**UNIVERSIDADE FEDERAL DO CEARÁ**

**CENTRO DE TECNOLOGIA**

**DEPARTAMENTO DE ENGENHARIA DE TRANSPORTES**

**PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TRANSPORTES**

**KAIO GEFFERSON DE ALMEIDA MESQUITA**

# MÉTODO DE IDENTIFICAÇÃO DOS PADRÕES DE USO E LOCAIS DE EMBARQUE A PARTIR DO BIG DATA DE TRANSPORTE PÚBLICO: UMA ABORDAGEM BASEADA EM MACHINE LEARNING

**FORTALEZA**

**2023**

KAIO GEFFERSON DE ALMEIDA MESQUITA

# MÉTODO DE IDENTIFICAÇÃO DOS PADRÕES DE USO E LOCAIS DE EMBARQUE A PARTIR DO BIG DATA DE TRANSPORTE PÚBLICO: UMA ABORDAGEM BASEADA EM MACHINE LEARNING

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia de Transportes do Programa de Pós-Graduação em Engenharia de Transportes do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia de Transportes. Área de concentração: Planejamento Urbano.

Orientador: Prof. PhD. Francisco Moraes de Oliveira Neto.

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

---

M544m Mesquita, Kaio Gefferson de Almeida.
Método de identificação dos padrões de uso e locais de embarque a partir do big data de transporte público : Uma abordagem baseada em machine learning / Kaio Gefferson de Almeida Mesquita. – 2023.
147 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Transportes, Fortaleza, 2023.
Orientação: Prof. Dr. Francisco Moraes de Oliveira Neto.

1. Bilhetagem. 2. Big Data. 3. Padrões espaço-temporais. 4. Aprendizado de Máquina. I. Título.
CDD 388

---

KAIO GEFFERSON DE ALMEIDA MESQUITA

# MÉTODO DE IDENTIFICAÇÃO DOS PADRÕES DE USO E LOCAIS DE EMBARQUE A PARTIR DO BIG DATA DE TRANSPORTE PÚBLICO: UMA ABORDAGEM BASEADA EM MACHINE LEARNING

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia de Transportes do Programa de Pós-Graduação em Engenharia de Transportes do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia de Transportes. Área de concentração: Planejamento Urbano.

Aprovada em: 29/06/2023.

BANCA EXAMINADORA

_____
Prof. PhD. Francisco Moraes de Oliveira Neto (Orientador)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Bruno Vieira Bertoncini
Universidade Federal do Ceará (UFC)

_____
Profª. Dra Cira Souza Pitombo
Universidade de São Paulo (USP)

A Deus.

Aos meus pais, amigos, familiares e professores. .

## AGRADECIMENTOS

A Deus, por todas as bênçãos que ocorreram nestes últimos anos que é difícil de explicar. Quero agradecer também por ter me dado força e coragem em momentos que sucumbi.

Aos meus pais, André e Sílvia, que são meu maior tesouro e inspiração. Não conheço pessoas que encaram a vida de tão bom grado e destreza do que os senhores. Sou grato por tantos ensinamentos e por tanto amor. Que me tirem tudo, menos o conhecimento. Amo vocês.

Aos meus irmãos Andrey e Jaqueline por sempre me apoiarem nas decisões mais inusitadas e por vibrarem comigo a cada conquista. Amo vocês.

A minha namorada Ellen por caminhar comigo cada passo desse árduo ciclo. Sem você esse sonho não se tornaria realidade. Minha eterna gratidão. Te amo.

Aos meus amigos João Carlos, Reinaldo, Davi, Neto e Mônaco por tornarem essa caminhada mais leve e suportável. Obrigado de coração, amo meu cafofo.

Aos meus amigos Lucas Sousa, Gabriel, Gescilam, Israel, Jonas, Lucas Moreira, Nilson, Matheus, João Lucas, Altanízio, Beliza, Aldaianny, David, Diego e Lira. Uns mais próximos, outros mais distantes, mas os ciclos fazem parte da vida, e tenho plena consciência da importância de cada um nesta trajetória. Amo vocês.

Ao meu orientador e amigo Prof. Moraes, por toda paciência, ensinamentos e destreza nas palavras para acalmar os ânimos. Quero agradecer também por sempre enxergar um potencial em meu trabalho acima do que eu enxergava. Muito Obrigado por tudo!

Aos Professores Bruno Bertoncini e Cira Pitombo por participarem com tanto entusiasmo das discussões e estarem disponíveis à contribuir com o desenvolvimento desta dissertação. Obrigado.

Ao Programa de Pós-graduação em Engenharia de Transportes – PETRAN e aos professores do departamento, por todo o conhecimento compartilhado.

A todos os amigos da Imtraff, mas em especial ao Fred, Igor, Kleberson, Fabrício, Luan e Thiago Reis por acreditarem no meu potencial e abrirem as portas em momentos tão difíceis, além de me darem suporte todas as vezes que precisei. Muito Obrigado!

A todos aos meus alunos da Unifametro que indiretamente me mostraram uma nova paixão: a de lecionar e estar dentro da sala de aula.

“Na incerteza, os indivíduos criam instintos inovadores. Na rotina, padrões repetitivos. (Thimer).“

## RESUMO

Nos últimos anos dados automáticos de tarifação (bilhetagem eletrônica) e de localização automática de veículos têm sido explorados para apoiar o planejamento e análise do sistema de transporte público. Contudo, existem vários desafios na utilização destes dados massivos, como a identificação dos locais de embarque e desembarque em sistemas abertos em redes multimodais e do tipo tronco-alimentadas, através do padrão de uso do sistema. O objetivo deste trabalho é desenvolver um método de identificação dos locais de embarque das viagens utilizando aprendizado de máquina a partir do *Big Data* do Sistema Integrado de Transporte Público de Fortaleza (SIT-FOR), para padrões habituais de uso do sistema. Como objetivos específicos, tem-se: (i) consolidar uma estrutura de gerenciamento do *Big Data* do SIT-FOR; (ii) identificar a partir de métodos de mineração de dados os padrões habituais de uso do sistema; e (iii) analisar a partir de modelagem supervisionada como os padrões habituais podem auxiliar na identificação dos locais de embarque. Acredita-se que os padrões recorrentes, espaciais ou temporais de uso do sistema, permitam identificar os atributos das viagens nos dados. Assim, os dados do *Big Data* foram tratados e integrados numa única base de dados relacional. Os padrões de uso foram identificados a partir de uma técnica de agrupamento (*K-means*), permitindo avaliar como diferentes atributos influenciam na formação de cada grupo. A partir dos padrões identificados, diferentes modelos supervisionados (*Naive Bayes, Random Forest, Rede Neural*) foram aplicados para prever a probabilidade de um usuário validar ao embarcar na primeira viagem do dia. Como resultados, foi possível identificar 4 padrões habituais de uso caracterizados por aspectos temporais e espaciais. Por fim, dentre os modelos supervisionados segregados por grupos, obteve-se um melhor desempenho (acurácias entre 0,58 e 0,67) com o *Random Forest*. Os resultados da modelagem indicaram principalmente que a segmentação dos dados em padrões habituais melhorou o desempenho dos modelos, corroborando a hipótese de que a compreensão dos diferentes padrões de uso pode apoiar a identificação de atributos das viagens nos dados de bilhetagem.



