Extração de assinaturas semânticas da mobilidade de usuários

Germano B. dos Santos[1], **Fabrício A. Silva**[1]

[1]Instituto de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa - Campus Florestal (UFV – CAF)
Rodovia LMG 818, km 6, 35.690-000 – Florestal – MG – Brasil

germano.santos@ufv.br, fabricio.asilva@ufv.br

Resumo. *A rotina de deslocamentos dos seres humanos é um indicador de suas preferências e pode ser expressa por uma rede de localizações em um espaço-tempo. Neste sentido, o estudo da mobilidade visando compreender o perfil de comportamento de usuários móveis é importante para tornar campanhas publicitárias mais assertivas, por exemplo. Entretanto, poucos estudos utilizam de dados de GPS em grande volume para definir o contexto da mobilidade da sociedade. Portanto, o presente trabalho utiliza grande volume de dados provenientes de GPS para extrair a assinatura semântica da mobilidade de usuários móveis para entender o contexto dos deslocamentos durante e após a pandemia de COVID-19. Os resultados sugerem que os usuários tiveram os padrões de mobilidade alterados após as restrições terem sido amenizadas, com a adesão de um modelo híbrido de trabalho e estudo.*

1. Introdução

O aumento do uso de dispositivos móveis possibilitou o estudo dos padrões de mobilidade de uma pessoa. Segundo os autores de [Barbosa et al. 2018], entender como e por qual razão um humano se movimenta, seja em uma rotina de trabalho ou por lazer, é relevante, por exemplo, para modelar a difusão de vírus em comunidades e predizer o tráfego de uma cidade em diferentes contextos. Esses autores também afirmam que pessoas não se movimentam aleatoriamente, pois possuem preferência por locais seguindo diferentes atributos. Portanto, existem modelos que visam explicar a mobilidade de usuários através de atributos como distância, o dia da semana e horário, que influenciam a decisão de um usuário em relação a qual local visitar [Song et al. 2010, Pappalardo et al. 2016].

A partir de registros geolocalizados em um determinado intervalo de tempo, é possível identificar o comportamento de um usuário em relação a sua mobilidade. Devido ao ritmo de vida da sociedade, existem padrões de deslocamento como trabalho-casa e casa-trabalho-lazer, descritos em [Ma et al. 2022] como assinaturas do estilo de vida de um usuário. Essa assinatura é definida por meio de *motifs* [Schneider et al. 2013] semânticos utilizando as categorias das localizações, nomeadas por [Psyllidis et al. 2022] como ponto de interesse, que um indivíduo visita, como bar, escola, casa, comércios, e assim traça-se um perfil do comportamento desse usuário.

Para definir esse perfil de mobilidade do usuário, podem ser usadas diferentes fontes de dados coletados por meio de GPS ou de registros de chamada *Call Detail Records (CDRs)*. Os atuais trabalhos focam em extrair o perfil do usuário utilizando dados de registros de chamada, o que pode levar a imprecisão da localização. Em zonas rurais e em

cidades menores, a densidade de torres de telefonia é baixa e assim, as localizações registradas não representam exatamente o local da chamada. Além disso, os estudos não focam em suportar o grande volume de dados que atualmente é um empecilho da implementação dos trabalhos científicos em ambientes de produção.

Nesse sentido, esse trabalho visa resolver os problemas mencionados com a implementação e integração à biblioteca SENDAS (Scalable ENrichment for mobility DAtaSets) de uma solução para extração de assinatura semântica de usuários móveis, aderente a dados de localização GPS e a grandes volumes de dados. Para isso, foi necessário implementar dois algoritmos da literatura: um para inferir lugares de interesse a partir de pontos de paradas de uma trajetória e outro que classifica semanticamente esses pontos, chamados de Pontos de Interesse. Com isso, o presente estudo validou as implementações com a avaliação da assinatura semântica da mobilidade de usuários em 2021 e em 2022 com a finalidade de verificar se houve mudança de comportamento da rotina de deslocamento após as restrições da COVID serem amenizadas.

As principais contribuições são:

- Implementação do algoritmo de identificação de POI, em um ambiente de larga-escala;
- Implementação do algoritmo de classificação semântica de POI, em um ambiente de larga-escala;
- Integração de algoritmos de métricas de mobilidade ao SENDAS;
- Comparação da assinatura semântica da mobilidade de usuários em 2021 e em 2022, com análises sobre as mudanças de comportamento.

O restante deste texto está organizado da seguinte forma. Na Seção 2, são apresentados os trabalhos relacionados. A Seção 3 descreve os dados utilizados e os algoritmos implementados, enquanto a Seção 4 apresenta as estatísticas extraídas das assinaturas semânticas da mobilidade dos usuários. Na Seção 5 são discutidos os resultados de forma analítica. Por fim, na Seção 6 são apresentadas as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Entender como as pessoas se comportam nas dimensões de espaço e tempo e por que interagem diferentemente conforme o tipo do estabelecimento que visitam é um desafio importante e atual [Psyllidis et al. 2022].

