UM ESTUDO SOBRE O IMPACTO DA PANDEMIA DA COVID-19 NO TRÁFEGO DE REDES RESIDENCIAIS

Mariana Corrêa Ribeiro

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Edmundo Albuquerque de Sousa e Silva

Rio de Janeiro
Janeiro de 2023

UM ESTUDO SOBRE O IMPACTO DA PANDEMIA DA COVID-19 NO TRÁFEGO DE REDES RESIDENCIAIS

Mariana Corrêa Ribeiro

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientador: Edmundo Albuquerque de Sousa e Silva

Aprovada por: Prof. Edmundo Albuquerque de Sousa e Silva
Prof. Donald Fred Towsley
Prof. Geraldo Bonorino Xexéo

RIO DE JANEIRO, RJ – BRASIL
JANEIRO DE 2023

Corrêa Ribeiro, Mariana

Um estudo sobre o impacto da pandemia da COVID-19 no tráfego de redes residenciais/Mariana Corrêa Ribeiro. – Rio de Janeiro: UFRJ/COPPE, 2023.

XI, 47 p.: il.; 29, 7cm.

Orientador: Edmundo Albuquerque de Sousa e Silva

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2023.

Referências Bibliográficas: p. 43 – 47.

1. Redes de residênciais. 2. COVID-19. 3. PARAFAC. I. Albuquerque de Sousa e Silva, Edmundo. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Aos meus avós Berthier Ribeiro-Filho e José Reinaldo Corrêa, que mesmo não estando mais nesse plano, se fazem presentes em minha vida. E aos meus pais que sempre me apoiaram e me guiaram em todas as minhas decisões.

Agradecimentos

Gostaria de agradecer a minha família, em especial aos meus pais Rosa Malena A. Corrêa e Berthier Ribeiro-Neto que sempre acreditaram, me incentivaram e nunca mediram esforços para me dar todo o suporte no que eu precisasse. Ao meu orientador, prof. Edmundo de Souza e Silva, pela orientação, incentivo, paciência e por todos os ensinamentos e conselhos que permitiram o meu crescimento profissional e pessoal. A minha amiga Ananda Streit por ter me ensinado, ajudado e apoiado durante esses anos. A prof. Rosa Leão e a todos os demais professores e alunos do laboratório LAND que se dispuseram a ouvir e contribuir para o desenvolvimento desta dissertação. Por fim, a todos aqueles que ajudaram de alguma forma para a minha formação acadêmica e para a realização deste trabalho.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UM ESTUDO SOBRE O IMPACTO DA PANDEMIA DA COVID-19 NO TRÁFEGO DE REDES RESIDENCIAIS

Mariana Corrêa Ribeiro

Janeiro/2023

Orientador: Edmundo Albuquerque de Sousa e Silva

Programa: Engenharia de Sistemas e Computação

Em Março de 2020, a Organização Mundial de Saúde (OMS) declarou a pandemia da COVID-19. Medidas de confinamento foram implementadas em vários países ao redor do mundo com o objetivo de conter a propagação do vírus. Grande parte da população mundial foi obrigada a utilizar a rede de Internet doméstica para trabalho, educação e outras atividades, o que causou uma drástica mudança na rotina das pessoas e consequentemente, alteraram os padrões de tráfego das redes residênciais. Essa dissertação tem o objetivo de estudar as mudanças ocorridas em decorrência da pandemia da COVID-19 nos padrões de tráfego das redes residênciais. Analisamos o tráfego de várias residências espalhadas em 15 cidades no estado do Rio de Janeiro por vários meses durante anos de 2020, 2021 e 2022. Utilizamos informações de tráfego residencial fornecidas por um Provedor de Serviços de Internet (ISP) de médio porte (Gigalink) para comparar o tráfego antes e depois do início da quarentena em 15 cidades do estado do Rio de Janeiro. Aplicamos decomposição tensorial, clusterização e classificação para identificar os perfis de tráfego residencial. Descobrimos que 20% das residências mudaram seus perfis diários imediatamente após o confinamento. Também comparamos os perfis de tráfego com os dados de mobilidade do Google. Nossos resultados indicam que é possível inferir a adesão das populações das cidades às medidas de confinamento usando métricas de tráfego simples, que não comprometem a privacidade dos usuários.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

THESIS TITLE

Mariana Corrêa Ribeiro

January/2023

Advisor: Edmundo Albuquerque de Sousa e Silva

Department: Systems Engineering and Computer Science

In March 2020, the World Health Organization (WHO) declared the COVID-19 pandemic. Lockdown measures were implemented in several countries around the world aiming to contain the spread of the virus. A large part of the world population was forced to use the home Internet network for work, education and other activities, which caused a drastic change in the routine of people and consequently altered the traffic patterns of residential networks. The objective of this dissertation is to analyze the effect of COVID-19 pandemic in the patterns of residential networks. We analyzed the traffic of several residences spread across 15 cities in the state of Rio de Janeiro for several months during the years of 2020, 2021 and 2022. We use residential traffic data provided by a mid side Internet Service Provider (ISP) call Gigalink to compare the traffic before and after the start of quarantine in the 15 cities of the Rio de Janeiro state. We apply a tensor decomposition, clustering and classification to identify residential traffic profiles. we found that 20% of residences changed their daily profiles after confinement. We also compare traffic profiles with mobility data from google. Our results indicate that it is possible to infer the adherence of city populations to confinement measures using simple traffic metrics, which do not compromise user privacy.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A pandemia causada pela COVID-19 é um fenômeno global que provocou drásticas mudanças nos padrões de comportamento de bilhões de pessoas. Os primeiros casos suspeitos de COVID-19 aconteceram na Ásia e foram notificados no final de 2019. Em março de 2020, a Organização Mundial de Saúde (OMS) declarou a pandemia da COVID-19 e como consequência diversos países ao redor do mundo adotaram medidas de isolamento e distanciamento social com o objetivo de conter a propagação do vírus. Enquanto alguns países adotaram políticas rigorosas com restrições de mobilidade até controlarem a disseminação do vírus, outros países adotaram medidas mais brandas, resultando em uma taxa de contaminação relativamente alta, por longos períodos. Devido às medidas de confinamento, grande parte da população mundial foi obrigada a utilizar a rede de internet doméstica para trabalho, educação e outras atividades, o que causou uma mudança radical na rotina das pessoas e o que consequentemente, gerou alterações nos padrões de tráfego das redes residênciais.

Relatórios e artigos têm observado alguns dos efeitos dessas mudanças no tráfego da Internet [2–5]. Os estudos de Feldmann `et al.` [6] e [4] relatam um aumento de 15-20% no tráfego da internet na Europa e nos Estados Unidos da América, uma semana após o decreto da pandemia da COVID-19. Geralmente, o crescimento esperado nessas áreas é de 30% ao ano. Já no Brasil, o Comitê Gestor de Internet do Brasil (CGI) e a Algar telecom [7, 8] relataram um aumento de 30% no tráfego de internet em um perído de duas semanas. Em contrapartida, as redes de ensino ao redor do mundo sofreram quedas significativas em seu volume de tráfego [4, 9]. No Brasil foi reportada uma queda de 27,3% [10].

Com o prolongamento do confinamento, a tecnologia se tornou uma ferramenta fundamental para ajudar os usuários a aliviar o estresse e a ansiedade causados pela pandemia. As platarformas de vídeo tiveram um crescimento astronômico em poucos meses [11, 12] o que afetou a qualidade de serviço percebida pelos usuários e gerou uma redução no desempenho da rede. No Brasil, por exemplo, o tempo de visualização de um vídeo e o envolvimento do público aumentaram em 20% e

25%, respectivamente [13]. A fim de evitar um maior congestionamento na rede, as grandes plataformas online de vídeo (e.g. YouTube, Netflix e Amazon) reduziram a qualidade de transmissão [14].

Inevitavelmente, a pandemia da COVID-19 acelerou a depedência da sociedade em relação à Internet e é de suma importância analisarmos as mudanças e os problemas que ocorreram na rede, a fim de estarmos melhor preparados para eventos futuros. Nesse trabalho estudamos o tráfego das redes residênciais. Um dos desafios deste estudo é obter uma análise sem utilizar dados pessoais identificáveis (PII) dos usuários e nenhuma informação sobre objetos solicitados pelos usuários ou informações contidas no cabeçalho dos pacotes. Usamos apenas a taxa de bits de upload e download coletados nos roteadores residencias. Alguns trabalhos analisam as mudanças do tráfego das classes de aplicativos [4, 13, 15] através da inspeção profunda de pacote e/ou consideram padrões pré determinados para caracterizar o fluxo de tráfego. Esse tipo de análise não é o foco do nosso estudo. Além disso, diferentemente dos outros trabalhos, nosso foco é analisar os "perfis"de tráfego dos usuários e não o tráfego total em determinados pontos da rede.

Utilizamos dados de tráfego de download e upload a cada minuto de roteadores residenciais localizados em 15 cidades no estado do Rio de Janeiro. Os dados são coletados em parceria com um servidor de Internet brasileiro de médio porte. Nosso objetivo é estudar os efeitos da pandemia da COVID-19 no tráfego residencial dessas cidades pela perspectiva dos perfis residênciais.

Utilizamos a metodologia de Streit `et al.` (2019) [16] onde os autores propõem um método para reduzir a dimensionalidade dos dados através de um algoritmo de decomposição de tensores chamado PARAFAC . Isso nos permite extrair padrões de tráfego durante diferentes intervalos de tempo. A partir dos resultados da metodologia, identificamos o tráfego residencial em diferentes perfis de uso diário.

Objetivos. Abordamos as seguintes questões: **(a)** Que mudanças ocorreram nos padrões de tráfego depois que as políticas de confinamento foram adotadas? Essas mudanças foram diferentes entre as 15 cidades? **(b)** Quais mudanças ocorreram no tráfego após o relaxamento das medidas de confinamento? **(c)** Seria possível correlacionar medidas de mobilidade com os perfis de tráfego residencial?

Contribuições. As principais contribuições estão resumidas abaixo.

• *Uma metodologia que permite a análise de eventos.*

Apesar do foco desse trabalho ser o estudo das mudanças ocorridas em decorrência da pandemia da COVID-19 nos padrões de tráfego das redes residências, utilizamos uma metodologia que também pode ser empregada no gerenciamento de rede, como por exemplo estudar o impacto de eventos (e.g feriados, copa do mundo, etc).

• *Perfis residenciais e o impacto do confinamento sobre esses perfis.* Analisamos

um conjunto de dados que contém informações não sensíveis do tráfego da rede de residências (taxa de bits) de 15 cidades no estado do Rio de Janeiro e utilizamos a metodologia proposta em [16] para identificar os perfis diários. Estudamos as mudanças que ocorreram nesses perfis antes e depois do confinamento. Acompanhamos a evolução desses perfis por vários meses durante a pandemia e comparamos a evolução dos perfis de tráfego das 15 cidades onde coletados dados. Para a análise, um grande conjunto de dados de tráfego residencial foi coletado com granularidade de um minuto. (Desconhecemos outro estudo utilizando dados com essa granularidade coletados em milhares de residências.)

• *Períodos onde ocorreram mudanças nos perfis residenciais devido ao confinamento.* Estimamos os períodos de tempo em que ocorreram as mudanças nos perfis residenciais usando uma *abordagem Frequentista*. Essa análise é útil para a gerência e o planejamento da rede visando mitigar efeitos adversos causados por mudanças repentinas nos perfis de tráfego da rede.

• *Perfis residenciais e dados de mobilidade.* Comparamos a evolução dos perfis de tráfego residencial com os dados de mobilidade do Google nas quatro cidades mais populosas dentre as 15 cidades onde coletamos os dados. Encontramos uma forte correlação entre os dois dados, nas quatro cidades analisadas. Portanto, os perfis que obtivemos podem ser usados como uma métrica de aderência ao confinamento nas cidades de forma bastante simples, com pouca informação e sem recorrer à utilização de dados de mobilidade provenientes do celular.

Além dessa introdução, o texto está organizado da seguinte forma: O Capítulo 2 descreve os trabalhos relacionados ao tema e tem como objetivo apresentar os estudos que já foram feitos. O Capítulo 3 detalha os conceitos básicos que são necessários para a compreensão deste trabalho. O Capítulo 4 apresenta a metodologia utilizada em nosso trabalho. Descrevemos os passos necessários para descobrir e caracterizar os perfis de tráfego. Assim como, a coleta e as medições do conjunto de dados usado nesse estudo. O Capítulo 5 detalha a metodologia utilizada e apresenta as análises feitas durante a pandemia da COVID-19 em 15 cidades do estado do Rio de Janeiro através dos perfis de tráfego residênciais obtidos por meio da aplicação da metodologia. Por fim, o Capítulo 6 apresenta as conclusões e trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

O impacto da pandemia da COVID-19 no ecosistema da Internet tem sido estudado e apresentado em trabalhos acadêmicos e relatórios. Trabalhos de Feldmann `et al.` [6] e [4] analisaram as mudanças que ocorreram no tráfego na Europa e nos Estados Unidos devido a pandemia da COVID-19 em diferentes classes de serviços da Internet. Foi constatado que houve um aumento de 15-20% no tráfego logo após o início do confinamento. Como comparação, geralmente, o crescimento anual esperado nessas áreas é de 30%. Embora não utilizamos dados de inspeção profunda de pacote em nossa análise, os resultados corroboram com os nossos.

Em [13] os autores apresentam uma visão geral sobre as mudanças no comportamento da Internet durante a pandemia da COVID-19 na rede de borda do Facebook. Segundo os autores, houve um alto crescimento do tráfego em todo o mundo no início da pandemia. Eles também relataram um aumento de 20% no tráfego de vídeo e 25% na média de horas de vídeo assistidas no Brasil, embora as taxas médias de sessões ruins (vídeos que demoram a iniciar, com travamentos frequentes ou resolução ruim) estivessem entre 5-6%. O trabalho de Candela `et al.` [17] estuda o impacto causado pela pandemia na latência da Internet de cidades europeias. Foram relatadas relevantes alterações na rede durante o período da tarde devido às mudanças nas rotinas dos usuários.

O estudo [9] descreve as mudanças que ocorreram no tráfego da rede do campus da universidade de Politecnico di Torino devido ao aumento das atividades remotas durante a pandemia da COVID-19. A universidade criou uma solução interna. Os resultados mostram que o tráfego de entrada na universidade diminuiu drasticamente, enquanto o tráfego de saída dobrou devido à plataforma de aprendizado online. Na mesma universidade, [18] mostrou que durante a semana a taxa de bits total excedeu 1 Gbit/s devido ao aumento de aulas online durante a pandemia, que representam um terço do tráfego. Já nos finais de semana, observou-se um alto tráfego de 750 Mbit/s. Apesar de não haver aula online nos finais de semana, os autores acreditam que esse aumento no tráfego está relacionado com os download de pales-

tras e materias didáticos feitos pelos alunos. Estes artigos tem pouca semelhança com essa dissertação, mas apresentam os desafios enfrentados pelas universidades durante a pandemia para reduzir o fluxo de pessoas no campus e continuar suas atividades acadêmicas.

Outros relatórios usam dados de tráfego para entender os efeitos das medidas de distanciamento social durante a pandemia da COVID-19. O Google fornece informações sobre o fluxo de pessoas em áreas públicas e privadas usando dados de GPS dos celulares [19]. Já a Apple informa a tendência de mobilidade dos usuários [20]. Há também a startup brasileira In Loco [21] que cruza dados de GPS, Wi-Fi e Bluetooth para rastreamento geográfico. Eles avaliam taxas efetivas de distanciamento social com base nos movimentos espaço-temporais dos consumidores.

Em [22] os autores comparam o comportamento da COVID-19 nas cidades de Manaus e Fortaleza, analisando as medidas legais dos governos locais e os níveis de isolamento. Eles utilizaram os dados de mobilidade do Google [19] para calcular o índice de Permanência Domiliciliar (IPD). O trabalho de [23] analisa dados de uma operadora de rede móvel do Reino Unido. Eles quantificaram as mudanças na mobilidade dos usuários e o impacto no uso e no desempenho da rede de telefonia celular durante a pandemia. Os autores afirmaram que a redução da mobilidade foi mais significativa em áreas urbanas densamente povoadas do que em áreas rurais e que as medidas de confinamento implementadas pelo governo do Reino Unido mudaram drasticamente os padrões de mobilidade dos usuários e o tráfego da rede. Em [24], os dados de Wi-Fi de universidades em Cingapura e nos EUA são analisados durante o confinamento. Os resultados mostram que a mobilidade dos usuários não diminuiu conforme o esperado, embora existam menos pessoas no campus, levando à conclusão de que apenas um confinamento severo diminuiria a mobilidade das pessoas dentro das universidades.

