

# Análise dos Padrões Espaço-Temporais de Validação em Sistemas de Transporte Público: Uma Abordagem Baseada em Mineração de Dados

Kaio G. de A. Mesquita, Francisco M. de O. Neto, Vandeyberg N. de Souza,  
Luan P. de H. Barros

Universidade Federal do Ceará – CE, Brasil

{kaio,moraes,vandeyberg}@det.ufc.br, luanpablodehb@alu.ufc.br

**Abstract.** *Understanding the spatio-temporal patterns present in the behavior of users of a public transport system is essential for a more assertive planning about the needs of the population. In this sense, the present article aims to diagnose the patterns of temporal and spatial displacements of users of the public transport network in Fortaleza, which is configured as open, tap-on and trunk-powered, based on usage data generated prior to pandemic. For this, validation data from November 2018 were used. As a method of diagnosing the patterns, Machine Learning and Statistical Learning models were mainly used in an approach based on Data Mining. As a result, it was possible to observe that the users' first daily validations have a higher percentage of both spatial and spatio-temporal patterns when compared to the last daily validations. In addition, users rarely present spatial and temporal patterns simultaneously, requiring individual attention to both.*

**Resumo.** *A compreensão dos padrões espaço-temporais presentes no comportamento dos usuários de um sistema de transporte público é fundamental a um planejamento mais assertivo acerca das necessidades da população. Nesse sentido, o presente artigo tem como objetivo diagnosticar os padrões de deslocamentos temporal e espacial dos usuários da rede transporte público de Fortaleza, o mesmo configurado como aberto, tap-on e tronco-alimentado, apoiando-se em dados gerados previamente à pandemia. Para tal, foram utilizados dados de validações de novembro de 2018. Como método de diagnóstico dos padrões, os modelos de Machine Learning e Statistical Learning foram utilizados numa abordagem baseada em Data Mining. Como resultados, foi possível observar que as primeiras validações diárias dos usuários possuem um maior percentual de padrão tanto espacial quanto espaço-temporal quando comparadas às últimas validações diárias. Além de que dificilmente os usuários apresentam padrão espacial e temporal simultaneamente, sendo necessário uma atenção individual para ambos.*

## 1. Introdução

Ortúzar e Willumsem (2011) discutem o papel do planejamento de transportes, onde a oferta do sistema deve ser condizente com a demanda, e, portanto, compreender o

padrão de deslocamentos dos usuários e como essa demanda varia no tempo pode auxiliar na proposição de um planejamento da oferta (Mesquita *et al.*, 2017). Esses padrões podem variar com a influência de fatores como: tipo de atividade realizada pelo usuário, rede de Transporte Público (TP), conhecimento da rede pelo usuário e seus hábitos. Ao conhecer esses padrões e a distribuição na demanda é possível realizar análises e tomar decisões mais assertivas na oferta e nível de serviço do sistema, como: definir as zonas mais produtoras e atratoras de viagens, prever variações ao longo do dia e entre dias, definir custos futuros, propor valores de tarifas adequados às necessidades dos usuários e dispor informações de qualidade em tempo real para os mesmos, amparando-se nos sistemas da informação disponíveis (Hora *et al.*, 2017).

A análise da demanda de sistemas de transporte público avançou significativamente a partir do final da década de 1990, quando os sistemas de pagamento com cartão inteligente foram incorporados aos Sistemas de TP em cidades como Washington D.C. e Tóquio, também conhecidos como *Automated Fare Collection* – AFC (Sistema de Bilhetagem Eletrônica – SBE), permitindo o pagamento da tarifa (i.e., validação da viagem) através de *Smart Cards* e equipamentos de leitura instalados nos veículos (Zhao *et al.*, 2007; Pelletier *et al.*, 2011; Munizaga e Palma, 2012). A Tarifação evoluiu de sistemas fechados, em terminais físicos, para sistemas abertos (com possibilidade de validar em paradas da rede fora dos terminais de integração), garantindo maior acessibilidade do usuário. Além do SBE, muitas cidades no mundo vêm adotando também sistemas *Automatic Vehicle Location* (AVL), compostos por *Global Positioning System* (GPS), para localização em tempo real dos veículos, tendo um aspecto logístico, mas que atualmente vem sendo utilizado para compreensão da variabilidade da oferta e da demanda do TP. Outros sistemas da informação citados na literatura e que ganharam notoriedade na última década foram a Especificação Geral de Feeds de Trânsito (GTFS) e contadores automáticos (APC). Em muitos sistemas urbanos, incluindo-se o de Fortaleza, espera-se uma variabilidade nos padrões de deslocamento (itinerários, horários, destinos e transferências) que pode afetar a forma de utilização do sistema e de pagamento da tarifa. Em especial, os fatores que podem contribuir são o conhecimento sobre o sistema (usuários regulares), nível de serviço (lotação e atrasos), o propósito, o uso de outros modos, dentre outros aspectos (Mesquita; Neto, 2021).

Cheng *et. al* (2021) propuseram um método no qual caracterizaram e descreveram duas categorias de usuários, os regulares e os irregulares, apontando que o comportamento na escolha do trajeto poderia e deveria ser modelado de maneiras distintas. Dessa forma, o que configura um padrão de deslocamentos usualmente são as viagens que saem de uma zona de origem para uma zona de destino com elevada frequência, muitas vezes através de um mesmo conjunto de linhas, com tipos de usuários, horários e motivos que se repetem em uma larga escala temporal. Essa caracterização, embora se mostre essencial para definição do método de modelagem e compreensão da demanda, é muitas vezes negligenciada. Vale ressaltar que usuários regulares são diferentes de usuários que apresentam padrões de deslocamentos bem definidos (Cats e Ferranti, 2022). Embora essa diferença não esteja clara na literatura, neste trabalho será testado a hipótese de que mesmo o usuário, não sendo regular no que diz respeito ao uso do sistema, pode apresentar um padrão ou conjunto de padrões em pelo menos um dia específico da semana, por conta do seu tipo de atividade.