O estudo de [Barbosa et al. 2018] discute diversos modelos de mobilidade que tentam predizer comportamento a partir dos eventos coletados de usuários móveis. [Song et al. 2010] propõem um novo modelo de mobilidade nomeado como EPR (*Exploration - Preferential Return*) que visa definir o padrão de visitas de um usuário a partir da probabilidade da exploração de um novo local e a probabilidade de retorno a um local previamente descoberto.

Além disso, os autores de [Barbosa et al. 2018] também discutem métricas que podem ser utilizadas para entender os padrões de visitas de um usuário. O estudo de [Pappalardo et al. 2015] apresenta o k-raio de giro e o trabalho [Pappalardo et al. 2016] desenvolve a entropia não correlacionada ao tempo.

Os modelos e métricas são importantes para entender o contexto de mobilidade de um usuário móvel, porém é necessário compreender a interação entre uma rede de

locais e suas categorias. O trabalho de [Ma et al. 2022] caracteriza um local visitado a partir do código do estado americano (NAICS) que define tipos de estabelecimentos. A partir disso, os autores calculam *motifs* semânticos com o intuito de definir uma assinatura do estilo de vida dos usuários. Nesse sentido, o estudo de [Cao et al. 2019] caracteriza usuários com os *motifs* para entender se existe alguma relação entre a distância e o padrão de deslocamento. O trabalho de [Capanema et al. 2019] infere a semântica da região que o usuário visita, definindo intervalo de inatividade de usuários para identificar se a região pode ser classificada como casa, trabalho ou outra atividade como lazer.

No entanto, os trabalhos que exploram técnicas de identificação de um ponto de interesse e métricas de mobilidade não possuem suporte ao grande volume de dados [Montoliu et al. 2013, Jordahl et al. 2020, Pappalardo et al. 2019, Graser 2019]. Além disso, [Cao et al. 2019, Xiong et al. 2021] utilizam dados provenientes de registros de chamadas (CDRs) que possuem imprecisão quando comparado às localizações geradas por GPS. A densidade das torres de telefonia é baixa em pequenas cidades e em zonas rurais, influenciando na precisão da localização registrada, visto que a chamada é direcionada à torre mais próxima. Além disso, um deslocamento dentro da cobertura de uma mesma torre celular não é detectado, mas pode representar um comportamento relevante para algum usuário.

Portanto, o presente trabalho tem o intuito de integrar as métricas de mobilidade definidas em [Pappalardo et al. 2015, Pappalardo et al. 2016] e o algoritmo de inferência da semântica de ponto de interesse [Capanema et al. 2019] ao SENDAS, um *framework* de alta escalabilidade, implementado utilizando SPARK, um ambiente que suporta grande volume de dados. Além disso, esse trabalho utiliza dados coletados por meio de GPS para extrair as assinaturas semânticas da mobilidade de usuários.

3. Metodologia

Nessa seção estão descritos os métodos utilizados para extrair a assinatura da mobilidade do usuário.

3.1. Pontos de Parada

[Montoliu et al. 2013] definem dois conceitos para identificar lugares de interesse dos usuários de acordo com suas visitas: pontos de parada e região de parada. Um ponto de parada $sp_i,\ sp_i \in SP$ é um agrupamento de pontos $p_i,\ p_i \in P$. Para que o agrupamento seja válido, é necessário que p_{i+1} esteja a no máximo D_{max} metros de distância de p_i e tenha uma diferença de tempo de no mínimo T_{min} e no máximo T_{max}. P é o conjunto de registros de um usuário u_i e U é o conjunto de usuários presentes na base de dados.

Para identificar os pontos de parada dos usuários, neste trabalho, os parâmetros D_{max}, T_{min} e T_{max} foram definidos como 200 metros, 10 minutos e 8 horas, respectivamente. O parâmetro T_{min} foi definido como 10 minutos para retirar pequenas paradas em uma trajetória, por exemplo, como uma parada devido a trânsito ou semáforos; o parâmetro T_{max} foi definido como 8 horas, pelo fato do ciclo circadiano dos seres humanos, em geral, as pessoas permanecem no trabalho por 8 horas. Por fim, o parâmetro D_{max} é igual a 200 metros pelo fato de ser o menor valor considerado por [Montoliu et al. 2013] que produz os melhores resultados para essa identificação. Um algoritmo que já havia sido implementado e integrado ao SENDAS foi utilizado para determinar o conjunto de pontos

(a) (b)

Figura 1. (a) Regiões de Parada sem filtro por área; (b) Regiões de Parada com filtro de $0,01km^2$. Essa área representa quadrados de $100m \times 100m$.

de parada SP de cada usuário. Formalmente, os pontos de paradas podem ser definidos como uma tupla $sp = \langle id, t_{start}, t_{end}, lat, long \rangle$. O primeiro campo define o identificador do usuário, o t_{start} e o t_{end} representam, respectivamente, o tempo de chegada e o tempo de saída daquela parada. Além disso, existe a representação geográfica do ponto, especificados por latitude (lat) e longitude ($long$).

