Efeito do confinamento causado pela pandemia Covid-19 nos perfis de tráfego residencial

Ananda Streit[1], **Mariana Corrêa Ribeiro** [1], **Rosa M. M. Leão**[1], **Edmundo de Souza e Silva**[1].

[1]Programa de Engenharia de Sistemas e Computação
Universidade Federal do Rio de Janeiro (COPPE/UFRJ) – Rio de Janeiro, RJ – Brasil

{agstreit,marianacr,rosam,edmundo}@land.ufrj.br

Abstract. *Most countries have been implementing lockdown measures to mitigate the spread of the Covid-19 virus. Home officing, distance learning, and home entertainment have become commonplace, affecting Internet home traffic. We analyze the domestic Internet traffic from 13 cities in the state of Rio de Janeiro for several months during 2020 since the social isolation took effect on March 16 2020. We use residential traffic data provided by an Internet Service Provider to compare traffic immediately before and after the start of quarantine in cities in the state of Rio de Janeiro. We use tensor decomposition, clustering and classification to identify distinct residential traffic profiles. We find that 20% of residences changed their daily profiles imediatelly after the lockdown. We also compare traffic profiles against Google's mobility data. Our results indicate that it is possible to assess the adherence of the cities population to confinement measures using very simple traffic metrics, which do not compromise users' privacy.*

Resumo. *A maioria dos países no mundo tem implementado medidas de confinamento para mitigar a propagação do Covid-19. Trabalho remoto, ensino à distância e entretenimento doméstico se tornaram comuns, afetando o tráfego residencial da Internet. Analisamos o tráfego residencial de 13 cidades no estado do Rio de Janeiro durante vários meses do ano de 2020, incluindo o dia de início da quarentena, 16 de março de 2020. Utilizamos informações de tráfego residencial fornecidas por um Provedor de Serviços de Internet para comparar o tráfego imediatamente antes e depois do início da quarentena em cidades do estado do Rio de Janeiro. Utilizamos decomposição tensorial, clusterização e classificação para identificar 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.*

1. Introdução

A pandemia da Covid-19 mudou o comportamento diário das populações ao redor do mundo desde a implementação de políticas de distanciamento social, que estão entre as formas mais eficazes de mitigar a rápida propagação da doença. Diferentes países e cidades adotaram diferentes políticas de confinamento durante períodos distintos de tempo.

Enquanto algumas regiões adotaram políticas rigorosas por várias semanas e suspenderam as restrições de mobilidade após controlarem a disseminação do vírus, outras regiões adotaram medidas menos rigorosas, resultando em uma taxa de contaminação relativamente alta por longos períodos. Até que uma grande parte da população mundial esteja vacinada e imunizada contra a Covid-19, as políticas de confinamento continuarão a ser empregadas, afetando os padrões de mobilidade das pessoas. Como consequência, as populações mudaram suas rotinas diárias e a forma como utilizam a Internet, o que por sua vez, teve um grande impacto no tráfego residencial.

Conforme relatado em Cloudflare (2020), entre o início de janeiro e o final de março de 2020, as principais cidades do mundo tiveram um grande aumento no tráfego da Internet. Os exemplos incluem Nova York (33%), Paris (23%) e Berlim (11%). Um estudo detalhado realizado recentemente por Feldmann *et al.* (2021) observou que o volume de tráfego da Internet aumentou entre 15-20% em uma semana, em um conjunto de pontos de medições localizados na Europa e nos EUA, durante o pico do confinamento. Por outro lado, redes de instituições de educação experimentaram quedas significativas nos volumes de tráfego nos dias de semana após o confinamento [Feldmann et al. 2021, Favale et al. 2020].

Outros trabalhos investigaram a utilização de aplicações específicas. No Brasil, por exemplo, o tempo de visualização de um vídeo e o envolvimento do público aumentaram em 20% e 25%, respectivamente [Böttger et al. 2020]. O relatório da Sandvine (2020), com base em informações de grandes operadoras de rede fixa em todo o mundo, indicou que a demanda por serviços como YouTube, Netflix e Facebook aumentou mais de 12%, 9% e 6%, respectivamente, entre 1º de fevereiro e 19 de abril de 2020. Até mesmo o tráfego de BitTorrent, um aplicativo popular no início deste século, cresceu aproximadamente 3%.

Essas mudanças repentinas e sem precedentes nos padrões de tráfego afetaram a qualidade de serviço percebida por usuários em todo o mundo, e problemas de congestionamento foram relatados [Now 2020]. A fim de evitar maior deterioração da qualidade dos serviços oferecidos, o YouTube, Netflix, Facebook e a Amazon reduziram sua qualidade de transmissão de vídeo [CNBC 2020].

Não há dúvida de que a sociedade aumentou sua dependência da Internet e, como tal, é necessário estudar detalhadamente todas as mudanças que impactaram os serviços de Internet, a fim de nos prepararmos para futuros imprevistos e melhorar a gestão da rede. Nosso trabalho é um dos esforços nesse sentido.