Embora na literatura existam trabalhos com foco na predição, os mesmos não apresentam uma preocupação em definir e identificar os padrões e os fatores que os afetam, surgindo a necessidade de aprofundar-se em como os sistemas funcionam, além de avaliar o comportamento dos usuários no espaço e no tempo durante uma série de validações. Assim, este trabalho parte da premissa de que a compreensão do Sistema de TP em análise, desde sua contextualização aos padrões “invisíveis” presentes nos dados, pode auxiliar na modelagem da reconstrução das viagens e gerenciamento do Sistema de TP. Em Fortaleza, alguns percalços existentes no sistema, que dificultam as análises desses padrões são: (i) não se sabe o real local de embarque nem o local de destino; (ii) não há informação se a validação configura uma transferência ou atividade curta; (iii) não se sabe o motivo da viagem; e (iv) não há estudos que se preocupem com o tratamento adequado dos dados antes das análises exploratórias (Mesquita; Neto, 2021). Perante o exposto, apresenta-se a questão motivadora do trabalho: Como definir e identificar os padrões espaço-temporais em Sistemas abertos (possibilidade de validação e transferência fora de um terminal de integração), *tap-on* (validação logo após o embarque, mas não no desembarque) e tronco-alimentadores (linhas trocais que interligam os terminais ao centro comercial, alimentadas por linhas que ligam os bairros aos terminais) de Transporte Público?

O objetivo deste trabalho é identificar o padrão de deslocamento temporal e espacial dos usuários da rede de Transporte Público de Fortaleza, configurado como *tap-on*, aberto e tronco-alimentador. Portanto será contextualizado e definido o que configura um usuário como regular e irregular, bem como os tipos de padrões encontrados para validação das hipóteses levantadas, variando principalmente por tipo de usuário (influência da atividade que será realizada) e horário frequente do deslocamento, além da hipótese de que existe um centroide de validação que representa os pontos frequentes de acesso ao sistema pelo usuário e que os mesmos podem ser tanto identificados quanto utilizados para reconstruir a cadeia de viagens. Para viabilizar o diagnóstico, técnicas de mineração de dados utilizando *k-means*, algoritmos de aprendizado de máquina, bibliotecas *scikit-learn*, *numpy*, *pandas* e *tensorflow* foram reformuladas para o tipo de problema citado.

Dessa forma, a estruturação do trabalho está dividida em mais 4 seções, sendo elas: (i) Mineração de dados para análise da demanda de transporte público; (ii) Proposta metodológica para diagnóstico dos padrões; (iii) Aplicações e resultados; e (iv) Considerações finais.

## 2. Data Mining para Análise da Demanda de Transporte Público

É conhecido na literatura que quando se parte de um conjunto de dados gerados por sistemas automáticos e deseja-se compreender as características do sistema para a reconstrução das viagens, etapas como tratamento dos dados, inferência do destino, diferenciação entre transferências e atividades, agregação da informação pontual em informação zonal, e validação do método são explorados separadamente (Chu e Chapleau, 2008; Chen *et al.*, 2016; Zhao *et al.*, 2007; Kurauchi e Schmöcker, 2016; Hussain *et al.* 2021).

A análise dos padrões de validação não é citada como parte integrada dos métodos de análise da demanda de TP nestes estudos. Li *et al.* (2018) fizeram uma revisão da literatura sobre métodos para reconstrução das viagens em sistemas abertos

de TP e classificaram esses modelos como de probabilidade, encadeamento de viagem e de aprendizado de máquina. Na literatura, existe um grande leque de trabalhos utilizando o encadeamento de viagens, baseados no pendularismo das viagens (a primeira validação do dia é considerada a origem, enquanto a última validação é considerada próxima ao destino da primeira viagem) (Mesquita *et al.*, 2017; Arbex; Cunha, 2020).

O estudo de Trépanier *et al.* (2007) sugeriu melhorias nas premissas do pioneiro método de encadeamento de viagens proposto por Barry *et al.* (2002), obtendo-se uma taxa de inferência dos destinos correta em 66% (melhorando pouco mais de 10% em relação ao original), mas como desvantagem teve uma alta taxa de dados descartados (13%), pois estes detinham apenas uma validação diária, e uma vez que não se sabe os padrões, não se pode inferir uma rota deste tipo de usuário. Cats e Ferranti (2022) propuseram um estudo para avaliar os padrões temporais de mobilidade usando dados de *smart card* no sistema *tap-on/tap-off* de TP em Estolcomo, Suécia. Se ampararam em técnicas de classificação (*k-means*) de padrões e classificação hierárquica, usando um modelo de mistura Gaussiana. Dessa forma foi possível encontrar 10 padrões de deslocamento desde viajantes regulares em horário pico, até viajantes regulares durante a madrugada. Porém, nesse estudo não ficam claras as características que definem usuários regulares e irregulares, além do objetivo não ter foco na mobilidade urbana.

## 2.1. Modelos de regressão, classificação e clusterização

Essa nova proposta para reconstrução das viagens, coloca a compreensão dos padrões de deslocamentos à frente da modelagem de previsão. A utilização de técnicas de *Machine Learning* (ML) e *Statistical Learning* se mostram superiores às tradicionais de reconstrução de viagens, por não necessitarem de várias regras de previsão e suposições, sendo esta segunda uma estrutura do ML de análise funcional para lidar com problemas de inferência estatística de encontrar uma função preditiva baseada em dados. Em resumo é o processo de extração da informação de um conjunto de dados, com foco na descoberta de propriedades desconhecidas dos mesmos. Alguns modelos de *data mining* desenvolvidos são regras de associação, classificação, *clustering*, padrões sequenciais e padrões de similaridade (Géron, 2019). Em contrapartida os métodos de aprendizado de máquina têm foco na predição, baseado em características conhecidas, podendo ser *supervisionado* (com dados de treinamento), *não-supervisionado* (Agrupamento de dados, sem *inputs* de treinamento) ou por Reforço (mudança do ambiente por estímulos externos). Por fim, os modelos de aprendizado de máquina assumem métodos probabilísticos, porém contendo conjuntos de treinamento, teste e validação dos dados, de modo que seja possível inferir os melhores coeficientes e compatibilidade entre as variáveis de forma otimizada em relação aos outros modelos. Suas desvantagens ocorrem pela dependência de grandes contingentes de dados, alto nível de processamento em máquina e possibilidade de sobreajuste (*overfitting*) dos modelos, ou seja, a perda da capacidade de generalização para novos conjuntos de dados (Cats e Ferranti, 2022).