3.2. Região de Parada

Uma região de parada é calculada a partir do agrupamento de pontos de paradas, gerando, ao final, um retângulo mínimo que contém todos esses pontos. Portanto, uma região é definida como $r = \langle id, minbox, lat_{cm}, long_{cm} \rangle$, sendo id, o identificador do usuário, $minbox$, o retângulo mínimo que contém todos os pontos agrupados da região e o centroide da região representado por latitude (lat_{cm}) e longitude ($long_{cm}$).

Para realizar o agrupamento, foi utilizado o algoritmo DBSCAN [Ester et al. 1996] que desconsidera ruídos. A aplicação desse tipo de agrupamento, com a finalidade de identificar regiões de parada, já foi aplicado em estudos como [Montoliu et al. 2013, Capanema et al. 2019]. Para a integração do algoritmo ao SENDAS, foi utilizada a implementação do algoritmo em ambiente paralelo descrito em [He et al. 2014]. Além disso, o código, acessível via GitHub[1], foi adaptado para ser possível integrá-lo ao SENDAS.

É notável no exemplo da Figura 1(a) o resultado do algoritmo de agrupamento, sem realizar nenhum tratamento adicional. No entanto, para extrair a assinatura semântica de um usuário é necessário que as regiões tenham características homogêneas. Não é possível ter certeza da homogeneidade com uma região com área grande conforme a Figura 1(a). Portanto, após a identificação das regiões, retiramos aquelas que possuíam área maior que $0,01km^2$. Esse valor foi definido pelo fato de a distância máxima de varredura do DBSCAN (definida pelo parâmetro *eps*) ser igual a $100m$. O resultado das regiões de interesse após a aplicação do filtro é apresentado na Figura 1(b).

3.3. Classificação das Regiões de Parada

Após extrair as regiões de parada de cada usuário e aplicar os devidos filtros, é possível inferir a semântica da região. Nesse estudo, assume-se três tipos de localizações que um

[1]https://github.com/irvingc/dbscan-on-spark

usuário pode visitar: casa, trabalho e outro. Para classificar uma região é necessário definir os horários que o usuário mais visita aquele local. Logo, a partir da definição de [Capanema et al. 2019], uma região é definida como Casa, se é uma área que possui visitas frequentes, em horário específico, de 20:00 às 8:00; uma região é definida como Trabalho se contempla a maioria das visitas do usuário no intervalo de 9:00 às 19:00. É importante ressaltar que se o dia do registro não é dia de semana, a região não é classificada como Trabalho. Além disso, existirá apenas 1 região classificada como Casa e 1 região classificada como Trabalho. A definição da região de parada, portanto, pode ser reformulada para $r = \langle id, minbox, lat_{cm}, long_{cm}, class \rangle$, em que *class* indica a categoria da região.

Os pontos de paradas podem ser classificados a partir da relação espacial com as regiões de parada definidas previamente. Se um ponto de parada está contido em uma região de parada, possuirá a mesma categoria. Assim, a tupla do ponto de parada pode ser reescrita como $sp' = \langle id, t_{start}, t_{end}, lat, long, class \rangle$.

3.4. Assinatura Semântica

Para extrair a assinatura semântica da mobilidade de um usuário, é necessário analisar os padrões de visita, caracterizados por *motifs*. Um *motif* é um subgrafo direcionado recorrente que representa as visitas em um intervalo de tempo. Pode ser classificado em dois tipos: canônico e semântico. O *motif* canônico é rotulado a partir do identificador atribuído às visitas a diferentes lugares de interesse. O *motif* semântico é rotulado a partir da semântica do local visitado. Formalmente, um *motif* pode ser escrito como $H(V, A)$, onde $V(x, y, l)$ é o conjunto de pontos visitados, representados por latitude (x), longitude (y) e um rótulo l. A é o conjunto de arcos ordenado temporalmente, que representa os deslocamentos de um local V_i a outro V_{i+1}.