Nosso foco é voltado ao tráfego residencial. A maioria dos trabalhos anteriores avaliam as mudanças de tráfego em diferentes classes de serviços de Internet (por exemplo, [Feldmann et al. 2021, Böttger et al. 2020]). Eles se concentram na inspeção profunda de pacotes e/ou consideram padrões pré-determinados para classificar os fluxos de tráfego. Embora nosso trabalho compartilhe do objetivo geral de caracterizar o tráfego, usamos apenas medidas coletadas em roteadores residenciais. O conjunto de dados que usamos não contém informações sobre os objetos solicitados pelos usuários nem qualquer outra informação contida no cabeçalho dos pacotes. Usamos apenas as taxas de bits de upload e download coletadas nos roteadores residencias.

Através de uma parceria com um servidor de Internet de médio porte, coletamos

dados de tráfego de download e upload a cada minuto de roteadores residenciais localizados em 13 cidades no estado do Rio de Janeiro. Um dos nossos objetivos é estudar os efeitos da pandemia da Covid-19 no tráfego residencial dessas cidades.

Utilizamos o PARAFAC [Bro 1997], um algoritmo de decomposição de tensores eficiente e interpretável, para reduzir a dimensionalidade dos dados. Isso nos permite extrair padrões de tráfego durante diferentes intervalos de tempo. A partir dos resultados do PARAFAC, é realizada a clusterização e a classificação do 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? **(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.

• *Identificação de perfis de tráfego de redes 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 13 cidades no estado do Rio de Janeiro e identificamos perfis diários utilizando técnicas não supervisionadas. Estudamos as mudanças que ocorreram nesses perfis antes e depois do confinamento. Acompanhamos a evolução desses perfis por vários meses durante a pandemia. 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.)

• *Instantes onde ocorreram mudanças nos perfis residenciais devido ao confinamento.* Estimamos os instantes de tempo em que ocorreram as mudanças nos perfis residenciais e quantificamos a variação na fração de residências associadas a cada perfil, usando uma abordagem Bayesiana. 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 13 cidades onde coletamos os dados. Surpreendentemente 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.

Este trabalho está organizado da seguinte forma. A Seção 2 descreve o conjunto de dados que utilizamos e a metodologia proposta. Nossos resultados são apresentados na Seção 3. Os trabalhos relacionados são apresentados na Seção 4 e as nossas principais conclusões estão resumidas na Seção 5.

2. Conjunto de dados e Metodologia

2.1. Conjunto de dados

O conjunto de dados utilizado em nossa análise foi obtido de um provedor de Internet de médio porte, localizado no estado do Rio de Janeiro, através de um projeto cooperativo de pesquisa que também inclui uma startup incubada na universidade (UFRJ). Os roteadores residenciais do provedor executam um software baseado no OpenWRT, desenvolvido pela startup para gerenciamento de rede. O software inclui coleta de métricas obtidas através de medições ativa e passiva.

Neste trabalho usamos a taxa de bits de download e upload coletadas a cada minuto em mais de 5.000 roteadores residenciais espalhados em 13 cidades. (O conjunto de dados é anônimo e não inclui nenhuma informação sensível do usuário, do provedor ou a localização do roteador, exceto o bairro/cidade ao qual o roteador pertence.) Os dados analisados compreendem amostras de 10 de fevereiro a 6 de setembro de 2020 (aproximadamente sete meses). Além disso, usamos um mês de amostras de 2019 (entre 19 de agosto a 22 de setembro de 2019) para elaboração de um modelo de referência. As restrições de mobilidade no estado começaram em 16 de março de 2020. Portanto, o conjunto analisado inclui dados antes e depois do início do confinamento no estado do Rio de Janeiro.

Como os dados de 2020 a serem avaliados (volume de tráfego) poderiam possuir alterações em consequência das medidas de distanciamento durante a pandemia, adotamos um modelo de referência obtido a partir de medições realizadas em 2019 (antes da pandemia). Esse modelo apresenta os mesmos padrões que outros modelos PARAFAC obtidos para diferentes períodos entre 2018 e 2019, incluindo períodos de férias. A similaridade entre os padrões desses modelos é avaliada pelo *Tucker Congruence Coefficient* (TCC) [Lorenzo-Seva and Ten Berge 2006].

Na Figura 1, cada um dos eixos X e Y mostra a data de início de coleta de dados de um período de uma semana. Para cada semana, um modelo PARAFAC foi obtido, perfazendo um total de 21 semanas entre julho de 2018 e setembro de 2019. Na figura, os retângulos em verde representam pares de modelos que apresentam padrões similares e, em vermelho, pares de modelos que não são similares (baixa congruência). Note que obtivemos o mesmo padrão para quase a totalidade dos 210 pares de modelos. Por exemplo, o modelo correspondente à semana iniciada em *2018-07-16* é similar a todos os outros modelos semanais obtidos de *2018-8-20* até a semana iniciada em *2019-09-16*. Somente os modelos das semanas iniciadas em *2019-01-14* e *2019-05-27* não são semelhantes, entre todos os 210 pares de modelos. Esse resultado mostra que o modelo de referência representa o padrão de tráfego da rede doméstica dos usuários desse provedor durante um longo período de tempo, mais de um ano neste caso.