Capítulo 3

Conceitos básicos

Este capítulo apresenta alguns conceitos que servem como base para o entendimento do texto. A Seção 3.1 explica o método PARAFAC, que foi utilizado para a obtenção de fatores que descrevem os dados originais de forma mais compacta. O PARAFAC foi usado na Seção 4.2. A Seção 3.2 descreve o método de clusterização hierárquica utilizado para agrupas amostras semelhantes, a clusterização hierárquica foi usada na Seção 4.2. A Seção 3.3 apresenta conceitos básicos sobre Árvore de Decisão que é um método usado para criar classificadores, esse método também foi usado na Seção 4.2.

3.1 PARAFAC: Método de decomposição de tensores [1]

Um tensor é uma generalização de matrizes para dimensões maiores e ele pode ser tratado como uma matriz multidimensional. Por exemplo, o tensor de ordem 1 é um vetor, o de ordem 2 é uma matriz e tensores de ordem 3 ou superiores são chamados de tensores de ordem superior.

O PARAFAC é um método de decomposição de tensores que tem como objetivo reduzir a dimensionalidade de dados multidimensionais. Ao lidar com dados de alta dimensão, muitas vezes é útil reduzir a dimensionalidade, ou seja, projetar os dados em um subespaço de dimensão inferior, para capturar informações relevantes dos dados, o que consequentemente pode facilitar a interpretação do modelo. Basicamente, o PARAFAC decompõe um tensor em um conjunto de fatores (variáveis latentes) e cargas (*loadings*) que descrevam os dados de forma mais compacta em comparação com o conjunto original. O método é muito utilizado como uma ferramenta exploratória para detecção de estruturas (fatores) subjacentes interpretáveis que expliquem as correlações entre o conjunto de variáveis originais. Muitas vezes, variáveis observadas têm padrões semelhantes de respostas, estando todas associa-

das a um fator que não é diretamente medido. As cargas expressam a relação entre cada variável e cada um dos fatores subjacentes detectados.

Inicialmente o PARAFAC foi construído para ser aplicado em um tensor de terceira ordem, que tem o formato de um cubo e possui 3 dimensões, e posteriormente foi estendido para tensores de ordens superiores [25]. Para simplicar a explicação, a discussão desse capítulo será restringida ao modelo PARAFAC para o tensor de ordem 3, porém a maioria dos resultados é válida para tensores de qualquer ordem. A decomposição do PARAFAC para um tensor $\mathcal{X}$ de ordem 3 ($\mathcal{X} \in R^{I \times J \times K}$) é dada por:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk}, \tag{3.1}$$

onde x_{ijk} é um elemento do tensor $\mathcal{X}$, R é o número de fatores e a_{ir}, b_{jr} e c_{kr} são as cargas (*loadings*) do fator r correspondentes aos modos A, B e C (também chamadas de matrizes de carga). Já e_{ijk} são os residuais, que representam os dados que o modelo não consegue explicar. A Figura 3.1 mostra a representação gráfica da equação 3.1 para um modelo de 2 fatores (R=2).

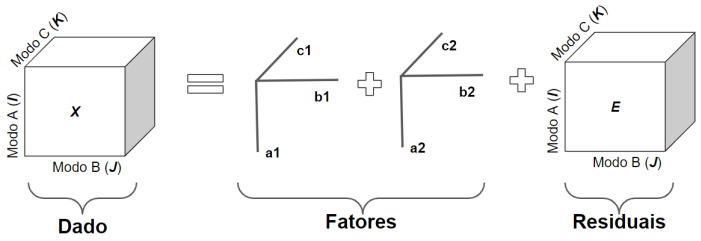


Figura 3.1: Representação gráfica do modelo PARAFAC de 2 fatores para o tensor de ordem 3

O PARAFAC utiliza o algoritmo de mínimos quadrados alternantes (*Alternanting Least Squares* - ALS [26, 27]) para calcular as cargas que minimizam a soma dos quadrados dos resíduos. A Figura 3.2 mostra um pseudo-código do PARAFAC-ALS. O algoritmo inicializa os modos B e C e continuamente fixa 2 modos para estimar o conjunto de parâmetros desconhecidos do terceiro modo (do passo 2 ao passo 4), até satisfazer um dos critérios de convergência ou até que não ocorram mudanças nas estimativas. O símbolo $\odot$ representa o produto de Khatri-Rao, também chamado de produto Kronecker em colunas [26]. Dada as matrizes $B \in R^{J \times R}$ e $C \in R^{K \times R}$, o produto Khatri-Rao é definido como $B \odot C = [b_1 \otimes c_1 \; b_2 \otimes c_2 \; ... \; b_r \otimes c_r]$ e o resultado é uma matriz de tamanho $(JK) \times R$. Como exemplo, se definirmos $J = K = 2$, $R = 3$, $B = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}$ e $C = \begin{pmatrix} g & h & i \\ j & k & l \end{pmatrix}$, então $B \odot C = \begin{pmatrix} ag & bh & ci \\ aj & bk & cl \\ dg & eh & fi \\ dj & ek & fl \end{pmatrix}$. Já $X_{(1)}$, $X_{(2)}$ e $X_{(3)}$ são os tensores desdobrados do primeiro, segundo e terceiro modo que têm tamanho $I \times JK$, $J \times IK$ e $K \times IJ$, respectivamente. Desdobrar um tensor consiste em

transformá-lo em uma matriz. Para faciltar o entendimento, considere um tensor $\mathcal{X}$ de ordem 3 de tamanho $2 \times 2 \times 2$ dividido em duas matrizes 2×2 que chamaremos de X_1 e X_2, como mostrado na Figura 3.3.

PARAFAC - ALS

Entrada: ***X*** e o número de fatores (***R***)
Saída: ***A***,***B*** e ***C***

passo 1: Inicializar **B** e **C**
While (enquanto o critério de convergência não
é satisfeito ou enquanto ocorram mudanças em ***A***,***B*** e ***C***)

passo 2:

$$Z = C \odot B$$
$$A = X_{(1)}Z(Z^TZ)^{-1}$$

passo 3:

$$Z = C \odot A$$
$$B = X_{(2)}Z(Z^TZ)^{-1}$$

passo 4:

$$Z = B \odot A$$
$$C = X_{(3)}Z(Z^TZ)^{-1}$$

Figura 3.2: Pseudo-código do PARAFAC

Agora considere um modelo PARAFAC de um componente para o tensor $\mathcal{X}$. Este modelo também pode ser representado por dois modelos bilineares (Figura 3.4). Para gerar o desdobramento, são concatenadas as matrizes X_1 e X_2 para formar um modelo bilinear que também pode representar o modelo do tensor $\mathcal{X}$, como mostrado na Figura 3.5. Este desdobramento é feito tanto para o modo A quanto para os modos B e C no PARAFAC-ALS.

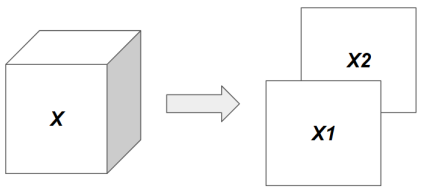


Figura 3.3: Tensor $\mathcal{X}$ de tamanho $2 \times 2 \times 2$ representado por X_1 e X_2, duas matrizes 2×2

Além da matriz de dados X, o algoritmo do PARAFAC recebe como entrada o critério de convergência e o número de fatores (R). Um critério de convergência geralmente utilizado é a análise da diferença do ajuste (*fit*) do modelo entre duas iterações. Se ela for menor que um valor escolhido (e.g., 10^{-6}), o algoritmo termina o processo.

Já a escolha do número de de fatores (R) pode ser feita através do *Screen plot* que analisa o percentual da variância explicada utilizada para medir a discrepância entre o modelo e os dados, ou seja, é a diferença entre variância total do modelo e a variância do erro [28]. A variância explicada é calculada da seguinte maneira:

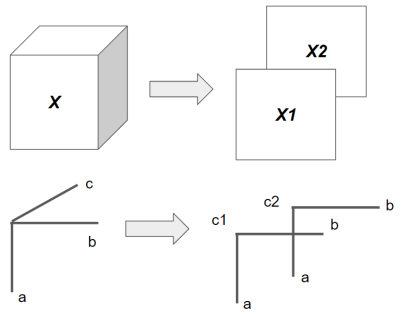


Figura 3.4: Decomposição PARAFAC trilinear que pode ser representada por um modelo de $\mathcal{X}$ ou por dois modelos de matrizes de tamanho 2×2 (X_1 e X_2)

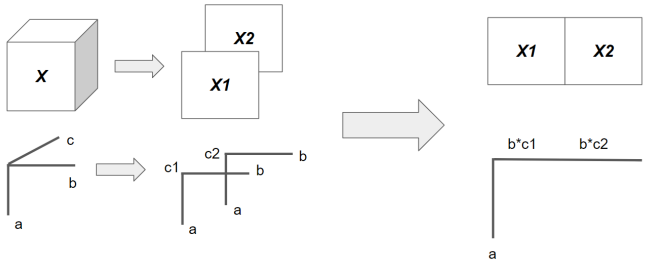


Figura 3.5: Princípio do desdobramento aplicado ao tensor $\mathcal{X}$ e ao seu correspondente modelo PARAFAC de um componente

$$EV = 1 - \frac{\sum E^2}{\sum X^2}, \tag{3.2}$$

onde $\sum E^2$ é a soma dos quadrados do resíduos (E) e $\sum X^2$ representa a soma dos quadrados dos dados (X). Quanto maior o percentual da variância, mais provável que o valor de R analisado esteja correto. Entretanto, grandes valores de R podem levar ao *overfitting*. Outro método utilizado para determinar o número de fatores é o *Split Half Validation* (SV) [29] junto com o *Tucker Congruence Coefficient* (TCC) [30], que tem como vantagem avaliar se a solução é única e generalizável para outro conjunto de dados similar [31].

O *Split Half Validation* divide as amostras de um dos modos (A, B ou C) de forma aleatória em quatro grupos ($G1$, $G2$, $G3$ e $G4$) com o propósito de testar a similaridade entre os modelos PARAFAC dos subconjuntos independentes ($G1$ + $G2$) vs. ($G3$ + $G4$) e ($G1$ + $G3$) vs. ($G2$ + $G4$). Para cada uma das validações de similaridade têm-se dois modelos, m_1 e m_2. A similaridade entre os modelos é medida pelo TCC utilizando o vetor de cargas dos outros 2 modos.

Para simplificar, vamos restringir a explicação a seguir para o modo A. A equação 3.3 é um exemplo do TCC para aos modos B (ϕ_b) e C (ϕ_c), quando as amostras do modo A são utilizadas pelo SV para formar os 4 grupos, onde $1 \leq r \leq R$.

$$\phi_b(r) = \frac{\sum_{j=1}^{J} b_{jr}^{m_1} b_{jr}^{m_2}}{\sqrt{\sum_{j=1}^{J}(b_{jr}^{m_1})^2 \sum_{j=1}^{J}(b_{jr}^{m_2})^2}} \quad \phi_c(r) = \frac{\sum_{k=1}^{K} c_{kr}^{m_1} c_{kr}^{m_2}}{\sqrt{\sum_{k=1}^{K}(c_{kr}^{m_1})^2 \sum_{k=1}^{K}(c_{kr}^{m_2})^2}} \tag{3.3}$$

O valor de ϕ é calculado para cada fator e para cada um dos dois modos (B e C). Como os modelos são gerados a partir de subconjuntos diferentes do modo A, não se calcula ϕ para este modo. O objetivo do método é descobrir se os fatores latentes (fatores inferidos pelo modelo) são similares em $\boldsymbol{b_r}$ e $\boldsymbol{c_r}$ mesmo com amostras diferentes em A. Quanto mais próximo de 1 é o valor de ϕ, mais semelhantes são as cargas dos fatores. O artigo de Lorenzo-Seva e Ten Berge [30] sugerem um valor mínimo de $0,95$ para que os dois vetores possam ser considerados similares. Caso nenhuma das duas comparações apresentem o mesmo padrão de carga (baixa congruência), pode-se concluir que o conjunto de dados é inadequado e não pode ser representado pelo PARAFAC para este tipo de análise, ou que o modelo não apresenta solução única para o critério de convergência e número de fatores (R) escolhido.

3.2 Clusterização hierárquica aglomerativa

A clusterização hierárquica aglomerativa [32, 33] é uma abordagem muito conhecida para definir grupos (*clusters*) de objetos com base na sua similaridade. O algoritmo utiliza uma abordagem *bottom-up*, ou seja, ele começa tratando cada objeto como um *cluster* próprio. Em seguida, pares de *clusters* similares são sucessivamente mesclados (aglomerados) à medida que se sobe na hierarquia, até que todos os objetos pertençam a um único *cluster*. O resultado é um diagrama que mostra a relação hierárquica entre os objetos, chamada de dendrograma. Ele tem o formato de uma árvore hierárquica (como mostrado na Figura 3.6) e sua principal vantagem é permitir a visualização dos *clusters* gerados em cada nível da árvore. Desse modo, não é necessário determinar o número de *clusters* antes de se executar o algoritmo, como acontece no *K-Means* [34], por exemplo, que também é um algoritmo de clusterização.

Um dos métodos mais utilizados para analisar a similaridade entre os *clusters* é chamado de variação mínima de Ward [35]. Basicamente ele considera a análise do *cluster* como um problema de análise de variância ao invés de usar métricas de distância ou medidas de associação. Em cada etapa, o método encontra o par de *clusters* que leva ao menor aumento na variância total dentro do *cluster*, após a mesclagem. Este aumento é a soma das distâncias quadradas entre as amostras e o

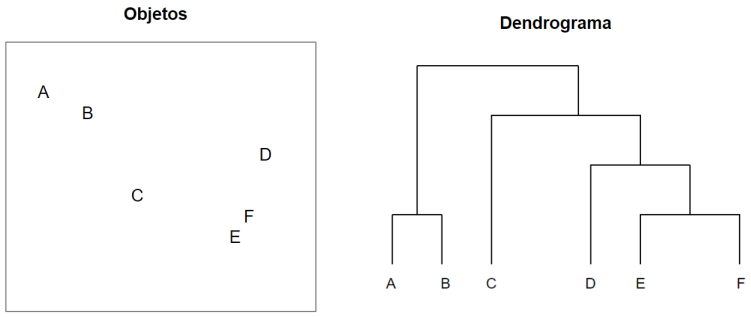


Figura 3.6: Objetos e Dendrograma

centróide do *cluster*, definida como [36]:

$$E_k^2 = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{r=1}^{R} (a_{irk} - \overline{a}_{rk})^2, \tag{3.4}$$

onde a_{irk} é o valor associado à variável r da amostra i que pertencem ao *cluster* k, n_k é o número de amostras no *cluster* k e $\overline{a}_{rk} = \sum_{i=1}^{n_k} \frac{a_{irk}}{n_k}$.

3.3 Árvore de decisão

A árvore de decisão [37, 38] pertence à família de algoritmos de aprendizado supervisionado e ela é muito utilizada para resolver problemas de regressão (predição) e classificação, além de estar entre os algoritmos de inferência mais populares. O objetivo do método é utilizar exemplos para criar um modelo generalizável que consiga analisar um conjunto de dados e classificá-los em classes conhecidas. A entrada consiste em uma matriz do conjuto de dados e o conjunto de classes conhecidas e a saída é o mapeamento desses dados para as classes. Para fazer o mapeamento, o algoritmo cria uma árvore com 3 tipos de nós diferentes, sendo eles:

- O nó raiz que representa toda a população ou amostra
- O nó de decisão (interno) que representa basicamente uma condição (critério de divisão, ou seja, teste) para as possíveis escolhas daquele nível
- O nó folha (terminal) que possui o resultado final, ou seja, a classe que aquele dado pertence.

