CONEXÃO UNIFAMETRO 2024
XX SEMANA ACADÊMICA
ISSN: 2357-8645

RECONSTRUÇÃO DOS LOCAIS DE EMBARQUE A PARTIR DA BASE DE DADOS DE TRANSPORTE PÚBLICO UTILIZANDO INTELIGÊNCIA ARTIFICIAL

Kaio Gefferson de Almeida Mesquita
Docente - Centro Universitário Fametro - Unifametro
kaio.mesquita@professor.unifametro.edu.br

Área Temática: Tecnologia em Engenharia de Tráfego e Transporte
Área de Conhecimento: Ciências Tecnológicas
Encontro Científico: V Encontro de Experiências Docentes

RESUMO

O objetivo desse artigo foi apresentar um método de modelagem utilizando técnicas de aprendizado supervisionado e não supervisionado para identificar os locais onde os usuários embarcam no Sistema Integrado de Transporte Público de Fortaleza. A metodologia inclui etapas de extração, transformação e carregamento de informações provenientes dos dados de bilhetagem eletrônica, GPS, GTFS e cadastro de usuários, análises exploratórias e modelagem através de aprendizado de máquina. Além disso, foi definido atributos relacionados aos padrões de uso como frequência das validações e intervalos temporais e utilizados na modelagem supervisionada da distância de validação (variável alvo). De acordo com os resultados, verificou-se que a previsão do local de embarque teve alto desempenho nos modelos supervisionados, sendo a Floresta Aleatória o algoritmo que obteve os melhores indicadores de precisão com uma acurácia de 92%. Posteriormente, a partir da análise de agrupamentos, foram identificados dois tipos diferentes de usuários: os frequentes e os esporádicos. Apresentou-se divergências nos padrões de validação entre eles. O uso de big data e aprendizado de máquina é essencial para melhorar a gestão e operação dos sistemas de transporte público, como apresentado pela importância dos atributos temporais e espaciais na inferência precisa dos locais de embarque.

Palavras-chave: Bilhetagem Eletrônica; Big Data; Padrão de deslocamento; Aprendizado de máquina.

INTRODUÇÃO

Arbex e Cunha (2020) afirmam que as pesquisas domiciliares são normalmente utilizadas para identificar os padrões de deslocamento. Por outro lado, realizar essas coletas manualmente demanda um tempo e custos significativos. Na década de 1990, sistemas de pagamentos com cartões inteligentes chamados Sistemas de Bilhetagem Eletrônica (SBE) foram adotados em algumas cidades. Esses sistemas permitiram a coleta da tarifa usando cartões eletrônicos e leitores instalados nos veículos de forma automática. Entretanto, uma vez que o SBE não tem a finalidade de coletar dados de viagens, é preciso realizar várias análises para

ISSN: 2357-8645

inferir informações sobre atributos das viagens como origem, destino e propósito, além de obter matrizes OD em diferentes níveis espaciais agregados (Kurauchi e Schmöcker, 2017; Hussain *et al*. 2021). É importante destacar que a validação nem sempre acontece durante o embarque, especialmente em sistemas em que a tarifa também é cobrada por meio de catracas no veículo ou através de validadores distribuídos dentro dele (Zhao, 2023; Arbex e Cunha, 2020). Nas cidades brasileiras, há um espaço de acomodação entre a porta de embarque e a catraca, o qual pode impactar na validação do comportamento em diversas condições de ocupação veicular (Arbex e Cunha, 2020). Em Fortaleza, estudo de caso desse artigo, o SBE é também do tipo aberto, em que o pagamento pode ser efetuado durante o percurso. Algumas características operacionais do sistema e de configuração da rede podem afetar o comportamento de validação. Primeiro, o sistema permite integração tarifária dentro de um intervalo de duas horas entre validações, o que pode induzir a realização de validações próximo ao destino das viagens. Segundo, a configuração da rede é tronco-alimentada, em que linhas alimentadores conectam os bairros a sete terminais de integração física, de onde partem linhas troncais para área central de comércio e serviços da cidade, localizada ao norte do município.

Diante do exposto, o objetivo deste artigo é desenvolver um método de modelagem para identificar e reconstruir os locais de embarque, a partir de atributos sobre o modo de uso do Sistema Integrado de Transporte Público de Fortaleza (SIT-FOR) utilizando técnicas supervisionadas de aprendizado de máquina. É válido destacar que foi utilizado um Big Data com informações deste sistema e modelagem não-supervisionada para avaliar uma possível formação de grupos dentre esses dados e verificação do comportamento da variável alvo (apresentada posteriormente) nos grupos encontrados (Shalit *et al*., 2022; Tang *et al*., 2023).

