimento aos métodos de medição baseado em sessão em curso.

5.4 Explorando Dependência Espacial

Constatada a importância de se considerar a esparsidade da base, em um contexto em que a correlação temporal tem importante papel, nesta seção avalia-se o papel da correlação espacial na construção dos modelos de predição. Mais especificamente, os estudos realizados nesta seção consideram o conjunto de 15 medições consecutivas, realizadas por um cliente vídeo *streaming*, como sendo a unidade básica das bases, e todos os tratamentos realizados consideram tal unidade.

5.4.1 Dependência Inter-Sessões

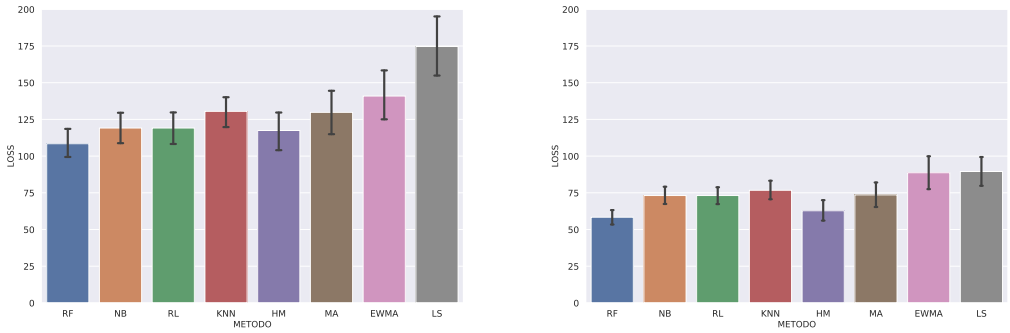
A primeira tentativa de verificar a importância da dependência espacial nas bases considerou cada sessão indivisível. Para isso, extraiu-se aleatoriamente 30% das sessões para o teste dos modelos e o restante da base foi usado para treinamento. Essa extração considerou a esparsidade da base do tratamento anterior, ou seja, para três períodos do dia, i.e. 00:00∼10:00hs, 10:00∼18:00hs e 18:00∼24:00hs, extraíram-se proporcionalmente as sessões que foram usadas para validação dos modelos.

Acredita-se que os métodos de AM, nos tratamentos anteriores, seguiram basicamente a mesma ideia do método baseado em seção em curso. Dado que, em análises passadas a base de dados usada mostrou ser esparsa, no sentido de haver em certos períodos do dia da semana há pouco ou nenhum dado para os métodos criarem modelos de predição mais precisos para este período e, ao retirar as instâncias para o subconjunto de teste, os esparsamentos aumentam.

Assim resolveu-se usar durante o treino e o teste, sessões completas de forma aleatória, i.e., selecionar aleatoriamente nos períodos citados e independente do período do ano. Dessa maneira os métodos tiveram maior chance de identificar os padrões da vazão e sazonalidade em sessões completas. O uso de sessões completas durante a fase de treinamento e teste impede a ausência de medições importantes na identificação de padrões nas sessões, enquanto tais medições nunca foram apresentadas na fase de teste.

A racionalidade desse tratamento é a apreensão dos padrões de interação nas sessões, enquanto se evita a concentração de sessões em um período do dia na fase de

teste. Tal concentração pode produzir um viés na validação dos modelos de AM se um padrão de um período específico for melhor apreendido na fase de treinamento. A Figura 11 mostra os resultados obtidos conforme o tratamento da base.



(a) RMSE (b) MAE

Figura 11 – Dependência Inter-Sessão e seus Efeitos na Predição de Vazão: RMSE e MAE

Fonte: Produção própria

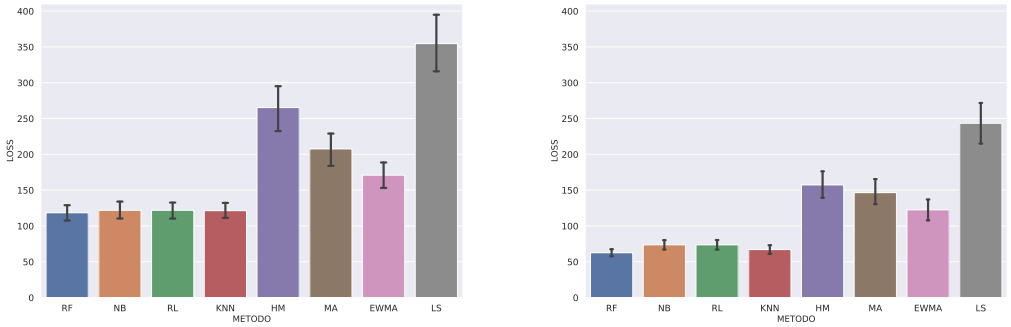
A conclusão geral é que os métodos de AM apresentam um melhor desempenho, comparados a resultados anteriores. Essa melhora é uma indicação de que os padrões que se repetem nas sessões estão sendo apreendidos com maior distinção. Além disso, os resultados apresentam intervalo de confiança menor que aqueles verificados anteriormente, indicando maior precisão dos métodos. Visando melhorar a relação entre erro e confiança, realizou uma nova abordagem na manipulação dos dados, que será descrito na próxima seção.