2.2. Tensores

O método de decomposição tensorial PARAFAC é a base da nossa análise para extrair padrões de tráfego de residências. Estamos interessados em perfis residenciais diários, portanto nosso *dataset* é composto de um conjunto de séries temporais, onde cada série representa o tráfego 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. O modo i identifica uma única residência (anônima) em um determinado dia,

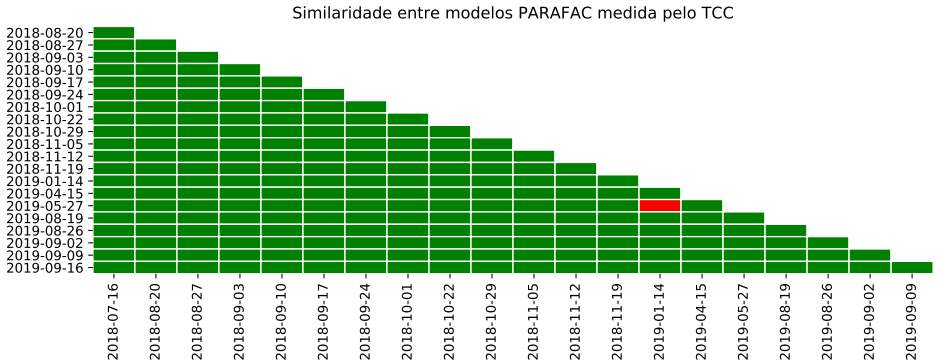


Figura 1. Similaridade dos modelos PARAFAC (5 fatores) obtidos para diferentes semanas entre 2018 e 2019.

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 a taxa de bits de 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} ($0 \leq i < I, 0 \leq j < 1440, k = 0, 1$).

A coleta de dados pode possuir interrupções devido a diversas razões, podendo gerar séries temporais com várias amostras ausentes. A fim de mitigar problemas causados por um grande número de amostras ausentes em uma série temporal, removemos 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). Depois de remover essas séries temporais, são definidos dois tensores: um para elaboração do modelo de referência (período de um mês em 2019) e outro para análise dos perfis de tráfego antes e após o confinamento (período de sete meses durante 2020). Em 2019, 2.873 roteadores residenciais estavam coletando medidas. Esse número aumentou para 5.092 roteadores em 2020. Observe que o número de séries temporais 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 com o número de amostras ausentes segundo o critério indicado acima, são removidas do conjunto de dados. Retirando as séries com o número de amostras ausentes segundo o critério definido acima, temos que o tensor definido para o modelo de referência tem tamanho igual a $58.048 \times 1440 \times 2$ e o tensor definido para a análise dos perfis de tráfego tem tamanho $782.743 \times 1440 \times 2$.

2.3. Metodologia

A Figura 2 mostra os principais componentes da nossa metodologia, proposta em Streit *et al.* (2019). As principais etapas consistem em utilizar o método PARAFAC para extrair as cargas (*loadings*) de cada RD e, em seguida, usar técnicas de clusterização e classificação considerando como *features* as cargas obtidas dos RDs.

Análise fatorial: decomposição tensorial

Um dos principais objetivos da decomposição de tensores é fatorar uma matriz multidimensional (o *tensor*) em conjuntos de fatores (variáveis latentes) e cargas (*loadings*), para descrever os dados de forma condensada.

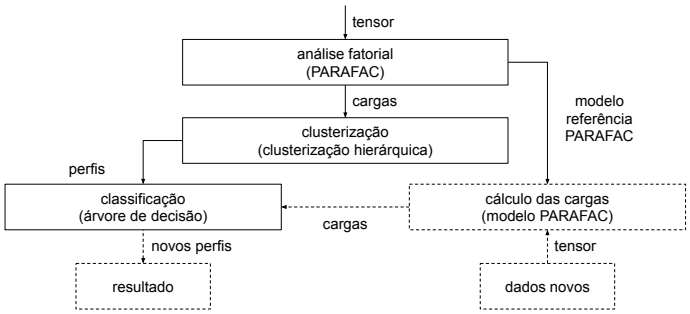


Figura 2. Metodologia para obtenção dos perfis residenciais.

Seja $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ um conjunto de dados tridimensionais (tensor) tendo x_{ijk} como um de seus elementos. Entre as inúmeras técnicas para decompor um tensor de três modos em fatores, o PARAFAC e o Tucker são dois dos métodos mais populares. O PARAFAC garante a obtenção de uma solução única sob certas condições. Já a decomposição de Tucker não garante solução única, exceto sob condições muito mais fortes do que as definidas para o PARAFAC. Por esta razão, escolhemos o método PARAFAC para decomposição tensorial [Harshman and Lundy 1984, Kruskal 1983].

A decomposição PARAFAC de um tensor tridimensional $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ pode ser obtida da seguinte maneira:

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

onde R é o número de fatores e a_{ir}, b_{jr} e c_{kr} são as cargas a serem determinadas para o fator r. a_{ir} é a carga relativa ao RD$_i$, 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 objetivo do método é calcular as cargas que minimizam a soma dos quadrados dos residuais, usando o algoritmo dos Mínimos Quadrados Alternantes (*Alternating Least Squares - ALS*) [Bro 1997].

