Modelo de Predição de Escolha de Rotas baseado em Traces de Mobilidade Veicular

Augusto C.S.A. Domingues[1], **Letícia Pinto**[1],
Fabrício A. Silva[2], **Rosângela H. Loschi**[1], **Antonio A. F. Loureiro**[1]

[1]Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, MG

[2]Universidade Federal de Viçosa, *campus* Florestal, Florestal, MG

Resumo. *O estudo da mobilidade veicular e o seu impacto no desenvolvimento das cidades é um tópico importante de pesquisa, contribuindo para o cenário dos Sistemas de Transporte Inteligentes (ITS). Neste trabalho, introduzimos um modelo logit multinomial (MNL) para predizer qual rota um usuário irá tomar dado um conjunto de rotas possíveis entre pontos de origem e destino e outras informações contextuais. Para isso, definimos e aplicamos uma abordagem estatística para entender como os veículos se comportam em um cenário urbano considerando os efeitos do tráfego. O modelo é aplicado em dois rastros reais de mobilidade veicular e os resultados mostram que ele é capaz de capturar as influências dos fatores existentes, obtendo predições superiores em comparação a dois baselines. Os resultados indicam que o modelo pode ser aplicado para o preenchimento de grandes lacunas espaciais, como por exemplo na geração de dados sintéticos porém realísticos de trajetórias para rastros de origem-destino.*

1. Introdução

Redes móveis ad hoc (MANETs) são redes compostas por entidades que se movem em um dado espaço enquanto se comunicam entre si. Desta forma, a comunicação está diretamente relacionada à mobilidade, i.e., ela depende da distância física entre a origem e o destino, que regem a qualidade do sinal entre os pontos. O comportamento de mobilidade é definido em grande parte pela própria entidade (e.g., uma pessoa portando um *smartphone*) e, portanto, considerando as particularidades de cada indivíduo, não é possível predizer completamente os movimentos [Teixeira et al. 2021]. Essa falta de conhecimento sobre a mobilidade de um usuário pode impor barreiras quanto ao correto funcionamento da rede [Cotta et al. 2017], o que pode reduzir a experiência do usuário.

Muitos estudos na literatura abordam o problema do comportamento de mobilidade através de análises utilizando rastros (*traces*) de mobilidade, como aqueles contendo trajetórias de viagens de táxi [Yuan et al. 2010] e *logs* de redes WiFi [Kotz et al. 2009]. Esses rastros de mobilidade são alternativas interessantes para análises do mundo real, considerando que ainda existem custos e esforços elevados para implantar e testar redes reais utilizando as tecnologias disponíveis atualmente. Entretanto, devido a diferentes questões, como a inacurácia de sensores de localização, esses dados podem conter erros [Celes et al. 2017], implicando em representações incorretas dos comportamentos sensoriados dos indivíduos. Essa informação enviesada, por sua vez, pode reduzir a utilidade dos rastros gerados por essas entidades.

Buscando aumentar a qualidade de estudos baseados em simulações, investigamos neste trabalho as características de mobilidade veicular em contextos urbanos consi-

derando fatores como velocidade, duração das viagens e as variações no trânsito durante horários de pico. De maneira geral, as principais contribuições deste trabalho são:

1. Um modelo logit multinomial (MNL) para predizer qual rota um usuário irá tomar dado um conjunto de rotas possíveis entre um ponto de origem O e um ponto de destino D e outras informações contextuais, como tempo e tráfego.
2. Escolha das variáveis que compõem o modelo com base na caracterização de dados de trajetórias veiculares reais.
3. Validação do modelo proposto em dois conjuntos de dados de rastros reais.

O restante do trabalho está organizado da seguinte maneira. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 introduz o conjunto de dados utilizado e a metodologia de análise dos mesmos. A análise considerando as características das viagens e a detecção de mudanças no trânsito é construída e apresentada na Seção 4. Em seguida, introduzimos e validamos um modelo para predizer seleções de rotas em rastros de mobilidade na Seção 5. As considerações finais são feitas na Seção 6.

2. Trabalhos Relacionados

Há um número extensivo de trabalhos na literatura que utilizam rastros de mobilidade em suas pesquisas. São inquestionáveis as contribuições que esses conjuntos de dados trazem para a comunidade ao reduzir o esforço necessário em diferentes áreas como, por exemplo, simulações de mobilidade [Hess et al. 2016, Hoteit et al. 2014, Li et al. 2015], caracterização de tráfego [Liu et al. 2012b], otimização de redes [Cotta et al. 2017], entre outros tópicos. Adicionalmente, esses conjuntos servem como fonte fundamental para a geração de rastros sintéticos através de modelos de mobilidade [Yoon et al. 2006], permitindo a construção de cenários complexos e não disponíveis atualmente.

Por outro lado, não é novidade que os rastros de mobilidade encontrados na literatura possuem vários problemas. Existem falhas relacionadas ao processo de coleta, que são inerentes ao dispositivo sensor, como também limitações de armazenamento e envio, restrições de privacidade, entre outros. Essas questões geram lacunas espaciais e temporais nos dados, afetando os resultados das simulações [Silva et al. 2015]. Assim, dada a importância da qualidade dos rastros de mobilidade, diversos estudos propõem diferentes soluções para o enriquecimento dos dados coletados, através de técnicas como a interpolação [Hoteit et al. 2014] e a calibração. Entretanto, nos casos em que as lacunas compreendem a maior parte dos deslocamentos, como nos rastros que só possuem o par origem-destino, pouco tem sido feito.

Uma das abordagens mais simples – em termos de complexidade computacional – para o enriquecimento de rastros de mobilidade é através da interpolação de pontos das trajetórias. Essa técnica utiliza a informação de dois ou mais pontos para gerar um ponto intermediário que aproxima a localização real da entidade [Hoteit et al. 2014, Hoteit et al. 2016]. Apesar de ser capaz de gerar dados similares àqueles das trajetórias reais, essas soluções não consideram a malha viária existente, o que afeta a qualidade dos rastros enriquecidos.

Outra abordagem mais robusta é através da construção de sistemas de calibração que consideram fatores contextuais como a malha viária, dados históricos e comportamentos de mobilidade. Su et al. [Su et al. 2015] apresentam um sistema de calibração

multi-critério baseado em pontos-âncora, enquanto Celes et al. [Celes et al. 2017] utilizam dados históricos para gerar um conjunto de pontos de referência. Apesar do uso de sistemas de calibração produzir dados realistas para o preenchimento de lacunas, essas soluções dependem altamente da existência de dados históricos e dados enriquecidos. Finalmente, outros estudos [Liu et al. 2012a, Chen et al. 2011, Zhu and Levinson 2015] consideram a caracterização de escolhas de rotas em relação às suas otimalidades, i.e., se elas seguem o caminho ótimo, como por exemplo, o mais curto.