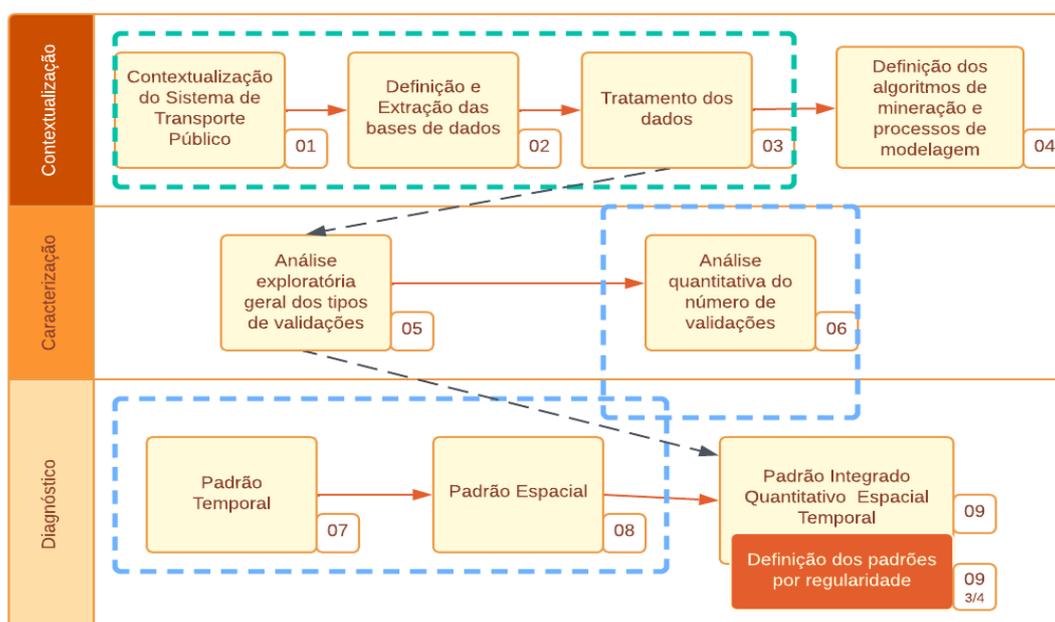
## 2.2. Algoritmos para Mineração de dados

Os algoritmos de mineração podem ser classificados de acordo com sua tarefa, ou propósito particular, variando suas implementações e adequando-se a novas finalidades, como é o caso destes estudos. Dentre as principais técnicas estão: (i) Árvore de decisões

- Baseada em estágios de decisões (nós) e na separação de classes e subconjuntos de forma hierárquica (Ex.: CART, CHAID e ID-3); (ii) Redes Neurais – Modelos inspirados na fisiologia do cérebro, no qual o “conhecimento” é fruto do mapa de conexões neurais, representadas por funções probabilísticas, em grande parte, e nos pesos dessas conexões (Ex.: Perceptrons, MLP e CNN); (iii) Algoritmos Genéticos – Métodos de otimização, inspirados na teoria da evolução, em que a cada nova geração, soluções melhores tem mais chances de ter “descendentes” pois estão mais adaptadas (Ex.: AGS, *Genitor* e *GA-Nuggets*); (iv) Conjuntos Fuzzy – Forma de lógica multivalorada, na qual os valores verdade podem ser qualquer número real entre 0 (falso) e 1 (verdadeiro), distanciando-se da lógica booleana (Ex.: *K-means* e FCMdd) (Fayyad et al., 1996). Além de alguns dos algoritmos citados, este trabalho também utilizou técnicas de Processamento de Linguagem Natural (PNL), Regras de Associação de Dados Estruturados, Normalização de Banco de Dados, Espacialização de bilhetagem com dados de GPS e ETL (*Extract, Transform and Load*). Vale destacar o algoritmo de *k-means* sendo fundamental para clusterização das validações, compreendendo um processo de partição dos elementos de um banco de dados em conjuntos ou *clusters*, de uma maneira que os registros, que são semelhantes, fiquem agrupados diferenciando dos demais dos outros subconjuntos. Nesta tarefa, não existem classes pré-definidas, mas pode-se definir o número de *clusters* que serão verificados, bem como os parâmetros de verificação, neste caso coordenadas.

### **3. Método de Diagnóstico dos Padrões Espaço-Temporais de Validação**

A proposta deste trabalho está ligada diretamente aos modelos citados (com ênfase em *Machine Learning* e *Statistical Learning*), porém para uma nova abordagem, baseada em *Data Mining*. Neste tópico será apresentado o método para diagnóstico dos padrões de validação espaço-temporais empregado neste trabalho, porém não se abstendo de padrões referentes a quantidade de validações por dia específico da semana ou por questões vinculadas às atividades. A Figura 1 resume o método nos seus aspectos técnicos e fenomenológicos no que concerne às micro interações das etapas e que podem se reorganizar de forma independente, dependendo dos propósitos futuros desta pesquisa.



**Figura 1: Método para diagnosticar padrões de validação espaço-temporais do TP.**

O método global, proposto para mineração de padrões, é composto pelas macro etapas Contextualização, Caracterização e Diagnóstico. Durante o processo de contextualização foram apresentadas as características do sistema de transporte público de Fortaleza, assim como informações sobre integração e tarifação (**Etapa 01**), além da disposição das linhas e paradas na rede. Posteriormente foram apresentados os dados disponibilizados pela Empresa de Transporte Urbano de Fortaleza (ETUFOR), compondo dados do ano de 2018, porém afunilando para o mês de novembro, uma vez que é um mês típico, com baixa influência de sazonalidade, e sem influência do período de pandemia (**Etapa 02**). No total 20 bases de dados foram extraídas, ordenadas e agrupadas (segundo critérios de normalização de bancos de dados estruturados, garantindo-se que não haja redundâncias) para o posterior tratamento (**Etapa 03**). Dentre as bases de dados estão a base de bilhetagem, GPS, GTFS, cadastro de usuários, *shape files* de linhas e paradas, *shape files* de terminais, códigos de identificação dos veículos entre as bases de bilhetagem e GPS, entre outras. Perante as principais transformações, estão a utilização de modelos de Processamento de Linguagem Natural (NPL) para compatibilização dos nomes das linhas em diversas bases, extração espacial de coordenadas da base de GPS fora do zoneamento estabelecido e mineração das validações em união ao GPS utilizando o horário, data, linha, sentido e identificador do veículo entre as bases (Braga, 2019), processo que deteve um alto custo computacional pois trata-se de um Big Data de Transporte Público (BD-TP) (Han et al., 2011; Kurauchi; Schmocker, 2016). Por fim, finalizando a primeira macro etapa, foram definidos os algoritmos (**Etapa 04**) para extração dos padrões de regularidade do sistema.