O algoritmo de treinamento constrói a árvore recursivamente, de cima para baixo, de forma a identificar o caminho (o nó) mais relevante para a classificação das amostras disponíveis em cada um dos nós da árvore. Em resumo, um dado é classificado começando no nó raiz da árvore, em seguida ele passa por um nó de decisão que

calcula algum resultado com base nos critérios de divisão, onde cada resultado possível está associado a uma das subárvores. Quando uma folha é eventualmente encontrada, o dado é classificado em uma classe. O responsavél pela decisão mais apropriada de dividir ou não os nós da árvore de decisão é chamado de índice de Gini [39]. Ele calcula a probabilidade de um elemento aleatoriamente escolhido ser classificado de forma incorreta por um determinado nó interno. O coeficiente de Gini é definido como:

$$G = 1 - \sum_{i=1}^{N} (p_i)^2, \tag{3.5}$$

onde p_i é a probabilidade de um elemento pertencer a uma classe i e N é o número de classes. O valor de Gini está sempre entre 0 e 1, onde 0 denota que todos os elementos pertencem a uma determinada classe (ou a divisão é pura) e 1 indica que os elementos estão distribuídos aleatoriamente em várias classes. O objetivo é gerar nós internos minimamente heterogêneos (impuros) a partir do processo de divisão, ou seja, à medida que o algoritmo percorre a árvore (de forma descendente) o coeficiente de Gini em cada um dos nós resultantes deve alcançar o menor valor possível dentre todas as possibilidades de divisão. Quando apenas existirem amostras de uma única classe em um nó (nó folha), o coeficiente atinge seu valor mínimo (zero).

Capítulo 4

Metodologia

Neste capítulo, explicamos a metodologia proposta em Streit *et al.* [16, 40]. A Figura 4.1 mostra seus principais componentes. A primeira etapa consiste em coletar os dados dos roteadores residênciais. A seção 4.1 detalha como é implementado o ambiente de medição necessário para a coleta das métricas de download e upload das residências. Em seguida, a Seção 4.2 descreve a metodologia utilizada neste trabalho. Primeiramente apresentamos o tensor de três modos que é definido a partir dos dados coletados, em seguida usamos o método de decomposição tensorial PARAFAC para extrair características relevantes dos dados, chamadas de cargas (*loadings*). A próxima etapa consiste em agrupar (*clusterizar*) as cargas (*loadings*) obtidas pela decomposição tensorial do PARAFAC. Por fim, um classificador é treinado, utlizando as cargas (*loadings*) e os grupos obtidos na *clusterização*, com o objetivo de aprender os padrões de tráfego e futuramente classificar novas séries temporais.

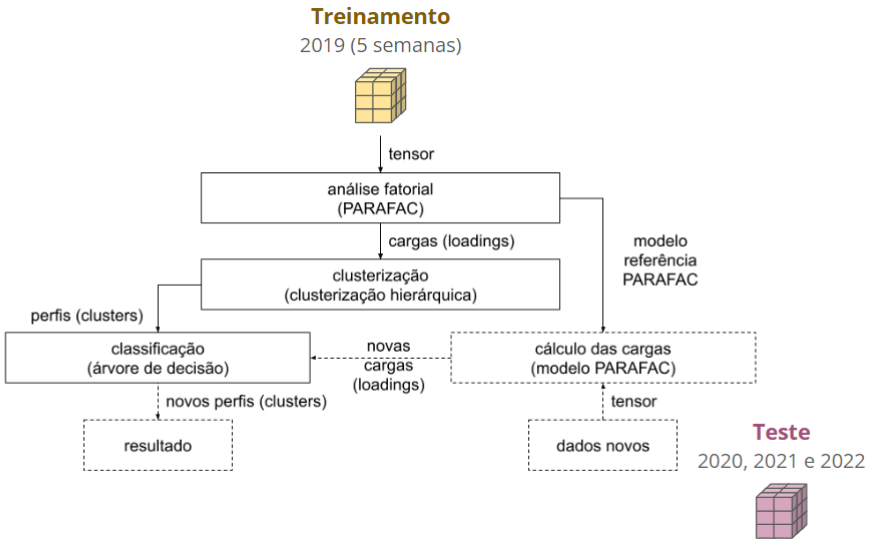


Figura 4.1: Metodologia para obtenção dos perfis residenciais

4.1 Medição e coleta do conjunto de dados

Utilizamos um conjunto de dados coletados em parceria com a Gigalink que é um provedor de serviço de Internet (ISP) de médio porte localizado no estado do Rio de Janeiro. Participamos de um projeto cooperativo de pesquisa que inclui o ISP e a Anlix, uma startup incubada na universidade (UFRJ). Os roteadores residenciais do provedor executam um software baseado no OpenWRT [41], versão do Linux adaptada para sistemas embarcados, desenvolvido pela startup para coletar e enviar informações para o servidor. O software inclui coleta de métricas obtidas através de medições ativas (latência e perda) e passivas (número de bytes e pacotes enviados e recebidos nos roteadores).

Os nossos dados são formados por contadores de bits de download e upload coletados a cada minuto em mais de 5.000 roteadores residenciais espalhados em 15 cidades entre o período de 10 de fevereiro de 2020 a 30 de janeiro de 2022. O conjunto de dados é anônimo e não inclui nenhuma informação sensível do usuário, do provedor ou a localização do roteador, exceto o cidade ao qual o roteador pertence. É importante enfatizar que não é o objetivo deste estudo identificar/classificar os tipos de dispositivos conectados nas redes residênciais.

A quarentena no estado do Rio de Janeiro começou dia 16 de março de 2020. Portanto, o conjunto analisado inclui dados antes e depois do início da quarentena no estado do Rio de Janeiro.

Para referência, a Tabela 4.1 apresenta informações sobre as quatro maiores cidades mais populosas dentre as quinze que possuímos em nosso conjunto de dados, de acordo com o último censo de 2010 [42]. Como não temos acesso aos roteadores de toda a população, analisamos se o número de roteadores do nosso conjunto de dados era grande suficiente para representar o número de residências em cada cidade. Assumimos que a população de roteadores dométicos a que temos acesso é selecionada aleatóriamente em relação ao número total de residências de cada cidade. Em oito das quinze cidades do nosso conjunto de dados, o conjunto de roteadores amostrados em cada cidade foi suficientemente grande para obter uma precisão do tamanho da amostra da população de residências de pelo menos 90% de intervalo de confiança e $\pm 10\%$ de erro. Para calcular o tamanho ideal da amostra que representa o número de residências de cada cidade, utilizamos a fórmula de Cochran definida como:

$$n_0 = \frac{Z^2 * p * q}{e^2}, \tag{4.1}$$

onde n_0 é o tamanho da amostra necessária, Z^2 é o z-valor para o nível de confiança desejado (no nosso caso, 90%), e é a margem de erro, p é a proporção estimada de um atributo que está presente na população (caso o valor de p seja desconhecido,

Tabela 4.1: Informação da população e número de residências das 4 cidades mais populosas

Cidades	População	Densidade (Hab/Km^2)	Número de Residências
Niterói	487.562	3.640,8	169.162
Rio das Ostras	105.676	461,38	34.644
Campos dos Goytacazes	463.731	115,16	142.418
Nova Friburgo	182.082	195,07	63.569

assume-se a máxima variabilidade que é 0.5), e q é igual a $1 - p$ [43].

4.2 Detalhamento da metodologia

• *Tensor*: Em nosso estudo utilizamos dados multidimensional e estamos interessados em perfis residenciais diários. Nosso *dataset* é composto por um conjunto de séries temporais, onde cada série representa o tráfego de download ou de upload de uma determinada residência para um determinado dia.

Definimos um tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ de três modos para representar o conjunto de dados, como mostrado na Figura 4.2. O modo i identifica uma única residência (anônima) em um determinado dia, chamaremos esse par residência-dia de RD. O número de RDs em nosso conjunto de dados é denotado por I. O modo j identifica os minutos em um intervalo de um dia (24h) e k as métricas de interesse. As métricas consideradas são a taxa de bits de download ($k = 0$) e upload ($k = 1$) por minuto. O valor da métrica de interesse k para o RD i durante o minuto j é denotado por x_{ijk} ($0 \leq i < I$, $0 \leq j < 1440$, $k = 0, 1$). Nota: Não seria possível organizar esses dados em uma matriz (tensor de 2 modos), sem perder as relações entre as métricas analisadas (RDs $\times minutos \times$ download ou upload)).

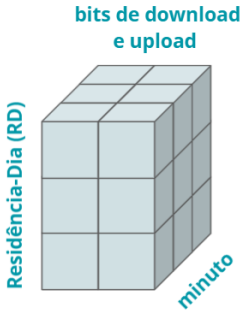


Figura 4.2: Tensor de três modos

A coleta de dados pode possuir intermitência devido a diversas razões podendo gerar séries temporais com várias amostras ausentes. Retiramos do conjunto de dados as séries temporais que têm mais de 10 amostras consecutivas ausentes ou menos de 70% do número máximo de amostras possíveis para o dia (ou seja, menos de 0.7×1440 amostras) com o objetivo de suavizar os problemas causados por um grande número de amostras ausentes em uma série temporal. Depois de remover essas séries temporais, são definidos dois tensores: um para a elaboração do modelo de referência contendo 5 semanas de 2019 (chamaremos de conjunto de treinamento) e outro para a análise dos perfis de tráfego antes e após a quarentena com os novos dados de 2020, 2021 e 2022 (chamaremos de conjunto de teste). Em 2019, 2.873 roteadores residenciais estavam coletando medidas. Esse número aumentou para 6.223 roteadores em 2020 e 5.546 roteadores em 2021. Observe que o número de séries temporais coletadas para cada um dos períodos não é igual ao número total de roteadores multiplicado pelo número de dias de cada período, pois as séries temporais que possuem amostras ausentes segundo os critérios indicados acima, são removidas do conjunto de dados. Além disso, os roteadores podem ser desligados ou começarem a medir durante nosso período de coleta de dados. Logo, os conjunto de roteadores de onde as séries temporais são obtidas, varia com o tempo. O conjunto de dados de referência tem um total de 58.048 séries temporais de residências-dia e uma média diária de 1.658 séries temporais. Já o conjunto de dados a serem analisados possuem 2.306.516 *RDs* e a média diária de 3.613 séries temporais.

• *Análise fatorial (PARAFAC)*:

Conforme foi apresentado na Seção 3.1, um dos principais objetivos da decomposição de tensores é fatorar uma matriz multidimensional (o *tensor*) em um conjunto de fatores (variáveis latentes) e cargas (*loadings*) para descrever os dados de forma condensada.

A decomposição PARAFAC de um tensor tridimensional $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ pode ser obtida através da equação 3.1. Em nossa metodologia, a carga relativa ao RD_i é a_{ir}, b_{jr} é a carga relativa ao minuto j e c_{kr} é a carga relativa à métrica k. Os residuais são denotados por e_{ijk}. O número de fatores (R) foi determinado utilizando os métodos *Split-Half Validation (SV)* [29] junto com o *Tucker Conguence Coefficient (TCC)* [30]. Esta é uma etapa importante pois grandes valores de R podem levar ao *overfitting* [29, 31].

Inicialmente, aplicamos o PARAFAC no conjunto de dados de 2019 para obtenção do modelo de referência (treinamento). Em seguida, utilizamos os mesmo valores de b_{jr} e c_{kr} do modelo de referência (2019) para calcular os novos valores de a_{ir} relativas aos RDs de 2020, 2021 e 2022 (Figura 4.1).

Como os dados de 2020 a 2022 a serem avaliados (volume de tráfego) podem possuir alterações em consequência das medidas de distanciamento durante a pan-

demia, é necessário verificar se o modelo de 2019 é generalizável para representá-los. Com esse propósito, contruímos modelos independentes gerados semanalmente em períodos distintos, incluíndo périodos de férias, entre os anos de 2018 e 2022 e utilizamos o método do *Tucker Congruence Coefficient* (TCC) [30] para verificar a similaridade dos fatores destes modelos PARAFAC, ou seja, para comparar os modelos PARAFAC gerados. Como explicamos na Seção 3.1, o TCC é usado para avaliar se a solução (modelo) é única e generalizável para outro conjunto de dados similar através da similaridade de fatores. Comparamos todos os fatores dos modos j e k, enquanto o modo i variava, de todos os modelos que foram parametrizados (entre 2018 e 2022). Caso algum fator seja diferente entre dois modelos, isso indica que houve uma mudança no padrão dos modos, de uma semana para outra.

Na Figura 4.3, cada um dos eixos X e Y mostra a data de início dos dados coletados ao longo de um período de uma semana. Para cada semana, um modelo PARAFAC foi obtido, perfazendo um total de 123 semanas (modelos) entre julho de 2018 e janeiro de 2022. Na figura, os retângulos em verde representam pares de modelos que apresentam padrões similares e, em vermelho, pares de modelos diferentes (baixa congruência).

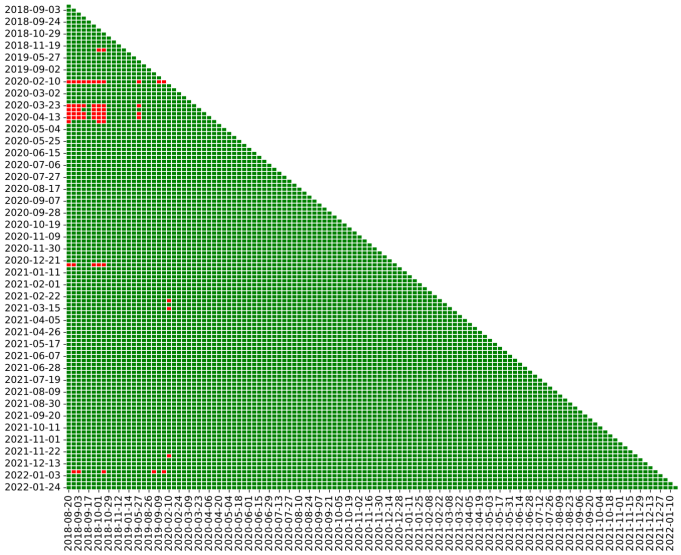


Figura 4.3: Similaridade dos modelos PARAFAC (5 fatores) obtidos para diferentes semanas entre 2018 e 2022.

Por exemplo, o modelo correspondente à semana que começa em *2021-03-15* é semelhante a quase todos os outros modelos semanais obtidos de *2018-08-20* até *2021-01-24*. O único modelo dissemelhante é o referente à semana que começa em *2020-02-10*. Já o modelo associado a semana *2020-12-21* é semelhante a todos os outros modelos semanais. Por outro lado, os modelos de agosto, setembro e outubro de 2018 não são similares aos modelos de março e abril de 2020, muito provavelmente devido a implementação das medidas de confinamento que ocorreram em Março de

2020 no estado do Rio de Janeiro. Além disso, esses 3 meses de 2018 possuem fatores dissemelhantes aos fatores do modelo associado a última semana de 2020, o que pode ter ocorrido devido à temporada de férias natalinas.

No entanto, obtivemos o mesmo padrão para quase todos os 7.503 pares de modelos. Este resultado mostra que o modelo de referência (de *2019-08-19* a *2019-09-22*) representa o tráfego padrão dos roteadores domésticos do ISP durante um longo período de tempo, mais de dois anos neste caso.

• *Clusterização e Classificação* :

Após obter os modos A, B e C para o modelo de referência de 2019, utilizaremos as cargas a_{ir} relativas aos RDs para procurar padrões de tráfego domésticos e consequentemente encontrar os *clusters* (perfis de tráfego residenciais). Cada entrada a_{ir} indica a importância (peso) da residência-dia i no fator latente r. Como a clusterização é uma técnica não supervisionada, ou seja, não possui rótulos prévios para determinar os padrões de tráfego, um desafio é interpretar os fatores latentes e extrair informações importantes dos *clusters*. Usamos o algoritmo de clusterização hierárquica aglomerativa (*agglomerative hierarchical clustering*) para identificar os perfis e encontrar o significado para cada um deles. Este método tem a vantagem de não exigir o número inicial de *clusters* como dado de entrada.

Com o objetivo de facilitar a classificação das séries temporais de 2020 a 2022, utilizamos as cargas (*loadings*) de cada RD e os perfis obtidos na *clusterização* dos dados de 2019 (referência) para treinar a árvore decisão, ou seja, as cargas são as entradas e os perfis (as classes) são as saídas do classificador. Esse treinamento é feito para ensinar o classificador a identificar as cargas e classificá-las nos perfis (*clusters*) encontrados na etapa da clusterização. Essa árvore treinada é então usada para classificar os novos RDs do conjunto de dados de 2020, 2021 e 2022.

Capítulo 5

Resultados

Neste capítulo apresentamos os resultados obtidos através da aplicação da metodologia proposta no Capítulo 4. Detalhamos os passos necessários para a caracterização dos perfis de tráfego residenciais. Na Seção 5.1 descrevemos uma simples análise do tráfego residencial, como motivação para o nosso trabalho, e apresentamos os perfis de tráfego residencias encontrados a partir da metodologia utilizada. Na Seção 5.2 analisamos a evolução dos perfis residenciais durante a pandemia da COVID-19. Na Seção 5.3 apresentamos os instantes de tempo que ocorreram as mudanças nos perfis de tráfego residenciais e a variação na fração de residências associadas a cada perfil residencial. Na Seção 5.4 comparamos o perfis de tráfego residencias com os dados de mobilidade do Google. E por fim, na Seção 5.5 apresentamos os impactos da quarentena nos padrões de tráfego residencias durante os dias de semana e finais de semana.