Existem diversas aplicações para o conhecimento do comportamento de escolha de rotas. Primeiro, ele permite o entendimento de padrões de mobilidade, o que pode fornecer *insights* sobre a formação de congestionamentos entre outras questões relacionadas ao tráfego [Yuan and Li 2021]. Segundo, dados os pontos de origem e destino, é possível prever as rotas a serem seguidas pelos motoristas, o que é uma informação relevante na disseminação oportunista de dados. Por fim, os padrões de mobilidade extraídos podem ser aplicados na geração de dados de mobilidade sintéticos, que por sua vez podem ser usados em diversas simulações [Harri et al. 2009].

De maneira geral, os trabalhos discutidos aqui buscam melhorar a qualidade de rastros de mobilidade através da geração de pontos intermediários inexistentes devido a diversas falhas. Diferente deste trabalho, esses estudos focam no preenchimento de pequenas lacunas, onde técnicas simples de interpolação e sistemas de calibração produzem resultados satisfatórios. Assim, as soluções existentes não são adequadas para preencher trajetórias completas, i.e., quando somente os pontos de origem e destino estão disponíveis. [Liu et al. 2012a] é o único estudo que trata de grandes lacunas e assume que as trajetórias sempre seguem o caminho mais curto, o que pode não representar o comportamento real dos motoristas. Em nossa abordagem, consideramos a probabilidade da ocorrência de desvios no caminho mais curto, incorporando ações tomadas pelos motoristas devido a fatores externos como congestionamentos ou incidentes de trânsito.

3. Preparação dos Dados

Nesta seção, apresentamos os passos para extrair as características de mobilidade de um trace de mobilidade veicular. Primeiro, apresentamos o conjunto de dados que será utilizado, junto com as definições formais de seus registros. Em seguida, os dados são preparados, o que inclui etapas de filtragem e limpeza, como também o enriquecimento com outros dados contextuais. Esses passos são aplicáveis a qualquer trace de mobilidade veicular. Na verdade, os únicos requisitos necessários são em relação à escala do trace, visando representar o comportamento da população como um todo, e a granularidade dos pontos sensoriados, visando garantir que os processos aplicados terão precisão suficiente.

3.1. O Trace de Táxis de São Francisco

O trace de Táxi de São Francisco [Piorkowski et al. 2009] é um conjunto de dados contendo trajetórias de táxis em São Francisco, EUA[1], coletado em 2008. Ele contém informações sobre as localizações dos veículos, amostradas periodicamente através de um sensor GPS embarcado. Adicionalmente, há um indicador que aponta se um passageiro está sendo transportado, i.e., se uma viagem está ocorrendo. Em suma, há um total

[1]De acordo com o INRIX *Global Traffic Scorecard* (http://inrix.com/scorecard/), São Francisco está entre as dez cidades mais congestionadas dos Estados Unidos (2020).

de 442.718 viagens, com média de 17.690 viagens por dia e um total de 535 veículos monitorados durante um período de 25 dias (Fig. 1).

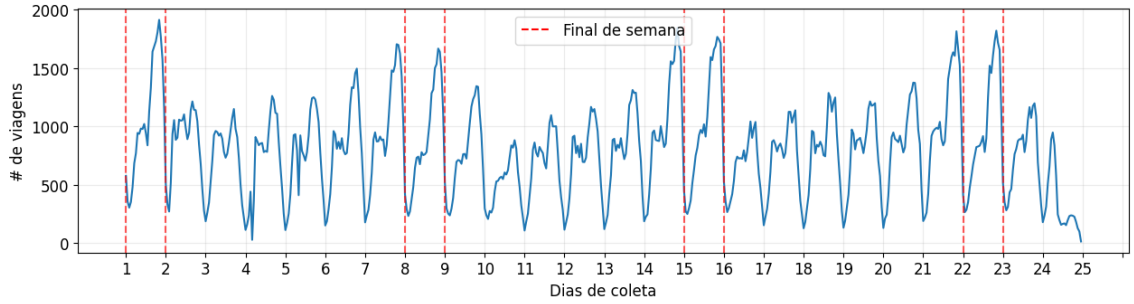

Figura 1. Distribuição do número de viagens durante o período de coleta

3.2. Definições

Cada registro no trace é definido por uma tupla $r = \langle u, t, l, v \rangle$, onde u é o id único do veículo, t é o *timestamp* da coleta, l é a localização reportada pelo dispositivo GPS e v é o indicador de viagem ($v = 1$ se existe passageiro, $v = 0$ caso contrário). O conjunto de todos os registros no trace é definido como S e estão ordenados de forma crescente pelo *timestamp*. Assim, para dois registros r_i e r_j com $i < j$, $r_i.t \leq r_j.t$ (neste trabalho, assumimos a notação $r_i.t$ como o valor da variável t contida dentro da tupla r_i).

Definimos o conjunto de registros de um mesmo veículo X como $S_{\{X\}} = \{r | r.u = X\}$. Uma viagem de um táxi é uma sequência de registros sequenciais (com *timestamp* crescente) no qual o primeiro registro é a origem, e o último o destino. A partir de agora, o operador $[n]$ indica um registro na posição n em uma viagem. Assim, dado um táxi X, para todos os registros $S_{\{X\}}$, uma viagem começa no registro K se $S_{\{X\}}[K].v = 1$ e $S_{\{X\}}[K-1].v = 0$. Similarmente, uma viagem termina no registro Y se $S_{\{X\}}[Y].v = 1$ e $S_{\{X\}}[Y+1].v = 0$. Assim, uma viagem t é uma sequência de registros $r_0, r_1, \cdots, r_{n-1}$, onde o primeiro registro r_0 é o embarque, e o último registro r_{n-1} é o desembarque. Finalmente, $T_{\{X\}}$ representa o conjunto contendo todas as viagens do táxi X.