5.4.2 Dependência Intra-Sessões

O par de protocolo HTTP/TCP foi utilizado para transferência dos dados que formam a base. A ação dos mecanismos de controle de congestionamento do TCP é caracterizada por um crescimento rápido da vazão, no início de cada sessão, seguido por uma certa estabilização na parte final da sessão. Esse comportamento sugere a existência de correlação entre requisições consecutivas, nos dois estados por que passam cada sessão.

Para verificar a existência da correlação dos eventos dentro de uma sessão e a sua importância na definição dos modelos, extraiu-se 1/3 das medições de cada sessão para teste e o restante foi usado para treinamento dos modelos. Os subconjuntos de treinamento

e teste foram construídas agrupando as medições de cada sessão em triplas consecutivas, e para cada uma dessas triplas retirou-se as duas primeiras medições para a base de treino e a terceira para base de teste. A Figura 12 mostra o resultado obtido para os diferentes modelos quando as bases são formadas conforme aqui descrito.



(a) RMSE (b) MAE

Figura 12 – Dependência Intra-Sessão e seus Efeitos nas Predições: RMSE e MAE

Fonte: Produção própria

Nessa manipulação os métodos de AM, além de ter a percepção das últimas interações entre cliente-servidor da sessão sob predição, eles são treinados com histórico de sessões entre o mesmo cliente-servidor, ou no mínimo com histórico de sessões cliente-servidor com características similares. Pois, percebeu-se que durante o treino quando há informações de sessões que compartilham com a sessão sob predição, elas têm melhor desempenho de precisão. Assim, nessa manipulação de dados, as predições são tomadas em base na percepção da sessão sob predição e de observações de sessões com as mesmas características, gerando previsões com melhor precisão.

A conclusão geral é que os métodos foram capazes de identificar os padrões da vazão com maior precisão. Os métodos baseados em médias e *última amostra - LS* apresentaram menor precisão o que é explicado pelo tratamento realizado para produzir a base de teste. A ausência de 2/3 das medições em cada sessão impactou a estimativa baseada em histórico recente, que é o caso dos métodos baseado em média, e para o caso do método baseado na última amostra (LS), a ausência de duas medidas, que estão correlacionadas ao valor que se deve estimar, reduziu sua precisão.

5.5 Estabilidade da Sessão: Vazão realizável

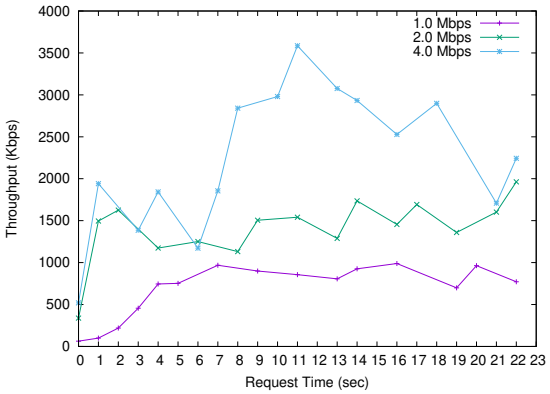
Nesta sessão apresenta-se conceito de vazão realizável de uma conexão estabelecida para transferir segmentos de um fluxo de vídeo que usa taxa de bit adaptável para casar o recurso de transmissão e qualidade do fluxo, seguidamente, é apresentado os resultados obtidos derivados desta abordagem.