**Palavras-chave**: Bilhetagem; Big Data; Padrões espaço-temporais; Aprendizado de Máquina.

## ABSTRACT

In recent years automatic fare pricing (electronic ticketing) and automatic vehicle location data have been exploited to support public transportation system planning and analysis. However, there are several challenges in using this massive data, such as the identification of boarding and alighting locations in open systems in multimodal and trunk-fed networks through the system usage pattern. The objective of this work is to develop a method to identify the boarding locations of trips using machine learning from the Big Data of the Integrated Public Transportation System of Fortaleza (SIT-FOR), for habitual patterns of system use. As specific objectives, we have: (i) to consolidate a management structure for the SIT-FOR Big Data; (ii) to identify through data mining methods the usual patterns of system use; and (iii) to analyze through supervised modeling how the usual patterns can help in the identification of boarding locations. It is believed that recurring patterns, spatial or temporal patterns of system usage, allow the identification of travel attributes in the data. Thus, the Big Data data was processed and integrated into a single relational database. The usage patterns were identified from a clustering technique (K-means), allowing to assess how different attributes influence the formation of each group. From the identified patterns, different supervised models (Naive Bayes, Random Forest, Neural Network) were applied to predict the probability of a user validating when boarding the first trip of the day. As results, it was possible to identify 4 habitual usage patterns characterized by temporal and spatial aspects. Finally, among the supervised models segregated by groups, a better performance (accuracies between 0.58 and 0.67) was obtained with Random Forest. The modeling results mainly indicated that segmenting the data into habitual patterns improved the performance of the models, supporting the hypothesis that understanding different usage patterns can support the identification of travel attributes in ticketing data.



**Keywords**: Smart card; Big Data; Spatio-temporal patterns; Machine Learning.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| AFC | Automated Fare Collection |
| AFP | Automatic Fare Payment |
| APC | Automatic Passenger Counter |
| APTS | Sistemas Avançados de Transporte Público |
| ANTP | Associação Nacional de Transporte Público |
| ATIS | Advanced Traveler Information System |
| AVL | Automatic Veicle Location |
| BD-TP | Big Data de Transporte Público |
| BDR | Bancos de Dados Relacionais |
| CAD | Computer Aided Dispatch |
| CPU | Unidades de Processamento Central |
| CSV | Comma Separated Values |
| ETL | Extract Transform Load |
| ETUFOR | Empresa de Transporte Urbano de Fortaleza |
| GIS | Sistemas de Informações Geográficas |
| GPS | Global Positioning System |
| GPU | Unidade de Processamento Gráfica |
| GTFS | General Transit Feed Specification |
| HDFS | Hadoop Distributed File System |
| IBGE | Instituto Brasileiro de Geografia e Estatística |
| IPEA | Instituto de Pesquisa Econômica Aplicada |
| ITS | Sistemas Inteligentes de Transportes |
| JSON | Java Script Notation |
| LGPD | Lei Geral de Proteção de Dados |
| LLE | Locally Linear Embedding |
| MAE | Erro Médio Absoluto |
| MTD | Distância Máxima de Transferência |
| MTE | Ministério do Trabalho |
| MTT | Tempo Máximo de Transferência |
| NBC | Classificador Bayesiano Ingênuo |
| NTU | Associação Nacional das Empresas de Transportes Urbanos |
| PASFOR | Plano de Acessibilidade de Fortaleza |
| PCA | Análise de Componentes Principais |
| PNL | Processamento de Linguagem Natural |
| RFM | Random Forest Model |
| RSME | Erro Quadrático Médio |
| SBE | Sistema de Bilhetagem Eletrônica |
| SEFIN | Secretaria de Finanças do município de Fortaleza |
| SGBDR | Sistemas Gerenciadores de Bancos de Dados Relacionais |
| SGD | Gradiente Descendente Estocástico |
| SIT-FOR | Sistema Integrado de Fortaleza |
| STPP | Sistema de Transporte Público de Passageiros |
| TCP | Transmission Control Protocol |
| TCM | Trip-Chaining Method |
| T-SNE | T-distributed Stochastic Neighbor Embedding |

**SUMÁRIO**

# 1 INTRODUÇÃO

Tradicionalmente os padrões de viagem são identificados em Pesquisas Origem/Destino (O/D), ferramentas mais comuns para se obter informações sobre a mobilidade de uma cidade, realizadas em campo, com a população da área de estudo, por intermédio da amostragem dessa população. Informações estas como motivo, origem/destino, modo e frequência de viagens. As desvantagens de tais pesquisas são seu alto custo para serem conduzidas e o elevado tempo para sua conclusão. Também é conhecido que todo processo de amostragem não consegue englobar devida e representativamente toda a população, podendo ser a utilização de dados massivos, um complemento ao processo tradicional de obtenção de informações da demanda. Uma das vantagens é a possibilidade de obtenção de dados socioeconômicos do indivíduo (e.g. renda, composição familiar, posse de veículos, etc.). Esse tipo de estudo é realizado de modo geral na maioria das capitais brasileiras a cada 10 anos (Guerra *et al*., 2014). A cidade de Fortaleza (Brasil) por exemplo, teve uma nova pesquisa realizada em 2019, o Plano de Acessibilidade Sustentável de Fortaleza (PASFOR), com o intuito de aumentar a eficiência e melhorar o sistema de transportes (Nordeste, 2019), sendo a última pesquisa realizada anterior à esta, apenas em 1996, caracterizando mais de 20 anos sem que ocorresse uma nova atualização das condições de mobilidade da cidade (Sousa *et. al.* 2019).