Usamos dois métodos para determinar o número de fatores R do PARAFAC: *Split-Half Validation (SV)* [Harshman 1984] e *Tucker Conguence Coefficient (TCC)* [Lorenzo-Seva and Ten Berge 2006]. Essa escolha é importante pois valores grandes de R podem levar ao *overfitting* [Stedmon and Bro 2008, Harshman 1984].

Na metodologia, inicialmente, aplicamos o PARAFAC ao conjunto de dados de 2019 para obtenção do modelo de referência. Este modelo será usado para o cálculo das cargas a_{ir} relativas aos RD$_i$ de 2020 (Figura 2).

Clusterização e classificação das séries temporais

Após a obtenção do modelo de referência, a próxima etapa consiste no agrupamento das séries temporais de tráfego residencial com base nas cargas a_{ir} relativas aos RD$_i$ de 2019, ou seja, as cargas obtidas para o modelo de referência. Usamos o algoritmo de clusterização hierárquica aglomerativa (*agglomerative hierarchical clustering*) que tem a vantagem de não exigir o número inicial de *clusters* como dado de entrada. A partir dos *clusters* obtidos, uma árvore decisão é treinada. Essa árvore é então usada para classificar os RDs do conjunto de dados de 2020.

3. Resultados

3.1. Analise do tráfego residencial em período pré e pós quarentena

Inicialmente, analisamos o volume de tráfego residencial (download e upload) coletado por todos os roteadores residenciais que possuem o software de medição (mais de 5.000 roteadores) para os meses de março a setembro de 2020. O objetivo dessa análise simples é observar apenas se houve alguma mudança no volume total de tráfego por minuto das residências devido a quarentena. A Figura 3 mostra a *mediana do tráfego* de download e upload por minuto entre 9 e 23 de março de 2020. (Nota: a quarentena foi iniciada em 16 de março de 2020 no estado do Rio de Janeiro). Comparando o pico da *mediana do tráfego* uma semana antes (9 a 15 de março) com o pico da *mediana* na semana imediatamente seguinte ao início da quarentena (16 a 23 de março), observamos que houve um aumento de 50% e 300% nas medianas, respectivamente.

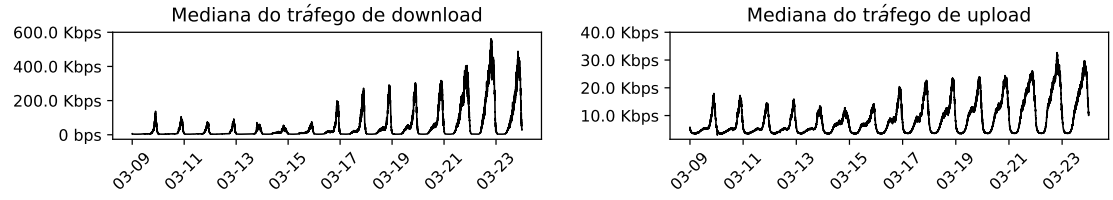


Figura 3. Mediana de download e upload do tráfego residencial por minuto.

Em seguida, usamos a metodologia da Seção 2.3 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 observar que a denominação *perfil de tráfego* por nós usada não é uma simples média ou mediana diária ou outra estatística simples. O que se busca é achar de forma não supervisionada se há características ou padrões (os *perfis*) semelhantes nas séries temporais correspondentes ao tráfego diário residencial de download e upload, simultaneamente. E ainda se é possível agrupar as residências dentro desses padrões, e o número adequado de grupos, se for possível achá-los. Além disso o que teria acontecido com os perfis após a quarentena. O primeiro passo é a obtenção do modelo de referência PARAFAC usando os dados de um período de 2019. O modelo foi validado com cinco fatores ($R = 5$), com variância explicada igual a 97,80%. A Figura 4 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.

Cabe ressaltar 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 a mediana do tráfego por dia (Figura 3).

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.

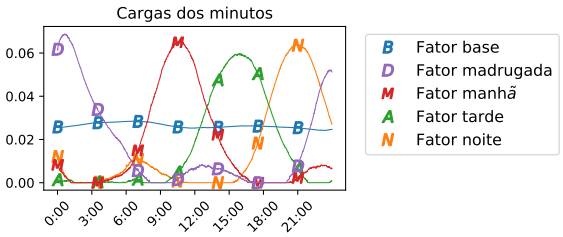
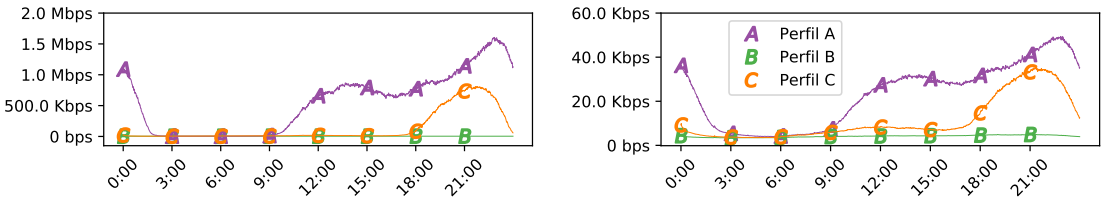


Figura 4. Modelo de referência PARAFAC.