METODOLOGIA

O método deste trabalho está representado na Figura 1, sendo dividido em quatro macro etapas: (i) Extração, transformação, carregamento dos dados e delimitação da arquitetura do banco de dados; (ii) Definição dos atributos relacionados aos padrões de uso; (iii) Modelagem supervisionada da distância de validação em rota; e (iv) Análise dos agrupamentos com modelagem não-supervisionada. Foram utilizadas quatro bases de dados do ano de 2018 para este estudo. Utilizou-se dados de 6 meses típicos em 2018. A principal base utilizada foi a de Bilhetagem Eletrônica que contém todos os registros diários de validação dos usuários, incluindo a hora de validação e linha utilizada por cada usuário. Dados da programação do serviço ofertado pelo SIT-FOR em 2018 no formato GTFS – General Transit Feed Specification

ISSN: 2357-8645

– foram utilizados para identificar os locais de embarque para os usuários do cadastro com endereço válido e que usaram o sistema em 2018. As localizações geográficas a cada 30 segundos dos veículos da frota durante a operação por meio de equipamentos de GPS – *Global Positioning System* – foram utilizadas para identificar as coordenadas geográficas das validações no SBE, já que os dados de GPS e do SBE não são integrados em Fortaleza.

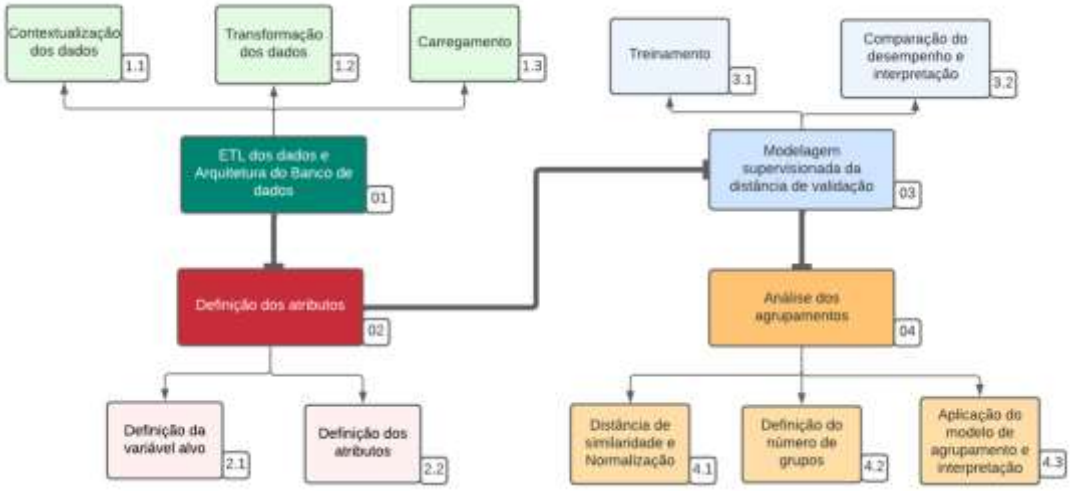


Figura 1 – Proposta metodológica
Fonte: Autores

A principal informação do Cadastro dos Usuários para esta pesquisa foi o endereço dos usuários. Vale ressaltar, que devido a confidencialidade dos dados de cadastro, informações pessoais dos usuários não foram fornecidas pelo órgão gestor do sistema. Os endereços completos atualizados em 2018 foram georreferenciados utilizando uma API do Google Maps. Como os endereços registrados podem estar desatualizados, considerou-se como endereços válidos aqueles cujas distâncias euclidianas à parada mais próxima da linha mais frequente na primeira validação do dia do usuário sejam menores do que 1000 m, assumindo-se como distância máxima que um usuário estaria disposto a caminhar.

A principal hipótese deste trabalho é que os atributos relacionados ao modo de uso do sistema podem auxiliar na identificação de características dos deslocamentos, como o local de embarque (Shalit *et al*., 2022; Tang *et al*., 2023). Para verificar a hipótese central deste trabalho foi definida a variável target de distância de validação em rota, sendo definida como a distância em rota em quilômetros entre a parada mais próxima da residência (considerando a linha mais frequente na primeira validação do dia) e o centroide de validações considerando 6 meses de uso (junho, julho, agosto, setembro, outubro e novembro de 2018).