3.3. Enriquecimento dos Dados

O próximo passo contempla o enriquecimento das viagens, o que consiste na adição de informações da malha viária ao trace. Para isso, usamos o *Open Street Maps* (OSM), que fornece informações sobre cada segmento de via em um dado local. Com esses dados em mãos, realizamos o *map matching* [Newson and Krumm 2009] das viagens, i.e., dado um segmento de pontos em uma trajetória, os mesmos são mapeados para as respectivas vias na malha, resultando em medidas mais precisas. Formalmente, para cada viagem $t = \{r_0, r_1, \cdots, r_{n-1}\}$, produz-se um pareamento $M = \{m_0, m_1, \cdots, m_{n-1}\}$, de tal forma que para um índice qualquer i, m_i é o ID (vindo do OSM) da via equivalente à localização reportada em r_i ($r_i.l$). Para quaisquer pareamentos consecutivos i e $i+1$ tal que i e $i+1$ não pertencem à mesma porção da via, existe uma interseção $\langle m_i, m_{i+1} \rangle$ entre eles na malha viária.

3.4. Métricas para a Caracterização de Viagens

Para cada viagem, definimos sua duração como a diferença entre o *timestamp* relatado no último e no primeiro registro, como visto na Eq. 1. Podemos definir também a sua

distância total (Eq. 2) como a distância par a par entre os registros, considerando a fórmula de haversine como a função de distância d. Dados ambos, a Eq. 3 apresenta a velocidade média da viagem. Uma outra métrica mais robusta, chamada de otimalidade do caminho [Domingues et al. 2018] (Eq. 4), refere-se à diferença da trajetória seguida pelo veículo do ponto de origem até o ponto de destino, e a trajetória de caminho mais curto entre os dois pontos (*MinDist*). Assumimos o caminho mais curto como a trajetória que conecta os dois pontos com a menor distância possível. Essa trajetória é obtida do OSM e considera os comprimentos e interseções das vias, que servem como entrada para o algoritmo de Dijkstra, resultando nos segmentos de via que compõem a trajetória.

$$\text{Dur}(T) = T[r_{n-1}].t - T[r_0].t \quad (1) \qquad \text{Dist}(T) = \sum_0^{n-2} d(\, T[r_i].l \,,\, T[r_{i+1}].l \,) \quad (2)$$

$$s_{avg}(T) = \frac{\text{Dist}(T)}{\text{Dur}(T)} \quad (3) \qquad P_{opt}(T) = \frac{\text{Dist}(T)}{\text{MinDist}(\, T[r_0].l \,,\, T[r_{n-1}].l \,)} \quad (4)$$

Como definido acima, uma viagem segue o caminho mais curto quando a distância total percorrida da origem até o destino é a mesma da obtida usando um algoritmo de caminho mais curto (dado um *threshold* de variação ϵ que contabiliza as flutuações na distância causadas por inacurácias do GPS). Viagens que não seguem o caminho mais curto possuem desvios. A distribuição da otimalidade permite observar os efeitos de congestionamentos e padrões temporais e espaciais de mobilidade.

3.5. Filtragem e Limpeza dos Dados

Dado o trace, primeiramente são removidas as viagens que ocorrem fora dos limites da cidade de São Francisco. Assim, focamos a nossa análise na caracterização e modelagem do comportamento de viagens urbanas e diárias. Em termos de dimensões espaciais (Fig. 2a), removemos todas as viagens com comprimento inferior a 2 km ou superior a 35 km. Considerando a dimensão temporal (Fig. 2b), removemos as viagens que duraram menos de 5 minutos ou mais de 2 horas. Considerando a velocidade média das viagens (Fig. 2c), removemos aquelas com valores acima de 90 km/h. Os limites inferiores apresentados visam filtrar as viagens curtas, ou que foram canceladas. Os limites superiores visam filtrar as viagens com múltiplos destinos ou destinos fora dos limites da cidade, comportamentos inesperados dos motoristas, ou até mesmo problemas com o dispositivo coletor. Nenhum desses cenários é de nosso interesse nesta caracterização.

4. Conhecendo os Dados de Mobilidade Veicular

Nesta seção, analisamos o trace de São Francisco buscando entender os padrões de mobilidade existentes e a sua relação com aspectos temporais e espaciais. Para isso, extraímos as métricas definidas na Seção 3 – velocidade média, duração e otimalidade de trajetória – e as usamos para explicar a ocorrência de mudanças nas condições de tráfego. Essas mudanças nos permitem entender melhor o processo de tomada de decisão dos motoristas em relação à escolha de rotas, conhecimento fundamental na construção de sistemas de transporte inteligentes (ITS) e de outras soluções de mobilidade cientes de contexto.

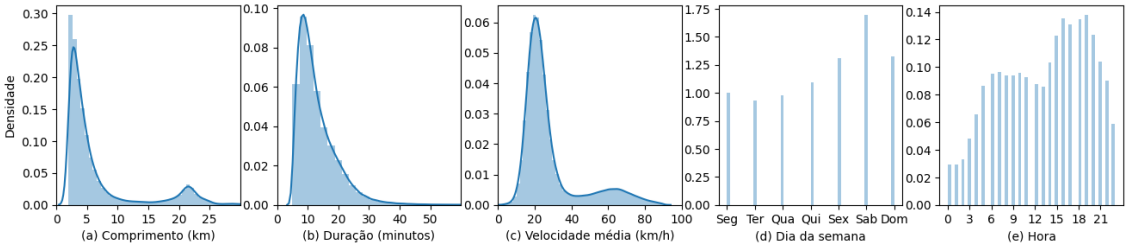

Figura 2. Comprimento das trajetórias (a), duração das viagens (b), velocidade média das viagens (c), distribuição semanal (d) e temporal (e) para o trace de São Francisco

4.1. Características Espaço-temporais

A distribuição de viagens sobre o tempo pode dizer muito sobre os fluxos de tráfego de uma cidade, e.g., intervalos com um número maior de viagens levam a um número maior de veículos nas vias, aumentando assim o tráfego. Esses intervalos podem ser relacionados à hora do *rush* entre outros cenários, como grandes eventos. Ao identificar os intervalos com número elevado de viagens, podemos usar essa informação para justificar o comportamento dos motoristas em relação às condições de trânsito.

A Fig. 1 mostra uma periodicidade clara do número de viagens em cada semana, com um comportamento similar também para cada dia, confirmando que o fluxo de trânsito não é um processo aleatório, mas sim um processo gerado por propósitos bem definidos, como os trajetos diários entre casa e trabalho (ou casa e escola). Vemos também como os finais de semana são responsáveis por um aumento no número de viagens, indicando que outras atividades também afetam o trânsito, como ir a restaurantes, shoppings, entre outros. Assim, é importante considerar também essas atividades irregulares.