Aplicamos o algoritmo de reconhecimento de *motifs* canônicos com o conjunto de pontos de paradas SP, sem usar a semântica presente em sp'. Portanto, para definir qual é a origem e o destino da visita, foi preciso identificar regiões onde os pontos se localizavam. Essa região pode ser definida por diferentes tesselações, como, por exemplo, a geometria S2[2], H3[3] ou GeoHash. Neste estudo foi utilizado o H3. Essa tesselação dispõe, em um plano, hexágonos com áreas proporcionais à resolução[4] definida pela ferramenta. Após aplicar o H3 às bases de dados com resolução igual a 10, que representa hexágonos de área $0,01km^2$, a tupla sp é transformada em $sph = \langle id, t_{start}, t_{end}, lat, long, h3hash_{orig}, h3hash_{dest} \rangle$. Assim, adiciona-se dois campos para representar a origem e o destino do deslocamento. Esses *hashes* únicos serão transformados em letras do alfabeto grego para respeitar a rotulação canônica.

A construção de *motifs* semânticos se assemelha ao dos *motifs* canônicos, porém o identificador será a semântica (i.e., atributo *class*) do ponto de parada sp'. A tupla sp' será transformada em $sph' = \langle id, t_{start}, t_{end}, lat, long, class_{orig}, class_{dest} \rangle$, adicionando-se, portanto, dois campos para representar a origem e o destino semânticos do ponto de parada. Por exemplo, o usuário possui um ponto de parada sp' classificado como casa e, após determinado Δt, tem um ponto de parada caracterizado como trabalho, logo, a

[2]https://s2geometry.io/

[3]https://github.com/uber/h3

[4]https://h3geo.org/docs/core-library/restable

origem será casa e o destino o trabalho.

4. Resultados

Nesta seção, um estudo de caso é elaborado para avaliar as implementações descritas na seção anterior. Os dados utilizados são descritos e caracterizados. Além disso, os resultados da extração de assinatura de mobilidade dos usuários são apresentados.

4.1. Dados Utilizados

Neste trabalho foram utilizados dados reais, disponibilizados por uma empresa parceira, de 8.869 usuários, coletados em 2021, de 1 de abril a 30 de maio, e em 2022, de 1 de outubro a 30 de novembro. Os dados de 2021 são compostos por 2.866.904 registros e os de 2022 contêm 509.578 registros, cada um deles representados por uma tupla $l = \langle id, lat, long, timestamp \rangle$. O primeiro campo id representa o identificador único dos usuários. A latitude, longitude e *timestamp* definem um evento geolocalizado. A coleta dos dados foi realizada a partir de uso de um aplicativo móvel que registra, periodicamente, um evento de um usuário através do GPS. Para seguir a Lei Geral da Proteção de Dados (LGPD), os usuários foram anonimizados e os identificadores criptografados.

4.2. Caracterização da Base

Primeiramente, é importante caracterizar os usuários a fim de analisar o contexto de mobilidade. Para isso, foram utilizadas três métricas: k-raio de giro, que indica o raio de giro a partir dos k locais mais visitados por um usuário; entropia não correlacionada ao tempo, uma métrica de mobilidade baseada na função de probabilidade de um usuário visitar uma localização dentre todos os distintos locais já visitados por ele; e o raio de giro, que indica a distância característica de um usuário.

A entropia não correlacionada ao tempo pode ser descrita, formalmente, como:

$$E_{enc}(u) = -\sum_{j=1}^{N_u} p_u(j) log_2 p_u(j) \tag{1}$$

A Equação 1, o N_u é o número de localizações distintas do usuário u e $p_u(j)$ é a probabilidade de u ter visitado um local j. Portanto, a entropia não correlacionada ao tempo estima o grau de exploração de um usuário u, se a entropia é alta, o usuário visita muitas localizações diferentes, caso contrário, possui visitas em lugares já conhecidos.

O raio de giro é calculado por:

$$r_g(u) = \sqrt{\frac{1}{n_u} \sum_{i=1}^{n_u} dist(r_i(u) - r_{cm}(u))^2} \tag{2}$$

O k-raio de giro é definido por:

$$r_g^{(k)}(u) = \sqrt{\frac{1}{n_u^{(k)}} \sum_{i=1}^{k} dist(r_i(u) - r_{cm}^{(k)}(u))^2} \tag{3}$$