Na Macro Etapa 2 (Caracterização), estão dispostas as análises exploratórias das validações (**Etapa 05**), bem como informações sobre localização da residência de usuários cadastrados e linhas mais frequentes, pois parte-se da premissa de que existe uma influência do tipo de linha seja ela alimentadora, troncal, convencional ou complementar, uma vez que realizam atividades operacionais específicas segregando os

possíveis padrões (Mesquita; Neto, 2021). Dando início as análises dos padrões, foi estabelecido uma caracterização quantitativa do número de validações por dia útil e por semana (**Etapas 06**).

Por fim, a última macro etapa (Diagnóstico) que compõe o diagnóstico dos padrões espaço-temporais, está dividida entre mineração por intermédio do *k-means* das diferenças espaciais dos centroides de validação até as respectivas validações que o compõem, e mineração do desvio horário de validações (**Etapas 7 e 8**). Neste trabalho foi necessária uma modelagem em linguagem *python* e SQL utilizando as bibliotecas *Pandas* para análise dos dados, *Scikit-Learn*, *TensorFlow* e *Numpy* para modelagem do agrupamento das validações em classes e *PyMySQL* para armazenamento e consulta. As análises foram divididas em primeiras, últimas e validações intermediárias. Nos dados de validação para cada dia do mês de novembro (média de 1.080.000 validações diárias) foram agrupados e ordenados por ID do *smarcard*, dia e horário da validação, respectivamente. Utilizando *arrays* diários e condicionais, foram separadas as primeiras e últimas validações de cada usuário. Dentro de cada *array* encontrava-se um dicionário contendo como chave o identificador do cartão e como valor as coordenadas da validação. O mesmo processo foi repetido para as últimas validações e concomitante para os momentos das primeiras e últimas validações. No total foram necessários 21 *arrays* compondo os dias úteis, repetidos entre as quatro categorias iniciais (Primeira validação – Espacial e Temporal, Última Validação – Espacial e Temporal), totalizando 84 *arrays*. Os dados foram salvos no banco de dados relacional, compondo 118 mil usuários válidos com validações passíveis de serem analisadas (primeiras, últimas e intermediárias), dentre os 330 mil disponíveis na base de dados. Vale destacar que o processo de transformação de um vetor bidimensional de coordenadas em um vetor unidimensional de distâncias é na realidade um processo de redução de dimensionalidade, realizado por transformações lineares algébricas.

A importância de se analisar as validações intermediárias deve-se ao fato de os padrões comportamentais sofrerem variação pelo modo como o usuário lida com as características da oferta, do local onde ele reside e local ao qual está se deslocando, dessa forma, usuários com padrões idênticos de origem e destino, não necessariamente realizam o mesmo trajeto, implicando em diferença no número de transbordos e passagens por terminais. Desse modo, as validações restantes que não se configuraram como primeira ou última (diária) foram agrupadas por ID, dia e horário. Validações com latitude e longitude nulas foram desconsideradas, uma vez que para alguns casos não foi possível identificar o veículo pela base de GPS.

Posteriormente, após as análises segregadas para cada usuário foi agregado por consulta ao banco de dados os valores por dia útil do mês de novembro, no que diz respeito aos padrões quantitativos, espaciais e temporais (**Etapas 09**). Para cada usuário foi realizado uma análise de categorização de quais padrões e conjunto de padrões ele se adequava. Por fim, uma análise global é apresentada e compõem o produto final desse trabalho, sendo base de análises subsequentes para modelagem da cadeia de viagens de sistemas abertos, *tap-on* e tronco-alimentados a partir dos padrões mais recorrentes de deslocamento (**Etapas 9<sub>3/4</sub>**). Dessa forma, algumas etapas compõem partes específicas que podem ser avaliadas de forma distinta, como as Etapas 1, 2 e 3 que representam o aspecto de compreensão e transformação dos dados. As Etapas 6, 7 e 8 representam o diagnóstico dos diversos tipos de padrões, enquanto existe uma forte influência da etapa de tratamento sobre os padrões dispostos da Etapa 9, pois o modo como os dados são

tratados e armazenados rege 80% do esforço desse tipo de análise e conseqüentemente influencia o produto final do trabalho (Géron, 2019), relação essa representada pelas setas pretas na Figura 1.

#### **4. Aplicação e Resultados**

Em Fortaleza, o sistema é aberto, possibilitando a validação durante o percurso e *tap-on*. A rede segue uma distribuição tronco-alimentadora, dessa forma as linhas alimentadoras levam a demanda dos bairros aos terminais e as linhas troncais coletam essa demanda e levam às regiões centrais, onde existe forte concentração das atividades e comércios. O Sistema Integrado de Fortaleza (SIT-FOR) tem quase a totalidade das rotas com pagamento da tarifa através de *smart card*, sendo o pagamento quando em dinheiro tratado diretamente com o motorista e sem a possibilidade de recebimento de troco pelo usuário, processo gradual de inovação do sistema. O mesmo contava com mais de um milhão de validações diárias (pré-pandemia), 279 linhas regulares e 22 linhas complementares que cobrem toda a cidade, de acordo com os dados do GTFS de 2021, com uma rede de aproximadamente 5650 km de extensão e média de 11 km por linha, com 14 empresas gerenciando as linhas regulares e 320 cooperados gerenciando as linhas complementares (Mesquita; Neto, 2021). A cidade de Fortaleza atualmente, possui sete terminais fechados integrados com controle de tempo de percurso por GPS e dois terminais abertos não-integrados no centro da cidade, com pouco mais de 5000 pontos de parada distribuídos na rede e uma frota aproximada de 2700 veículos (Braga, 2019). A rede opera com um valor de tarifa inteira de R\$3,90, e com meia passagem no valor de R\$ 1,80. Desde sua criação, a rede é integrada operacional e fisicamente. Desde 2013 foi incorporado a integração temporal em todo o sistema, possibilitando a realização de um número ilimitado de transferências em qualquer ponto da rede, em um período de até duas horas.