Para investigar a evolução das viagens com o passar do tempo, analisamos a autocorrelação (utilizando o coeficiente de Pearson ρ) de algumas das características das viagens por hora e por dia (Fig. 3). Para o primeiro caso, apresentamos a autocorrelação

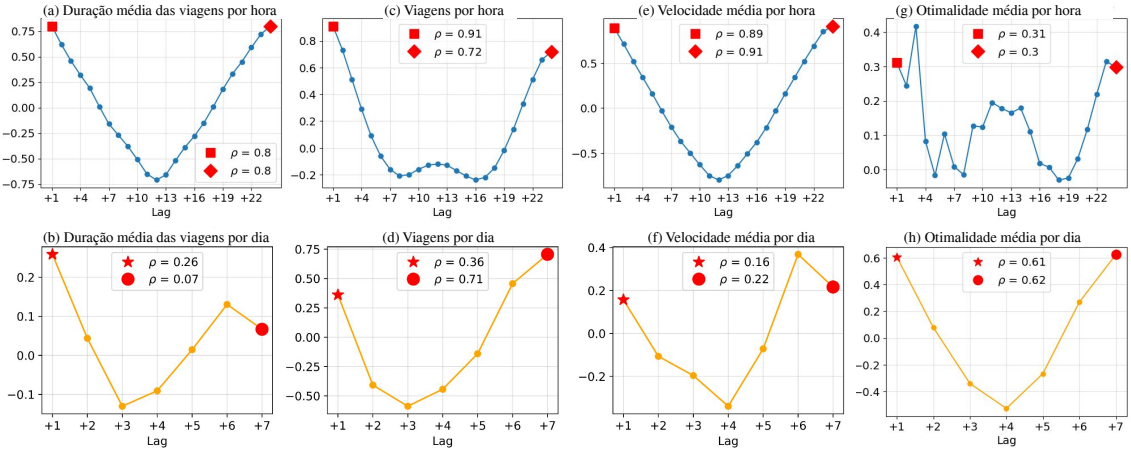

Figura 3. Autocorrelação da duração das viagens por hora (a) e dia (b), do total de viagens por hora (c) e dia (d), para a velocidade média por hora (e) e dia (f) e para a otimalidade do caminho por hora (g) e dia (h).

variando de *lag*-1 (i.e., entre a hora atual h e a hora seguinte, $h + 1$) até o *lag*-24 (entre a

hora atual h e a mesma hora no dia seguinte, $h + 24$). Para a variação diária, apresentamos a autocorrelação a partir do *lag*-1 (entre o dia atual d e o dia seguinte, $d + 1$) até o *lag*-7 (entre o dia atual d e o mesmo dia na semana seguinte, $d + 7$).

Considere o número de viagens por hora e por dia (Figs. 3c e 3d, respectivamente). Para o primeiro, há uma correlação positiva alta ($\rho = 0{,}91$) para o *lag*-1, indicando como o comportamento do fluxo de tráfego se estende sobre o tempo. De maneira similar, há também uma autocorrelação positiva ($\rho = 0{,}72$) para o *lag*-24, reiterando a existência de uma periodicidade no fluxo. Considerando as duas horas seguintes (*lag*-1 e *lag*-2) e as duas anteriores (*lag*-23 e *lag*-24), temos uma autocorrelação média $\rho = 0{,}75$, ou seja, observar intervalos recentes e futuros (i.e., as horas seguintes porém em um dia anterior) pode ser útil na predição do tráfego em um dado tempo. Para o agrupamento diário, existe uma correlação positiva para o *lag*-1 ($\rho = 0{,}36$) e para o *lag*-7 ($\rho = 0{,}71$), indicando como esses valores também podem ser usados para estimar o tráfego em um dado dia.

4.2. Detectando mudanças no fluxo

Em seguida, investigamos como detectar mudanças no fluxo, como as que acontecem devido a hora do *rush*, acidentes, eventos em larga escala e proximidade de pontos-de-interesse (PoI). Para isso, exploramos as três métricas de mobilidade definidas na Seção 3: duração da viagem (Eq. 1), velocidade média (Eq. 3) e otimalidade da trajetória (Eq. 4).

4.2.1. Duração da Viagem

A Fig. 4b mostra a duração das viagens agregada por dia e hora. As viagens mais longas ocorrem durante os períodos da manhã e do início da tarde, especialmente durante os finais de semana, o que, com exceção do período da noite, suporta os pontos levantados sobre o número total de viagens (Figs. 2d e 2e). É válido ressaltar que este conjunto de

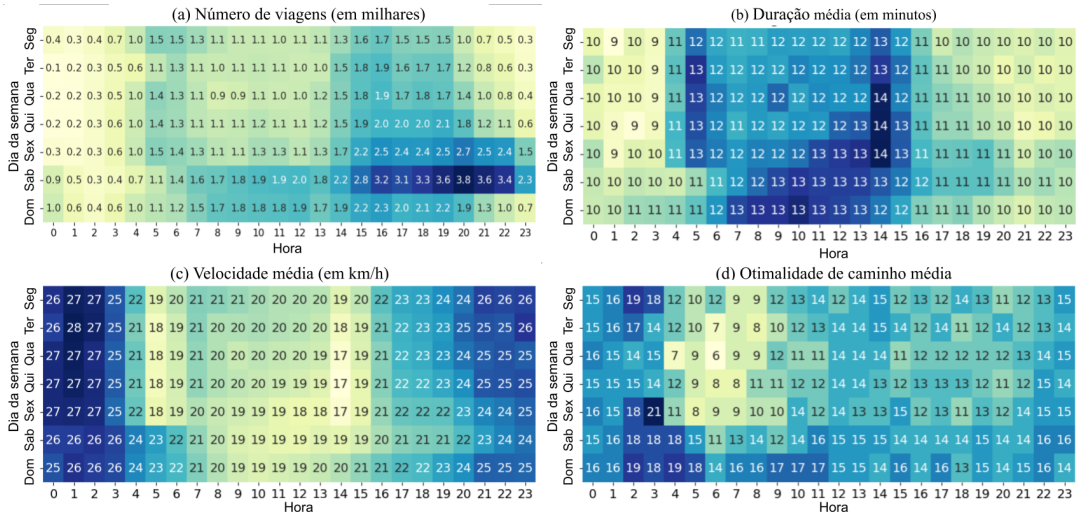

Figura 4. Características das viagens agregadas por dia e hora

dados pode não reproduzir de forma exata o comportamento esperado. Táxis são uma parte dos veículos de uma cidade e se movem de acordo com seus usuários, que podem

preferir outros meios de transporte durante certos horários. Portanto, os pontos levantados aqui podem não refletir o comportamento de veículos particulares, por exemplo.