5.1 Tráfego residencial durante o período pré e pós quarentena

O tráfego de rede é a quantidade de dados que se movem por uma rede, saindo de uma origem e chegando a um destino, em um determinado ponto do tempo. Os dados transferidos são compostos por pacotes, que são as unidades fundamentais de transferência de dados da rede. Monitorar e controlar o tráfego de rede é uma maneira de evitar problemas de congestionamento e manter os serviços sempre funcionando, diminuindo lentidões ou interrupções na transmissão de dados, monitorando onde está ocorrendo perda de pacote e evidenciando quaisquer desvios indesejáveis. Obviamente, o tráfego nos canais de comunicação de uma rede tem um impacto direto na qualidade dos serviços de rede (QoS) e na qualidade de experiência dos usuários (QoE), além de ser essencial para o planejamento e gerenciamento de uma rede. Logo, entender o que acontece no tráfego é fundamental para que se possa

planejar o futuro de forma a manter uma boa QoS e QoE. Pela primeira vez, um evento de grande magnitude impactou de forma drástica o uso da internet. A pandemia da COVID-19 não só mudou o estilo de vida das pessoas, mas também afetou a forma que os usuários utilizam a internet e como resultado, houve um aumento sem precedentes no tráfego da rede, principalmente nas plataformas de videoconferência e de *streaming*. Estudar e entender o que aconteceu é de suma importância para estarmos mais preparados para eventos futuros.

Analisamos o volume de tráfego (download e upload) dos roteadores residenciais que possuem o software de medição que o ISP parceiro coleta as informações. São mais de 5.000 roteadores coletando número de bits de download e upload a cada minuto. Nossos estudos foram feitos durante o período entre março de 2020 a janeiro de 2022.

O objetivo é fazer uma análise simples para verificar se ocorreu alguma mudança no volume de tráfego durante a quarentena, que se iniciou no dia 16 de março de 2020. Os estudos sobre os perfis de tráfego serão feitos após essa primeira análise sobre o volume de tráfego. Com o intuito de diferenciarmos os aplicativos de vídeoconferência, *streaming* e jogos online de outras aplicações como email, *browser* e etc, fizemos uma breve análise em relação às taxas mínimas de upload e download necessárias para participar de uma videochamada ou assistir a um vídeo, uma vez que jogos online precisam de taxas muito maiores. É válido enfatizar que para plataformas de *streaming*, como *Netflix* e *Prime vídeos*, o importante é a taxa de download posto que o usuário está recebendo informações de um servidor. Já para plataformas de videoconferência, jogos online e etc, tanto a taxa de upload quanto a de download são importantes, visto que o usuário está recebendo e enviando informações para um servidor.

De acordo com a Google [44], a taxa mínima de download necessária para assitir um vídeo no Youtube é 70 Kbps. Em outras plataformas de vídeo são requeridas taxas de download maiores [44–47]. Em relação às plataformas de videoconferência, são necessários no mínimo 150 Kbps de upload e de download para chamadas com áudio e vídeo no *teams*, plataforma de videoconferência da *Microsoft* [49]. Já em outras plataformas, as taxas são maiores [48–51]. No entanto, para videoconferências com compartilhamento de tela e vídeo em miniatura, as taxas mínimas necessárias de upload e download variam entre 50-150 Kbps no *Zoom* [51]. Com o intuito de sermos conservadores, definimos em 50 Kbps, tanto para upload quanto para download, as taxas mínimas necessárias para que uma aplicação seja considerada de vídeoconferência ou de vídeo.

A fim de facilitarmos o entendimento da análise a seguir, como exemplo, ilustramos na Figura 5.1 a série temporal da taxa de upload de uma residência durante 1440 minutos (1 dia). A linha vermelha indica uma taxa de 50 Kbps. Podemos ob-

servar que na maior parte do tempo, a residência utiliza menos de 50 Kbps, ou seja, é uma taxa considerada baixa e que geralmente é atribuída a aplicações estáticas como *browser*, email e etc. Uma vez que a residência passa desse valor limite e apresenta picos, muito provavelmente ela está utilizando aplicações que necessitam de uma quantidade maior de dados de upload como videoconferências e jogos online. Observe que a residência passou 8 horas (480 minutos) utilizando aplicações que requerem altas taxas de upload. (Note que para *streaming* teríamos que analisar o gráfico de download).

Nessa análise, queremos verificar se houve um aumento no tráfego das redes residências durante a quarentena devido às medidas de isolamento social. Para isso, comparamos a média do número de minutos que uma residência permanece acima de um valor limite (chamaremos de γ) durante o período anterior e posterior à quarentena (16 de março de 2020). Nós queremos entender se as residências estavam utilizando mais serviços de *streaming*, videoconferências, etc. no período pós-quarentena.

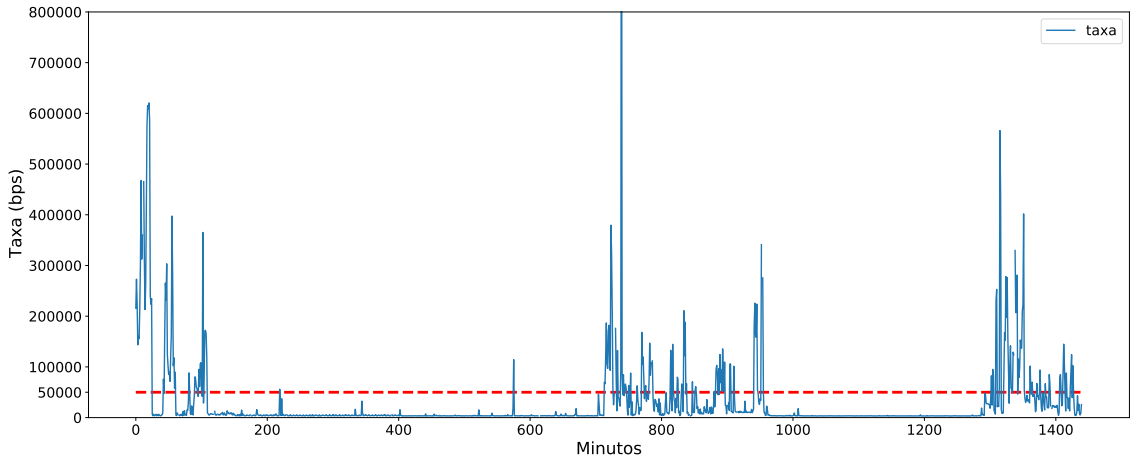


Figura 5.1: Taxa de upload de uma residência durante 1 dia (1440 minutos)

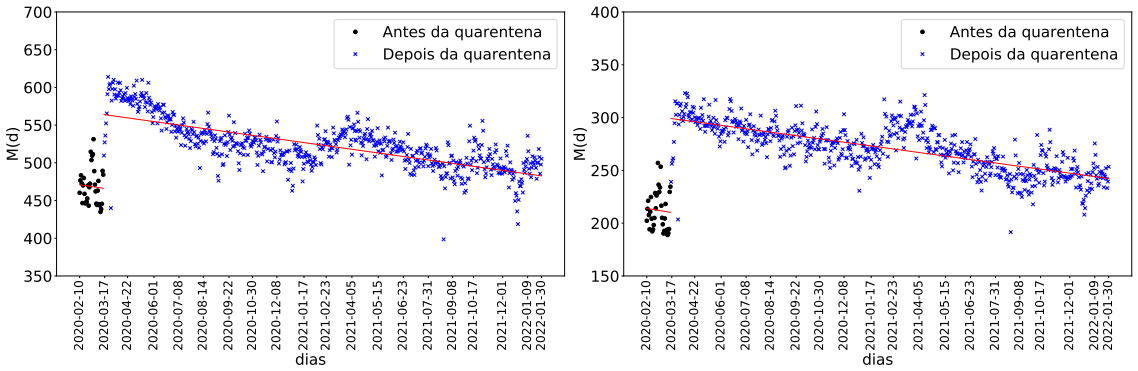
Definimos:

- R_d : número de residências cujas séries temporais de tráfego foram obtidas em um determinado dia d
- *Intervalo* : intervalo de 1 minuto
- *Intervalo Cheio* : intervalo cuja taxa de download/upload está acima de um valor γ limite
- *Intervalo Vazio* : intervalo cuja taxa de download/upload está abaixo de um valor γ limite

- *Período Cheio* : composto por n intervalos cheios que estejam entre 2 intervalos vazios, onde $n \geq 1$.
- *Série Temporal* : taxa de bits de download/upload de uma residência por um dia (1440 minutos). Também chamada de série residência-dia (RD)
- I^{rd} : variável aleatória que representa o número de intervalos cheios de uma série RD, isto é, da residência R no dia D
- L^{rd} : variável aleatória que representa o número de períodos cheios de uma série RD, isto é, da residência R no dia D
- X_i^{rd} : variável aleatória que é igual ao tamanho do i-ésimo período cheio de uma série RD, isto é, da residência R no dia D
- $M^d = \frac{\sum_r I^{rd}}{R^d}$: é a média diária do número de intervalos cheios por residência para cada dia d
- $P^d = \frac{\sum_r L^{rd}}{R^d}$: é a média do número de períodos cheios por residência para cada dia d

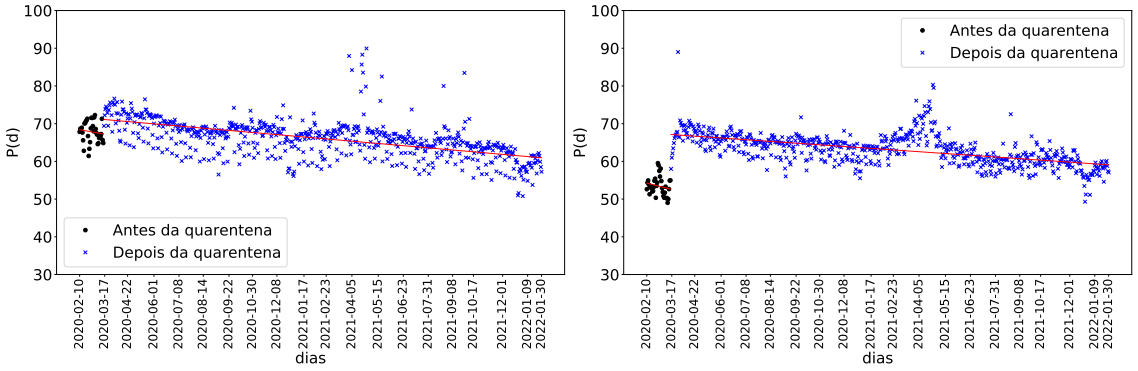
Na Figura 5.2, ilustramos M^d, para $\gamma = 50$ Kbps, antes e depois da quarentena (16 de março de 2020 no estado do Rio de Janeiro) para taxas de download (5.2a) e upload (5.2b). As linhas vermelhas representam a regressão linear dos períodos pré e pós quarentena. A inclinação da reta de M^d é de -0.090 depois da quarentena. Já para download, a inclinação da reta é de -0.129, pós quarentena. É interessante notar que a média do número de intervalos cheios aumentou 33% em relação à taxa de download e 50% em relação à taxa de upload em questão de dias. Em ambas as figuras, também houve um pequeno crescimento no ínicio de 2021 (fevereiro), o que pode ter ocorrido devido à segunda onda da COVID-19 no Brasil. No entanto, se observamos a linha da regressão linear para ambos os períodos, ela indica que ao longo do tempo a média do número de intervalos cheios diminuiu. Já na Figura 5.3, que representa P^d, enquanto a média do número de períodos cheios de upload (5.3b) obteve um aumento de 40%, em relação ao período pré e pós quarentena, para download (5.3a) não houve nenhum crescimento significativo. Além disso, ambos os gráficos apresentaram um decrescimento ao longo do tempo para a média do número de períodos cheios. A inclinação da reta de P^d de upload é de -0.012 no período posterior à quarentena. Em relação à taxa de download, a inclinação da reta é de -0.016 pós quarentena. Esses resultados indicam que houve um crescimento significativo na média do número de intervalos cheios para cada dia d (M^d) das redes residênciais e consequentemente no tempo em que os usuários ficaram conectados e utilizaram aplicações que exigem altas taxas de tráfego, durante o ínicio da pandemia da COVID-19. Porém, se acompanharmos essas médias

ao longo do tempo, em ambos os casos, tanto para P^d quanto para M^d, houve um decrescimento praticamente linear ao longo do tempo. É impotante enfatiza que P^d e M^d representam intervalos cheios e não número de pacotes ou bytes por minuto. Por isso os gráficos de upload e download são semelhantes.



(a) Download (b) Upload

Figura 5.2: Média do número de intervalos cheios para cada dia d, para γ igual a 50 Kbps



(a) Download (b) Upload

Figura 5.3: Média do número de períodos cheios para cada dia d, para γ igual a 50 Kbps

Em seguida, usamos a metodologia da Seção 4 para verificar se o tráfego residencial diário pode ser agrupado em *perfis* distintos e avaliar as possíveis mudanças nesses perfis após a quarentena. É importante frisar que buscamos encontrar características ou padrões (os *perfis*) semelhantes nas séries temporais correspondentes ao tráfego diário residencial de download e upload, simultaneamente, utilizando uma abordagem não supervisionada. Também examinamos a possibilidade de agrupar as residências dentro desses padrões, e encontrar o número adequado de grupos, se for possível achá-los. Além disso, comparamos os perfis antes e depois da quarentena para entender se houve alguma mudança durante esse período. Primeiramente é necessário obter o modelo de referência PARAFAC usando os dados de 2019. O mo-

delo foi validado com cinco fatores ($R = 5$), com variância explicada igual a 97,80% (Figura 5.4).

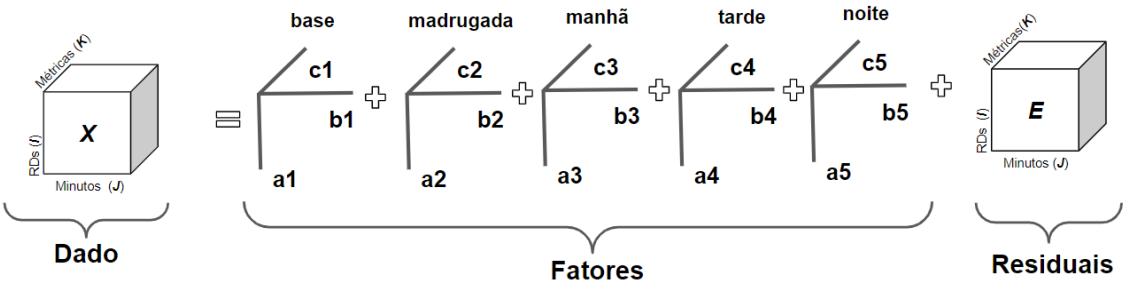


Figura 5.4: PARAFAC aplicado ao conjunto de dados de tráfego

A Figura 5.5 mostra as cargas para cada minuto e para cada um dos cinco fatores em um intervalo de 24 horas. Existem quatro fatores que estão claramente associados ao tráfego intenso em períodos distintos: madrugada, manhã, tarde e noite. Além disso, há um fator base que provavelmente está relacionado as taxas de bits relativamente baixas durante todo o dia.

Ressaltamos que nosso modelo não supervisionado permitiu identificar diferentes características de uso da Internet durante um dia e associar cada RD a essas características. Um determinado perfil é um conjunto de características, o que possibilita a classificação das residências segundo os perfis encontrados. Essa análise não seria possível usando estatísticas simples como análise do volume de tráfego (Figura 5.3).