##### **4.1. Extração, Tratamento e Compatibilização dos Dados**

As bases utilizadas neste trabalho compõem o GTFS, Bilhetagem, GPS e cadastro de usuários, além de outras complementares. Os arquivos do GTFS utilizados foram as rotas, viagens, *shape*, horário das paradas e paradas. A base de bilhetagem é composta pelo identificador do cartão, linha, sentido, número do carro, data, hora e tipo do cartão. A base de GPS contém coordenadas, data, hora e identificador do veículo. Para o cadastro, os dados se referem a identificação do usuário (nome e idade), endereço da residência, nome da empresa solicitante e endereço da empresa solicitante. Os dados utilizados constituem 20 bases de dados em formato csv, txt, JSON e shp. Para a formulação do modelo relacional para o banco, separou-se os dados em três grupos: I – Dados de cronograma (GTFS); II - dados coletados passivamente por equipamentos nos veículos; e III – Dados complementares. O Grupo I é composto pelos cinco arquivos do GTFS, já o Grupo II corresponde aos dados de Bilhetagem e GPS de novembro de 2018. Esse ano foi escolhido por ser pré-pandemia, e por conter dados suficientes de todas as bases. O Grupo III corresponde aos dados de cadastro dos usuários, shapes de terminais e de zoneamento, e dicionário com código dos veículos (ligação entre os identificadores nas bases de GPS e Bilhetagem) construído a partir de dados de anos anteriores. Todos os dados (excetuando os dicionários e os shapes de terminais que foram formulados pelos autores) foram cedidos pela ETUFOR, responsável pelo controle, regulação e fiscalização.

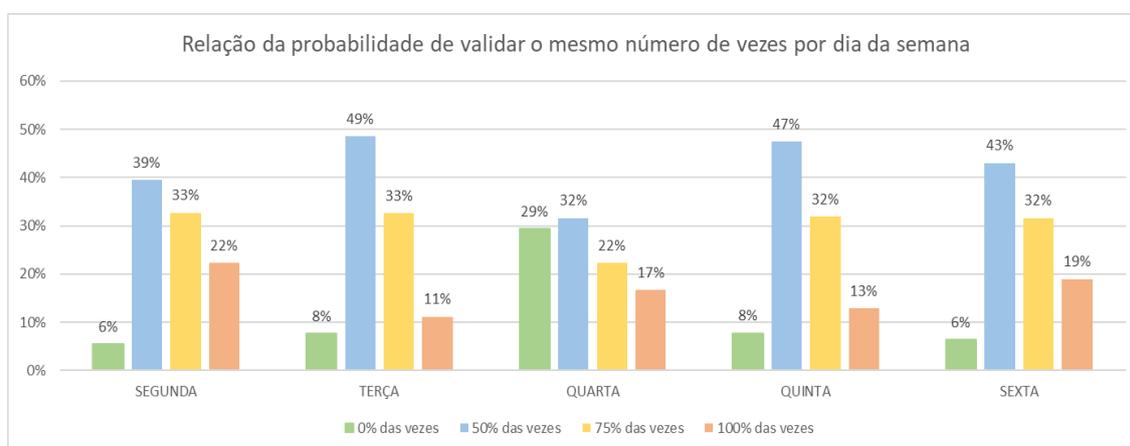
Após a criação do modelo relacional, definiu-se a prioridade relacional e, conseqüentemente, a mesma ordem deve ser mantida para criação das tabelas e carregamento no banco de dados. Essa ordem relacional favorece a normalidade do banco de dados (conjunto de regras que visa reduzir a redundância). Para todas as bases foi feita uma limpeza nos dados, excluindo valores nulos, duplicados e incoerentes. Os dados de cadastros dos usuários, dicionário, bilhetagem e GPS, após a transformação, foram agrupados em um único arquivo cada. Uma vez que os dados de bilhetagem não continham identificação do local de validação, foi preciso utilizar o dicionário que correlaciona o identificador do veículo no arquivo do GPS com o número do carro no arquivo da bilhetagem. Dessa forma, para cada validação identificou-se a coordenada, utilizando como parâmetros o código do veículo, linha, sentido, hora e data da viagem. Já na base de cadastros, também foi necessário identificar as coordenadas relacionadas ao endereço das residências dos usuários, utilizando a linguagem R e a API do Google *maps*.

#### **4.2. Análise Exploratória das Validações**

Na etapa de caracterização foi realizado uma análise exploratória dos dados, obtendo-se mais de um milhão de validações médias diárias (2018), sendo em torno de 20% correspondente a pagamentos em dinheiro, ou seja, sem a utilização de um *smartcard*. Perante esses registros em torno de 328 mil usuários utilizam o sistema diariamente, o equivalente a 12% da população de Fortaleza no ano em questão, e apenas 47% desses usuários estão com cadastro válido no sistema da ETUFOR, sendo possível obter 110 mil registros confiáveis sobre o local exato de residência desses usuários, conforme método apresentado na Seção 3. Dentre esses usuários 76% detém a linha mais frequente de utilização do dia como a mesma da primeira validação do dia, e 27% tem a primeira validação do dia próximo aos terminais. Também se avaliou a média de validações diárias, semanal e mensal de cada usuário durante o mês de novembro, identificando-se que 27% dos mesmos validam uma única vez diariamente, enquanto que 72% validam até duas vezes e 86% até 3 vezes. Enquanto que nas validações semanais, considerando que o usuário necessita de pelo menos duas validações para fechar uma cadeia de viagens diária e considerando pelo menos uma integração por sentido de viagem, tem-se um limite aceitável de até 20 validações semanais, representando 97% dos usuários. Foi possível notar que o número de validações semanais detém picos maiores no início e fim do mês, mostrando uma tendência dos usuários a utilizarem mais o TP nessas 2 semanas. Avaliando-se em uma escala mensal, 52% apresentaram até 30 validações mensais e 90% até 52 validações mensais (dias úteis apenas).