Em seguida, analisamos a autocorrelação da duração média das viagens por hora e por dia (Figs. 3a e 3b, respectivamente). Para o primeiro caso, observamos os mesmos valores de correlação para o *lag*-1 e para o *lag*-24 ($\rho = 0{,}8$), indicando a influência do tráfego existente. De fato, quando um congestionamento começa, aumentando a duração esperada das viagens, espera-se que os seus efeitos se propaguem nas próximas horas, dado o tempo que leva para os veículos em excesso chegarem a seus destinos. Mais ainda, a alta correlação com as horas anteriores indica a progressão do tráfego até o momento em que atinge um pico. Por outro lado, a autocorrelação por dia apresenta valor considerável para o *lag*-1, indicando a similaridade entre dias próximos na semana.

Adicionalmente, para melhor entender a relação entre as características, computamos a correlação entre o tamanho das trajetórias das viagens e as suas durações, obtendo um valor de $\rho = 0{,}72$. Isso levanta a hipótese que parte do aumento (ou decréscimo) na duração de uma viagem é causado por um aumento (ou decréscimo) no tamanho de sua trajetória.

4.2.2. Velocidade média

A velocidade média pode servir como um *proxy* das condições de tráfego, i.e., velocidades mais baixas podem indicar sinal de congestionamentos ou acidentes. A Fig. 4c apresenta a distribuição da velocidade média agregada por dia e por hora. Os valores mais altos ocorrem durante a noite e a madrugada, de 21 h até às 3 h, indicando o efeito do menor número de veículos nas vias (Fig. 4a). De forma complementar, os menores valores ocorrem durante a manhã e no fim da tarde, o que novamente pode ser explicado pelo aumento no número de viagens nesses períodos (Fig. 4a). De fato, existe uma expressiva correlação negativa entre a velocidade média das viagens e o número de viagens em um dado tempo ($\rho = -0{,}46$), o que reforça esse ponto.

Considerando os efeitos duradouros do tráfego, analisamos a autocorrelação da velocidade média das viagens por hora e por dia (Figs. 3e e 3f, respectivamente). Existe uma alta correlação positiva entre duas horas consecutivas ($\rho = 0{,}89$), e quando consideramos as duas horas anteriores e seguintes ($\rho = 0{,}68$) indicando que a velocidade média em um intervalo pode ser estimada usando informações do passado recente.

Tipicamente, as horas dos *rush* em São Francisco ocorrem durante a manhã (de 7 h até às 9 h) e no fim da tarde (16 h até às 18 h), de segunda a sexta[2]. Dada essa informação, vemos que a métrica de velocidade média obtém sucesso parcial em comunicar o comportamento esperado do fluxo de tráfego, pois apesar dos valores mais altos estarem contidos nesses intervalos, eles acontecem em horários anteriores (entre 5 h e 6 h e entre 14 h e 15 h). Isso pode ser explicado pela existência de vias exclusivas para táxis [**?**], o que significa que suas velocidades podem ser menos afetadas pelo tráfego ao redor nos horários de pico.

[2]De acordo com o TomTom *Traffic Index* (https://www.tomtom.com/traffic-index/san-francisco-traffic/)

4.2.3. Otimalidade de Trajetória

A detecção de trajetórias que não seguem o caminho ótimo é um tópico importante de pesquisa devido a suas aplicações, como o gerenciamento de tráfego [Zhu and Levinson 2015, Chen et al. 2011], detecção de fraude [Lu et al. 2019] e privacidade de localização. A seguir, analisamos a métrica de otimalidade de trajetória.

A Fig. 4d apresenta a otimalidade de trajetória média, agregada por dia e por hora. O valor de 15, por exemplo, indica que na média as viagens durante aquele intervalo tiveram desvios em suas trajetórias que elevaram o seu comprimento em 15% em relação à trajetória ótima. Algumas observações podem ser feitas: primeiro, o período da manhã durante os dias úteis possui as menores médias, com trajetórias próximas às suas versões ótimas. Por outro lado, é durante a noite e a madrugada que os maiores desvios ocorrem, incluindo os finais de semana também. Isto combina diretamente com os valores de duração média e velocidade média vistos nas Figs. 4b e 4c, respectivamente, indicando que os motoristas podem escolher caminhos mais longos buscando se beneficiar de vias menos ocupadas ou vias com limites de velocidade maior. Finalmente, existe uma homogeneidade no comportamento durante a tarde e noite nos finais de semana, indicando que esses períodos foram menos sensíveis a variações de tráfego.

Em seguida, analisamos a autocorrelação da otimalidade média das viagens agregada por hora e por dia (Figs. 3g e 3h, respectivamente). Para o primeiro, existe uma correlação positiva suave nas próximas três horas ($0{,}31$, $0{,}24$ e $0{,}42$ para *lag*-1, *lag*-2 e *lag*-3, respectivamente), indicando que as motivações para a ocorrência de desvios na rota ótima tendem a se arrastar com o tempo. Por fim, a autocorrelação em intervalos diários apresenta resultados interessantes. A existência de uma forte correlação positiva para o *lag*-1 e para o *lag*-7 aponta para a presença de fatores contextuais que periodicamente afetam um dado dia como também o dia seguinte, e.g., tarefas e eventos rotineiros.

4.3. Discussão

As análises feitas aqui permitiram entender os efeitos de fatores contextuais no comportamento do tráfego, como os aumentos no número de viagens e na sua duração, ou a redução na velocidade média das viagens e até mesmo o aumento no número e no tamanho de desvios tomados em uma viagem. Além disso, vimos como a influência desses eventos se propaga com o tempo e tende a se repetir no futuro próximo. Assim, é possível tirar vantagem desses comportamentos para construir sistemas cientes-de-contexto que explorem tal conhecimento. Também vimos que apesar do conjunto de dados de rastros de táxis não representar de forma completa o comportamento de mobilidade de um centro urbano, ele é capaz de refletir os acontecimentos através da análise de suas características.

5. Modelo de Seleção de Rotas

Nesta seção, introduzimos e validamos um modelo de seleção de rotas capaz de predizer a rota a ser tomada por um motorista, dado um conjunto de rotas possíveis entre um par de pontos de origem e destino, e outras informações contextuais, como data e fluxo de tráfego. Para isso, utilizamos os conhecimentos extraídos da Seção 4 para definir as variáveis que vão afetar as probabilidades de cada rota ser considerada a escolhida para um dado contexto. É válido destacar que o modelo pode ser usado para enriquecer rastros de mobilidade com trajetórias altamente esparsas, i.e., trajetórias onde somente os pontos de origem e destino são conhecidos.