Identificamos três clusters, cada um representando um perfil residencial com características de tráfego distintas. A Figura 5 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. Mediana do tráfego por minuto para cada perfil residencial.

Conforme mencionado na Seção 2.3, o resultado do algoritmo de clusterização é usado para treinar uma árvore de decisão. Em seguida são obtidos os loadings a_{ir} para os RDs de 2020 usando o modelo de referência. Por fim, os perfis de tráfego dos RDs de 2020 são obtidos a partir da classificação gerada pela árvore de decisão.

3.2. Evolução 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. Dividimos o período analisado em dois intervalos: pré-quarentena e pós-quarentena.

A Figura 6 mostra a fração de residências por dia que está associada a cada perfil para o período de 10 de fevereiro a 6 de setembro de 2020 para 13 cidades do Rio de Janeiro. As linhas verticais pretas indicam a data de início da quarentena e os dias subsequentes do início das fases de relaxamento da quarentena. As linhas azuis horizontais mostram a fração do Perfil A, uma semana antes e uma semana depois de 16 de março de 2020.

Em 9 de março, 14,5% das residências estavam associadas ao Perfil A e, no dia de início da quarentena essa fração passou para 23%, um aumento de 58%. Apenas uma semana depois (23 de março), o percentual foi ainda maior e atingiu 32%, um aumento de 120% em relação a 9 de março. No entanto, semanas depois, o Perfil A mostrou uma tendência de queda.

Por outro lado, se compararmos a porcentagem de residências no Perfil B no dia de início da quarentena e no dia 23 de março em relação ao dia 9 de março, houve uma queda de 15% e 35%, respectivamente. O Perfil C teve um aumento de 10% entre 9 e 23 de março, permanecendo aproximadamente constante nas semanas seguintes.

Esses resultados mostram como as residências mudaram de um perfil de tráfego para outro após a quarentena. Podemos também observar na Figura 6 que a fração de residências que pertencem ao Perfil A (alto uso de Internet) está constantemente diminuindo, e pela figura as fases de relaxamento não alteraram a taxa de diminuição.

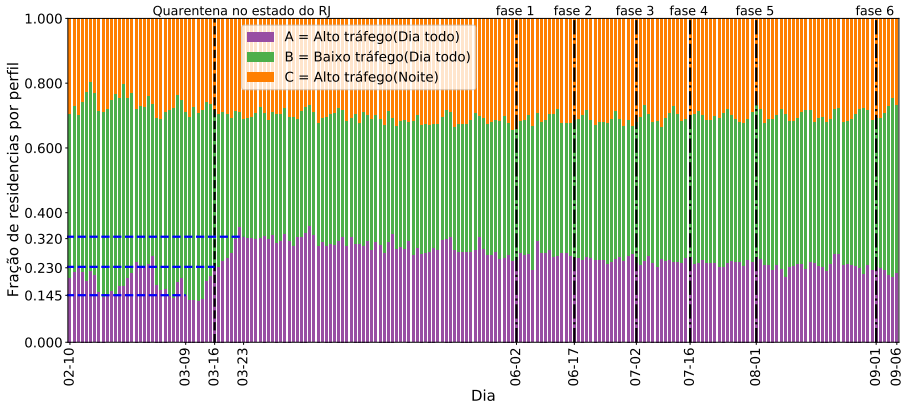


Figura 6. Porcentagem de residências em cada perfil por dia.

Visando acompanhar a evolução dos perfis durante a quarentena, calculamos a média da fração diária de cada perfil para o período de 4 semanas antes da quarentena: de 23 de fevereiro a 15 de março de 2020. Chamamos esse período de *período de referência* (W1-W4). Em seguida calculamos a média da fração diária de cada perfil para seis períodos consecutivos de 4 semanas pós-quarentena. O primeiro período inicia em 16 de março e o último período termina em 30 de agosto de 2020. Esses períodos estão referenciados na Figura 7 como W5-W8, W9-W12, ..., W25-W28.

O gráfico da Figura 7 ilustra a diferença entre a média obtida para cada um dos seis períodos e o período de referência W1-W4. Pode-se notar que para o primeiro período (W5-W8), logo após o início da quarentena, ouve um aumento de 13% para o Perfil A (uso intenso da Internet) e uma diminuição de 16% para o Perfil B (pouco uso da Internet). É interessante notar que a fração de residências que mudou do Perfil B para os Perfis A ou C após o início da quarentena, diminuiu ao longo dos meses. Por exemplo, comparando-se o período de W5-W8 com W25-W28, pode-se observar que a fração de residências associadas ao Perfil A diminuiu de 13% para 5%. Já para o Perfil B, ocorreu o oposto, a fração aumentou em 6%. Esses resultados indicam que as residências estão lentamente voltando aos seus perfis de pré-quarentena.