5.5.1 Vazão realizável: conceito

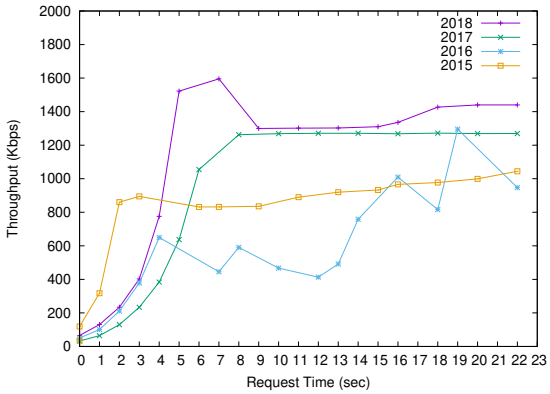
Esse estudo baseia-se nos resultados apresentados na Seção 4.2 quando se verificou que as sessões com duração de até 22 segundos são a maioria absoluta na base. A implicação direta desse resultado é que as conexões foram capazes de transportar o fluxo completo antes da sua total reprodução, pelo menos oito segundos antes da reprodução. O conteúdo entregue tem nível de qualidade compatível com a vazão do canal, garantido pela dinâmica do algoritmo de adaptação da taxa.

Na Figura 13a, a vazão das sessões com duração inferior a 22 segundos foram plotadas para os planos que dominam a base de dados: 1,0, 2,0 e 4,0 Mbps. O algoritmo usado para adaptar a taxa de bits está representando medições de sessões, ou seja, a análise do último segmento acessado define a taxa de bits do próximo. Essa abordagem baseia-se em decisões de curta memória e expõe os efeitos da partida lenta e da prevenção de congestionamento do TCP na camada de aplicação. O padrão de rápido crescimento da vazão dura poucas iterações que é seguido por um padrão de oscilações, que expõe a tentativa contínua do aplicativo em harmonizar a taxa de bits e a vazão à conexão de sessão avaliada, e como a mecânica de prevenção de congestionamento do TCP reage a essas tentativas.

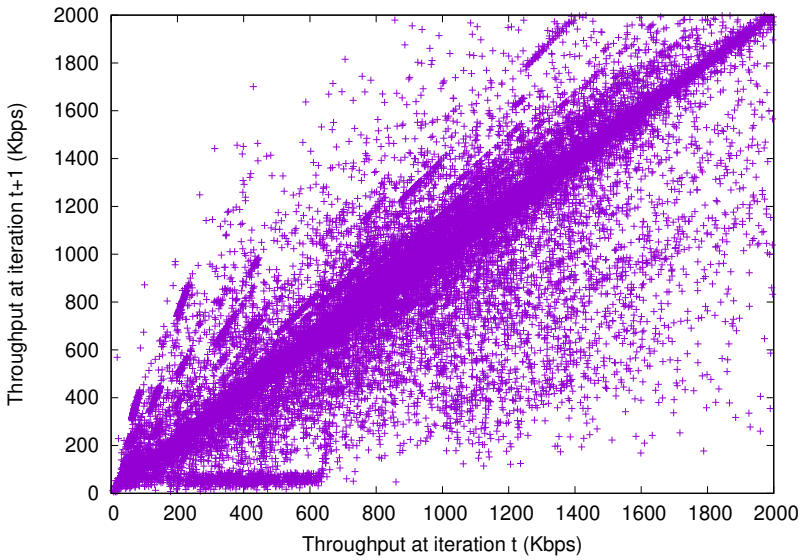
Na Figura 13b, mostram-se exemplos de sessões construídas com o plano de 2,0 Mbps. Esses exemplos foram coletados ano a ano, ao longo do mesmo horário, de 20:00 às 21:00. A duração das sessões avaliadas é longa o suficiente para expor o nível de congestionamento enfrentado pelas conexões, a vazão medida caiu no primeiro terço da sessão pelo menos uma vez, em sua maioria. Além disso, no último terço da sessão, o limite superior da vazão medida de fato sustentou-se. Por exemplo, a vazão da sessão de 2016 evoluiu de um indício de estado de congestionamento antecipado, i.e., a queda acontece



(a) Plano de Dado



(b) Correlação Temporal



(c) Correlação Espacial

Figura 13 – Medindo a Estabilidade da Vazão

Fonte: Produção própria

após quatro iterações de transmissão, para o limiar superior, conforme as solicitações aproximam-se do último terço da sessão.

A correlação espacial da vazão medida também foi considerada, com uma distância entre eventos iguais a um passo (Figura 13c). O padrão surgido é que solicitações consecutivas compartilham vazões semelhantes e essa similaridade se torna clara à medida que a vazão medida se aproxima da vazão teórica dos planos de dado considerados. Em outras palavras, as infraestruturas avaliadas conseguiram manter a taxa de transferência que foi atingida no último terço da sessão, tal padrão identificado foram a base para a seleção dos dados neste experimento.