Conforme apontado, os métodos de obtenção de informação tradicionais através de pesquisas podem não ser representativos da população, surgindo a problemática abordada neste trabalho, sobre a proposição inadequada da oferta de transporte público pelo fato de não se conhecer em detalhes como os usuários se deslocam (Zhao *et al*., 2007; Pelletier et al., 2011; Munizaga e Palma, 2012). Para alguns sistemas, não é possível identificar diretamente como os usuários se deslocam, por isso, o esforço em entender esses padrões e o que os impactam é importante do ponto de vista do fenômeno de deslocamento por transportes. Dessa forma, a utilização de dados massivos pode complementar e avançar nos estudos de compreensão dos padrões de deslocamento. Portanto, este trabalho visa avançar sobre os estudos de análise dos padrões e modelagem do local de embarque com foco em contribuir com o planejamento operacional.

No final da década de 1990, os sistemas de pagamento com cartão inteligente foram incorporados em algumas cidades, como *Washington D.C*. (EUA) e Tóquio (Japão), também conhecidos como *Automated Fare Collectio*n – 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, ocorrendo em terminais físicos, garantindo segurança ao usuário e menor tempo de caminhada, para sistemas abertos (com possibilidade também de validar em paradas da rede fora dos terminais de integração), garantindo maior acessibilidade do usuário no sistema. Essa nova tecnologia logo se espalhou por outras cidades e se tornou um dos meios mais importantes de cobrança de tarifa, onde em muitos países emergentes, esta continua sendo sua principal funcionalidade. Em cidades como Chicago e Santiago, em que o sistema de Transporte Público (TP) tem alta penetração, 90% e 97% respectivamente, a obtenção de informação sobre as validações é bastante facilitada. Isso se deve ao fato de que esses dados são de baixo custo, fácil reprodução e alto nível de desagregação. Além do SBE, muitas cidades no mundo vêm adotando também sistemas *Automatic Veicle 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 demanda por transportes públicos. Outros sistemas da informação citados na literatura e que ganharam notoriedade na última década foram a *General Transit Feed Specification* (GTFS), contadores automáticos (APC), câmeras para identificação dos usuários com Inteligência Artificial e Sistemas de Informações Geográficas (GIS).

Uma questão importante, é que os Sistemas de Bilhetagem Eletrônica não foram originalmente projetados para coletar dados das viagens e sim para arrecadar as tarifas, por isso, uma série de etapas são necessárias para reconstruir os atributos das viagens, já que várias informações não estão presentes nos dados (e.g., origem, destino, pontos de embarque/desembarque, tempo de espera, transferências, rotas, propósito da viagem etc.). 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 dos atributos, etapas como tratamento dos dados, inferência do destino, diferenciação entre transferências e atividades são explorados, separadamente (Chu e Chapleau, 2008; Chen *et al*., 2016; Zhao *et al*., 2007; Kurauchi e Schmöcker, 2016; Li *et al*., 2018; Hussain *et al.* 2021). É importante destacar que nestes trabalhos a reconstrução dos atributos se refere a cadeia de deslocamentos, não priorizando aspectos como o motivo, por exemplo, e que nesta dissertação está interligado à reconstrução dos atributos que compõem os deslocamentos dos

usuários. Também deve-se apontar que os dados de bilhetagem e GPS de alguns sistemas de transporte público, permitem conhecer características das viagens, tais como tempo, distância. Tendo como principais vantagens a possibilidade de obtenção de uma grande série temporal de validações para um mesmo usuário ou conjunto de usuários, porém ainda tendo como limitação defeitos nos processos de extração, tratamento e armazenamento, além da falta de informações importantes (motivo e destino, por exemplo).

A maioria das validações nos Sistemas Integrados de Transportes Públicos em regiões emergentes é feita apenas na entrada do veículo, citando um caso contrário em um país com baixa taxa de desemprego e alto índice de mobilidade, na cidade de Cingapura, que tem validadores na entrada e saída dos veículos (Liu *et al.*, 2019). Outro agravante é que frequentemente a validação na entrada não é realizada de imediato durante o embarque do usuário no veículo, devido a fatores relacionados ao comportamento dos usuários e as características operacionais da rede, como lotação (Zhao, 2004). Arbex e Cunha (2020), apresentaram recentemente a existência da problemática de não se saber o real local de embarque para analisar acessibilidade do transporte público de São Paulo (Brasil), embora não tenha sido proposto um método para contornar essa problemática, assumindo a validação como o próprio local de embarque.

No caso de Fortaleza, o sistema é aberto, possibilitando o pagamento/validação da viagem durante o percurso e *tap-on* (validação depois do embarque, mas não no desembarque). 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 até as regiões centrais, onde existe forte concentração das atividades e comércios. O Sistema Integrado de Fortaleza (SIT-FOR) tem quase totalidade das rotas com pagamento da tarifa através de *smart card* e algumas poucas rotas aceitam pagamento em dinheiro, processo gradual de inovação do sistema. Alguns percalços existentes no sistema, são que não se sabe o local que realmente ocorreu o embarque e nem mesmo o local de destino, não existe informação se a validação realmente configura uma transferência ou atividade curta, não se sabe o motivo da viagem, não se têm dados sobre os usuários não rastreados (que pagam em cédula) e não se têm informações sobre transbordos em terminais. Outro ponto importante é que no modelo de encadeamento admite-se que se conhece o exato local de embarque, não sendo verdade para sistemas abertos. Portanto é necessária uma modelagem para ajustar os reais locais de embarque, pois não se tem certeza destes pontos.

Dessa forma, a ***lacuna central deste trabalho configura o fato de que em muitos sistemas de transporte público, se assume que o local de embarque é o mesmo do local de validação (coleta da tarifa), o que pode resultar em erros na estimação dos destinos ou reconstrução dos atributos das viagens*** (Zhao, 2004; Arbex e Cunha, 2020; Hussain et al. 2021). A identificação do local de embarque é um atributo importante para o planejamento de transportes, pois não saber como os usuários se deslocam, podem impactar negativamente na proposta de uma oferta do sistema condizente com a demanda.

Espera-se que a integração de um grande contingente de dados das mais variadas bases de TP e tratados devidamente, forneçam resultados mais precisos. Portanto, uma gama de dados é necessária e podem ser obtidos tanto pelas pesquisas de campo realizadas por órgãos gestores, quanto por Sistemas da Informação (SI) nos veículos.