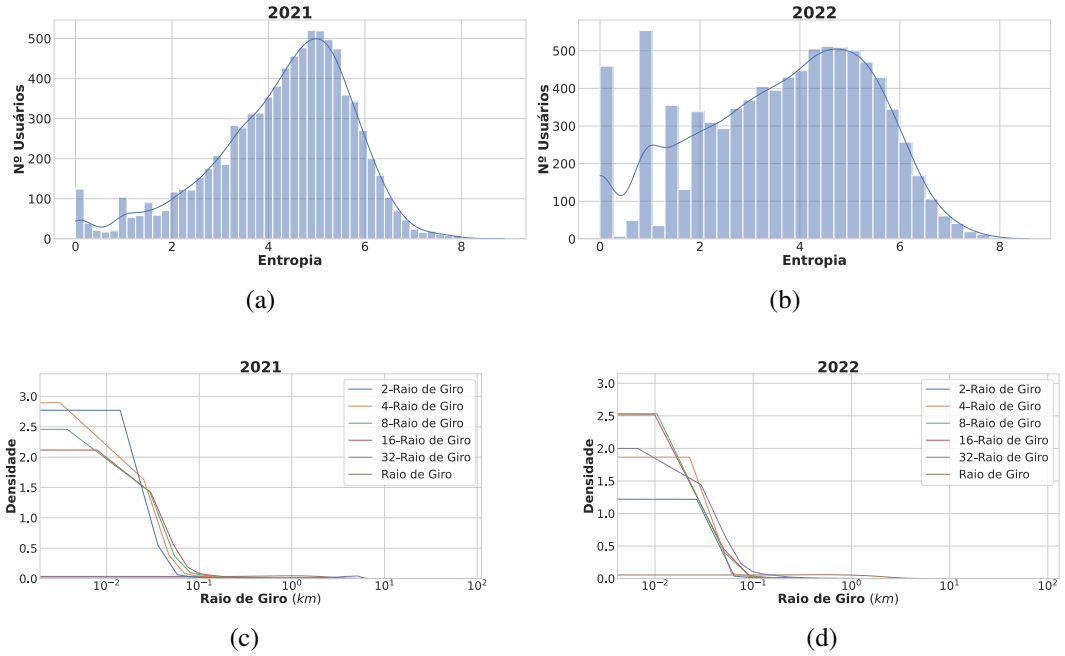


Figura 2. (a) Distribuição da entropia não correlacionada ao tempo no ano de 2021; (b) Distribuição da entropia não correlacionada ao tempo no ano de 2022; (c) Distribuição do raio de giro e k-raio de giro no ano de 2021; (d) Distribuição do raio de giro e k-raio de giro no ano de 2022.

As Equações 3 e 2 são semelhantes. A diferença entre os dois cálculos é o k que determina as k localizações mais frequentes de um usuário u. Nesse sentido, o $r_i(u)$ representa todas as localizações registradas por um usuário u, já o $r_g(u)^k$ representa as k localizações mais frequentes registradas por u. O r_{cm} é o centro de massa de u e o r_{cm}^k é o centro de massa considerando as k localizações mais frequentes.

As Figuras 2(a) e 2(b) apresentam a entropia dos usuários. É possível perceber que poucos usuários possuem baixa ou uma alta entropia. Apesar de as figuras apresentarem concentração dos usuários entre os valores 3 e 6, a entropia do ano de 2022, em média, é menor se comparado ao de 2021. Isso se deve ao fato de aproximadamente 1000 usuários não possuírem pelo menos 10 eventos em 2022. A coleta dos dados foi realizada por um agente instalado em um aplicativo, logo se o usuário desligar o GPS os registros das localizações não são geradas, afetando a coleta de dados.

Para complementar o estudo obtido pela interpretação das distribuições de entropia, realizou-se o cálculo do k-raio de giro com k sendo igual a $2, 4, 8, 16$ e 32. Analisando as Figuras 2(c) e 2(d), é notável que a Lei da Potência está presente, ou seja, poucos usuários apresentam raio de giro alto e muitos usuários apresentam raio de giro baixo. Além disso, quanto maior o k, maior o k-raio de giro, pelo fato de k ser um parâmetro que controla as localizações mais frequentes. Assim, há uma maior aproximação da trajetória real do usuário quando k aumenta. Utilizando o conceito dessa métrica, é possível classificar usuários em *explorers* e *returners* em relação ao k, conforme apresentado na Tabela 1.

Para construir a Tabela 1, foi utilizada a definição de [Pappalardo et al. 2015]. Se

Ano	k^5	*Returners*	*Explorers*
2021	2	0	8869
2021	4	0	8869
2021	8	0	8869
2021	16	0	8869
2021	32	6738	2131
2022	2	5	8864
2022	4	5	8864
2022	8	5	8864
2022	16	5	8864
2022	32	710	8159

Tabela 1. Quantidade de usuários classificados em Returners e Explorers

o k-raio de giro for menor que o raio de giro dividido por 2, o usuário é considerado *explorer*, caso contrário é classificado como *returner*. É possível observar através dessa tabela e dos gráficos relacionados a entropia e raio de giro, que um usuário pode ter seu padrão de mobilidade explicado com 32 localizações mais frequentes em 2021. Esse comportamento não ocorre em 2022, nesse caso, é provável que tenham uma mudança em suas rotinas, com viagens mais distantes do centro de massa ou com muitas viagens próximas ao centro de massa, mas que não podem ser explicadas por apenas 32 localizações.