Todas estas evidências apresentadas aqui foram alicerce para a criação do conceito de vazão realizável e a metodologia empregada nestes experimentos. Então, o que é a vazão realizável (ou taxa realizável)?. A taxa realizável é um valor predito a partir de um modelo preditor de AM. Esse modelo preditor é treinado sob dados históricos de sessões de acesso de vídeos do passado, i.e., históricos de vazão e os fatores (característica de infraestruturas e temporais) que podem afetar a taxa de transferência de um arquivo de vídeo durante uma sessão cliente-servidor. Neste trabalho utilizou-se quatro algoritmos clássicos de AM para gerar os modelos preditores e realizar os experimentos. A ideia de utilizar esses quatro algoritmos foi de analisar e avaliar seus comportamentos diante do problema e, futuramente, usar o que melhor se adaptou ao problema para realizar as predições.

Essa taxa realizável, por si só, não visa determinar a vazão durante uma sessão de vídeo, mas sim, utilizar esse valor para ajudar na tomada de decisão dos algoritmos de adaptação, i.e., a taxa realizável irá acrescentar uma informação à mais, além do *buffer* e a vazão do último segmento baixado, no momento em que os algoritmos decidem a qualidade do próximo segmento a ser baixado. Essa abordagem visa melhorar a qualidade de experiência dos usuários ao acessar um vídeo *streaming*, que pode tornar possível entregar conteúdos com qualidade que melhor expresse a vazão do canal desse usuário.

5.5.2 Vazão realizável: resultados

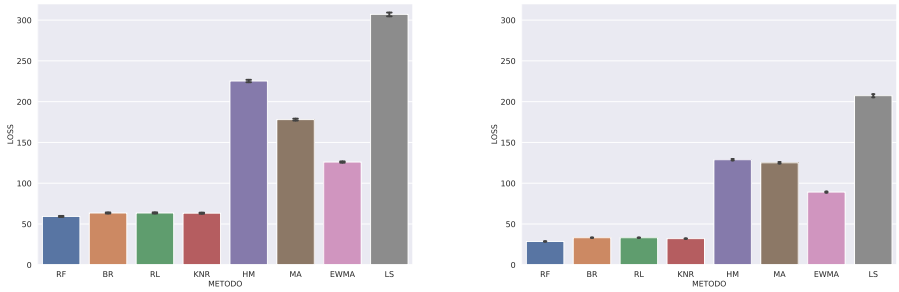
Nessa avaliação, consideraram-se às cinco últimas medições de cada sessão, tendo-se a seguinte questão para ser respondida: qual a vazão realizável de um canal? A resposta a essa questão, a partir de sessões de duração intermediária, i.e., não caracterizadas como sessões elefantes e muito menos ratos [35] tem aplicação, por exemplo, na oferta de anúncios em plataformas de vídeo sob demanda. Em tais plataformas o anúncio deve ser oferecido antes da vinculação de um conteúdo selecionado pela audiência, e uma estratégia para apresentação é a disponibilização antecipada da propaganda no dispositivo da audiência. Determinar um limiar de tempo máximo para que tal propaganda seja disponibilizada é uma ação que depende do conhecimento da taxa realizável do canal que serve a audiência.

Os testes dos métodos foram conduzidos com a técnica de Validação Cruzada (do inglês, *Cross-Validation*) k-folds, com $k = 5$. A redução da base de dados para um terço do

seu tamanho original foi a principal motivação para uso dessa técnica. Com o propósito de verificar o erro para cada um dos planos, a análise dos resultados dos experimentos seguiu separadamente: *i*) utilizando somente os planos de 1,0 Mbps e *ii*) utilizando somente os planos de 2,0 Mbps.

As abordagens utilizaram todos os algoritmos descritos na Seção 2.5. O experimento (*i*) foi a que apresentou menores erros de predição dos métodos de AM, como mostra a Figura 14. As métricas mostram que os erros dos modelos estão menos dispersos, isso é graças à vazão das medições serem menores que as presentes nos outros planos e a vazão das sessões flutuar sempre na mesma faixa.

No experimento (*ii*), o erro segue padrão similar ao obtido no experimento (*i*), como visto na Figura 15. Entretanto, devido ao plano ser de grau acima do anterior, esses erros também são de magnitude superior.