Por fim, realizou-se a verificação da existência de padrão dado o número de validações e a espacialização por tipo de linha e ordem da validação. Neste trabalho o que se considera como padrão está relacionado a frequência de um ato em pelo menos um dos dias úteis da semana em mais de 50% dos casos analisados de um mesmo usuário. Dessa forma a Figura 2, demonstra que os usuários estão mais suscetíveis a seguir um padrão nas segundas e sextas-feiras (considerando as classes de 75% e 100%), e foi evidenciado que 87% dos usuários apresentam uma frequência do número de validações idênticas acima de 75% em pelo menos um dia da semana, ou seja, é possível evidenciar pelo menos um padrão relacionado a frequência de utilização para a

quase totalidade dos usuários caso considere-se que o padrão de deslocamento pode se modificar por dia, e não apenas por localização e horário conforme esperado.



**Figura 2: Relação da probabilidade de validar a mesma quantidade de vezes por dia.**

### 4.3. Diagnóstico do Padrão Temporal e Espacial

Dando prosseguimento ao diagnóstico, foram realizadas análises temporais e espaciais de cada usuário, separados entre as primeiras, últimas e validações intermediárias, com a finalidade de verificar os possíveis padrões que auxiliem a compreensão da variabilidade da demanda, dado os diferentes horários do dia. Porém seguindo o vetor de deslocamento nas horas pico, onde os usuários se deslocam das regiões periféricas para o centro comercial pela manhã e detêm forte concentração de validações no período da hora pico da tarde, quando se deslocam dos centros comerciais para as regiões periféricas. Fator explicado pela segregação urbana de grandes cidades. Foram avaliadas as linhas alimentadoras e troncais para *smartcards* de estudantes e vale transporte, ambos validaram o apontamento anterior. A Figura 3 representa a concentração de validações por linhas alimentadoras e vale transporte, apresentando concentração de manchas na região periférica durante o início do dia e concentração no centro comercial próximo ao terminal do Papicu (Nordeste) e Centro comercial de Messejana (Sudeste) no pico da tarde.

Conforme verificou-se nas análises anteriores, há uma influência direta do dia útil da semana em relação ao padrão, portanto as análises espaço-temporais também foram realizadas seguindo essa lógica. Para cálculo das distâncias espaciais (Figura 4a), conforme apresentado no método, foi verificado a distância de validação entre o centroide e as respectivas validações que o compõem, distâncias acima de 1000m (Mesquita e Moraes, 2017) foram desconsideradas e o centroide foi recalculado por um processo iterativo para evitar interferência de *outliers*. Avaliando-se as classes de usuários de Vale Transporte, Estudantes e Gratuidade, os mesmos apresentaram distâncias médias inferiores a 600m para primeiras e últimas validações, com a gratuidade possuindo maior variação (585m) e o Vale Transporte menor (528m), contribuindo para validação da hipótese de que o motivo do deslocamento (atividade) influencia nos padrões, e principalmente, quanto mais rigoroso o horário do tipo de atividade, menor será a variação dos locais de validação, sendo possivelmente mais prático de inferir os reais locais de embarque e desembarque dos mesmos. Vale destacar

que 60% dos usuários têm distâncias efêmeras de até 490m, inferiores até da distância média entre paradas da rede de TP de Fortaleza (500m). Também foi realizada a análise da distribuição das distâncias de validação por *cluster* das validações de cada usuário (Figura 4b). Essa distância configura o local de embarque e o local de validação, considerado como a parada da linha mais frequente do usuário e mais próxima de sua residência (local de embarque). As maiores distâncias de validação se concentram próximas aos terminais (pontos vermelhos no mapa), porém os estudantes apresentaram menores distâncias nesses casos (pontos azuis), evidenciando que esses usuários tendem a validar logo em seguida ao embarque.