4.3. Assinatura Semântica

A partir dos métodos definidos na Seção 3.4, os *motifs* diários canônicos e semânticos foram calculados. A Figura 3 apresenta a frequência de cada tipo de *motif*. Cada padrão de deslocamento é definido como uma letra, para o caso do padrão canônico, por exemplo, 'ABC', é caracterizado pelo início da trajetória em uma célula H3 A, há um deslocamento de A para C, passando por B. Esse padrão rotulado de forma canônica pode ser representado por diversos *motifs* semânticos como 'OWO', 'WHW', 'HOH', entre outros, em que 'H' indica a *Casa* do usuário, 'W' o *Trabalho*, e 'O' um local classificado como *Outro*. É importante notar que nas figuras dos padrões canônicos não existe uma sequência com apenas 1 letra, diferente dos padrões semânticos. Isso se deve ao fato das células H3 terem uma resolução reduzida de $0,01km^2$. Então, diariamente, todos os usuários possuem deslocamentos maiores que um hexágono de $0,01km^2$.

Ao comparar as Figuras 3(a) e 3(b) nota-se uma grande semelhança nas frequências de *motifs* de cada ano. Portanto, o deslocamento entre regiões não sofreu mudanças significativas. No entanto, ao analisar os padrões semânticos, notam-se algumas mudanças. Em 2021, os três padrões mais frequentes têm uma sequência com 2 letras, que representam 2 deslocamentos, enquanto nos padrões de 2022, somente um dentre os três mais frequentes apresentam sequência com 2 letras. No caso, 'WO' que poderia representar um deslocamento do trabalho para um restaurante ou para outra atividade de lazer.

Com essas diferenças entre os padrões mais frequentes, é interessante conhecer os *motifs* mais frequentes de cada usuário e observar se mudaram ao longo desse intervalo de tempo, após as restrições impostas pela COVID-19. A Figura 4 mostra a transição dos *motifs* mais frequentes do ano 2021 para 2022. Analisando o padrão canônico (Figura

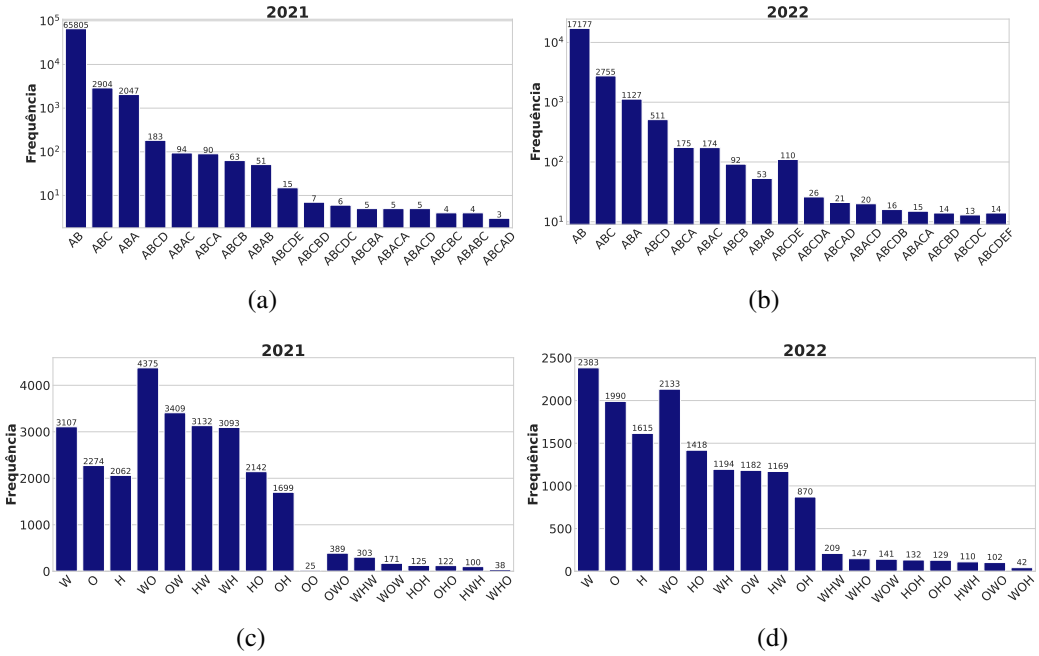


Figura 3. (a) Frequência dos *motifs* canônicos em 2021; (b) Frequência dos *motifs* canônicos em 2022; (c) Frequência dos *motifs* semânticos em 2021; (d) Frequência dos *motifs* semânticos em 2022.