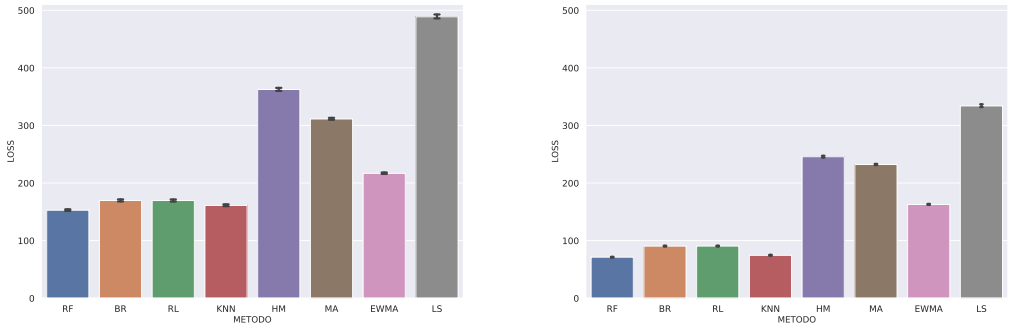


(a) RMSE da base 2015 à 2018 (b) MAE da base 2015 à 2018

Figura 14 – Estabilidade da Vazão nos Planos de 1,0 Mbps: RMSE e MAE

Fonte: Produção própria

Mediu-se a precisão das predições de diferentes perspectivas, evitando-se que tendências, características das métricas de desempenho utilizadas, fossem tomadas como únicas. Para tal, além das métricas RMSE e MAE, uso-se a métrica R^2 na avaliação dos modelos. Essa métrica é usada com frequência na avaliação de modelagens feitas com técnicas AM para regressão. Entre os algoritmos AM avaliados, nos diferentes planos, o RF teve o melhor desempenho, com um valor de R^2 igual a $0,94$ e $0,88$, nos experimentos *i*) e *ii*), respectivamente. Esse desempenho do RF deve-se a forma como os dados são manipulados para evitar *overfitting* e as características da base, i.e., reduzido tamanho e esparsidade. Outras modelos de AM apresentam desempenho equivalente, mas estão



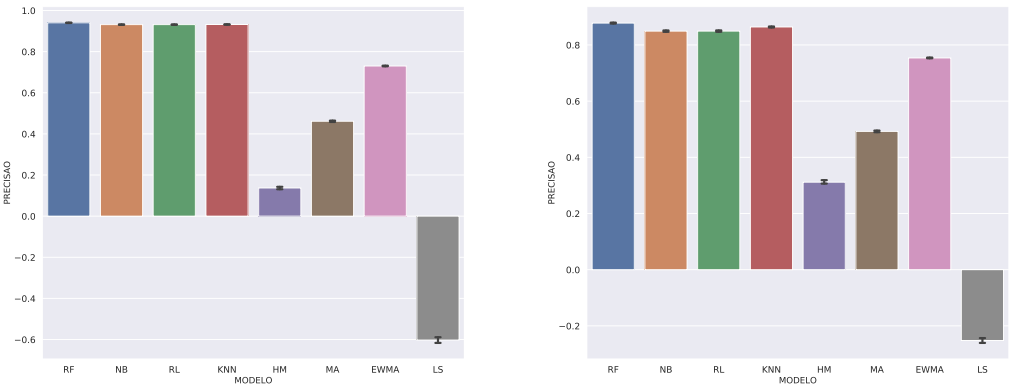
(a) RMSE da base 2015 à 2018 (b) MAE da base 2015 à 2018

Figura 15 – Estabilidade da Vazão nos Planos de 2,0 Mbps: RMSE e MAE

Fonte: Produção própria

abaixo do alcançado pelo RF.

Os métodos baseados em média móvel e na última amostra apresentam erros que são até 100% maiores que os verificados quando os modelos baseados em AM são utilizados para realizar a predição. Entre aqueles modelos, o EWMA apresenta os melhores desempenhos, que pode ser explicado devido à ponderação para cada dado com sua redução exponencial da importância na medida que a idade dos dados avaliados avança. Dessa forma, as medidas mais recentes da vazão têm peso maior, o que faz todo sentido, pois elas caracterizam melhor série até aquele momento.



(a) Planos de 1,0 Mbps da base de 2015 à 2018 (b) Planos de 2,0 Mbps da base de 2015 à 2018

Figura 16 – Estabilidade da Vazão: R^2

Fonte: Produção própria

O uso da última vazão registrada se mostrou uma abordagem inadequada. O Algoritmo LS apresentou os valores mais altos de RMSE e MAE, e apresentou um R^2 negativo, indicando que a sua inadequação na modelagem. Esse comportamento é atribuído a esparsidade dos dados, i.e., a última amostra da vazão que o modelo dispõe nem sempre representa a vazão do momento. Além disso, os dados dos planos de 2,0 Mbps indicam ser mais estáveis que os dados dos planos teóricos de 1,0 Mbps, visto que a precisão dos métodos baseados em média móvel e última amostra são maiores neste plano.