Além da cobrança de tarifa, os cartões inteligentes também coletam continuamente o comportamento dos passageiros. Desse modo o tamanho dos dados pode se tornar tão grande que pode superar a capacidade de processamento dos meios convencionais, adicionando-se ao volume, a velocidade de extração, a variedade dos dados e o valor da informação, compondo o chamado *Big Data* de Transporte Público (BD-TP) (Dumbill, 2012; Kurauchi; Schmocker,2016; Han *et al*., 2022). Dessa forma, é essencial tratar o *Big Data* quanto a inconsistências devido ao erro humano (e.g. passar o cartão no validador mais de uma vez) e de equipamentos (e.g. não identificar as coordenadas por causa de interferências de túneis e prédios). A explosão de dados gerados contribui para o desafio de como armazená-los e gerenciá-los. Necessitando, portanto, de novas formas de análise e modelos condizentes com as características dos dados. Dessa forma, apresenta-se a primeira questão de pesquisa: ***(i) Como tratar e estruturar dados massivos de sistemas de coleta de dados automáticos do transporte público que viabilize análise sobre o comportamento da demanda?***

De acordo com Habib & Weiss (2014), a maioria das viagens que ocorrem no Transporte Público é realizada por viajantes pendulares. Assim, infere-se que a estabilidade das viagens reflete à longo prazo, movimentos característicos dos viajantes no cotidiano influenciados pelas áreas de atividades e ambiente construído. Além disso, algumas pesquisas foram implementadas do ponto de vista da variabilidade de viagens, mas estudos que consideram os padrões espaciais (localidade das validações e linhas utilizadas) e temporais (horários e frequência de uso) de uso do sistema para identificar atributos das viagens nos dados de Big Data ainda são incipientes (Cui *et al.,* 2014; Cui & Long, 2015).

No entanto, o rápido progresso da inteligência de transportes forneceu uma fonte de dados favorável para o estudo de identificação de comportamento de viagem e análise das características dos passageiros durante um longo período (Cui, 2014). Cheng *et. al* (2021) propuseram um método onde 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 formas distintas. 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.

Um aspecto essencial do sistema de transportes público de passageiros é a previsão dos deslocamentos que é uma questão significativa na área do planejamento de transportes, devido à sua importância operacional (Thiagara e Prakashkumar, 2021). São muitos avanços e aplicações inovadoras que têm sido introduzidos para um ambiente mais seguro e eficiente. Para tanto as implicações destas alterações na oferta, apenas são possíveis com a compressão de como a demanda de usuários por transporte público se desloca na rede. Cats *et. al* (2015) examinaram a distribuição espaço-temporal do fluxo de passageiros dos transportes públicos de Estolcomo (Suécia) para identificar e classificar centros de atividades baseados nestes dados de mobilidade. Lin *et. al* (2020) propuseram um estudo que avaliou como os usuários pendulares se deslocam na rede. A compreensão dos padrões de deslocamento pendular pode fornecer um apoio eficaz ao planeamento e à operação dos sistemas de transportes públicos. Ainda não foi verificado a relação dos atributos de validação no processo de deslocamento e na formação dos grupos de usuários do sistema, bem como outras variáveis relacionadas a operação ainda não foram testadas.

Os estudos em geral, não abordam a potencialidade da análise de padrões de uso do sistema para identificar atributos das viagens, como o local de embarque e desembarque. Além disso, nesses estudos também não é considerado atributos de oferta do sistema na análise dos padrões, como a estrutura da rede, as funções das linhas, e a influência de terminais de integração. Portanto de acordo com as discussões anteriores apresentam-se mais duas questões de pesquisa: (ii) ***como identificar os locais de embarque a partir da análise de padrões de uso do sistema com base no Big-Data de transporte público?*** (iii) ***Baseado nos padrões de uso do sistema de transporte público, como eles podem apoiar os métodos de reconstrução dos atributos?***

Li *et al.* (2018) fizeram uma revisão da literatura sobre estimativas de destinos para reconstrução das viagens em sistemas abertos e classificaram esses modelos como de probabilidade, encadeamento de viagem e de aprendizado de máquina. Os modelos de probabilidade assumem que existe um padrão entre os dados de entrada e que esse padrão é bem representado por uma distribuição estatística, a desvantagem desses métodos é que não são realizados estudos prévios para avaliar os tipos de padrões de viagens. Além desse modelo, tem-se um grande leque de trabalhos utilizando o encadeamento de viagens, baseados exclusivamente no pendularismo das viagens (a primeira validação do dia é considerada a origem, enquanto a última validação do dia é considerada próxima ao destino da primeira viagem), sendo este o método tradicional de reconstrução das viagens (Mesquita *et al.,* 2017; Arbex; Cunha, 2017). O método de encadeamento, é portanto, um método para identificação ou estimação dos destinos, de modo a fechar a cadeia de viagens. Esses modelos partem de premissas sobre o local de embarque ser o mesmo de validação e as transferências entre ônibus ocorrerem segundo um padrão de tempo máximo ou distância máxima, além do foco ser na modelagem dos destinos. Por fim, os modelos de aprendizado de máquina, assumem métodos probabilísticos, porém contendo conjuntos de treinamento, teste e validação dos dados. As variáveis de predição são geralmente zonas de origem, destino e/ou transferência, além do tempo de viagem. Suas desvantagens ocorrem pela sua 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 (Géron, 2019).

Recentemente, os pesquisadores estão mudando do método de encadeamento de viagem para técnicas de aprendizado de máquina supervisionado, para inferir a localização do desembarque, através de dados de treinamento em sistemas *tap-on/tap-off (validação no embarque e no desembarque),* uma vez que nesses sistemas se sabe exatamente onde ocorreu o início e fim das viagens. Essas técnicas embora tenham ganhado notoriedade na última década (2010-2020) pela utilização de dados massivos, existem desde a década de 1980. Elas são utilizadas principalmente por causa da alta previsibilidade dos métodos de aprendizado de máquina e para diminuírem as suposições do encadeamento de viagem. Espera-se que os atributos dos dados do *smartcard*, por estarem disponíveis de forma massiva em muitos sistemas, contribuam para resultados mais precisos. Jung e Sohn (2017) criaram uma arquitetura de aprendizado profundo com 4 camadas, sendo 2 internas (*Hidden Layers*) para

prever destinos de passageiros a partir de dados de *smart card*, porém continham dados de embarque e desembarque para treinar e validar o modelo, fato que facilita o processo. Assemi *et al.* (2020) propuseram uma metodologia envolvendo rede neural para inferir informações que o encadeamento de viagem não foi capaz de realizar, detendo também dados de embarque e desembarque dos passageiros. Outra aplicabilidade de algoritmos se refere ao *Random Forest.* Lin *et. al* (2020) previu a demanda de passageiros em diferentes horários do dia, o que permitiu às empresas de transporte planejarem a oferta de serviços de forma mais eficiente, porém sem caracterizar quais atributos mais impactaram nas previsões e como eles se relacionavam entre si. Desse modo, como avançar ou adaptar os métodos existentes para reconstruir os atributos das viagens em sistemas como o de Fortaleza, para identificar informações faltantes é de suma importância (Jolliffe, 2002; Halevy *et al*., 2009).