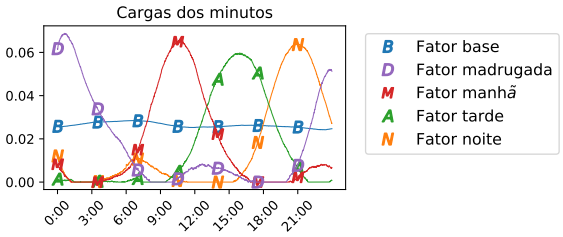
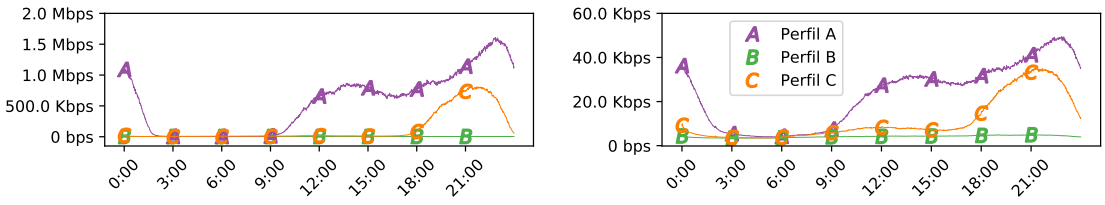


Figura 5.5: Modelo de referência PARAFAC.

Na próxima etapa, executamos o algoritmo de clusterização hierárquica aglomerativa usando como entrada os vetores de carga $\mathbf{a_i} = (a_{i1}, a_{i2}, a_{i3}, a_{i4}, a_{i5})$ de cada RD_i obtidos pelo modelo de referência. Uma vez que as cargas de cada RD_i podem variar em até três ordens de magnitude, aplicamos a normalização Min-Max.

Identificamos três clusters, cada um representando um perfil residencial com características de tráfego distintas. A Figura 5.6 mostra a mediana do tráfego por minuto para cada um dos três perfis residenciais para o período de 24h. O Perfil A representa residências que geram altas taxas de tráfego durante um longo período, de 9h às 2h, enquanto que as residências do Perfil B têm baixas taxas de tráfego durante 24h. Residências do Perfil C geram maior tráfego a noite, entre 18h e 23h59.



(a) Mediana do tráfego de download (b) Mediana do tráfego de upload

Figura 5.6: Mediana do tráfego por minuto para cada perfil residencial.

Conforme mencionado na Seção 4, utilizamos os *clusters* obtidos na clusterização e os *loadings* a_{ir} de cada RD_i do modelo de referência (2019) para treinar uma árvore de decisão (classificador).

Em seguida foram obtidos os loadings a_{ir} para os RDs de 2020, 2021 e 2022 usando o modelo de referência, para serem utilizados como entrada na árvore de decisão já treinada. Por fim, os perfis de tráfego dos RDs de 2020 a 2022 foram obtidos a partir da classificação gerada pela árvore de decisão.

Em seguida, apresentaremos a análise dos perfis no período de 10 de fevereiro de 2020 a 31 de janeiro de 2022 (conjunto de avaliação). Ele inclui amostras antes e depois do início da quarentena no estado do Rio de Janeiro.

5.2 Análise temporal dos perfis residenciais durante a pandemia COVID-19

O estado do Rio de Janeiro emitiu medidas de isolamento social em 16 de março de 2020 (dia do início da quarentena) e todas as cidades do estado foram afetadas. Até a primeira fase das medidas de relaxamento (2 de junho de 2020) no estado, todos os estabelecimentos comerciais e educacionais não essenciais estavam fechados. Embora o governo estadual não tenha restringido a circulação de pessoas nas ruas, ele reduziu os horários de ônibus e metrô e limitou a circulação de carros particulares de passageiros (por exemplo, uber e táxi), o que levou a uma queda drástica nas tendências de mobilidade nas ruas [20]. Em 2021, devido ao aumento de mortes pela COVID-19, o governador do estado do Rio de Janeiro decretou feriado de uma semana de 26 de março a 04 de abril com o objetivo de conter o avanço da COVID-19 nas cidades do estado. Os estabelecimentos foram obrigados a operar em horários reduzidos durante o feriado. Apesar de que no Brasil não houve uma quarentena como em outros países, utilizaremos a palavra quarentena para nos referirmos às medidas de isolamento implementadas. O período analisado será dividido em dois intervalos: antes e depois da quarentena.

A Figura 5.7 mostra a fração de residências por dia que está associada a cada

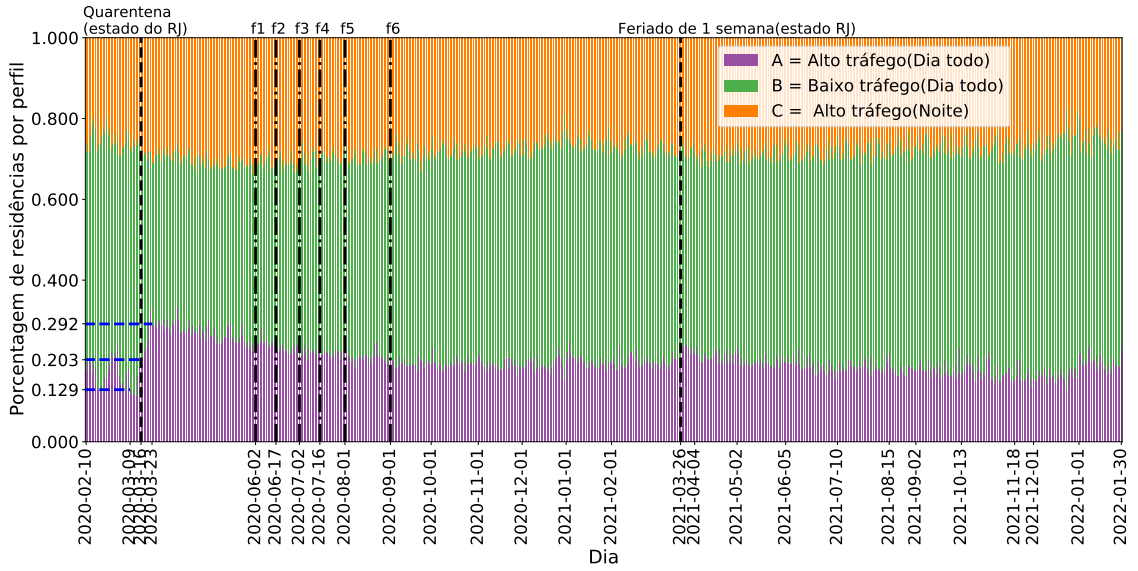


Figura 5.7: Porcentagem de residências em cada perfil por dia.

perfil de tráfego durante o período de 10 de fevereiro de 2020 a 31 de janeiro de 2022 em 15 cidades do estado do Rio de Janeiro. As linhas verticais pretas indicam: a data de início da quarentena (2020), os dias subsequentes do início das fases de relaxamento da quarentena (f1-f6) e o ínício do feriado de uma semana no estado do Rio de Janeiro (2021). As linhas azuis horizontais mostram a fração do Perfil A, uma semana antes (9 de março de 2020) e uma semana depois (23 de março de 2020) de 16 de março de 2020.

Em 9 de março (uma semana antes da quarentena), 12,9% das residências estavam associadas ao Perfil A (que corresponde a um alto uso da Internet, ver Figura 5.6) e, no dia do início da quarentena essa fração passou para 20,3%, um aumento de 57,3%. Apenas uma semana depois (23 de março), o percentual foi ainda maior e atingiu 29,2%, um aumento de 126,3% em relação a 9 de março. No entanto, semanas depois, o Perfil A mostrou uma tendência de queda, até o início de 2021, quando voltou a aumentar lentamente. Esse padrão de decrescimento e crescimento do Perfil A continuou ocorrendo durante 2021 e 2022. Após março de 2021, o percentual de residências associados ao Perfil A diminuiu e voltou a aumentar em janeiro de 2022.

O Perfil C (alto uso de internet à noite) se comportou de maneira semelhante ao Perfil A durante 2020 e 2021. A fração de residências associadas ao perfil C aumentou 14% de 9 a 23 de março de 2020 e, nas semanas seguintes, essa fração diminuiu lentamente, voltando a crescer em 2021. Obviamente, em contrapartida, o Perfil C teve uma tendência de queda entre março de 2021 e janeiro de 2022.

Por outro lado, se compararmos a porcentagem de residências no Perfil B (pouco uso da Internet) no dia do início da quarentena e no dia 23 de março em relação ao

dia 9 de março, houve uma queda de 13% e 33%, respectivamente. Esta tendência terminou no primeiro semestre de 2020 quando a porcentagem do Perfil B começou a aumentar. Entretanto, no início de 2021, o perfil voltou a diminuir novamente. E posteriormente, o Perfil B permaneceu em tendência de queda.

Esses resultados mostram como as residências mudaram de um perfil de tráfego para outro após a quarentena. Podemos também observar na Figura 5.7 que a fração de residências que pertencem ao Perfil A (alto uso de Internet) constantemente diminuiu durante 2020, e pela figura as fases de relaxamento não alteraram a taxa de diminuição e tiveram pouco ou nenhum efeito sobre os perfis de tráfego.

Visando acompanhar a tendência dos perfis durante a quarentena, calculamos a média da fração diária de cada perfil para períodos de 4 semanas antes e depois da quarentena. Chamaremos o perído de quatro semanas pré-quarentena (período entre 23 de fevereiro a 15 de março de 2020) de período de referência (W1-W4) e os vinte quatro períodos de quatro semanas subsequentes de período pós quarentena, que estão referenciados na Figura 5.8 como W5-W8, W9-W12, ..., W97-W100.

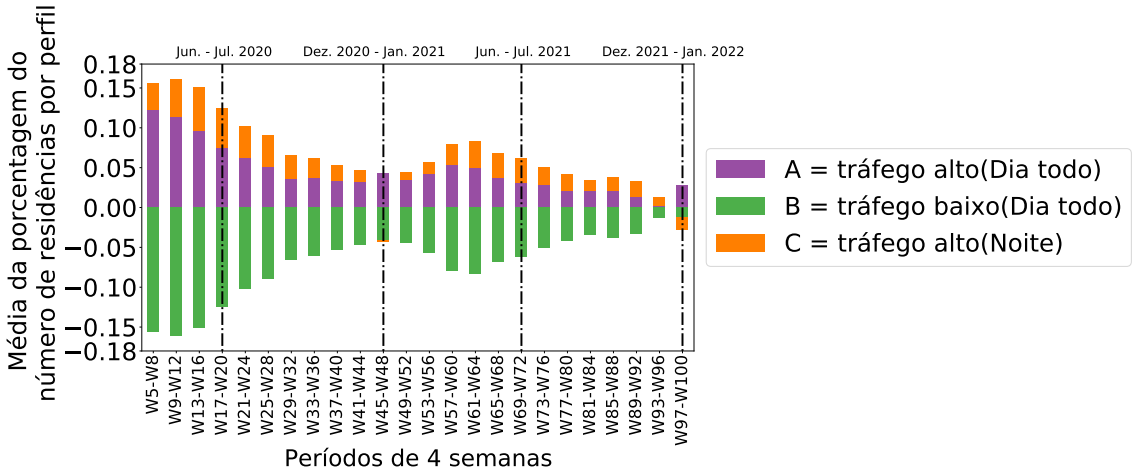


Figura 5.8: Diferença entre a fração de residências associadas a cada perfil antes e depois da quarentena.

O gráfico da Figura 5.8 ilustra a diferença entre a média obtida para cada um dos 24 períodos e o período de referência W1-W4. Note que temos 24 valores diferentes que representam cada um desses períodos pós-quarentena. Observe que, para cada perfil, se a média durante um período for igual a zero, então a média do período pós-quarentena é idêntica à do pré-quarentena. No primeiro período (W5-W8), logo após o início da quarentena, houve um aumento de 12% para o Perfil A (uso intenso da Internet) e um decrescimento de 16% para o Perfil B (pouco uso da Internet). Já o Perfil C (uso intenso de Internet à noite) teve um aumento de 4%. É interessante notar que a fração de residências que mudaram do Perfil B para os Perfis A ou C diminuiu após o início da quarentena. Em 2021, essa tendência mudou e a porcentagem de residências no Perfil B voltou a aumentar. Por exemplo, comparando o

período W5-W8 com W41-W44, pode-se observar que a fração de residências associada ao Perfil A diminuiu de 12% para 3%. O oposto ocorreu quando o perfil B é considerado. A fração aumentou 11%. Esses resultados indicam que as residências estavam lentamente retornando ao seus perfis pré-quarentena. No entanto, no início de 2021 houve uma mudança, quando comparamos W41-W44 a W57-W60, que causou um aumento no Perfil A e C de 2% cada e um decrescimento no Perfil B de 4%, o que coincidiu com o início de uma nova onda da pandemia da COVID-19 no estado do Rio de Janeiro. Essa alteração ocorreu mesmo que nenhuma medida adicional de isolamento tenha sido tomada. Observe que o feriado obrigatório de uma semana no Rio de Janeiro ocorreu durante as semanas de W57-W60 e, portanto, parece que a tendência de alta em ambos os Perfis A e C começou antes desse período, corroborando o comentário feito acima, indicando que as políticas de restrição têm pouco efeito sobre as pessoas, mas talvez as notícias do aumento do número de casos de COVID-19 tenham impactado o comportamento das mesmas. Outra observação interessante é que após março de 2021 (W61-W64), houve novamente uma tendência de queda tanto no Perfil A quanto no C, e uma tendência de alta no Perfil B. A "terceira onda"da COVID-19 no Brasil começou no ínicio de 2022 (W97-W100), e nesse período a tendência dos três perfis mudou novamente.

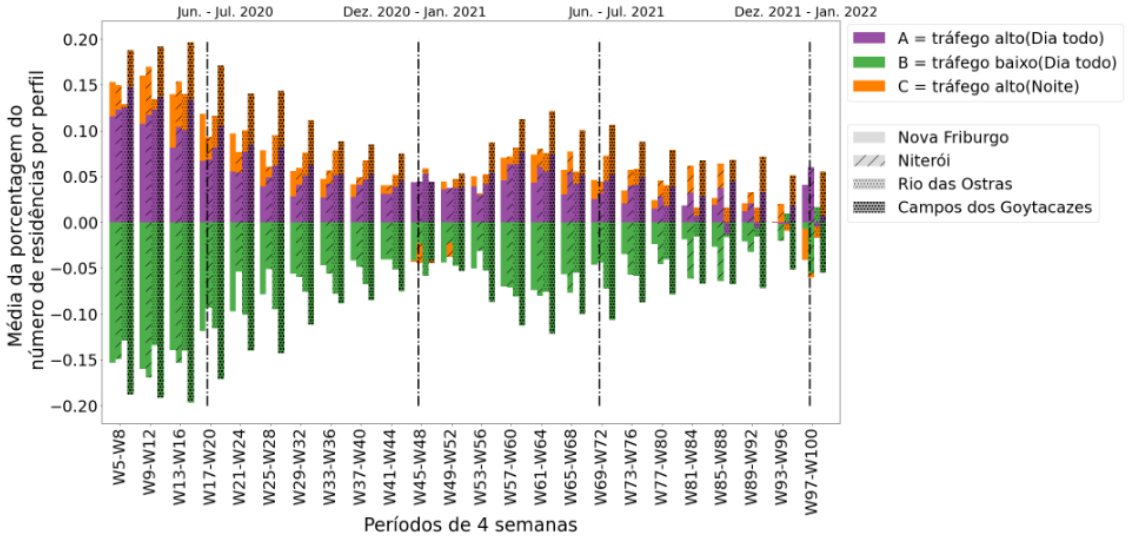


Figura 5.9: Diferença por cidade entre a porcentagem média das residências associadas a cada perfil antes e depois da quarentena

A seguir, comparamos a evolução dos perfis em diferentes cidades. A Figura 5.9 apresenta métricas semelhantes à Figura 5.8, mas ilustramos apenas as quatro cidades mais populosas em nosso conjunto de dados. Vale a pena mencionar que, apesar de serem cidades bastante diferentes, elas seguem um padrão semelhante. Até o início de 2021 (W45-W48), os Perfis A e C apresentavam tendência de queda. Já o Perfil B apresentou tendência de alta. Nas semanas entre W45-W48 e W61-W64,

essas tendências se inverteram, em todas as cidades, e após W61-W64 voltaram a diminuir. Além disso, em W97-W100 as tendências dos perfis mudaram mais uma vez, de forma semelhante à Figura5.8. Curiosamente, nas semanas W93-W97, a cidade de Nova Friburgo praticamente voltou aos níveis pré-quarentena, ou seja, a porcentagem de residências de cada perfil se tornaram próximos aos valores do período de referência (W1-W4). Esses resultados mostram que, embora cada cidade tenha seguido suas regras individuais de restrição, elas não causaram um impacto perceptível no tráfego residencial.