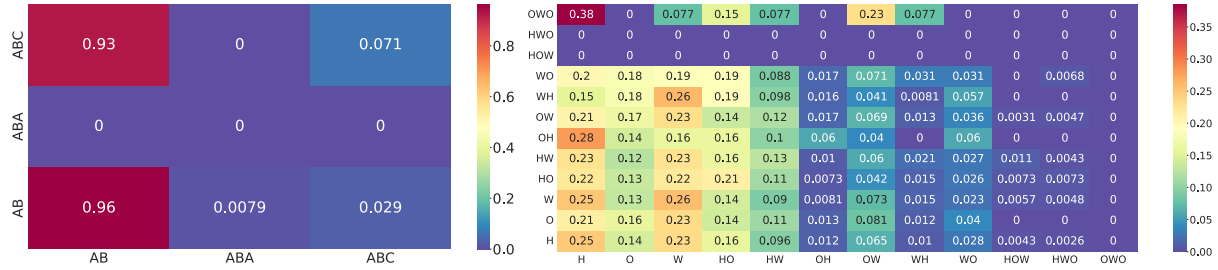
4(a)), é notável que existe apenas uma mudança significativa. 93% usuários que possuíam 'ABC' como mais frequente tiveram mudança no seu padrão de deslocamento para 'AB'.

No entanto, quando os *motifs* semânticos são analisados, é perceptível mudanças nos principais padrões. As diagonais em 4(b), que mostram mudanças significativas nos padrões mais frequentes, apresentam, em sua maioria, valores abaixo de 50%. Pode-se afirmar que os padrões tiveram uma transição instável. Apenas 25% dos usuários que possuíam o padrão 'H' como mais frequente em 2021, continuam com o mesmo padrão em 2022, a maioria dos usuários mudou seu padrão para 'O', 'H' e 'HO', por exemplo, o que mostra a instabilidade dos padrões de deslocamento nesse intervalo de tempo.

5. Discussão

A partir das análises dos resultados da extração da assinatura da mobilidade dos usuários, é possível entender os motivos da mudança do padrão de deslocamento dos usuários após as restrições da pandemia de COVID-19 serem amenizadas.

O raio de giro de 2022 é menor em comparação a 2021, porém existiram muitas mudanças semânticas em 2022. Esse resultado pode ser explicado pelo fato de usuários terem mudado o seu padrão de deslocamento após a pandemia de COVID-19. Durante a pandemia era necessário conhecer lugares próximos a sua casa/trabalho, diminuindo o seu raio de giro, para que o vírus não se espalhasse, observação também feita em [Iio et al. 2021]. Logo, usuários exploraram locais classificados como lazer, próximo a seus pontos de interesse. Esse comportamento pode ser reforçado pela análise da entropia e da tabela 1. A entropia dos usuários, em geral, diminuiu em 2022, o que demonstra que a previsibilidade de suas rotinas aumentou. No entanto, em 2022, para 32 localizações mais frequentes, apenas 710 usuários não foram classificados como exploradores, en-



(a) (b)

Figura 4. (a) Transição dos *motifs* canônicos mais frequentes em 2021 e 2022; (b) Transição dos *motifs* semânticos mais frequentes em 2021 e 2022. Eixo y representa os *motifs* frequentes de 2021, e o eixo x os de 2022. Os valores apresentados indicam a proporção de usuários que tinham um padrão como mais frequente em 2021 e mudou para outro padrão em 2022.

quanto para 2021, esse número é significativamente maior. Portanto, usuários tiveram seu padrão de deslocamento alterado, visitando muitos locais próximos à casa/trabalho.

Os *motifs* semânticos frequentes de 2021 e de 2022, possuem diferença. O padrão 'H' em 2021 é o oitavo mais frequente, já em 2022 é o quarto. Essa alteração pode ser entendida também como uma variação nos padrões de mobilidade do usuário. Quando se analisa essa mudança juntamente à matriz de transição, observa-se uma alteração nos padrões de trabalho para *home office* e educação à distância, como também observado em [Santana et al. 2022]. Apesar de o modelo híbrido e o trabalho remoto ter sido adotado durante a pandemia de COVID-19, esse modelo está ainda sendo utilizado em grande escala, mesmo após o espalhamento da doença ter diminuído.

Existe também a mudança para o padrão de 'W' que pode ser vista na segunda coluna da figura 4(b). Essa alteração também pode estar associada à mudança de hábito em relação ao trabalho remoto ou presencial. Durante a pandemia da COVID-19, foi necessário aderir ao trabalho em casa, porém a qualidade do ambiente, nem sempre favorece à produtividade. Características como qualidade do ar, qualidade de conexão, velocidade de comunicação entre os times podem influenciar a decisão de trabalhar em casa ou ir ao escritório [Umishio et al. 2022]. Após as restrições serem amenizadas, alguns trabalhadores puderam escolher qual ambiente trabalhar, sendo que muitos adotaram o modelo híbrido quando permitido pela empresa.