Com base nos dados disponíveis e nas hipóteses de estudo da revisão da literatura, foram definidos os atributos para modelagem da variável target e posterior identificação dos

ISSN: 2357-8645

agrupamentos que podem ajudar a entender os modos de uso do sistema. Acredita-se que bons preditores para a variável target devem incorporar aspectos temporais e espaciais, relacionados à frequência de uso do sistema e oferta da rede (Morency *et al*., 2007; Cats e Ferranti, 2022; Zhao *et. al*., 2023). Assim, definiram-se os seguintes atributos, conforme apresentado na Tabela 1. Vale ressaltar que atributos de nível de serviço relacionados a lotação não foram considerados neste estudo, devido a limitação nos dados.

Tabela 1 - Resumo das Descrição dos aspectos, atributos e unidades para modelagem

Atributo	Unidade	Aspectos	Definição
Atributos I1-I5 (FREQ _DIA): Frequência média diária de validações (Segunda à Sexta)	N° de validações/ dia	Frequência de uso do sistema	Estes atributos correspondem à média de validações por dia útil, sendo calculados separadamente para a amostra de usuários válidos do cadastro (com endereços válidos).
Atributo I6 (FRQ_VAL_TERMINAL): Validações próximas aos terminais	N° de validações/ dia	Oferta em termos de itinerário, conectividade da rede.	Este atributo corresponde ao número de vezes em média que o usuário utiliza o terminal. Este atributo é calculado como frequência média diária de validações na região formada pela zona que contém o terminal de integração e suas zonas vizinhas.
Atributos I7-I11 (INTERVALO _DIA): Intervalo temporal entre validações (Segunda à Sexta)	Horas	Tipo de atividade e Duração da atividade	Estes atributos correspondem aos intervalos médios em horas, entre as primeiras e últimas validações do dia para cada um dos dias da semana.
Atributos I12-I13 (FAIXA_HORARIA): Período de validação de pico e entre-pico para as primeiras e últimas validações	Categórica com 5 classes	Tipo de atividade e horário de realização das atividades	Estes períodos foram definidos conforme os volumes de validações horárias de um dia típico, nas seguintes classes: Classe 0 (fora pico) de 0h à 6h, Classe 1 (Pico da Manhã) de 7h à 9h; Classe 2 (Entre Pico) de 8h à 17h; Classe 3 (Pico da Tarde) de 17h à 19h; Classe 4 (Madrugada) de 19h à 0h.
Atributos I14-I17 (PROP_LINHA): Proporção de validações por tipo de linha (Alimentadora, Troncal, Convencional e Complementar).	Percentual	Aspectos operacionais de conectividade da rede e nível de serviço da oferta	A partir destes atributos busca-se evidenciar a influência de aspectos operacionais da tipologia das linhas do sistema no padrão de uso dos usuários. Ele é obtido pelo número de validações por tipo de linha em relação ao total de validações do período de 6 meses.

Fonte: Autor

Nesta etapa, especificou-se os modelos supervisionados para prever a distância de validação em rota (variável target) da parada mais próxima a residência ao centroide de validação. Conforme será disposto, dado que as distâncias de validação apresentam uma elevada variabilidade e que diferentes aspectos podem estar envolvidos numa rede de transportes que afetam esta variável, adotou-se os seguintes modelos supervisionados fundamentados em aprendizado de máquina: Árvore de Decisão (AD), Floresta Aleatória (FA) e Rede Neural (RD) (Géron, 2019).

RESULTADOS E DISCUSSÃO

Primeiramente se observou a distribuição da variável target, conforme apresentado na Figura 2. É possível observar uma distribuição do tipo assimétrica à direita, com média das distâncias em torno de 5,2 km, podendo assumir valores maiores chegando até 22 km. Essa distribuição mostra que a maioria dos usuários tende a validar distante do local de embarque,

ISSN: 2357-8645

já que 95% validaram a uma distância maior do 1km, correspondendo em média a mais do que 2 paradas do local de embarque. Possivelmente a lotação dentro dos veículos em horários de pico influencia nesse comportamento do local de validação ser mais distante do embarque, ou até mesmo uma necessidade de se esperar chegar próximo ao terminal para realizar uma integração.