5.6 Considerações Finais do Capítulo

Neste capítulo discutiu-se a experimentação utilizada para o estudo de predição de vazão e a avaliação dos métodos. A falta de estudos experimentais, como assim foi conduzido e descrito neste capítulo, justifica o empenho empregado nos experimentos realizados. Desta forma, apresentou-se um estudo experimental que teve como objetivo verificar formas de tratar dados de vazão, que são afetados por esparsamentos e pouco dados, e como esses tratamentos impactam os métodos de AM para regressão e baseados em sessão em curso para diferentes estratégias, justificado pelo objetivo específico 3.

6 CONSIDERAÇÕES FINAIS

Neste trabalho apresentou-se estudo conduzido para construção de modelos de predição de vazão de conexões estabelecidas entre clientes e servidores de conteúdo multimídia na Internet. A tecnologia usada nas transmissões adapta a taxa de bits do conteúdo às flutuações próprias de uma conexão fim-a-fim na Internet, visando a construção de sessões com a melhor qualidade de imagem. Os modelos construídos consideraram dados históricos de sessões e foram baseados em técnicas de aprendizagem de máquina. Especificamente, empregaram-se algoritmos Random Forest, naive bayes, KNN e Regressão Linear na construção dos modelos. A contribuição principal deste estudo é a definição da taxa realizável por conexões estabelecidas em planos de dados de baixa velocidade.

Conforme discutido no Capítulo 3, a predição da vazão de conexões a partir dos clientes é uma tarefa desafiadora. A maioria dos trabalhos são aplicados em cenários específicos evidenciando a necessidade de modelos de predição que considerem as particularidades das diferentes classes de clientes, evidenciando que a chave para uma previsão precisa é a descoberta de sessões históricas que compartilham características em comuns com a sessão sob predição. Os tratamentos realizados na base de dados de sessões é outra questão que requer atenção especialmente em contexto com poucos dados disponíveis. Identificar padrões que caracterizam uma nova sessão permite que provedores de conteúdo ofertem serviços que sejam percebidos pelos usuários como tendo a melhor qualidade. Adicionalmente, os provedores de serviços de conexão a Internet (ISPs) podem se valer dessa identificação para realizar otimização e planejamento dos recursos de sua rede.

Realizou-se estudos com a base de dados para identificar suas características. Os dados coletados em diferentes provedores de acesso e em diferentes regiões metropolitanas do país foram considerados nesse estudo. Avaliou-se a esparsidade dos dados considerando-se a escala dia da semana e a escala horários do dia. Finalmente, avaliou-se a esparsidade para os diversos planos de dados de maior presença na base, i.e. um, dois e quatro megabits por segundo.

Essas análises e o conhecimento prévio do problema permitiram a definição do conjunto de características mais relevantes para a tarefa de predição da vazão. Para tal, foram considerados quatro algoritmos de Aprendizagem de Máquina, i.e. Random Forest, Naive Bayes, Regressão Linear e KNN; três algoritmos baseados em média, i.e Média Aritmética, Média Harmônica e Média Móvel Exponencial Ponderada; e finalmente um algoritmo baseado na última amostra (LS). Os experimentos conduzidos abordam o treinamento dos modelos de predição a partir de uma base esparsa, portanto com uma quantidade limitada de dados de treinamento. Mostrou-se como os métodos podem ser treinados e como os dados podem ser manipulados para suprir essas restrições, assim como os impactos que cada uma das abordagens.

Primeira parte do estudo experimental, i.e. seções 5.3 e 5.4, abordou-se o problema da predição da vazão durante toda sessão, que tem duração, na grande maioria dos casos, inferior a 30 segundos. Os modelos de AM foram treinados apresentaram os melhores resultados em comparação aos demais métodos utilizados. Entretanto, a esparsidade da base é um fator limitador de desempenho, mas como se verificou posteriormente, as modelos criados estão limitados pelas características disponíveis na base. Também mostrou-se que as pontuações de R^2 alcançadas são mais altas nos métodos de AM do que os métodos que usam apenas informações da sessão em curso.