5.1. Definição

Formalmente, dados um ponto de origem o, um ponto de destino d, um dia w, uma intervalo de tempo t e um conjunto de variáveis contextuais $X = \{X_1, \cdots, X_n\}$, um motorista irá selecionar, a partir de um conjunto de rotas possíveis $R = \{r_1, \cdots, r_m\}$, aquela que ele acredita ser a melhor. Neste trabalho, representamos esta seleção através de um modelo que considera as seguintes variáveis para cada rota possível r:

- Otimalidade da trajetória – $\{X_1 \mid X_1 \in \mathbb{R} \text{ and } X_1 \geq 1\}$: o aumento relativo em distância da rota selecionada r em relação ao caminho mais curto entre o e d.
- Tipo do dia – $\{X_2 \mid X_2 \in \{1, 2, 3\}\}$: o tipo do dia no qual a viagem ocorrerá: 1, contemplando o final de semana; 2, os dias que precedem e antecedem o final de semana, i.e., sexta e segunda; e 3, os dias restantes.

Para as variáveis seguintes, assumimos um particionamento do tempo em intervalos. Isto permite a sumarização das características de todas as viagens que ocorreram no intervalo, o que pode ser usado como fonte de informação para os eventos nos próximos intervalos, como visto na Seção 4.1. O tamanho dos intervalos deve ser escolhido considerando a quantidade de viagens disponíveis e a granularidade na qual os eventos no passado afetarão aqueles no presente. Em nossas análises, consideramos intervalos com 8 h de duração: de 0 h às 8 h, de 8 h às 16 h e de 16 h às 0 h.

- Duração média em $t - 1$ – $\{X_3 \mid X_3 \in \mathbb{R} \text{ and } X_3 \geq 0\}$: a duração média das viagens que aconteceram no intervalo anterior na rota r.
- Número de viagens em $t - 1$ – $\{X_4 \mid X_4 \in \mathbb{Z} \text{ and } X_4 \geq 0\}$: o total de viagens que ocorreram no intervalo anterior na rota r.
- Duração média em t – $\{X_5 \mid X_5 \in \mathbb{R} \text{ and } X_5 \geq 0\}$: a duração média das viagens no intervalo atual na rota r (considerando o histórico completo de viagens).

Para representar o impacto dessas variáveis na escolha das rotas, construímos um modelo logit multinomial (MNL), uma solução utilizada na literatura para o problema da seleção de rotas e de análises de demanda de viagens [Vacca and Meloni 2015]. Entretanto, até o presente momento a sua aplicação se restringiu somente a pequenos conjuntos de dados, e a sua validação em grandes conjuntos de dados é uma nova contribuição. Adicionalmente, a escolha das variáveis do modelo com base na caracterização de dados de trajetórias reais também é uma novidade. Assim, baseado nessas variáveis, definimos, para cada rota r, a probabilidade p_r da mesma ser escolhida pelo motorista como:

$$p_r = \frac{\exp(X^T \beta)}{1 + \sum_{R-1} \exp(X^T \beta)} \tag{5}$$

onde $\beta = \{\beta_0, \beta_1, \cdots, \beta_n\}$ representa o conjunto de coeficientes que queremos estimar, e $X = \{X_1, X_2, \cdots, X_n\}$ o conjunto de variáveis observáveis que são usadas como entrada para o modelo. A partir disso, a rota selecionada será aquela com a maior probabilidade calculada. Dado o conjunto de rotas R entre o e d, a distribuição $Y_{w,t}$ de viagens entre R no dia w e no intervalo de tempo t pode ser definida como a distribuição multinomial:

$$Y_{w,t} = \text{Multi}(\{p_1, p_2, \cdots, p_m\}, N) \tag{6}$$

onde p_1 indica a probabilidade da rota r_1 ser escolhida (p_2 a rota r_2, e assim por diante), e N indica o número total de viagens ocorrendo no dia w no intervalo de tempo t.

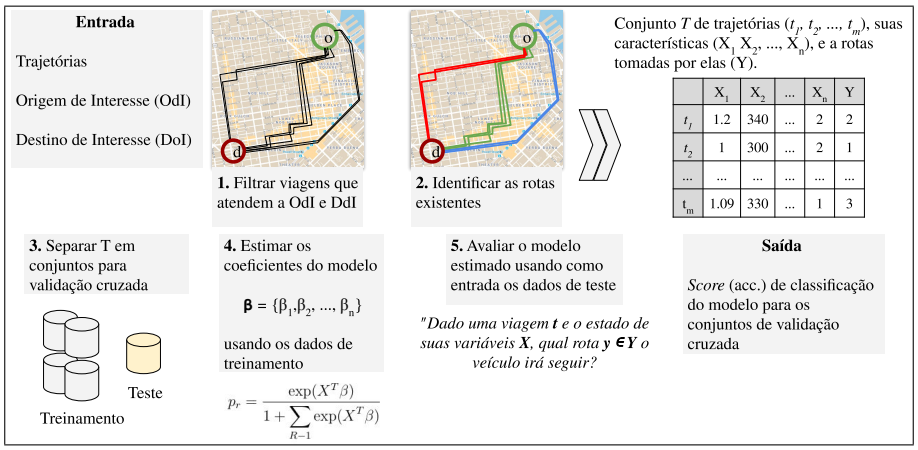

Figura 5. Metodologia do processo usado para validação do modelo

5.2. Validação

Para validar o modelo, extraímos um conjunto de viagens para verificar se o modelo pode replicar a distribuição de rotas existente utilizando somente as características das viagens. Além dos dados de São Francisco, utilizamos também o trace de táxis de Porto [Mendes-Moreira and Moreira-Matias 2015] – coletado em 2014 utilizando 442 veículos durante 1 ano – que possui processo de sensoriamento idêntico àquele utilizado em São Francisco. Ao utilizar esse segundo trace no processo de validação, avaliamos a robustez do modelo ao gerar distribuições de rotas para viagens com características de tráfego diferentes daquelas utilizadas para construir o modelo.