Dessa forma, pretende-se aplicar técnicas de modelagem de dados não-supervisionadas para identificar os padrões e para relacioná-los com o comportamento dos usuários (e assim identificar os locais de embarque). Portanto, como o local de embarque não é conhecido, espera-se que os padrões de uso do sistema auxiliem na identificação de atributos como a origem da viagem. A principal hipótese é que os usuários regulares do sistema tendem a repetir certos comportamentos na rede que podem ser identificados através de técnicas de mineração, e que podem auxiliar na identificação de características dos deslocamentos. Contudo, como identificar os padrões (que atributos espaciais e temporais utilizar!?) e como relacioná-los com o comportamento de embarque ainda precisa ser explorado, sendo as principais contribuições deste trabalho. Diante dessa explanação pergunta-se: (iv) ***qual o ferramental de modelagem adequado para identificar os locais de embarque com base nos padrões de uso do sistema?***

## 1.1 Objetivos e Hipóteses

O objetivo geral deste trabalho é desenvolver um método de identificação dos locais de embarque das viagens a partir do *Big Data* do Sistema Integrado de Transporte público, para os padrões habituais de deslocamento, utilizando como estudo de caso o sistema de Fortaleza (SIT-FOR). Os objetivos específicos estão dispostos a seguir:

1. Construir e disponibilizar um modelo relacional através de um Sistema de Gerenciamento de Banco de Dados para o Big Data – TP, integrando os diversos atributos das bases de dados.
2. Identificar os padrões de regularidade no uso do sistema a partir do Big Data – TP e analisar como os padrões habituais ou de regularidade podem auxiliar na identificação dos locais de embarque dos usuários.
3. Propor e avaliar métodos de identificação dos reais locais de embarques conforme os padrões identificados através de modelagem com aprendizado de máquina.

## 1.2 Estrutura da Dissertação

A dissertação é composta por 8 capítulos. O **capítulo 2** apresenta a revisão da literatura sobre dados massivos e obtidos passivamente através de equipamentos instalados nos veículos, assim como os métodos tradicionais para reconstrução dos atributos de deslocamento, baseado no pendularismo das viagens. Já o **capítulo 3** apresenta os métodos baseados em aprendizado de máquina. Ao final desse capítulo é discutido de forma objetiva os preceitos para modelagem relacional de dados. No **capítulo 4** está delimitado o método do trabalho explanado em 4 etapas, sendo elas: (i) estruturação do banco de dados; (ii) Análises Exploratórias das validações; (iii) Identificação dos padrões de uso do sistema; e (iv) Identificação do local de embarque. O **capítulo 5** apresenta a delimitação para construção do banco de dados, desde a apresentação das bases utilizadas, bem como o modelo relacional proposto e as validações sobre o método de tratamento. No **capítulo 6** estão descritas as análises exploratórias das validações em 3 níveis: (i) Espaciais e temporais agregadas; (ii) Análises das primeiras validações; e (iii) Análises a nível do indivíduo. No **capítulo 7** por intermédio da *clusterização* e dos padrões encontrados, foi exposto como se realizou a modelagem da probabilidade de validar ao embarcar. Por fim no **capítulo 8** são apresentadas as considerações finais sobre o método e resultados discutidos, além das proposições de trabalhos futuros utilizando modelos de aprendizado de máquina.

# 2 PADRÕES DE DESLOCAMENTO A PARTIR DO BIG DATA-TP

Para melhor nortear as discussões, a revisão da literatura deste trabalho foi dividida em dois capítulos, um que discuti o fenômeno de deslocamentos e os principais dados obtidos de forma passiva (**capítulo 2**) e outro com foco no ferramental de modelagem com aprendizado de máquina e modelo relacional de banco de dados (**capítulo 3**). Segundo Mesquita *et al.* (2017) os inputs dos sistemas de transportes são quaisquer insumos consumidos na produção do deslocamento, como pessoas, veículos ou combustível por exemplo. Enquanto os outputs são as pessoas ou produtos que foram transportados, existindo um agravante espacial e temporal nesse produto. Um detalhe que deve ser levantado, que ao contrário do sistema viário, o sistema de transporte público possibilita apenas entradas e saídas em pontos específicos da rede, como paradas de embarque/desembarque ou terminais de integração física.

O recorrente cenário das metrópoles brasileiras, consiste em uma queda expressiva na demanda por usuários que utilizam o transporte coletivo. Segundo a Associação Nacional das Empresas de Transportes Urbanos (NTU, 2018), em 5 grandes capitais brasileiras, entre os anos de 1997 e 2017, incluindo Fortaleza, houve uma redução de 35,6% dos usuários. A fim de coletar automaticamente a tarifa e absorver informações sobre a demanda, desde a última década, muitas cidades vêm adotando sistemas de bilhetagem eletrônica (*Smart Card*), de localização automática de veículos (*Global Position System* – GPS), e da delimitação da programação por intermédio do *General Transit Feed Specification* (GTFS). Além de permitir gerenciar o sistema de transporte público, estes subsistemas geram uma grande quantidade de dados que compõem o chamado *Big Data* do Transporte Público (BD-TP).

Pelletier *et al.* (2011) desenvolveram uma revisão da literatura sobre os usos de sistemas de *smartcard* para transporte público, na qual dividiram em três grupos: aplicações operacionais, táticas e estratégicas. Os estudos de nível operacional usam dados de cartão inteligente para medir indicadores de desempenho e oferta de trânsito; os estudos de nível tático se concentram em ajustes de serviço; e estudos de nível estratégico geralmente se relacionam ao planejamento de rede de longo prazo, previsão de demanda e comportamento de viagens, e que no caso desta dissertação será focado neste nível para reconstrução dos atributos dos deslocamentos.