Na segunda parte dos experimentos, seção 5.4.2, avaliou-se as predições de vazão para sessões de vídeos, considerando somente dados da prevenção de congestionamento, i.e., informações da vazão de quando a sessão está mais estável. Nesta abordagem, mostrou-se que os algoritmos de AM são mais precisos, quando comparados com os demais métodos e o erro gerado nas predições é menor. Esse resultado encontra aplicação na implementação de técnicas de adaptação da qualidade das sessões, uma vez que estabelece um limiar para a capacidade de transmissão do canal.

Como trabalhos futuros, pretende-se conduzir estudos para descoberta de novas características da base, assim como complementação da base a partir de outras fontes de dados, especificamente aquelas que registram eventos anormais que podem influenciar o desempenho das infraestruturas de comunicação. Além disso, com a continuidade do projeto de medição a base continua sendo atualizada permitindo que métodos mais sofisticados

possam ser empregados.

REFERÊNCIAS

1 CISCO, V. N. I. Global mobile data traffic forecast update, 2017–2022. *Cisco White paper*, 2019. Citado na página 17.

2 LI, Z.; KAAFAR, M. A.; XIE, G. Session throughput prediction for internet videos. *IEEE Communications Magazine*, IEEE, v. 54, n. 12, p. 152–157, 2016. Citado 2 vezes nas páginas 18 e 39.

3 KUA, J.; ARMITAGE, G.; BRANCH, P. A survey of rate adaptation techniques for dynamic adaptive streaming over http. *IEEE Communications Surveys Tutorials*, v. 19, n. 3, p. 1842–1866, thirdquarter 2017. Citado 3 vezes nas páginas 18, 20 e 25.

4 BOUTEN, N. et al. Dynamic server selection strategy for multi-server HTTP adaptive streaming services. In: *2016 12th International Conference on Network and Service Management (CNSM)*. [S.l.: s.n.], 2016. p. 82–90. Citado 2 vezes nas páginas 18 e 20.

5 HUANG, T.-Y. et al. Confused, timid, and unstable: Picking a video streaming rate is hard. In: *Proceedings of the 2012 Internet Measurement Conference*. New York, NY, USA: ACM, 2012. (IMC '12), p. 225–238. ISBN 978-1-4503-1705-4. Citado 2 vezes nas páginas 19 e 20.

6 CLAEYS, M. et al. Cooperative announcement-based caching for video-on-demand streaming. *IEEE Transactions on Network and Service Management*, v. 13, n. 2, p. 308–321, June 2016. ISSN 1932-4537. Citado na página 20.

7 THOMAS, E. et al. Application of sand technology in dash-enabled content delivery networks and server environments. *SMPTE Motion Imaging Journal*, v. 127, n. 1, p. 48–54, Jan 2018. ISSN 1545-0279. Citado na página 20.

8 SEUFERT, M. et al. A survey on quality of experience of http adaptive streaming. *IEEE Communications Surveys & Tutorials*, IEEE, v. 17, n. 1, p. 469–492, 2015. Citado na página 23.

9 AKHSHABI, S.; BEGEN, A. C.; DOVROLIS, C. An experimental evaluation of rate-adaptation algorithms in adaptive streaming over http. In: ACM. *Proceedings of the second annual ACM conference on Multimedia systems*. [S.l.], 2011. p. 157–168. Citado na página 24.

10 MILLER, K.; AL-TAMIMI, A.-K.; WOLISZ, A. Qoe-based low-delay live streaming using throughput predictions. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, ACM, v. 13, n. 1, p. 4, 2017. Citado na página 24.

11 REITER, U. et al. Factors influencing quality of experience. In: *Quality of experience*. [S.l.]: Springer, 2014. p. 55–72. Citado na página 25.

12 AMIN, R. A. A.; INDRAJIT, R. E. Analysis of effectiveness of using simple queue with per connection queue (pcq) in the bandwidth management (a case study at the academy of information management and computer mataram (amikom) mataram). *Journal of Theoretical and Applied Information Technology*, Journal of Theoretical and Applied Information, v. 83, n. 3, p. 319, 2016. Citado na página 26.

13 LIU, L. et al. Message dissemination for throughput optimization in storage-limited opportunistic underwater sensor networks. In: *2016 13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. [S.l.: s.n.], 2016. p. 1–9. Citado na página 27.

14 MUKHOTI, J. et al. Knowledge extraction from a time-series using segmentation, fuzzy matching and predictor graphs. In: IEEE. *Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on*. [S.l.], 2016. p. 1201–1208. Citado na página 27.