3.3. Períodos de mudança do perfil residencial

Na seção 3.2 foi possível identificar visualmente, através das Figuras 6 e 7, quando ocorreram mudanças nos perfis residenciais. No entanto, é importante identificar *automaticamente* quando houve mudanças nesses perfis. Utilizamos uma abordagem bayesiana [Smith 1975] e métodos de *Markov Chain Monte Carlo* (MCMC) visando identificar instantes de mudança dos perfis residenciais e quantificar a fração de residências por perfil

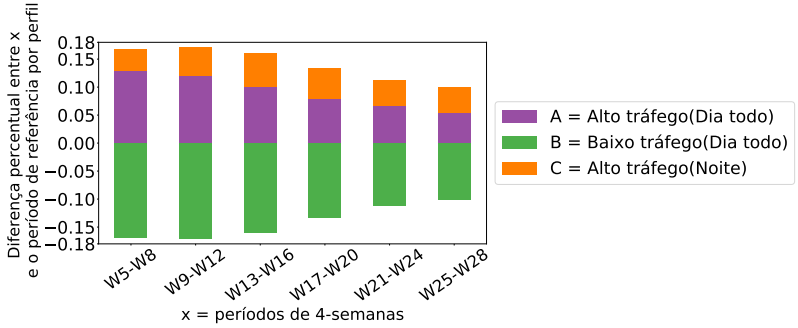
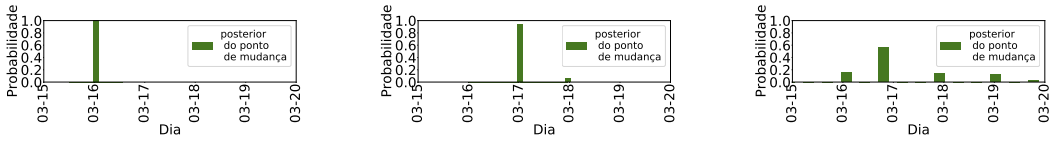


Figura 7. Diferença entre a fração de residências associadas a cada perfil antes e depois da quarentena.

antes e depois do instante de mudança. A biblioteca python PyMC3 [Salvatier et al. 2016] foi usada para obtenção dos resultados.

Seja τ uma variável aleatória que representa o instante de mudança e considere a *prior* de τ uma variável aleatória discreta Uniforme$(0, N)$, onde N é o período de observação. Definimos p_{ij}, $i = A, B, C$, $j = 1, 2$ como a probabilidade de uma residência ser associada ao perfil i no intervalo j, onde $j = 1$ para $0 \leq t \leq \tau$ (período antes do instante de mudança) e $j = 2$ para $t > \tau$ (período após o instante de mudança). Consideramos $Beta(1, 1)$ como *prior* para p_{ij}. A variável observável Y é o número de residências associadas ao perfil i em um dia. Y pode ser representada por uma variável aleatória Binomial com parâmetros M (número de residências associadas ao perfil i em um dia) e p_{i1} ou p_{i2}, dependendo do período analisado, antes ou depois do instante de mudança. As *posteriors* de p_{ij} e τ são obtidas a partir de métodos MCMC.

A Figura 8 ilustra a *posterior* de τ para os três perfis residenciais. O instante onde foi identificada uma mudança significativa na fração de residências associadas ao Perfil A ocorreu em 16 de março de 2020 com probabilidade 1 (data do início da quarentena no estado do Rio de Janeiro). Para o Perfil B, o instante de mudança com maior probabilidade (0.95), ocorreu em 17 de março de 2020, no dia seguinte ao início da quarentena. Por outro lado, a *posterior* de τ para o Perfil C indica que a mudança ocorreu no período entre 16 e 19 de março com alta probabilidade e que o dia mais provável para o instante de mudança foi 17 março de 2020.

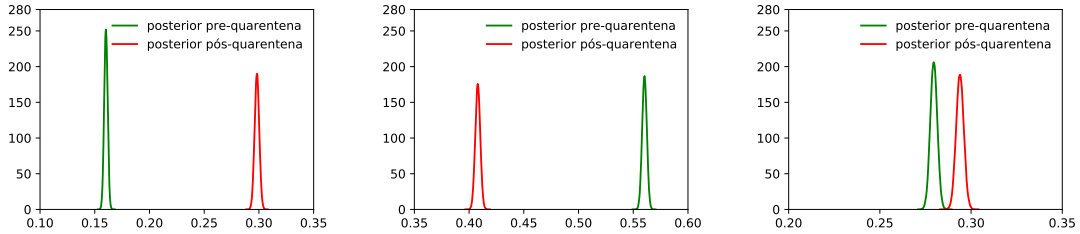


(a) Perfil A **(b) Perfil B** **(c) Perfil C**

Figura 8. *Posterior* do instante de mudança por perfil.

A Figura 9 ilustra a *posterior* da fração de residências por perfil para cada período de análise. A partir dessa figura, notamos que a média da *posterior* para o Perfil A após o instante de mudança (a data do início da quarentena) é aproximadamente o dobro da média da *posterior* antes da quarentena. Este resultado indica que houve um aumento significativo no número de residências associadas ao Perfil A (alto uso da Internet durante todo o dia) imediatamente após o início da quarentena. Já para o Perfil B ocorreu o oposto.

A média da *posterior* pré-quarentena é maior do que a média pós-quarentena, indicando uma queda no número de residências associadas ao perfil de baixo tráfego após o início da quarentena. O Perfil C teve comportamento semelhante ao Perfil A, mas o aumento no número de residências a ele associadas após a quarentena foi menos expressivo.



(a) Perfil A **(b) Perfil B** **(c) Perfil C**

Figura 9. *Posterior* da fração de residências por perfil.