Os principais temas e aplicações que utilizam dados de transporte público, incluem geração (estimativa ou reconstrução) de matriz origem e destino (Barry *et al.*., 2002; Munizaga; Palma, 2012; Nassir *et al.*; 2015); motivo da viagem (Chu e Chapleau, 2008; Lee; Hickman, 2013); mineração de padrão de viagens (Ma *et al.*, 2013; Morency *et al.*, 2007) e Identificação de melhorias do transporte público (Hussain *et al.*, 2020). Dessa forma, este capítulo busca sintetizar as ferramentas e estudos sobre dados massivos de transporte público, reconstrução dos atributos das viagens, os principais modelos de mineração de dados abordados na literatura, e um tópico exclusivo sobre modelagem relacional de dados, devido a necessidade neste estudo.

## 2.1 Sistemas de Informação do Transporte Público de Passageiros

Se tratando de problemas específicos em transportes de passageiros, um dos "remédios" essenciais é o uso de Sistemas Avançados de Transporte Público (APTS). Além disso, a Administração Federal de Trânsito dos EUA emitiu um relatório de Hwang *et al.* (2006) sobre os aspectos constituintes do APTS, que enfatizou a necessidade de integração do mesmo com Sistemas de Transporte Inteligentes (ITS). Os autores, enfatizaram a necessidade de fornecer melhores dados e padrões destes para: (1) melhorar o planejamento, manutenção, operações e gerenciamento de incidentes de trânsito, e (2) facilitar a coordenação, integração e interoperabilidade com fornecedores de transporte e organizações de segurança pública.

Os termos APTS comuns usados nesta descrição são *Automatic Vehicle Location* - AVL (Localização Automática de Veículos), *Automatic Passenger Counter* - APC (Balcão Automático de Passageiros), *Computer Aided Dispatch* - CAD (Despacho Auxiliado por Computador), *Automatic Fare Payment* - AFP (Pagamento Automático de Tarifas) e *Advanced Traveler Information System* - ATIS (Sistema Avançado de Informação ao Viajante).

Todos esses sistemas servem de base para um melhor entendimento da rede de transportes, podendo ser utilizados para produção de indicadores. A proposição desses indicadores é de extrema importância no processo de tomada de decisão, pois é através deles que os tomadores de decisão são capazes de analisar desempenho, eficiência, eficácia, e representar o funcionamento do sistema (Rabay, 2017). A partir das delimitações em Hwang

*et al.* (2006), Chapleu *et al.* (2011) classificaram os APTS em quatro fontes principais de informação, que possibilitam a extração de indicadores do sistema de transporte público:

- AFC (Automated Fare Collection): Sistema de Bilhetagem Eletrônica
- APC (Automatic Passenger Counting): Contagem Automática de Passageiros
- AVL (Automatic Veicle Location): Localização Automática de Veículos
- GIS (Geographic Informtion System): Sistema de Informações Geográficas

A utilização desses sistemas permite a extração de um grande contingente de dados sobre a operação e demanda. Pensando nessas considerações, nos tópicos seguintes serão apresentados os principais Sistemas de Informação utilizados no planejamento de transportes público.

### *2.1.1 Rastreamento da Frota – AVL*

O rastreamento da frota de veículos inicialmente se aplicava a transportes de carga por questões de segurança e controle do deslocamento, mas os princípios de planejamento e eficiência, que sustentam a logística, também se aplica ao transporte de passageiros, como exemplo temos os rastreadores de ônibus. Esses aparelhos funcionam por meio de radiofrequência ou monitoramento via satélite, que é o caso dos aparelhos GPS, podendo se comunicar com até 24 satélites para obter a informações de coordenadas geográficas dos veículos (latitude, longitude, e alguns casos altitude). DOD (2008) relata erros de 7,8m na posição estimada pelo GPS instalado para AVL. É esperado que o erro aumente também quando o veículo estiver próximo a edifícios, túneis, árvores e pontes. Outro meio de estimar a localização do veículo é pelo *Bluetooth*, que está presente em muitos aparelhos telefônicos, permitindo a identificação individual de cada equipamento, que transmite informações ao passar por sensores instalados nas vias, ambas as tecnologias permitindo coleta automatizada dos dados. A vantagem do rastreamento de rotas nessa crescente era tecnológica, é a produção de aplicações que permitem com que os passageiros vejam em tempo real a localização exata dos ônibus e transportes coletivos, dependendo do sistema com um atraso de alguns segundos para agrupamento da informação. O sistema de GPS dos ônibus de Fortaleza, por exemplo, realiza esse agrupamento a cada 30s.

Rossetti (1996) relata um sistema de monitoramento de trânsito baseado em identificação por radiofrequência que integra localização automática de veículos (AVL) com contagem automática de passageiros. Okunieff (1997) realizou uma revisão sobre alguns sistemas automáticos de localização (AVL) que incluíam postos de controle, pontos de rádio e leitura de hodômetros. Além disso outros autores identificaram o GPS como a tecnologia mais vantajosa entre os sistemas de rastreamento, por não requerer estruturas físicas e funcionar em quaisquer pontos. Algumas desvantagens são o sinal de GPS não alcançar lugares cobertos, devendo ser utilizado uma tecnologia complementar (Dessouky *et al.*, 1999).

Na cidade de Fortaleza, os dados de GPS de frota de ônibus e coletivos estão sendo utilizados cada vez mais em pesquisas. Braga (2019) corrigiu os dados de do Feed GTFS de Fortaleza através dos dados de GPS, como etapa inicial do seu trabalho em que analisou a variabilidade de indicadores da acessibilidade às oportunidades de trabalho e educação.