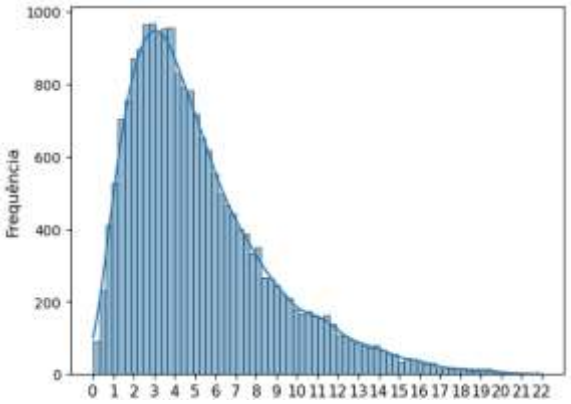


Figura 2 –Distribuição da distância em rota do local embarque ao cluster de validações (km)
Fonte: Autore

Tabela 2 - Indicadores de precisão dos modelos supervisionados

Modelo	Indicador	Resultado
Floresta Aleatória	MSE	0,0004
	RMSE	0,0206
	MAE	0,001
	R^2	0,99
Rede Neural	MSE	0,0021
	RMSE	0,0460
	MAE	0,0339
	R^2	0,97
Árvore de decisão	MSE	0,0008
	RMSE	0,0290
	MAE	0,0024
	R^2	0,99

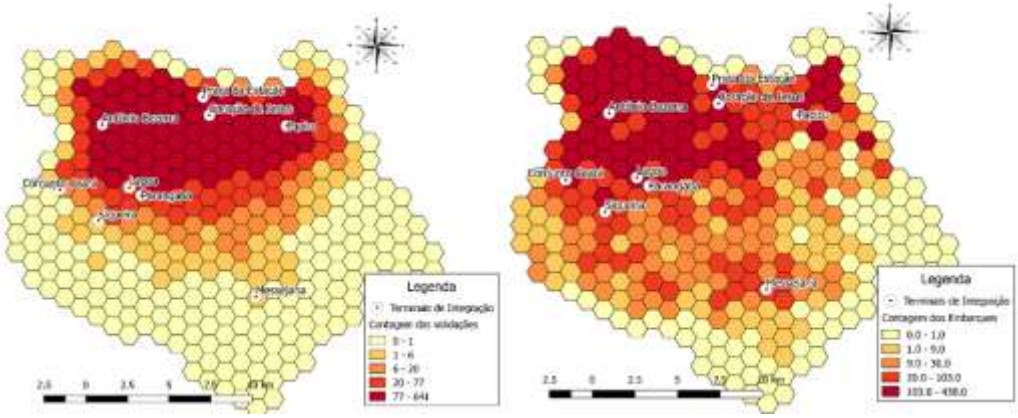
Fonte: Autor

Através da análise dos indicadores de desempenho dos modelos (Tabela 2) ficou evidente que os atributos se aderirem bem a variável *target* com R^2 acima de 0,95 indicando uma baixa variabilidade entre o modelado e os valores reais. Dessa forma há evidências que os atributos de frequência, intervalo temporal e uso do sistema auxiliam na identificação de informações faltantes de demanda, como o real local de embarque. Avaliando os indicadores de erro médio e erro quadrático médio o modelo de floresta aleatória obteve os menores erros indicando que a formulação por árvores de decisão possa ser mais adequada para o tipo de dado analisado, uma vez que não se conhece as relações entre as variáveis em estudo. Com este modelo, obteve-se um o erro médio absoluto de 10 m, o que indica um alto desempenho para prever o real local de embarque, considerando que a distância média entre paradas da rede é de até 550m.

ISSN: 2357-8645

Por fim, para o grupo de usuários de teste (30% da base, ou 6181 usuários) foi aplicado o modelo de Floresta Aleatória (por obter os melhores indicadores de desempenho) para prever a distância da primeira validação diária em rota. Para os casos em que a distância prevista ultrapassava a distância em rota disponível para realizar a subtração, foi considerado como origem a primeira parada da linha. A Figura 3 resume, portanto, todo o esforço aplicado na metodologia proposta neste trabalho. Ela apresenta em um primeiro momento os valores agregados por zona dos locais de validação, sendo estes considerados na maioria dos estudos de transporte público como sendo o real local de embarque. É possível visualizar que para essa amostra de usuários, existe uma maior concentração a noroeste da cidade indo no sentido do centro comercial. As regiões periféricas praticamente não apresentam influência neste mapa, o que é contraditório, pois a maioria dos usuários do transporte público de Fortaleza residem nas regiões periféricas da cidade, indicando que quase nenhum usuário valida ao embarcar. Enquanto no mapa a direita, considerando os locais de embarque previstos, ainda existe uma forte concentração a noroeste, porém os embarques estão mais distribuídos pela rede, chegando até as regiões periféricas, onde os usuários cativos do sistema podem realizar validações próximas as suas residências ou terminais de integração nessas regiões.