Os resultados obtidos nessa seção indicam que o início da quarentena teve um impacto grande na fração de residências associadas a cada perfil e que a mudança dos perfis ocorreu em um curtíssimo espaço de tempo (um ou dois dias). Uma fração significativa de residências migraram de um perfil de pouco uso da Internet para um perfil de uso intenso da Internet durante o dia todo.

3.4. Comparação de mudanças no perfil residencial com dados de mobilidade do Google

Nosso objetivo é comparar se a mudança no perfil de tráfego das residências que ocorreram após o início da quarentena estão correlacionadas com os dados de mobilidade do Google. Usamos dados do Google disponíveis publicamente [Google 2020]. Os dados permitem avaliar a eficiência do distanciamento social com base nos movimentos espaço-temporais dos dispositivos móveis dos usuários.

O Google estimou a mudança relativa entre o tempo que as pessoas passam em suas residências antes e depois da quarentena, e a mudança relativa entre o número de pessoas nos locais de trabalho antes e depois da quarentena. O valor de referência considerado para calcular a mudança relativa é a mediana do tempo que as pessoas passam em suas residências e a mediana do número de pessoas nos locais de trabalho para cada dia da semana no período de 03 de janeiro a 06 de fevereiro de 2020. Essas mudanças relativas foram calculadas para o período de 2 de março a 6 de setembro de 2020.

Pretendemos calcular a correlação entre as duas métricas de mudança relativa estimadas pelo Google e as duas métricas de mudança relativa que estimamos a partir dos perfis de tráfego residencial.

Definimos um vetor $\mathbf{b}_j(i)$, $i = 1, ..., 7$, $j = A, B$, onde cada elemento representa a mediana da fração de residências associadas ao perfil j para cada dia da semana. Esse vetor contém o valor da mediana calculada para um período de cinco semanas antes da quarentena, de 16 de janeiro a 19 de fevereiro de 2020. A mudança relativa é igual a $\mathbf{R}_j(i) = (\mathbf{a}_j(i) - \mathbf{b}_j(i))/\mathbf{b}_j(i)$, onde $\mathbf{a}_j(i)$ é a fração de residências associadas ao perfil j para o i-ésimo dia da semana durante o período de 2 de março a 6 de setembro de 2020.

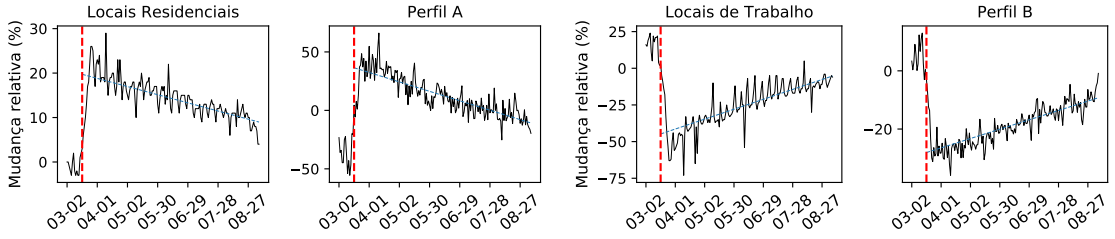
A Figura 10 mostra a mudança relativa das variáveis estimadas pelo Google e as variáveis que calculamos para a cidade de Nova Friburgo. Comparamos o tempo que

Tabela 1. Correlação entre os perfis de tráfego e as métricas de mobilidade do Google.

Cidade	Perfil A e Locais Residenciais		Perfil B e Locais de Trabalho	
	Coeficiente	p-valor	Coeficiente	p-valor
Nova Friburgo	0.88	7.87e-60	0.88	5.23e-59
Campos dos Goytacazes	0.84	5.07e-48	0.71	1.22e-28
Rio das Ostras	0.72	3.72e-30	0.61	4.87e-20
Niteroi	0.78	4.05e-37	0.73	1.05e-30

as pessoas passam nas residências com a fração de residências associadas ao Perfil A (tráfego intenso o dia todo), e o número de pessoas nos locais de trabalho com a fração de residências associadas ao Perfil B (baixo tráfego durante todo o dia). Todos os gráficos na Figura 10 mostram uma mudança repentina após o dia de início da quarentena, 16 de março de 2020, indicado na figura com uma linha tracejada vermelha.

Os gráficos indicam também que conforme medidas de relaxamento da quarentena são implementadas, as pessoas saem de casa com mais frequência e voltam ao trabalho, e com isso 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.



(a) Dados residenciais e Perfil A **(b) Dados de locais de trabalho e Perfil B**

Figura 10. Mudanças nos perfis residenciais e a mobilidade dos usuários na cidade de Nova Friburgo.

Usamos o coeficiente de correlação de Pearson para calcular a correlação entre as mudanças na fração de residências associadas a um determinado perfil e as métricas de mobilidade dos usuários estimadas pelo Google. Os resultados na Tabela 1 demonstram 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 e 6 de setembro de 2020 para quatro cidades consideradas no nosso estudo. Testamos a hipótese nula de que o coeficiente de correlação é igual a 0, com base no valor do coeficiente de correlação obtido no nosso estudo, e obtivemos valores muito baixos para o p-valor indicando a rejeição da hipótese nula (ver Tabela 1). Fizemos também um cálculo preliminar para obter o número médio de celulares conectados por mais de quatro horas por roteador doméstico durante o período normal 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.