A metodologia do processo de validação do modelo é apresentada na Fig. 5. Foram definidos pontos de origem e destino fixos para extrair viagens similares, i.e., viagens com pontos de origem e destino próximos. Definimos o ponto de origem o como uma região circular de raio fixo, a qual chamamos de Origem de Interesse (OdI). O mesmo foi aplicado para o destino d, o qual chamamos de Destino de Interesse (DdI). Assim, uma viagem t segue os pontos definidos se sua origem ($t[r_0]$) e destino ($t[r_{n-1}]$) estão dentro dos limites dos raios de OdI e DdI, respectivamente. Baseado em testes empíricos, o raio foi definido em 500 m. Assim, duas ou mais viagens são consideradas similares se elas compartilham ambas as regiões de OdI e DdI. As viagens t_i entre o e d que atendem a esse critério compõem o conjunto de viagens filtradas $F_{o,d}$ (passo 1). Então, considerando $F_{o,d}$, extraímos o conjunto de trajetórias em comum. Para isso, aplicamos a técnica de *Longest Common Subsequence* (LCSS) para agrupar trajetórias similares em rotas definidas. As diferentes rotas obtidas entre OdI e DdI compõem o conjunto de rotas $RO_{o,d}$ (passo 2). Em seguida, $RO_{o,d}$ é separado aleatoriamente em um conjunto de treino (80%) e teste (20%) (passo 3). Com o conjunto de treino, o modelo é ajustado assumindo distribuições *a priori* vagas para os coeficientes da regressão $\beta_i \sim N(0, \tau_\beta)$, com $\tau_\beta = 0.01$, onde N representa uma distribuição Normal (passo 4). O ajuste foi feito utilizando os softwares *JAGS* 4.3.0 e *RStudio* (pacote *R2jags*). A partir dos coeficientes estimados ($\hat{\beta}_i$) e das características observadas nos dados do conjunto de teste, a probabilidade de escolha de cada rota é obtida (passo 5). Por fim, a acurácia é computada como a porcentagem de predições corretas do modelo.

Para São Francisco, duas diferentes localidades foram selecionadas como OdI para investigarmos o comportamento em regiões com características diferentes. Em cada uma,

Tabela 1. Resultados da validação

Trace	Origem	Destino	# de viagens	# de rotas	Acc.	B1 Acc.	B2 Acc.
São Francisco	South of Market	Pacific Heights	1111	3	**92,72**	23,52	64,70
		The Castro	1426	3	**92,38**	87,75	87,75
		Nob Hill	1708	3	**91,90**	22,18	77,22
		Russian Hill	1161	3	**91,70**	29,16	37,50
		North Beach	1267	3	**87,19**	14,85	62,37
		Filmore	912	2	**86,26**	34,62	65,38
		Fisherman's Wharf	1632	2	**85,27**	27,43	72,57
		Haight-Ashbury	1772	3	**84,99**	21,73	43,47
	Financial District	Pacific Heights	1047	2	**88,74**	65,33	65,33
		Fisherman's Wharf	1069	2	**84,40**	26,39	73,61
		Marina District	908	3	79,93	17,39	**82,60**
		Russian Hill	872	3	**75,87**	19,14	65,95
Porto	Aeroporto de Porto	Centro da cidade	2012	4	**97,80**	6,75	59,45

selecionamos um conjunto de DdI. Para Porto, selecionamos como OdI o aeroporto da cidade, e como DdI o centro da cidade, dado o número considerável de viagens existentes entre ambos e o alto número de rotas disponíveis.

Além de apresentarmos os resultados de acurácia obtidos pelo modelo (Tabela 1), apresentamos também os resultados de duas outras abordagens como comparação. A primeira delas (*baseline* $B1$) assume que o motorista sempre seguirá a rota que apresenta o caminho mais curto. Apesar de ingênua em sua concepção, sabe-se que muitos motoristas têm como preferência a rota mais curta [Domingues et al. 2018, Liu et al. 2012a]. A segunda (*baseline* $B2$) assume que o motorista irá sempre selecionar a rota mais frequentemente usada, – considerando todas as viagens de todos os motoristas – por acreditar que a popularidade da rota é devido a sua otimalidade [Chen et al. 2011].

Para as viagens partindo da OdI na região do *South of Market*, 8 diferentes DdI foram considerados. Variando de $84{,}99\%$ a $92{,}72\%$ com uma média geral de 89% de acurácia, os resultados mostram que o modelo foi capaz de recomendar com sucesso a melhor rota a ser tomada dadas as variáveis contextuais de cada viagem. Como comparação, $B1$ apresentou acurácias entre $14{,}85\%$ e $87{,}75\%$, com uma média de $32{,}65\%$. Adicionalmente, $B2$ apresentou acurácias de $37{,}50\%$ até $87{,}75\%$, com média de $63{,}87\%$. Em seguida, analisamos os resultados das viagens que partem da OdI na região do *Financial District*, considerando quatro DdI diferentes. Aqui, as acurácias do modelo variam entre 75% a 88%, com uma média de $82{,}23\%$. O *baseline* $B1$ varia de $17{,}39\%$ a $65{,}33\%$, com acurácia média de $32{,}06\%$. O *baseline* $B2$, por sua vez, varia de $65{,}33\%$ a $82{,}60\%$, com acurácia média de $71{,}87\%$.

Para o trace de Porto, o modelo obteve uma acurácia de $97{,}8\%$, reiterando a robustez do mesmo em comparação aos *baselines*, que obtiveram acurácias de $6{,}75\%$ para $B1$ e $59{,}45\%$ para $B2$. Podemos atribuir essa diferença considerável ao fato das rotas encontradas possuírem grandes diferenças entre suas características, tornando mais fácil a identificação por parte do modelo das motivações para as escolhas feitas pelos motoristas.

Os resultados mostram que o modelo construído é capaz de capturar os fatores latentes do processo de seleção de rotas. Vimos como os resultados foram robustos, obtendo alta acurácia considerando os diferentes pares de OdI e DdI. É visível que, mesmo apresentando resultados satisfatórios em alguns cenários, ambos os *baselines* se comportam de forma inferior ao modelo. Por fim, apesar de demandar dados de treinamento para predizer rotas entre um par de OdI e DdI, é possível aplicar o modelo em rastros de origem-destino, estimando os parâmetros em um cenário com características semelhantes.