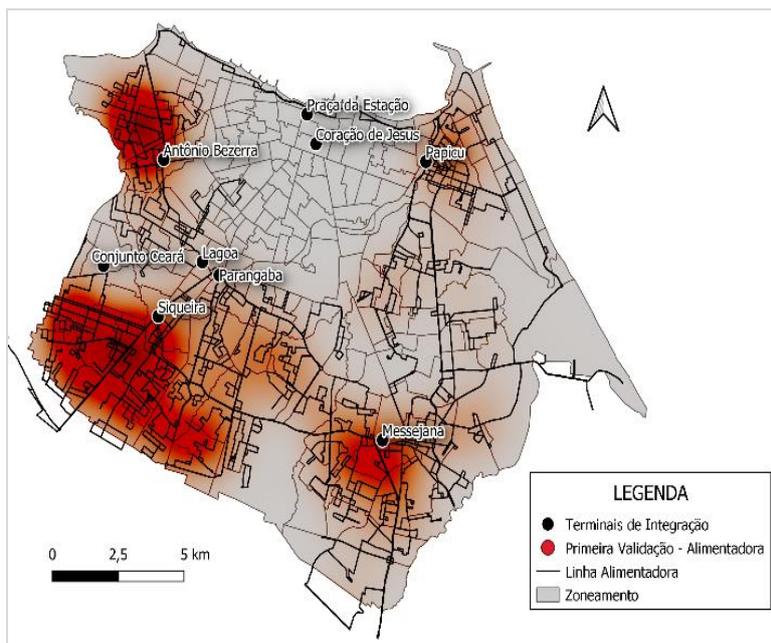


Figura 3: Mapa de Calor das primeiras validações em linhas alimentadoras

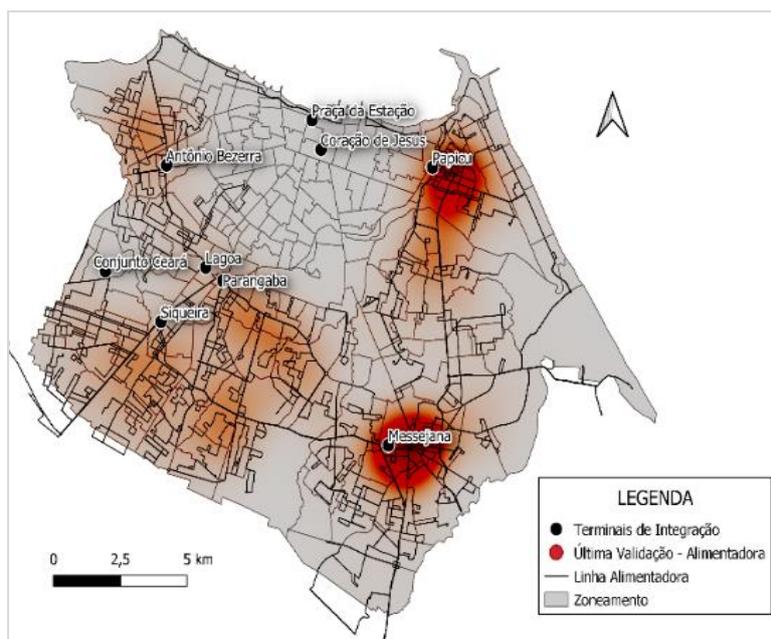
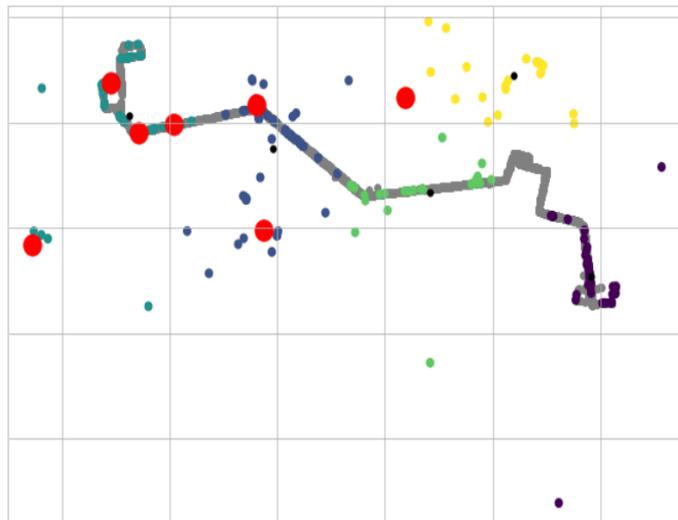
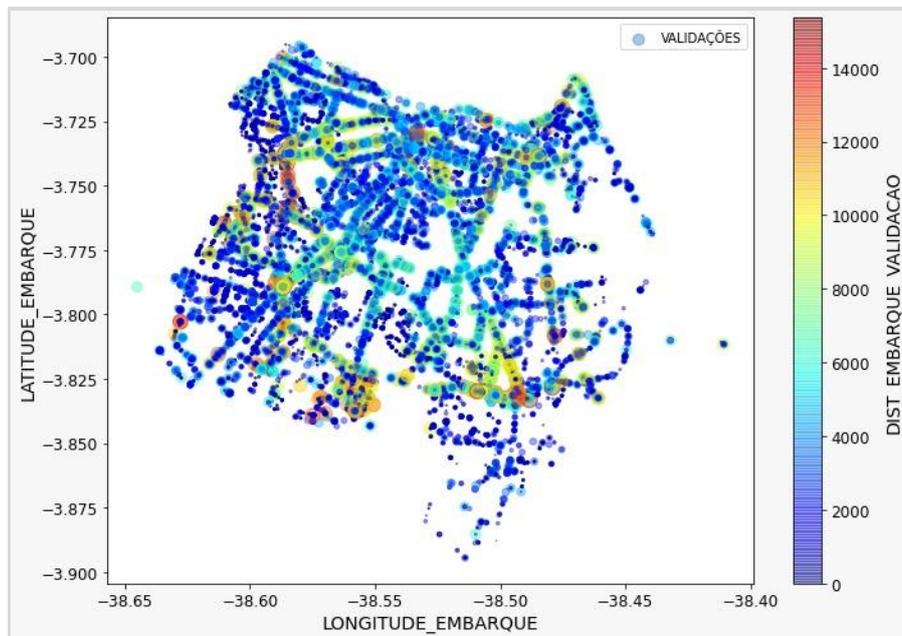


Figura 4: Mapa de Calor das últimas validações em linhas alimentadoras



**Figura 5: Clusterização das validações por usuário**



**Figura 6: Distância da primeira validação**

Por fim realizou-se análises das distâncias temporais entre as validações e dos horários médios por dia útil da semana que o usuário costuma validar, considerando um desvio temporal de até 60 min para as primeiras e últimas validações. O limite foi estabelecido pelo *headway* máximo dos veículos encontrado nos dados de GTFS. Quanto às distâncias temporais, as horas pico da tarde apresentaram valores superiores aos valores da hora pico da manhã, tendo forte concentração em até 40 s entre uma validação e outra pela tarde e de 25s pela manhã. Os horários entre picos apresentaram distâncias médias de até 100s entre validações. Vale destacar que essas médias duplicam nos finais de semana e feriados. Outra questão é que existem fatores nas horas pico que influenciam fortemente a distância temporal de validação, como à exemplo a lotação dos veículos nesses horários. Por fim, foi verificado que a curva do padrão espacial durante a semana assemelha-se a um arco convexo com picos de frequência nas

extremidades, enquanto que para os padrões temporais, se assemelha a uma curva côncava onde o pico se dá no centro da semana, mais especificamente nas quartas-feiras. Dessa forma há evidências de que todas as categorias de usuários avaliadas aparentam ter forte padrão espacial no início e final da semana, tendo maior flexibilidade temporal nesses dois extremos.