Figura 3 – Comparação espacial entre os (a) locais de validação e (b) locais de embarque



Fonte: Autor

Avaliando separadamente os grupos frequente (grupo 1) e esporádico (grupo 0), foi verificado por zona a diferença média ao se considerar a validação como real local de embarque, obtendo-se um erro agregado de 88% e 113%, para os grupos 0 e 1, respectivamente. Esse erro representa o quanto em média as viagens estão contabilizadas erradas por zona. O valor acima de 100% indica que em algumas zonas o erro chega a ser superior em mais de 2x do que foi considerado originalmente como local de embarque. Comparando os locais previstos em nível de zona com as zonas estimadas pela parada mais próxima a residência, o modelo obteve uma acurácia de 92%.

ISSN: 2357-8645

CONSIDERAÇÕES FINAIS

A análise dos locais de validação em Fortaleza mostrou que uma parcela considerável de usuários (mais de 90% da amostra) não valida no momento do embarque, o que impede a aplicação direta de técnicas de encadeamento para prever destinos e transferências das viagens, e que foi uma das hipóteses verificadas neste estudo. De modo geral a hipótese central do trabalho de que os atributos relacionados ao comportamento de uso de sistema, como frequência de uso por horário e tipo de linha, auxiliam na identificação de características do deslocamento e permitem prever os locais de embarque, foi verificada a partir do alto desempenho dos modelos de previsão (AD, FA e RN) adotados neste estudo. Os modelos indicaram que atributos relacionados a oferta do serviço, à regularidade de uso do sistema, e aos horários das atividades tem uma maior influência no comportamento de validação. Algumas direções para estudo futuros, podem ser a investigação de atributos relacionados ao nível de serviço das linhas no momento dos embarques, como a lotação dos veículos, assim como investigar atributos relacionados a características dos usuários ou das viagens.

REFERÊNCIAS

Arbex, R. O. e C. B. Cunha (2020) **Estimating the influence of crowding and travel time variability on accessibility to jobs in a large public transport network using smart card big data**, Journal of Transport Geography, v. 85, ISSN 0966-6923, https://doi.org/10.1016/j.jtrangeo.2020.102671.

Cats, O. e F. Ferranti (2022) **Unravelling individual mobility temporal patterns using longitudinal smart card data, Research in Transportation Business & Management**, v. 43, ISSN 2210-5395, DOI: https://doi.org/10.1016/j.rtbm.2022.100816.

Géron, A. (2019) Hands-On **Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems**. O'Reilly Media. v 1. 856 p.

Hussain, E.; A. Bhaskar e E. Chung (2021) **Transit OD Matrix Estimation Using Smartcard Data: Recent Developments and Future Research Challenges**. Transportat Research, v. 125. DOI: 10.1016/j.trc.2021.103044.

Kurauchi, F. e J. D. Schomocker (2017) **Public transport planning with smartcard data** (1ª ed.). Boca Raton: CRC Press.

Morency, C.; M. Trépanier e B. Agard (2007) **Measuring transit performance using smart card data. Presented at World** Conference on Transport Research, San Francisco, USA.

Shalit, N.; M. Fire e E. Ben-Elia (2022) **A supervised machine learning model for imputing missing boarding stops in smart card data**. Public Transport, v. 15, p. 287–319. DOI: https://doi.org/10.1007/s12469-022-00309-0.

Tang, T.; R. Liu; C. Choudhury; A. Fonzone e Y. Wang (2023) **Predicting hourly boarding demand of bus passengers using imbalanced records from smart-cards: a deep learning approach**. IEEE Transactions on Intelligent Transportation Systems, v. 24, n. 5, p. 5105 - 5119. DOI: 10.1109/TITS.2023.3237134.

Zhao, X.; M. Cui e D. Levinson (2023) **Exploring temporal variability in travel patterns on public transit using big smart card data**. Environment and Planning B: Urban Analytics and City Science, v. 50, n. 1, p. 198–217. DOI: https://doi.org/10.1177/23998083221089662.