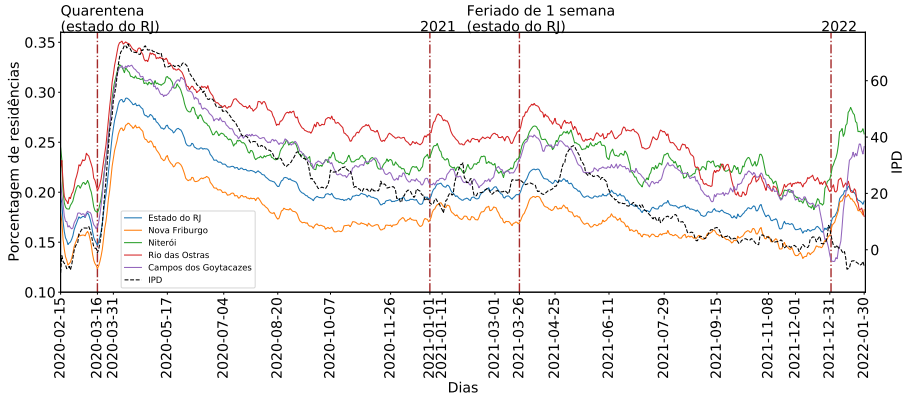
A Figura 5.10 ilustra a média movel da porcentagem diária das residências por perfil e o Índice de permanência domiciliar (IPD) publicado pela Fundação Oswaldo Cruz (Fiocruz [52]). O IPD é um índice relativo que visa comparar a efetividade das medidas de distanciamento social coordenadas pelo poder público em várias cidades. No nosso caso, utilizamos o IPD do estado do Rio de Janeiro. Um fato curioso é que na Figura 5.10a a tendência do Perfil A, em todas as cidades, é muito parecida com a tendência do IPD do estado. O oposto ocorre com o Perfil B (Figura 5.10b), onde as tendências são invertidas. Já o Perfil C (Figura 5.10c), não apresenta uma semelhança tão expresiva como a do Perfil A e B em relação ao IPD. Esses resultados indicam que os percentuais de residencias em cada perfil de tráfego parecem ser um índice da efetividade das medidas de distanciamento social.

5.3 Instantes de mudanças das distribuições de residênciais pelos perfis

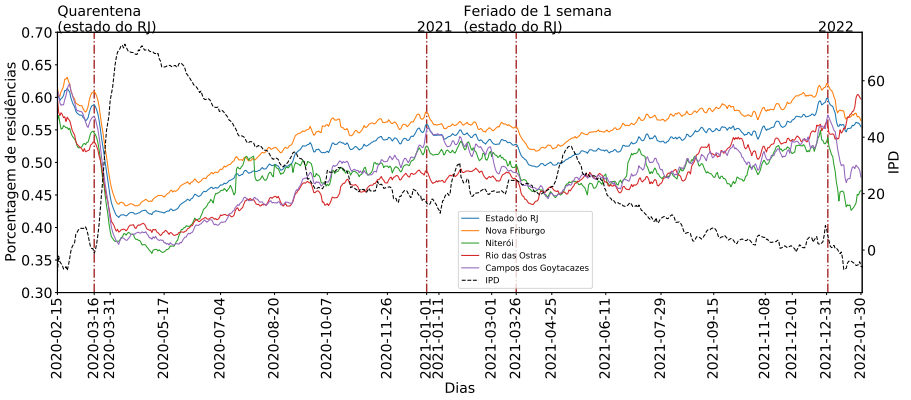
Na seção 5.2 foi possível identificar visualmente, através das Figuras 5.7 e 5.8, quando ocorreram mudanças na porcentagem do número de residências entre os perfis residenciais. No entanto, é importante identificar *formalmente* os instantes de mudanças dos percentuais de residências em cada perfil através de técnicas estatísticas adequadas.

Em termos gerais esse é um problema chamado de *Change Point Detection* (ponto de mudança estatística) em séries temporais. Existem inúmeros métodos para determinar pontos onde ocorrem mudanças das estatísticas em séries temporais [53]. Por exemplo, uma dessas técnicas é baseada em teste de hipótese entre duas alternativas: uma mudança ocorreu em um determinado momento ($H1$) ou nenhuma mudança ocorreu ($H0$) [53, 54]. As suposições que usaremos neste trabalho são feitas para permitir o uso de uma técnica simples, de baixo custo computacional, mas que se mostrou adequada para nosso problema.

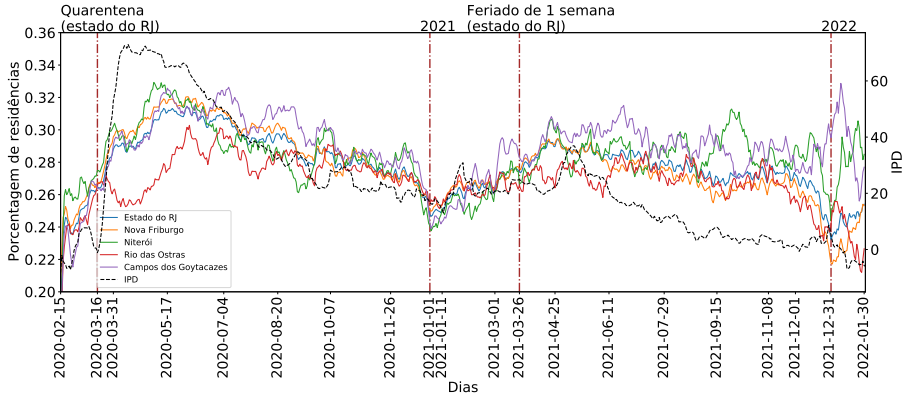
Utilizamos estimadores de máxima verossimilhança (MLE), uma análise frequentista, visando identificar instantes onde ocorreram mudanças no percentual de re-



(a) Perfil A



(b) Perfil B



(c) Perfil C

Figura 5.10: Média movel da porcentagem diária de residências por perfil e IPD

sidências em cada um dos perfis residenciais. Estamos interessados em localizar os instantes mais prováveis de mudança na fração de residências por perfil, podendo ser um ou mais instantes. Um dos nossos objetivos para determinar os possíveis instantes é encontrar as correlações destes com eventos relativos às políticas de isolamento social.

Para construir o modelo estatístico, usamos a suposição de que o número de residências é suficientemente grande para que a probabilidade de uma determinada residência aleatória pertencer a um dado perfil de tráfego não varie a medida que residências sejam amostradas sem reposição. Do total de residências, amostra-se N_i residências no dia i sendo N_i uma variável aleatória que representa o total de amostras (residências) do nosso conjunto de dados de um dia i.

Nosso algorítmo de clusterização determina o número de residências alocadas a cada perfil dentre as N_i residências medindo no dia. Supondo-se que o número total de residências é muito grande, que N_i são sorteadas, e que a probabilidade de se sortear uma residência pertencente a um determinado perfil é constante, então a probabilidade de que seja associado um determinado número de residências a cada perfil de uma amostra do dia i de tamanho N_i tem distribuição multinomial. Podemos então estimar os parâmetros da distribuição a partir das amostras diárias e ainda estimar se houve alguma mudança de distribuição durante um período. Dividimos o período de observação entre 2020-02-10 e 2022-01-09 em intervalos de 4 semanas ($\gamma = 28$ dias). Nosso objetivo é determinar se, para cada intervalo, houve mudança na probabilidade de uma residência aleatória pertencer a um determinado perfil. No total são 25 períodos de 4 semanas cada.

Seja D o número de períodos de 4 semanas ($D = 25$). Iremos supor que, em um intervalo de quatro semanas, existe no máximo um único ponto de mudança de tal probabilidade, para cada perfil. Essa suposição é feita para que se possa usar um modelo simples. Entretando, mais de um ponto de mudança pode ser encontrado pela análise de períodos consecutivos englobando todo intervalo de interesse. Seja τ a variável que indica o instante de mudança durante um período de observação, logo τ pode assumir valores de 1 a $\gamma = 28$.

Definimos p_{ij}, $i = A, B, C$, $j = 1, 2$ como a probabilidade de uma residência amostrada pertencer ao perfil i no primeiro sub-intervalo de um intervalo d ($1 \leq d \leq D$) [1,τ] (no caso de $j = 1$, período antes do instante de mudança) ou no segundo sub-intervalo de [τ, 28] (no caso de $j = 2$) . É importante observar que o modelo pode indicar que não houve nenhuma mudança estatística no intervalo [1,28] do período d. Neste caso, $p_{i1} \approx p_{i2}$ para qualquer valor de τ. Para estimar os valores de τ, utilizamos uma abordagem semelhante á *Markov Chain Monte Carlo* (MCMC), porém muito mais simples. Calculamos os valores de p_{i1} e p_{i2}, para cada valor que τ pode assumir ([1,28]). Em outras palavras, para cada valor de τ, as

probabilidades p_{i1} e p_{i2} são calculadas para maximizar a verosimilhança dos dados do período d (MLE). Seja $\mathcal{L}(\tau)$ a verossimilhança máxima do modelo para um dado valor de τ. (Note que p_{i1} e p_{i2} são os parâmetros que maximizam a verosimilhança $\mathcal{L}(\tau)$ para um determinado valor de τ.). Então,

$$\tau = \underset{x}{\operatorname{argmax}} \{\mathcal{L}(x)\} \tag{5.1}$$

$$\mathcal{L}(\tau) = \underset{\theta=p_{i1},p_{i2}}{\operatorname{argmax}} \{p(D|\theta)\} = \underset{p_{i1},p_{i2}}{\operatorname{argmax}} \{\prod_{t=1}^{\tau} {p_{i1}}^{n_{ti}}(1-p_{i1})^{N_t-n_{ti}} * \prod_{t=\tau+1}^{d} {p_{i2}}^{n_{ti}}(1-p_{i2})^{N_t-n_{ti}}\} \tag{5.2}$$

Calculamos os valores de τ, p_{i1} e p_{i2} utilizando janelas deslizantes de 28 dias com interseção, ou seja, 693 janelas. A primeira janela começa no dia 2020-02-10 e termina 28 dias depois. A segunda janela começa em 2020-02-11 e termina 28 dias depois, e assim por diante.

Onde θ representa os parâmetros do nosso modelo simples para um período d (p_{i1} e p_{i2}), N_t representa o número de residências no dia t e n_{ti} é o número de residênciais no dia t associados ao perfil i. Portanto n_{ti} será o número de amostras aleatórias de residências por perfil i para cada dia t, e N_t é o número de residências sorteadas no dia t.

Então:

$$\begin{aligned} \log(p(D|\theta)) = & \sum_{t=1}^{\tau}[n_{ti} * \log(p_{i1}) + (N_t - n_{ti}) * \log(1-p_{i1})] \\ & + \\ & \sum_{t=\tau+1}^{d}[n_{ti} * \log(p_{i2}) + (N_t - n_{ti}) * \log(1-p_{i2})] \end{aligned} \tag{5.3}$$

Em seguida, estimamos a máxima verossimilhança em relação a $\theta = \{p_{i1}, p_{i2}\}$, para cada valor de τ (períodos de 4 semanas):

$$p_{i1}(\tau) = \frac{\sum_{t=1}^{\tau} n_{ti}}{\sum_{t=1}^{\tau} N_t}$$

$$p_{i2}(\tau) = \frac{\sum_{t=\tau+1}^{d} n_{ti}}{\sum_{t=\tau+1}^{d} N_t}$$

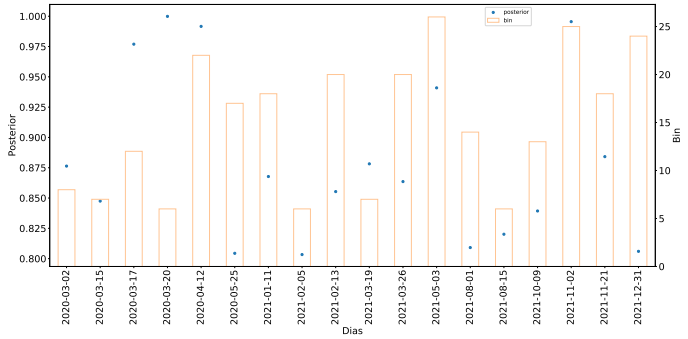
A Figura 5.11 ilustra os instantes de mudança de cada perfil residencial. O eixo x representa os dias que ocorreram os instantes de mudança, os dois eixos de y representam a probabilidade de um instante de mudança ter ocorrido em um determinado dia (posterior) e o número de vezes (janelas deslizantes) que um determinado dia foi

identificado como instante de mudança (Bins). Para gerar essa imagem, aplicamos três filtros, sendo eles: selecionamos apenas os dias que foram identificados como instantes de mudança em pelo menos 5 janelas deslizantes, a probabilidade de um determinado dia ser um instante de mudança tem que ser maior que 80% e que a diferença entre p_{i1} e p_{i2} seja pelo menos 5%. (A escolha de 80% e 5% é impírica, mas serve para ilustrar os casos onde a probabilidade de haver um ponto de mudança é alta.) É válido enfatizar que quanto maior a diferença entre p_{i2} e p_{i1}, maior é a chance de que o instante de mudança tenha ocorrido no determinado período. Na Figura 5.11a podemos observar que no início da quarentena (2020-03-16), o Perfil A apresentou três instantes de mudança (2020-03-15, 2020-03-17 e 2020-03-20) e que o dia mais provável entre eles foi dia 20 de março de 2020. Outro instante importante foi o dia 2020-04-12 que apresentou um aumento significativo no número de casos por COVID-19 no Brasil e isso pode ter impactado a rotina das residências. Além disso, no feriado de 1 semana do estado do RJ, o perfil A também apresentou instantes de mudança, nos dias 2021-03-19 e 2021-03-26. Entretanto em 2021, os períodos mais provavéis de instantes de mudança foram 2021-05-03 (período após a segunda onda da COVID-19) e 2021-11-21 (final do ano). Já no Perfil B (Figura 5.11b), no início da quarentena, o dia mais provável de mudança foi dia 2020-03-16 e no feriado de 1 semana no dia 2021-03-27 (um dia após o término do feriado). Em contrapartida, o Perfil C não apresentou um instante de mundança logo após o início da quarentena e nem no feriado de 1 semana. Isso é um indício que não houve grandes mudanças nesse perfil durante esses dois períodos.

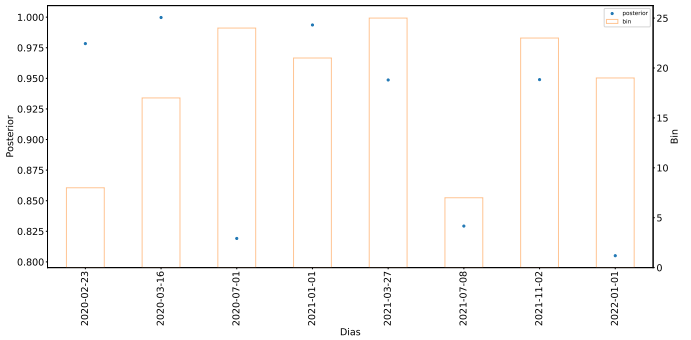
Os resultados obtidos nessa seção indicam que o início da quarentena teve um grande impacto na fração de residências associadas ao Perfil A e B. Uma fração significativa de residências migraram de um perfil de pouco uso da Internet (perfil B) para um perfil de uso intenso da Internet (perfil A) durante o dia todo. Porém, quase nenhuma mudança ocorreu no perfil de uso intenso de internet a noite (perfil C). Essas mesmas mudanças de perfis também foram identificadas no feriado de uma semana do estado do Rio de Janeiro.