15 ASADI, N.; ALAVIJEH, M. K.; ZILOUEI, H. Development of a mathematical methodology to investigate biohydrogen production from regional and national agricultural crop residues: A case study of iran. *International Journal of Hydrogen Energy*, Elsevier, v. 42, n. 4, p. 1989–2007, 2017. Citado na página 28.

16 BISHOP, C. M. *Pattern recognition and machine learning*. [S.l.]: springer, 2006. Citado na página 29.

17 FAREE, A.; WANG, Y.; LI, G. Modeling grain storage quality with linear regression. In: IEEE. *Computer and Communications (ICCC), 2017 3rd IEEE International Conference on*. [S.l.], 2017. p. 2904–2909. Citado na página 30.

18 BUŽIĆ, D.; DOBŠA, J. Lyrics classification using naive bayes. In: IEEE. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. [S.l.], 2018. Citado na página 32.

19 JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. Citado na página 33.

20 HALL, P. et al. Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 36, n. 5, p. 2135–2152, 2008. Citado na página 33.

21 HYNDMAN, R. J. Moving averages. In: *International encyclopedia of statistical science*. [S.l.]: Springer, 2011. p. 866–869. Citado na página 34.

22 ZHANG, W. et al. Samen-svr: using sample entropy and support vector regression for bug number prediction. *IET Software*, IET, v. 12, n. 3, p. 183–189, 2018. Citado na página 34.

23 ZHAO, H.; XU, Z.; CUI, F. Generalized hesitant fuzzy harmonic mean operators and their applications in group decision making. *International Journal of Fuzzy Systems*, Springer, v. 18, n. 4, p. 685–696, 2016. Citado na página 35.

24 LIU, Y.; LEE, J. Y. An empirical study of throughput prediction in mobile data networks. In: IEEE. *Global Communications Conference (GLOBECOM), 2015 IEEE*. [S.l.], 2015. p. 1–6. Citado 2 vezes nas páginas 37 e 38.

25 BUI, N.; MICHELINAKIS, F.; WIDMER, J. A model for throughput prediction for mobile users. In: VDE. *European Wireless 2014; 20th European Wireless Conference; Proceedings of*. [S.l.], 2014. p. 1–6. Citado na página 37.

26 WEI, B.; KANAI, K.; KATTO, J. History-based throughput prediction with hidden markov model in mobile networks. In: IEEE. *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*. [S.l.], 2016. p. 1–6. Citado na página 38.

27 KANAI, K. et al. A highly-reliable buffer strategy based on long-term throughput prediction for mobile video streaming. In: IEEE. *Consumer Communications and Networking Conference (CCNC), 2015 12th Annual IEEE*. [S.l.], 2015. p. 677–682. Citado na página 38.

28 JIANG, J.; SEKAR, V.; SUN, Y. Dda: Cross-session throughput prediction with applications to video bitrate selection. *arXiv preprint arXiv:1505.02056*, 2015. Citado 2 vezes nas páginas 39 e 55.

29 BASSO, S. et al. Measuring dash streaming performance from the end users perspective using neubot. In: *Proceedings of the 5th ACM Multimedia Systems Conference*. New York, NY, USA: ACM, 2014. (MMSys '14), p. 1–6. ISBN 978-1-4503-2705-3. Disponível em: <http://doi.acm.org/10.1145/2557642.2563671>. Citado 3 vezes nas páginas 41, 42 e 45.

30 LAB, M. M-lab. Último acesso em Junho, 2019. 2019. Disponível em: <https://www.measurementlab.net/data>. Citado na página 41.

31 NEUBOT. Neubot data analysis. Último acesso em Agosto, 2019. 2019. Disponível em: <http://streaming.polito.it/neubot/world.html>. Citado na página 42.

32 MAXMIND. https://www.maxmind.com/en/home. Último acesso em Junho, 2019. 2019. Citado na página 43.

33 Zhao, Z. et al. Lstm network: a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, v. 11, n. 2, p. 68–75, 2017. ISSN 1751-956X. Citado na página 55.

34 WANG, Y. et al. Traffic data reconstruction via adaptive spatial-temporal correlations. *IEEE Transactions on Intelligent Transportation Systems*, IEEE, n. 99, p. 1–13, 2018. Citado na página 55.

35 Hohn, N.; Veitch, D.; Abry, P. Cluster processes: a natural language for network traffic. *IEEE Transactions on Signal Processing*, v. 51, n. 8, p. 2229–2244, 2003. ISSN 1053-587X. Citado na página 63.