### *2.1.2 General Transit Feed Specification - GTFS*

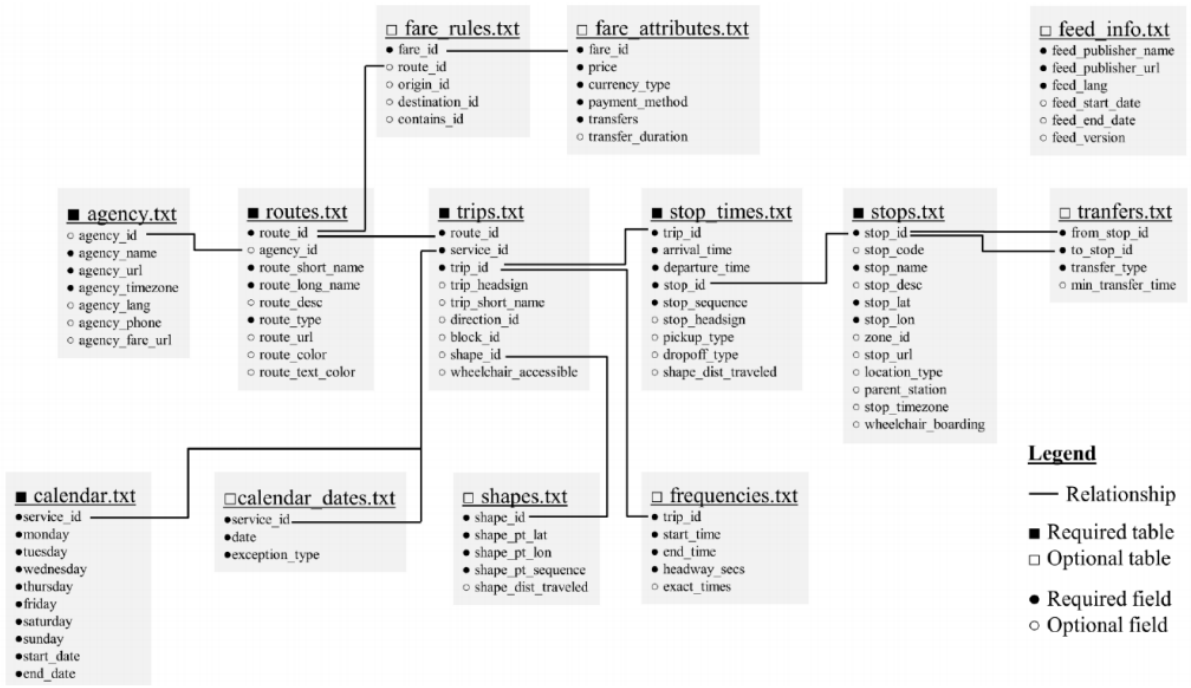
O Google foi a primeira empresa a criar uma plataforma para o processamento massivamente paralelo de dados, denominada *Google File System* (Google FS), posteriormente surgiram outras arquiteturas, como a *Hadoop Distributed File System* (HDFS). A Especificação Geral do Feed de Trânsito (*General Transit Feed Specification*), apresenta um formato comum de horários de transporte público e informações geográficas que possam estar associadas. A

Figura *1* a seguir descreve os arquivos que compõem o Feed, assim como suas relações. O padrão do Feed é composto prioritariamente por 13 arquivos, mas por dependências de órgãos gestores, não necessariamente todos estarão disponíveis.

O que se conhece atualmente como GTFS iniciou como um projeto secundário da Google criado em 2005. Nos Estados Unidos, local onde se criou o sistema, não havia nenhum padrão para programação de transporte público antes dos adventos do GTFS, em que empresas de gerenciamento de transportes tinham que dispor de dados em diferentes formatos (Developers, 2020). Por conta disso o formato público e grátis, assim como a disponibilidade dos horários, rapidamente fez com que desenvolvedores baseassem seu *software* relacionado a transporte nesse formato. O GTFS estabelece uma série de arquivos em formato CSV (*Comma Separated Values*) que, juntos, descrevem as paradas, viagens, rotas e informações de tarifas sobre o serviço de uma agência, e que de acordo com Wong (2013) é o padrão mais

usado para troca estática de dados de transporte público. Porém esse padrão não segue as regras de normalização para evitar redundância dos dados.

Figura 1 - Diagrama dos dados do Feed GTFS




Fonte: Wong, 2013.

### *2.1.3 Bilhetagem eletrônica*

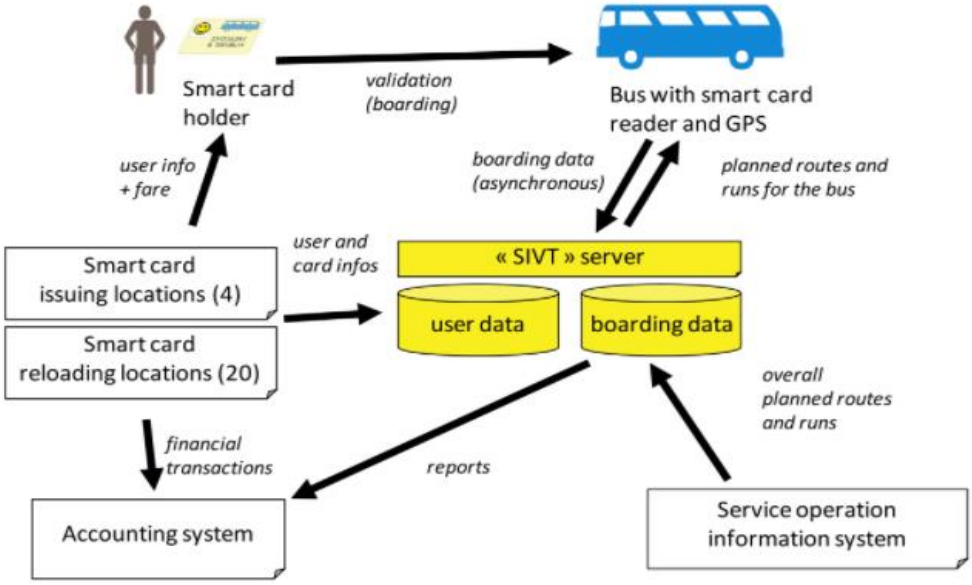
O conceito de Bilhetagem Eletrônica é bastante difundido como o pagamento do valor das passagens de forma eletrônica, utilizando dispositivos especiais, como cartão inteligente ou similares (ANTP, 2012). Mas, mais do que isso, é um forte ferramental para obtenção de informações sobre as condições de viagens dos passageiros. No contexto atual, a bilhetagem eletrônica tem evoluído para ser ainda mais segura e fornecer novos meios de pagamentos a usuários do transporte público por meio de bilhetes digitais. Com a evolução da tecnologia de *tockenização* e *Account Base Ticketing*, o cartão inteligente, pode ser substituído por qualquer meio de acesso ao transporte, seja um *smartphone*, relógio digital ou cartão de banco.

Quando o usuário apresenta seu cartão no validador é verificado se ele possui saldo (sendo esse acrescido antes da utilização do sistema em algum posto de recarga dos créditos, ou terminais), a identificação do cartão (sendo definida no momento do cadastro do usuário) e algumas informações relacionadas a rota e data/hora. Posteriormente o acesso é liberado e ao

final da operação quando o veículo é recolhido para a garagem, as informações do validador são coletadas e enviadas para o Sistema Central que armazena as informações (Freitas, 2015).