5.4 Perfis de tráfego residencial como indicadores de adesão à quarentena

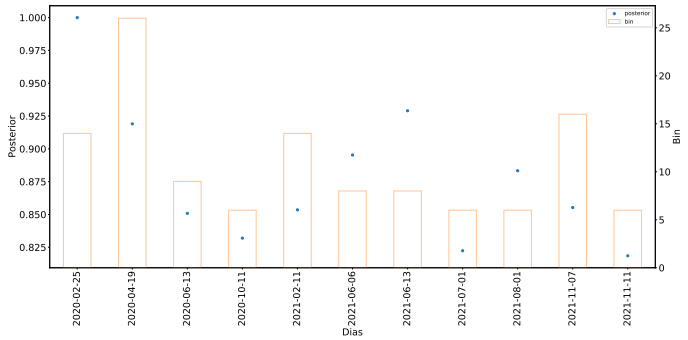
Utilizamos os dados recentemente disponíveis publicamente do Google [19] que mostram as tendências de movimento por região (países, estados e cidades) em diferentes categorias de lugares (por exemplo, locais de trabalho e residenciais) usando dados de GPS dos celulares dos usuários. Eles calculam a mudança relativa entre o número de usuários de cada dia da semana com os dados de referência do mesmo dia,



(a) Perfil A



(b) Perfil B



(c) Perfil C

Figura 5.11: Instantes de mudança por perfil

para cada categorias de lugares e região que charemos de categoria-região. Os dados de referência são o valor médio de um período de 5 semanas (3 de janeiro a 6 de fevereiro de 2020) para cada dia da semana - os dados de referência têm 7 valores diferentes. A mudança relativa é definida como

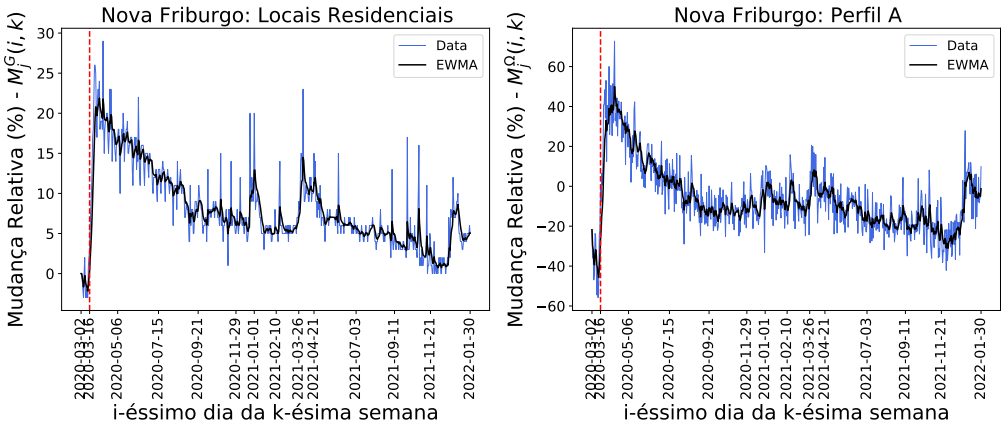
$$M_j^G(i) = \frac{(\mathbf{D}_j^G(i,k) - \mathbf{b}_j^G(i))}{\mathbf{b}_j^G(i)}, \tag{5.4}$$

onde $\mathbf{b}_j^G(i)$ = número de usuários no dia de referência i na categoria-região j e $\mathbf{D}_j^G(i,k)$ = número de usuários no dia i da semana k e da categoria-região j, onde $i = 1, .., 7$. Esses dados são fornecidos diariamente e sua publicação se iniciou dia 7 de fevereiro de 2020. O relatório tem como objetivo fornecer informações sobre as tendências de movimento dos usuários ao longo do tempo por categoria-região durante a pandemia COVID-19. De acordo com o Google, agentes de saúde pública estão usando esses dados para estudar a adesão da população em relação às políticas de combate a COVID-19.

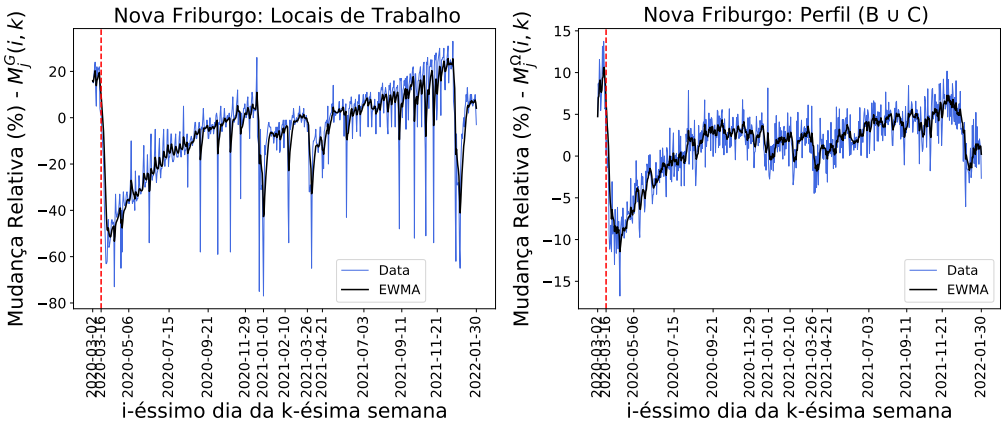
Com o objetivo de entender se as mudanças entre os perfis de tráfego estão relacionadas à adesão da população à quarentena, comparamos os dados do Google com os perfis de tráfego calculando a correlação entre as métricas de mudança relativa de locais residenciais e locais de trabalho (categoria de locais) estimados pelo Google com as métricas de mudança relativa que calculamos a partir dos perfis de tráfego residenciais. As regiões analisadas foram as 4 cidades mais populosas em nosso conjunto de dados e o período de análise foi entre 2 de fevereiro de 2020 a 31 de janeiro de 2022. A equação 5.4 também é usada para calcular a mudança reltiva em nossos dados, substituindo a letra G (dados do Google) pela letra Ω (nossos dados), definimos $\mathbf{D}_j^\Omega(i,k)$ como a fração de residências associadas ao perfil j na semana k para o i-ésimo dia da semana e o vetor $\mathbf{b}_j^\Omega(i)$ como os nossos dados de referência, para $i = 1, ..., 7$ e $j = A, (B \cup C)$, onde cada elemento deste vetor representa a mediana de um período de cinco semanas (16 de janeiro a 19 de fevereiro de 2020) da fração de residências associadas ao perfil j para cada dia da semana i.

A Figura 5.12 mostra a média móvel exponencialmente ponderada (EWMA) de 7 dias da mudança relativa das variáveis estimadas pelo Google e das variáveis que calculamos para a cidade de Nova Friburgo. Comparamos o tempo que as pessoas passam em suas residências com a fração de residências associadas ao Perfil A (tráfego intenso o dia todo), e o tempo que as pessoas passam nos locais de trabalho com a fração de residências associadas ao Perfil B ∪ C (baixo tráfego durante todo o dia e tráfego intenso durante à noite). Todos os gráficos na Figura 5.12 mostram uma mudança repentina após o dia de início da quarentena, 16 de março de 2020, ilustrado na figura com uma linha tracejada vermelha.

Os gráficos indicam também que conforme medidas de relaxamento da quaren-



(a) Dados residenciais e Perfil A



(b) Dados de locais de trabalho e Perfil (B ∪ C)

Figura 5.12: Mudanças nos perfis residenciais e a mobilidade dos usuários na cidade de Nova Friburgo.

tena foram implementadas, as pessoas começaram a sair de casa com mais frequência e voltaram ao trabalho, consequentemente a fração de residências associadas ao uso intenso da Internet diminui. Ressaltamos que, embora as variáveis estimadas pelo Google sejam obtidas a partir de dados de mobilidade das pessoas, e aquelas que calculamos estão relacionadas ao tráfego coletado de roteadores residenciais, elas mostram tendências quase idênticas e essa semelhança é surpreendente.

Devido à semelhança entre o EWMA das mudanças relativas mostradas na Figura 11, usamos o coeficiente de correlação de Pearson para calcular a correlação entre as médias moveis exponencialmente ponderadas (EWMA) de 7 dias da mudança na fração de residências associadas a um determinado perfil e das métricas de mobilidade dos usuários estimadas pelo Google. Os resultados na Tabela 5.1 demonstram uma forte correlação entre as mudanças que ocorreram na fração de residências associadas aos perfis de tráfego e as métricas de mobilidade do Google entre 2 de março de 2020 e 31 janeiro de 2022 para as quatro cidades consideradas no nosso estudo. Testamos a hipótese nula de que o coeficiente de correlação é igual

Tabela 5.1: Correlação entre os perfis de tráfego e as métricas de mobilidade do Google.

Cidade	Perfil A e Locais Residenciais		Perfil B ∪ C e Locais de Trabalho	
	Coeficiente de correlação	p-valor	Coeficiente de correlação	p-valor
Nova Friburgo	0.91	3.53e-156	0.92	5.26e-166
Campos dos Goytacazes	0.92	1.05e-168	0.88	3.69e-131
Rio das Ostras	0.86	5.78e-118	0.81	9.55e-97
Niteroi	0.90	1.49e-149	0.89	5.68e-138

a 0, com base no valor do coeficiente de correlação calculado no nosso estudo, e obtivemos valores muito baixos para o p-valor indicando a rejeição da hipótese nula (ver Tabela 5.1).

Como as fortes correlações mencionadas acima entre os perfis de tráfego e os dados de mobilidade do Google não implicam em uma relação de causa e efeito, fizemos uma análise usando dados de Wi-Fi das residências dos usuários com o objetivo de entender se o aumento do tráfego residencial e, consequentemente, do número de residências no Perfil A (alto uso de Internet), pode ser um indicativo da permanência dos usuários em casa. Essa análise parte do pressuposto de que, se o telefone celular estiver conectado ao roteador residencial, o usuário estará em casa; entendemos que as pessoas não costumam sair de casa sem seus telefones celulares.

Coletamos o número de bits por minuto de download e upload dos telefones celulares conectados à rede Wi-Fi dos usuários durante o período de 1 de fevereiro de 2020 a 31 de janeiro de 2022. É importante mencionar que só coletamos os dados dos telefones celulares quando eles estão conectados à rede doméstica, caso contrário, a medição não existe em nosso banco de dados. Consideramos o número de minutos que os telefones celulares dos usuários ficaram conectados à rede Wi-Fi entre 8:00 e 17:00 (horário padrão de trabalho) para verificar se o número de usuários em casa aumentou após a quarentena.

Definimos:

- R_d como o número de residências em um dia d
- N_{rd} como o número de telefones celulares que estavam conectados no dia d na residência r durante o período de observação (8 às 17 horas), dado que um celular esteja conectado por pelo menos 1 minuto ($N_{rd} > 0$)
- I_{cmrd} como a função indicadora que é igual a 1 se o telefone celular c está conectado no minuto m durante o período de observação (8 às 17 horas) na residência r no dia d e 0 caso contrário.
- $H_{rd} = \frac{\sum_{c=1}^{N_{rd}} \sum_{m=8*60}^{17*60} I_{cmrd}}{N_{rd}}$ como a variável aleatória que representa a média de minutos por telefone celular por residência por dia durante o período de

observação (8 às 17 horas), dado que $N_{rd} > 0$.

- $L_d = \frac{\sum_r H_{rd}}{R_d}$ como a média de minutos de cada telefone celular conectado à residência r no dia d durante o período de observação.

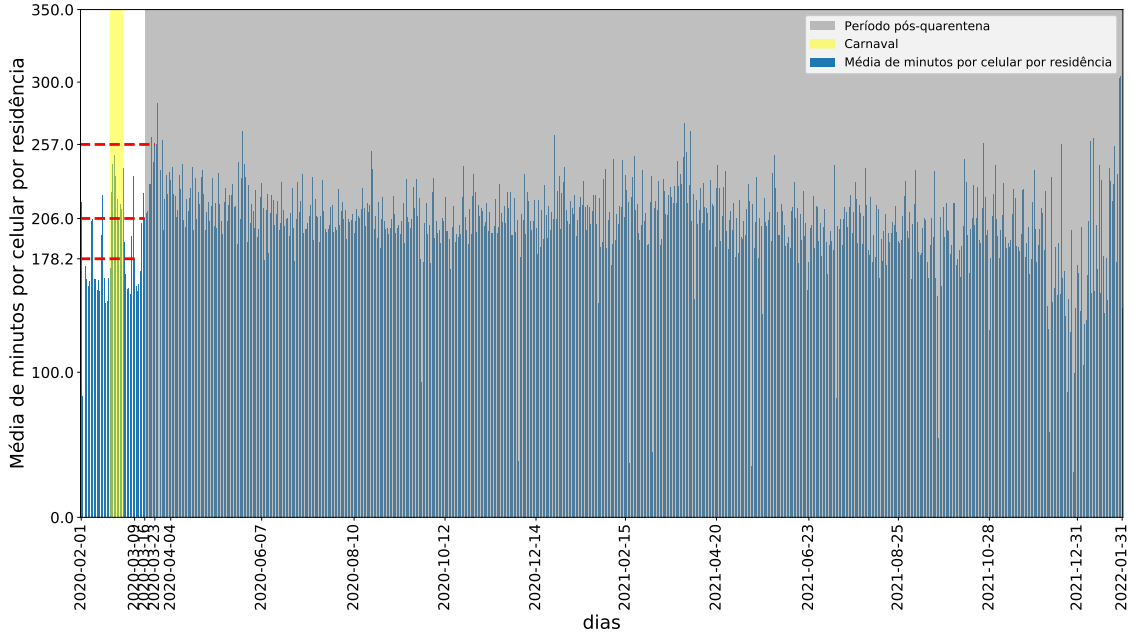
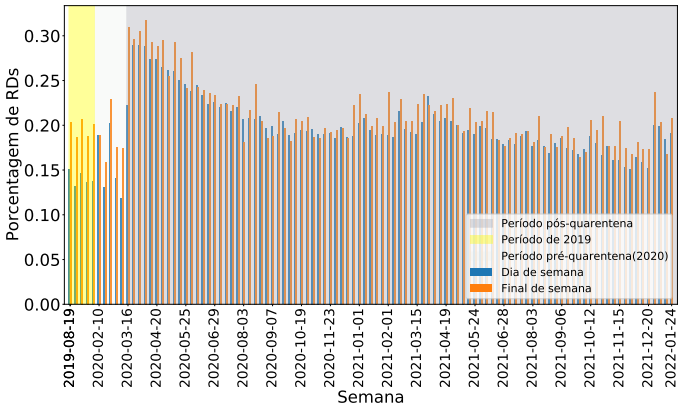


Figura 5.13: Média de minutos entre 8 às 17 horas por telefone celular por residência por dia (L_d)

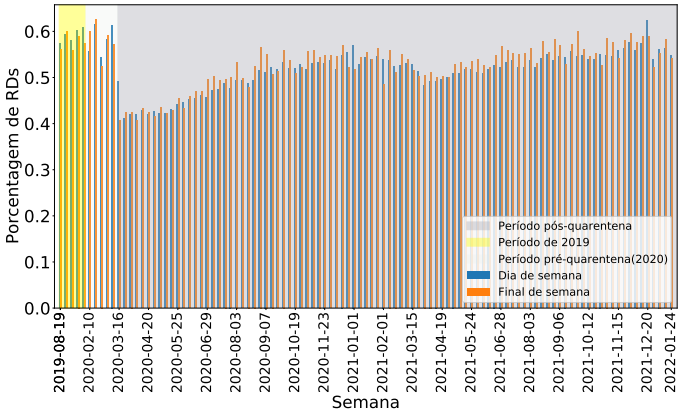
A Figura 5.13 ilustra L_d. As três linhas pontilhadas vermelhas representam o número médio de minutos por telefone celular por residência em 9 de março, 16 de março (dia de quarentena) e 23 de março de 2020. A área amarela corresponde ao carnaval em 2020 e a área cinza caracteriza o período pós-quarentena. O número médio de minutos por celular por residência em 9 de março (uma semana antes da quarentena) e 23 de março (uma semana após a quarentena) teve um aumento de 44% e, após este período permaneceu constante, até o final de 2021 e início de 2022 onde houve uma tendência de decrescimento seguida por uma tendência de crescimento. Além disso, o número médio de minutos por telefone celular por residência no período pré-quarentena é 13% menor do que no período pós-quarentena. Esses resultados mostram que o aumento do tráfego dos telefones celulares nas redes residênciais provavelmente significa que os usuários estarão em casa usando sua rede local, partindo da premissa que os usuários estão sempre com seus celulares. E essa análise corrobora a relação entre o aumento de tráfego residencial e os dados de mobilidade do Google.