4. Trabalhos Relacionados

Desde o surto da pandemia da Covid-19, a comunidade acadêmica vem estudando os impactos das medidas de confinamento no tráfego da rede. Um trabalho recente [Feldmann et al. 2021] analisa o efeito da mudança do tráfego que ocorreu na Europa e nos Estados Unidos. Foi constatado que houve um aumento de 15-20% no tráfego logo após o início do confinamento. Eles analisaram as mudanças do tráfego de diferentes classes de serviços da Internet. Embora não utilizemos dados de inspeção profunda de pacote em nossa análise, os resultados de Feldmann *et al.* (2021) corroboram com os nossos.

Em Böttger *et al.* (2020) os autores utilizaram os dados do Facebook para analisar as mudanças do tráfego e dos aplicativos da família do Facebook durante a pandemia da Covid-19 em todos os continentes. Os resultados mostram um alto crescimento do tráfego em todo o mundo no início da pandemia. Eles também relatam um aumento de 20% no tráfego de vídeo e 25% na média de horas de vídeo assistidas no Brasil.

O trabalho de Candela, Luconi e Vecchio (2020) estuda o impacto na latência da Internet de cidades europeias. Foram observadas importantes e relevantes alterações na rede durante o período da tarde devido às mudanças nas rotinas dos usuários.

Os efeitos das medidas de isolamento social têm sido analisados por diversas empresas, tais como o Google que fornece informações sobre o fluxo de pessoas em áreas públicas e privadas usando dados de GPS [Google 2020] e Apple que relata as tendências de mobilidade de usuários [Maps 2020].

5. Conclusão

Analisamos o impacto do confinamento sobre as características de tráfego residencial de 13 cidades do estado do Rio. 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.

Identificamos três perfis de tráfego residencial distintos usando um modelo de referência obtido com dados de 2019. 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 do isolamento, passando para o perfil diário caracterizado pelo uso mais intenso da Internet e por um período mais longo. Estimamos o dia mais provável da ocorrência das mudanças nos perfis e quantificamos a variação na fração de residências associadas a cada perfil, usando uma abordagem Bayesiana. Os resultados do modelo indicam que houve uma mudança significativa na média da fração de residências associadas a cada perfil e que a mudança ocorreu em um curtíssimo espaço de tempo (um ou dois dias). 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. 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.

Referências

Böttger, T., Ibrahim, G., and Vallis, B. (2020). How the internet reacted to covid-19 – a perspective from facebook's edge network. *ACM Internet Measurement Conference*.

Bro, R. (1997). Parafac. tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171.

Candela, M., Luconi, V., and Vecchio, A. (2020). Impact of the covid-19 pandemic on the internet latency: A large-scale study. *Computer Networks*, 182:107495.

Cloudflare (2020). The internet was #builtforthis. `https://www.cloudflare.com/builtforthis/insights/`.

CNBC (2020). 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`.

Favale, T., Soro, F., Trevisan, M., Drago, I., and Mellia, M. (2020). Campus traffic and e-learning during covid-19 pandemic. *Computer Networks*, 176:107290.

Feldmann, A., Gasser, O., Lichtblau, F., Pujol, E., Poese, I., Dietzel, C., Wagner, D., Wichtlhuber, M., Tapiador, J., Vallina-Rodriguez, N., et al. (2021). Implications of the covid-19 pandemic on the internet traffic. In *Broadband Coverage in Germany; 15th ITG-Symposium*, pages 1–5.

Google (2020). Covid-19 community mobility report. `https://www.google.com/covid19/mobility`.

Harshman, R. A. (1984). "how can i know if it's real?"a catalogue of diagnostics for use with three-mode factor analysis and multidimensional scaling. *Research methods for multimode data analysis*, pages 566–591.

Harshman, R. A. and Lundy, M. E. (1984). The parafac model for three-way factor analysis and multidimensional scaling. *Research methods for multimode data analysis*, 46:122–215.

Kruskal, J. (1983). Multilinear methods. In *Proc. Symp. Appl. Math*, volume 28, page 75.

Lorenzo-Seva, U. and Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2):57–64.

Maps, A. (2020). Mobility trends reports. `https://covid19.apple.com/mobility`.

Now, B. (2020). Internet speed analysis: Rural, top 200 cities april 5th – 11th. `https://broadbandnow.com/report/internet-speed-analysis-april-5th-11th/`.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55.

Sandvine (2020). The global internet phenomena report: Covid-19 spotlight. `https://www.sandvine.com/covid-internet-spotlight-report`.

Smith, A. (1975). A bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2):407–416.

Stedmon, C. A. and Bro, R. (2008). Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial. *Limnology and Oceanography: Methods*, 6(11):572–579.

Streit, A. G., Leão, R. M. M., de Souza, E., and Menasche, D. (2019). Descobrindo perfis de tráfego de usuários: uma abordagem não supervisionada. In *Anais Principais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 169–182. SBC.