#### **4.4. Diagnóstico Global dos Padrões e o Fenômeno de Deslocamentos**

Dessa forma evidenciou-se que 81,2% dos usuários apresentam algum tipo de padrão espacial apenas na primeira validação do dia e 65,2% na última validação. Enquanto que 66,9% dos usuários apresentam padrão espacial e temporal em pelo menos um dia útil da semana na primeira validação, porém um padrão espaço-temporal nas últimas validações ocorre em apenas 20,56% dos usuários, ou seja, para fechar a cadeia de viagem seria necessário se amparar em um padrão espacial ou temporal e não nos mesmos simultaneamente, para a grande maioria dos usuários. Técnicas de encadeamento de viagens podem ser utilizadas. Além disso, foram analisadas até o grau três das validações intermediárias e adicionadas as análises em questão. Dessa forma, 48%, 33% e 15% dos usuários apresentaram algum tipo de padrão nas primeiras, segundas e terceiras validações intermediárias, respectivamente. Sendo a grande maioria, em relação aos padrões espaciais, ou seja, os usuários tendem a realizar transbordos próximos ao mesmo local durante a semana, mas dificilmente em horários próximos.

### **5. Considerações Finais**

Portanto para esse trabalho, conclui-se que os estudos de demanda podem se beneficiar das análises dos padrões uma vez que se conhece quando e onde o usuário costuma validar na rede e dessa forma viabilizaria uma oferta mais assertiva ao sistema. Também se verifica que o conceito de frequência está intimamente ligado ao de padrão, no entanto, o inverso não pode ser afirmado, visto que os usuários podem apresentar um padrão em dia ou horário específicos devido a alguma atividade não muito recorrente, e mesmo assim não poder ser considerado um usuário frequente ou cativo do sistema e necessitar de uma modelagem diferenciada quando comparados aos usuários frequentes. Além disso, após as análises foi perceptível que para caracterização dos padrões dos usuários deve ser levado em conta, separadamente, os padrões espaciais e temporais, visto que dificilmente os usuários possuem os dois padrões simultaneamente.

Ao final do estudo, pôde-se alcançar o objetivo inicial proposto para o trabalho de diagnosticar o padrão de deslocamento temporal e espacial dos usuários da rede de TP de Fortaleza, os quais dentre os principais padrões encontrados estão os usuários cujas validações se concentram muito próximas ao centroide de validações no início e final do dia, mas seu percurso intermediário apresenta alto desvio espaço-temporal, assim como identificou-se usuários com padrões assertivos ao longo de todo o mês, assertividade essa correspondente aos usuários de vale transporte. Os estudantes apresentaram padrões de validação mais próximos dos terminais e com distâncias de validação mais curtas, uma vez que utilizam em grande parte a oportunidade de integração temporal presente no sistema. Padrões irregulares (em menor quantidade) também foram verificados no estudo. Para esses é necessário um estudo mais aprofundado, que pode se amparar em técnicas de *Machine Learning* para modelagem

da probabilidade de validar ao embarcar, classificação do local de embarque e distância de validação.

Pretende-se em trabalhos futuros utilizar os padrões encontrados para reconstruir a cadeia de viagens de transporte público, identificando através de uma série temporal os locais que os usuários costumam validar, bem como a modelagem do local de embarque e transferência em terminais.

## Referências

- Arbex, R. O., & da Cunha, C. B. (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*, n.85, ISSN 0966-6923, <https://doi.org/10.1016/j.jtrangeo.2020.102671>.
- Barry, J., Newhouser, R., Rahbee, A. and Sayeda, S. 2002. Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record: Journal of the Transportation Research Board*, (1817), p. 183-187.
- Braga, C. K. V. (2019) Big data de transporte público na análise da variabilidade de indicadores da acessibilidade às oportunidades de trabalho e educação. 108 f. Dissertação (Mestrado) - Curso de Engenharia Civil, Universidade Federal do Ceará, Fortaleza.
- Cats O.; Ferranti F. Unravelling individual mobility temporal patterns using longitudinal smart card data, *Research in Transportation Business & Management*, 2022,100816, ISSN 2210-5395.
- Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, v. 68, p. 285-299.
- Cheng, Z., Trépanier, M. & Sun, L. Probabilistic model for destination inference and travel pattern mining from smart card data. *Transportation* 48, 2035–2053 (2021). <https://doi.org/10.1007/s11116-020-10120-0>
- Chu, K.A., Chapleau, R., 2008. Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. *Transportation Research Record: Journal of the Transportation Research Board* 2063, 63–72.
- Fayyad, U. M.; Patesky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In: *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1996.
- 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
- Han, Jiawei; PEI, Jian; KAMBER, Micheline. *Data mining: concepts and techniques*. Elsevier, 2011.
- Hora, J. et al. (2017) Estimation of Origin-Destination matrices under Automatic Fare Collection: the case study of Porto transportation system. *Transportation Research Procedia*, v. 27, p. 664-671.

- Hussain E.; Bhaskar A.; Chung E. (2021) Transit OD Matrix Estimation Using Smartcard Data: Recent Developments and Future Research Challenges. 125th. Transportat Research. doi: 10.1016/j.trc.2021.103044.
- Kurauchi, F.; Schmocker, J. D. (2016) Public transport planning with smartcard data. 2016.
- Li, T., Sun, D., Jing, P., Yang, K. (2018). Smart card data mining of public transport destination: A literature review. Inf.
- Mesquita, H. C.; Amaral, M. J.; Carvalho, W.L;Matriz O/D com Base nos Dados do Sistema de Bilhetagem Eletrônica. Congresso Nacional de Pesquisa em Transportes - ANPET, Recife, 2017.
- Mesquita, K.G.A, Neto, F.M.O. Método de Identificação dos Embarques em Viagens Big Data de Transporte Público. 35º Congresso Nacional de Pesquisa em Transportes da ANPET, assíncrono, 2021.
- Munizaga, M.A. and Palma, C. (2012). Estimation of disaggregate multimodal public transport origin–destination matrix from passive smart card data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, 24, p. 9-18.
- Ortúzar, J. D.; Willumsen, L. G. *Modelling Transport*. 4th Edition ed. West Sussex, UK: Wiley, 2011.
- Pelletier, M.-P.; Trépanier, M.; Morency, C. Smart Card Data Use in Public Transit: A Literature Review. *Transportation Research Part C: Emerging Technologies*, v. 19, n. 4, p. 557–568, ago. 2011.
- Trépanier, M., Tranchant, N. and Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1), p. 1-14.
- Zhao, J.; Rahbee, A.; Wilson, N. H. (2007) Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, v. 22, n. 5, p. 376–387,. ISSN 10939687.