5.5 Impactos da quarentena nos padrões de tráfego residenciais

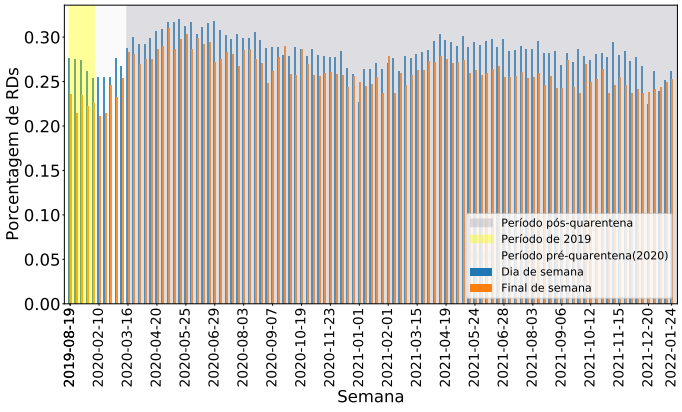
Nesta seção, comparamos a porcentagem de residências por dia de semana e finais de semana para cada perfil de tráfego residencial. Definimos R_{wij} como o número de residências na semana w no período i no perfil j, onde $i = 0$ representa os dias da semana (de segunda a sexta), $i = 1$ representa os dias de final de semana (sábado e domingo) e $j = A, B, C$. Calculamos $I_{wij} = \frac{R_{wij}}{\sum_{i=0}^{1} R_{wij}}$ como a porcentagem de residências por perfil por semana por período, ilustrado na Figura 5.14. As barras azuis indicam as porcentagens de residências por dias da semana ($i = 0$) e as barras laranja as porcentagens por final de semana ($i = 1$). A área amarela representa 5 semanas de 2019 entre 19 de agosto e 6 de setembro, e chamaremos de perído de referência. A área branca caracteriza as semanas de 2020 anteriores à quarentena, de 10 de fevereiro a 15 de março, e a área cinza indica as semanas após a quarentena durante os anos de 2020, 2021 e 2022. Na Figura 5.14a, quando comparamos os dados de 2019 com os dados de 2020, 2021 e 2022 após a quarentena, é interessante notar que as diferenças entre as porcentagens de dias de semana e finais de semana mudaram abruptamente. Essa diferença entre as porcentagens dos períodos (fim de semana e dia da semana) tornou-se menor após o dia da quarentena, o que significa que o número de residências no perfil A teve uma ligeira alteração entre os dias da semana e finais de semana, sendo um indicativo que os usuários estão constantemente em casa usando suas redes residênciais. Por outro lado, os perfis B e C não apresentaram variação percentual significativa entre os dias da semana e final de semana comparando os dados de 2019 com 2020, 2021 e 2022 pós-quarentena. As residências nestes perfis parecem ter mantido os seus hábitos usuais. Concluímos que a diferença da porcentagem de residências caracterizadas pelo perfil A (alto tráfego o dia todo) entre os dias de semana e finais de semana diminuiu após a quarentena e isso pode ser um indicativo que os usuários estavam passando mais tempo dentro de suas casas.



(a) Perfil A



(b) Perfil B



(c) Perfil C

Figura 5.14: Porcentagem de residências por semana (dias de semana e finais de semana)

Capítulo 6

Conclusões

Este estudo analisa o impacto que ocorreu no tráfego residencial de 15 cidades do estado do Rio de Janeiro durante a pandemia da COVID-19. Os dados foram fornecido por um provedor de Internet de médio porte e contém apenas taxas de bits de upload e download coletadas a cada minuto em roteadores domésticos, sem nenhuma informação dos cabeçalhos dos pacotes.

Utilizando a metodologia proposta em Streit *et al.* [16, 40] e identificamos três perfis de tráfego residencial distintos. O modelo de referência foi obtido com os dados de 2019 e ele foi utilizado para classificar os novos dados de 2020, 2021 e 2022. Acompanhamos os perfis identificados pelo modelo de referência por vários meses, incluindo um período anterior e outro posterior à quarentena, que começou dia 16 de março de 2020 no estado do Rio de Janeiro.

Mostramos que um percentual significativo de residências mudaram de perfil após o início da quarentena, passando para o perfil diário caracterizado pelo uso mais intenso da Internet e por um período mais longo. Similarmente, esse comportamento foi observado nas 4 cidades mais populosas de nosso conjunto de dados. Além disso, comparamos os dados disponibilizados pela Fiocruz e os perfis de tráfego residenciais e mostramos que os percentuais de residências em cada perfil de tráfego parecem ser um índice da efetividade das medidas de distanciamento social. Estimamos os dias mais prováveis que ocorreram mudanças nos perfis usando uma análise Frequentista. Os resultados do modelo indicam que houve uma mudança significativa na fração de residências associadas a cada perfil durante o início da quarentena. Esse resultado é importante para a gerência da rede pois permite identificar anomalias e mitigar efeitos adversos causados por mudanças repentinas nos perfis de tráfego residencial.

Encontramos uma forte correlação entre as mudanças nos perfis residenciais obtidos do modelo e os dados de mobilidade fornecidos pela Google. Fizemos um cálculo para obter o número médio de celulares conectados por mais de quatro horas por roteador doméstico durante o período de trabalho (entre 8:00 e 17:00 horas). Esse número médio praticamente dobrou após o início do confinamento. Essa análise

corrobora a relação entre o aumento de tráfego residencial e os dados de mobilidade. Conclui-se que os perfis obtidos e sua variação ao longo do tempo podem servir como parâmetro para estimar a eficácia de medidas de isolamento na população de uma região.

Além disso, comparamos as tendências dos perfis de tráfego durante os dias de semana e finais de semana utilizando os dados de 2019 (como referência), 2020 , 2021 e 2022. Concluímos que as residências associadas ao perfil diário caracterizado pelo uso mais intenso da Internet mudaram suas tendências de finais de semana após o início da quarentena.

Desta forma, os resultados obtidos nesse estudo mostram que os perfis de tráfego residenciais podem ser um parâmetro para estimar as medidas de isolamento de uma população. Outra aplicação importante deste trabalho é ajudar no planejamento e no gerenciamento da rede. Uma maneira de melhorar a utilização da rede, por exemplo, é detectar se as residências estão mudando de perfil. Ou seja, descobrir se a rede está sobrecarregada devido ao aumento do tráfego.

Em relação aos trabalhos futuros, seria interessante: **(a)** Diminuir o intervalo de coleta de dados de minutos para segundos e avaliar o método neste novo cenário; **(b)** Analisar os perfis de tráfego residencias em um período mais amplo com o objetivo de avaliar suas tendências ao longo do tempo; **(c)** Refazer toda análise levando em consideração os bairros de cada cidade e avaliar esse cenário; **(d)** Prever o impacto de feriados e outros eventos que alteram a permanência de usuários nas suas residências nos perfis de tráfego, com o intuito de ajudar no planejamento da rede.

Referências Bibliográficas

[1] HARSHMAN, R. A., LUNDY, M. E. "The PARAFAC model for three-way factor analysis and multidimensional scaling", *Research methods for multimode data analysis*, v. 46, pp. 122–215, 1984.

[2] NOKIA. "Network traffic insights in the time of COVID-19: June 4 update". https://www.nokia.com/blog/network-traffic-insights-in-the-time-of-covid-19-june-4-update/, 2020.

[3] SANDVINE. "The Global Internet Phenomena Report: COVID-19 Spotlight". https://www.sandvine.com/covid-internet-spotlight-report, maio 2020.

[4] FELDMANN, A., GASSER, O., LICHTBLAU, F., et al. "Implications of the COVID-19 Pandemic on the Internet Traffic". In: *Broadband Coverage in Germany; 15th ITG-Symposium*, pp. 1–5, 2021.

[5] FOR ECONOMIC CO-OPERATION, O., DEVELOPMENT. "The Global Internet Phenomena Report: COVID-19 Spotlight". https://www.oecd.org/coronavirus/policy-responses/keeping-the-internet-up-and-running-in-times-of-crisis-4017c4c9/, 2020.

[6] FELDMANN, A., GASSER, O., LICHTBLAU, F., et al. "Implications of the COVID-19 Pandemic on the Internet Traffic". In: *Broadband Coverage in Germany; 15th ITG-Symposium*, pp. 1–5. VDE, 2021.

[7] TELECOM, A. "Ações contra a COVID-19". https://algar2019.blendon.com.br/perspectivas/acoes-contra-a-covid-19/, 2019.

[8] ESTADO DE MINAS, J. "Com uso 30sinal". https://www.em.com.br/app/noticia/economia/2020/04/02/internas_economia,1135023/com-uso-30-maior-da-internet-operadoras-garantem-manter-sinal.shtml, 2020.

[9] FAVALE, T., SORO, F., TREVISAN, M., et al. "Campus traffic and e-Learning during COVID-19 pandemic", *Computer Networks*, v. 176, pp. 107290, 2020.

[10] DE ENSINO E PESQUISA (RNP), R. N. "Impactos da Pandemia do Covid-19 na Rede Nacional de Ensino e Pesquisa RNP". https://www.rnp.br/arquivos/documents/Impactos%20da%20Pandemia%20do%20Covid-19%20na%20Rede%20Nacional%20de%20Ensino%20e%20Pesquisa.pdf?opmYKPybAUY2ZwAShHnzL9wVQyqdr8zV=, 2020.

[11] YOUTUBE. "Watching the Pandemic". https://www.youtube.com/trends/articles/covid-impact/, 2020.

[12] AMAZON. "2020 Letter to Shareholders". https://www.aboutamazon.com/news/company-news/2020-letter-to-shareholders, 2020.

[13] BÖTTGER, T., IBRAHIM, G., VALLIS, B. "How the Internet reacted to Covid-19 – A perspective from Facebook's Edge Network", *ACM Internet Measurement Conference*, 2020.

[14] CNBC. "Facebook joins YouTube and Netflix in reducing video quality in Europe amid virus pandemic". https://www.cnbc.com/2020/03/23/coronavirus-facebook-to-reduce-video-streaming-quality-in-europe.html, 2020.

[15] DA SILVA, C. A. G., FERRARI, A. C. K., OSINSKI, C., et al. "The Behavior of Internet Traffic for Internet Services during COVID-19 Pandemic Scenario", *arXiv preprint arXiv:2105.04083*, 2021.

[16] STREIT, A. G., LEÃO, R. M. M., DE SOUZA, E., et al. "Descobrindo perfis de tráfego de usuários: uma abordagem não supervisionada". In: *Anais Principais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pp. 169–182. SBC, 2019.

[17] CANDELA, M., LUCONI, V., VECCHIO, A. "Impact of the COVID-19 pandemic on the Internet latency: A large-scale study", *Computer Networks*, v. 182, pp. 107495, 2020.

[18] TROPEA, M., DE RANGO, F. "COVID-19 in Italy: current state, impact and ICT-based solutions", *IET Smart Cities*, v. 2, n. 2, pp. 74–81, 2020.

[19] GOOGLE. "COVID-19 Community Mobility Report". https://www.google.com/covid19/mobility, 2020.

[20] MAPS, A. "Mobility trends reports". https://covid19.apple.com/mobility, 2020.

[21] LOCO, I. "Mapa brasileiro da COVID-19". https://mapabrasileirodacovid.inloco.com.br/pt/, 2020.

[22] BARRETO, I. C. D. H. C., COSTA FILHO, R. V., RAMOS, R. F., et al. "Colapso na Saúde em Manaus: o fardo de não aderir às medidas não farmacológicas de redução da transmissão da COVID-19", *Saúde em Debate*, v. 45, pp. 1126–1139, 2021.

[23] LUTU, A., PERINO, D., BAGNULO, M., et al. "A characterization of the covid-19 pandemic impact on a mobile network operator traffic". In: *Proceedings of the ACM Internet Measurement Conference*, pp. 19–33, 2020.

[24] ZAKARIA, C., TRIVEDI, A., CHEE, M., et al. "Analyzing the Impact of Covid-19 Control Policies on Campus Occupancy and Mobility via Passive WiFi Sensing", *arXiv preprint arXiv:2005.12050*, 2020.

[25] PHAN, A.-H., TICHAVSKỲ, P., CICHOCKI, A. "CANDECOMP/PARAFAC decomposition of high-order tensors through tensor reshaping", *IEEE transactions on signal processing*, v. 61, n. 19, pp. 4847–4860, 2013.

[26] ACAR, E., YENER, B. "Unsupervised multiway data analysis: A literature survey", *IEEE transactions on knowledge and data engineering*, v. 21, n. 1, pp. 6–20, 2008.

[27] BRO, R. "PARAFAC. Tutorial and applications", *Chemometrics and intelligent laboratory systems*, v. 38, n. 2, pp. 149–171, 1997.

[28] HOLLAND, S. M. "Principal components analysis (PCA)", *Department of Geology, University of Georgia, Athens, GA*, pp. 30602–2501, 2008.

[29] HARSHMAN, R. A. "How can I know if it's real?" *A catalog of diagnostics for use with three-mode factor analysis and multidimensional scaling*, pp. 566–591, 1984.

[30] LORENZO-SEVA, U., TEN BERGE, J. M. "Tucker's congruence coefficient as a meaningful index of factor similarity", *Methodology*, v. 2, n. 2, pp. 57–64, 2006.

[31] STEDMON, C. A., BRO, R. "Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial", *Limnology and Oceanography: Methods*, v. 6, n. 11, pp. 572–579, 2008.

[32] MURTAGH, F., CONTRERAS, P. "Algorithms for hierarchical clustering: an overview", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 2, n. 1, pp. 86–97, 2012.

[33] JAIN, A. K., MURTY, M. N., FLYNN, P. J. "Data clustering: a review", *ACM computing surveys (CSUR)*, v. 31, n. 3, pp. 264–323, 1999.

[34] HAMERLY, G., ELKAN, C. "Learning the k in k-means", *Advances in neural information processing systems*, v. 16, 2003.

[35] WARD JR, J. H. "Hierarchical grouping to optimize an objective function", *Journal of the American statistical association*, v. 58, n. 301, pp. 236–244, 1963.

[36] LEGENDRE, P. "and Legendre, L., Numerical Ecology, 3rd English ed". 2012.

[37] KOTSIANTIS, S. B. "Decision trees: a recent overview", *Artificial Intelligence Review*, v. 39, n. 4, pp. 261–283, 2013.

[38] PODGORELEC, V., KOKOL, P., STIGLIC, B., et al. "Decision trees: an overview and their use in medicine", *Journal of medical systems*, v. 26, n. 5, pp. 445–463, 2002.

[39] KINGSFORD, C., SALZBERG, S. L. "What are decision trees?" *Nature biotechnology*, v. 26, n. 9, pp. 1011–1013, 2008.

[40] STREIT, A., RIBEIRO, M. C., LEÃO, R. M., et al. "Efeito do confinamento causado pela pandemia Covid-19 nos perfis de tráfego residencial". In: *Anais do XXXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pp. 238–251. SBC, 2021.

[41] OPENWRT. "Projeto OpenWrt". `https://openwrt.org/pt-br/start/`, 2021.

[42] IBGE. "Census of Brazil". 2010. Disponível em: <`https://cidades.ibge.gov.br/brasil/rj/`>.

[43] ISRAEL, G. D. "Determining sample size", 1992.

[44] GOOGLE. "System requirements". `https://support.google.com/youtube/answer/78358?hl=en`, 2022.

[45] VIDEO (AMAZON, P. "Issues with Live Streams on Prime Video". `https://www.primevideo.com/help/ref=atv_hp_nd_cnt?nodeId=GP57SKQ7CB5DRS6F#:~:text=Prime%20Video%20recommends%20a%20minimum,on%20the%20bandwidth%20speed%20available.`, 2022.

[46] NETFLIX. "Internet connection speed recommendations". https://help.netflix.com/en/node/306#:~:text=A%20Standard%20or%20Premium%20Netflix,least%205%20megabits%20per%20second, 2022.

[47] NAM, H., KIM, K.-H., SCHULZRINNE, H. "QoE matters more than QoS: Why people stop watching cat videos". In: *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9. IEEE, 2016.

[48] ADHIKARI, V. K., GUO, Y., HAO, F., et al. "Measurement study of Netflix, Hulu, and a tale of three CDNs", *IEEE/ACM Transactions On Networking*, v. 23, n. 6, pp. 1984–1997, 2014.

[49] MICROSOFT. "Prepare your organization's network for Microsoft Teams". https://learn.microsoft.com/en-us/microsoftteams/prepare-network, 2022.

[50] GOOGLE. "Google Meet hardware requirements". https://support.google.com/a/answer/4541234?hl=en#:~:text=8.8.-,Outbound%20signals%20from%20a%20participant%20in%20all%20situations%20must%20meet,1.5%20mbps%20with%205%20participants, 2020.

[51] ZOOM. "Zoom system requirements: Windows, macOS, Linux". https://support.zoom.us/hc/en-us/articles/201362023-Zoom-system-requirements-Windows-macOS-Linux#:~:text=For%20high%2Dquality%20video%3A%201.0,%2C%204.0Mbps%20(49%20views), 2022.

[52] (FIOCRUZ), F. O. C. "Monitora COVID-19". https://bigdata-covid19.icict.fiocruz.br/, 2020.

[53] AMINIKHANGHAHI, S., COOK, D. J. "A survey of methods for time series change point detection", *Knowledge and information systems*, v. 51, n. 2, pp. 339–367, 2017.

[54] TARTAKOVSKY, A., NIKIFOROV, I., BASSEVILLE, M. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.