6. Conclusão

Neste trabalho, apresentamos uma abordagem genérica para preparar e caracterizar rastros de mobilidade veicular, visando entender quais são os fatores que impactam a mobilidade e como podemos identificá-los através de dados de mobilidade, e.g., viagens de veículos públicos como táxis. Especificamente, buscamos entender as variáveis que influenciam na seleção de uma rota por um motorista. Para isso, um modelo logit multinomial foi definido através de características extraídas das viagens. Os resultados da validação mostraram que o mesmo foi capaz de capturar os fatores que influenciam na decisão dos motoristas. Assim, o modelo proposto surge como opção interessante para o preenchimento de grandes lacunas espaciais em dados de trajetórias, através da geração de viagens sintéticas realísticas que levam em consideração fatores contextuais. Através desse preenchimento, rastros de mobilidade de baixa granularidade – como aqueles de origem-destino – podem ser enriquecidos, tornando-se fontes úteis no desenvolvimento e na simulação de tecnologias baseadas em mobilidade. Trabalhos futuros incluem a avaliação do modelo considerando características obtidas de fontes externas, como incidentes nas vias, e a validação em rastros de origem-destino e em outros cenários veiculares, por exemplo contendo veículos particulares, dadas as limitações do cenário de táxis.

Agradecimentos

Este trabalho contou com o apoio da CAPES, CNPq, Fapemig e Fapesp projetos #1524494-8 & 18/23064-8.

Referências

[Celes et al. 2017] Celes, C., Silva, F. A., Boukerche, A., de Castro Andrade, R. M., and Loureiro, A. A. (2017). Improving vanet simulation with calibrated vehicular mobility traces. *Transactions on Mobile Computing, 2017*, 16(12):3376–3389.

[Chen et al. 2011] Chen, Z., Shen, H. T., and Zhou, X. (2011). Discovering popular routes from trajectories. In *IEEE 27th Int. Conf. on Data Engineering*, pages 900–911.

[Cotta et al. 2017] Cotta, L., de Melo, P. O. V., and Loureiro, A. A. (2017). Understanding the role of mobility in real mobile ad-hoc networks connectivity. In *Symp. on Computers and Communications*, pages 1098–1103. IEEE.

[Domingues et al. 2018] Domingues, A. C., Silva, F. A., and Loureiro, A. A. (2018). Space and time matter: An analysis about route selection in mobility traces. In *Symp. on Computers and Communications, 2018*. IEEE.

[Harri et al. 2009] Harri, J., Filali, F., and Bonnet, C. (2009). Mobility models for vehicular ad hoc networks: a survey and taxonomy. *IEEE Communications Surveys & Tutorials*, 11(4).

[Hess et al. 2016] Hess, A., Hummel, K. A., Gansterer, W. N., and Haring, G. (2016). Data-driven human mobility modeling: a survey and engineering guidance for mobile networking. *ACM Computing Surveys*, 48(3):38.

[Hoteit et al. 2016] Hoteit, S., Chen, G., Viana, A., and Fiore, M. (2016). Filling the gaps: On the completion of sparse call detail records for mobility analysis. In *Proceedings of the Eleventh ACM Workshop on Challenged Networks*, pages 45–50.

[Hoteit et al. 2014] Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., and Pujolle, G. (2014). Estimating human trajectories and hotspots through mobile phone data. *Computer Networks*, 64:296–307.

[Kotz et al. 2009] Kotz, D., Henderson, T., Abyzov, I., and Yeo, J. (2009). CRAWDAD dataset dartmouth/campus (v. 2009-09-09). Downloaded from https://crawdad.org/dartmouth/campus/20090909.

[Li et al. 2015] Li, M., Ahmed, A., and Smola, A. J. (2015). Inferring movement trajectories from gps snippets. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 325–334. ACM.

[Liu et al. 2012a] Liu, S., Liu, C., Luo, Q., Ni, L. M., and Krishnan, R. (2012a). Calibrating large scale vehicle trajectory data. In *Proceedings of the 13th IEEE International Conference on Mobile Data Management*, pages 222–231. IEEE.

[Liu et al. 2012b] Liu, Y., Kang, C., Gao, S., Xiao, Y., and Tian, Y. (2012b). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of geographical systems*, 14(4):463–483.

[Lu et al. 2019] Lu, X., Cheng, W., Shen, Y., and Zhu, Y. (2019). Ladd: A length-adaptive approach to detecting taxi anomalous detours. In *25th International Conference on Parallel and Distributed Systems*, pages 141–144. IEEE.

[Mendes-Moreira and Moreira-Matias 2015] Mendes-Moreira, J. and Moreira-Matias, L. (2015). On learning from taxi-gps traces. In *Proceedings of the 2015th International Conference on ECML PKDD Discovery Challenge-Volume 1526*, pages 37–39.

[Newson and Krumm 2009] Newson, P. and Krumm, J. (2009). Hidden markov map matching through noise and sparseness. In *17th SIGSPATIAL Int'l Conf. on advances in geographic information systems*, pages 336–343. ACM.

[Piorkowski et al. 2009] Piorkowski, M., Sarafijanovic-Djukic, N., and Grossglauser, M. (2009). Crawdad dataset epfl-mobility (v. 2009-02-24). Downloaded from https://crawdad.org/epfl/mobility/20090224.

[Silva et al. 2015] Silva, F. A., Celes, C., Boukerche, A., Ruiz, L. B., and Loureiro, A. A. (2015). Filling the gaps of vehicular mobility traces. In *18th Int'l Conf. on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 47–54. ACM.

[Su et al. 2015] Su, H., Zheng, K., Huang, J., Wang, H., and Zhou, X. (2015). Calibrating trajectory data for spatio-temporal similarity analysis. *Intl' Journal on Very Large Data Bases*, 24(1):93–116.

[Teixeira et al. 2021] Teixeira, D. D. C., Viana, A. C., Almeida, J. M., and Alvim, M. S. (2021). The impact of stationarity, regularity, and context on the predictability of individual human mobility. *ACM Tran. on Spatial Algorithms and Systems*, 7(4):1–24.

[Vacca and Meloni 2015] Vacca, A. and Meloni, I. (2015). Understanding route switch behavior: An analysis using gps based data. *Transportation Res. Procedia*, 5:56–65.

[Yoon et al. 2006] Yoon, J., Noble, B. D., Liu, M., and Kim, M. (2006). Building realistic mobility models from coarse-grained traces. In *4th Int'l Conf. on Mobile Systems, Applications and Services*, pages 177–190. ACM.

[Yuan and Li 2021] Yuan, H. and Li, G. (2021). A survey of traffic prediction: from spatio-temporal data to intelligent transportation. *Data Science and Engineering*, 6(1):63–85.

[Yuan et al. 2010] Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., and Huang, Y. (2010). T-drive: driving directions based on taxi trajectories. In *18th SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems*, pages 99–108. ACM.

[Zhu and Levinson 2015] Zhu, S. and Levinson, D. (2015). Do people use the shortest path? an empirical test of wardrop's first principle. *PloS one*, 10(8):e0134322.