Apesar dessa análise ser fundamental para o entendimento do comportamento da mobilidade do usuário, não é claro definir o porquê ou como o usuário alterou sua rotina. É preciso ainda identificar mais categorias de locais, ou seja, explorar mais profundamente a semântica como em [Ma et al. 2022]. Com essa definição, é possível traçar um perfil mais assertivo dos padrões de comportamento móvel das pessoas.

6. Conclusão

Diante dos resultados apresentados, nota-se a importância do estudo da assinatura da mobilidade dos usuários para entender o contexto de sua rotina e preferências ao se deslocar. Além disso, o *framework* SENDAS se torna uma ferramenta com grande reusabilidade que pode ser usado para estudos futuros de mobilidade com grande volume de dados.

No entanto, a semântica é ainda um grande desafio. Portanto, como trabalhos futuros, pretende-se propor um novo método que infira o tipo do local classificado nesse trabalho como *Outro*. Assim, será possível ter maior assertividade em determinar a preferência do usuário por um tipo de local específico, como supermercado e comércios.

Referências

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., and Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, 734:1–74.

Cao, J., Li, Q., Tu, W., and Wang, F. (2019). Characterizing preferred motif choices and distance impacts. *Plos one*, 14(4):e0215242.

Capanema, C. G., Silva, F. A., and Silva, T. R. M. (2019). Identificação e classificação de pontos de interesse individuais com base em dados esparsos. In *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 15–28. SBC.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Graser, A. (2019). Movingpandas: efficient structures for movement data in python. *GIForum*, 1:54–68.

He, Y., Tan, H., Luo, W., Feng, S., and Fan, J. (2014). Mr-dbscan: a scalable mapreduce-based dbscan algorithm for heavily skewed data. *Frontiers of Computer Science*, 8(1):83–99.

Iio, K., Guo, X., Kong, X., Rees, K., and Wang, X. B. (2021). Covid-19 and social distancing: Disparities in mobility adaptation between income groups. *Transportation Research Interdisciplinary Perspectives*, 10:100333.

Jordahl, K., den Bossche, J. V., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur, A., Rey, S., Ren, C., Arribas-Bel, D., Wasser, L., Wolf, L. J., Journois, M., Wilson, J., Greenhall, A., Holdgraf, C., Filipe, and Leblanc, F. (2020). geopandas/geopandas: v0.8.1.

Ma, J., Li, B., and Mostafavi, A. (2022). Characterizing urban lifestyle signatures using motif properties in network of places. *arXiv preprint arXiv:2204.01103*.

Montoliu, R., Blom, J., and Gatica-Perez, D. (2013). Discovering places of interest in everyday life from smartphone data. *Multimedia tools and applications*, 62(1):179–207.

Pappalardo, L., Simini, F., Barlacchi, G., and Pellungrini, R. (2019). scikit-mobility: A python library for the analysis, generation and risk assessment of mobility data. *arXiv preprint arXiv:1907.07062*.

Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., and Barabási, A.-L. (2015). Returners and explorers dichotomy in human mobility. *Nature communications*, 6(1):1–8.

Pappalardo, L., Vanhoof, M., Gabrielli, L., Smoreda, Z., Pedreschi, D., and Giannotti, F. (2016). An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics*, 2(1):75–92.

Psyllidis, A., Gao, S., Hu, Y., Kim, E.-K., McKenzie, G., Purves, R., Yuan, M., and Andris, C. (2022). Points of interest (poi): a commentary on the state of the art, challenges, and prospects for the future. *Computational Urban Science*, 2(1):1–13.

Santana, C., Botta, F., Barbosa, H., Privitera, F., Menezes, R., and Di Clemente, R. (2022). Changes in the time-space dimension of human mobility during the covid-19 pandemic. *arXiv preprint arXiv:2201.06527*.

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., and González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246.

Song, C., Koren, T., Wang, P., and Barabási, A.-L. (2010). Modelling the scaling properties of human mobility. *Nature physics*, 6(10):818–823.

Umishio, W., Kagi, N., Asaoka, R., Hayashi, M., Sawachi, T., and Ueno, T. (2022). Work productivity in the office and at home during the covid-19 pandemic: a cross-sectional analysis of office workers in japan. *Indoor air*, 32(1):e12913.

Xiong, Q., Liu, Y., Xie, P., Wang, Y., and Liu, Y. (2021). Revealing correlation patterns of individual location activity motifs between workdays and day-offs using massive mobile phone data. *Computers, Environment and Urban Systems*, 89:101682.