A representação do funcionamento do Sistema de Bilhetagem Eletrônica está demonstrada na Figura 2, onde o *smartcard* é o meio de acesso ao sistema, que é validado dentro do ônibus na entrada (e saída, em alguns sistemas). Esses dados são enviados ao servidor central, onde serão armazenados, além de informações das linhas poderem ser retornadas ao veículo, tendo um banco específico para as validações e outro para os cadastros, garantindo a integridade da informação, onde apenas o código identificador do cartão é armazenado no banco de validação. Conforme apresentado, outros sistemas da informação podem estar ligados aos dados de validação, à exemplo o GPS.

Figura 2 - Exemplo de Sistema de Bilhetagem eletrônica




Fonte: Pelletier *et al.* (2011)

Porém essa é uma representação generalizada por Pelletier *et al.* (2011), e que não corresponde totalmente a realidade de Fortaleza. Ainda inexistindo uma compatibilidade entre os diferentes Sistemas de Informação, no sentido de normalização dos tipos de informação, padronização dos dados e integração das bases. Grande parte da integração das bases existentes, partiu da necessidade de outros estudos para utilização desses dados, não do próprio órgão gestor.

Enquanto os cartões inteligentes estão aumentando principalmente a conveniência para os viajantes, as operadoras valorizam em particular as taxas reduzidas de manuseio de dinheiro (Kurauchi; Schmocker, 2016). Os autores ainda citam que em muitos casos, outros dados podem ser coletados, como um identificador de veículo, o número da rota, a direção da

viagem e se a transação foi um trecho de viagem de origem ou uma transferência de um trecho de viagem anterior. Para grandes redes de trânsito, a operação de um único dia pode render dezenas de milhares ou milhões de transações de cartão inteligente.

No caso de Seul é algo a parte, pois normalmente não se tem esse gama de informações em um sistema de *smartcards*. Entre os que normalmente não estão disponíveis tem-se: (i) o objetivo da viagem, (ii) o nível de renda, (iii) o gênero, (iv) idade do viajante e (v) o local de desembarque. De forma contrária, as informações que mais aparecem são: (i) a hora no dia da validação, (ii) o tipo de tarifa (Que pode ser usado para inferir a faixa etária) e (iii) a frequência da viagem (por exemplo, diária, semanal, mensal, etc.).

Outro exemplo segue no Japão, que desde 2013, a maioria dos cartões inteligentes das principais operadoras públicas podem ser usadas em todo o país. Enquanto na Holanda essa realidade já está mais avançada, pois um único cartão pode ser utilizado em todo o país para viagens de longa e curta duração. Isso significa que, assim como acontece com os cartões de crédito, o total das tarifas de transporte acumuladas ao longo de um mês debitará da conta no próximo mês, uma espécie de cartões pós-pagos de transporte. A desvantagem de sistemas tradicionais de pós-pagamento para o usuário, é que ele exige dados pessoais do usuário e um aplicativo de qualificação para obter cartões. Destacam-se os descontos oferecidos em Londres, onde os bilhetes em papel podem ter o dobro do preço do pagamento pelo cartão *Oyster*. Finalmente, deve-se notar que em algumas cidades, como Santiago do Chile, é obrigatório que os usuários obtenham um cartão inteligente, uma vez que o pagamento em dinheiro não é mais possível, constando com validações tanto para embarque, quanto desembarque (*tap-on/tap-off*).

Algumas das vantagens de se utilizar essa tecnologia são: sua grande quantidade de dados sobre o comportamento dos passageiros com menor custo; analisar o comportamento agregado; analisar dados em nível pessoal para entender a variação de comportamento e combinar dados com outras informações. Porém também existem algumas desvantagens assim como: falta de informações; problemas de tratamento de big data e questões contratuais de privacidade. Dessa forma, relacionado ao tamanho crescente dos dados, existem problemas relacionados a big data. Como os cartões inteligentes coletam continuamente o comportamento diário dos passageiros, o tamanho dos dados pode se tornar tão grande que às vezes é difícil de manusear (Kurauchi; Schomocker, 2016).

A limpeza de dados requer informações sobre as fontes e tipos de erros para refinar os dados. Os erros mais comuns para dados de transportes são erros humanos e falha no equipamento. A falha do equipamento pode ocorrer por conta de relógios dessincronizados dos dispositivos ou pode ser devido também ao leitor do cartão estar inoperante em algum momento do dia. Erros humanos podem incluir, esquecer de retirar o cartão, tocar o cartão no local errado, tocar o cartão mais de uma vez, entre outros.

As falhas citadas podem introduzir os seguintes erros nos dados de *Smartcard* (Luo *et al.*, 2017): (i) Nenhum registro de tempo de embarque/desembarque ou local; (ii) O tempo registrado ou local de embarque é igual ao tempo ou local de desembarque; (iii) O tempo de embarque é anterior ao tempo de desembarque em uma mesma viagem;(iv) Identificador do cartão ausente; (v) transação em uma parada não rastreável.

## 2.2 Reconstrução dos Atributos dos Deslocamentos do Big Data-TP

Nos tópicos a seguir serão apresentados aspectos sobre reconstrução das viagens, sendo um dos principais atributos o local de embarque e desmebarquer, discutido em um tópico à parte e sendo este o ponto central dessa dissertação. Também é apresentado um breve descritivo sobre cadeia de viagens, de modo a apoiar tópicos subsequentes sobre a identificação de padrões e modelagem do real local de embarque.

### *2.2.1 Métodos de reconstrução de viagens e pesquisas de campo*

Neste tópico serão abordados estudos que buscaram reconstruir viagens utilizando diferentes aspectos, sejam operacionais ou do próprio usuário. Trabalhos como de Kurauchi e Schmocker (2016) e Pelletier *et al*. (2011) ajudam a entender melhor o cenário da utilização da tecnologia de Sistema da Informação dentro do meio acadêmico, uma vez que agregaram vários trabalhos sobre essa temática. Cabe citar também Barry *et al*. (2002) que obteve indicadores através da utilização de *smartcard*. Liu *et al*. (2019) replicaram o sistema de transporte público de Singapura através de dados de Smartcard, utilizando alocação direta no software PTV Visum. Nassir *et al*. (2018) propuseram um modelo de escolha de caminho recursivo baseado em dados de cartão inteligente. Chen *et al*. (2016) fizeram uma análise das promessas do *Big Data* e *Small Data* para caracterização no comportamento de escolha de viagens. No Brasil pode-se citar os trabalhos de Mesquita *et al*.(2017) e Arbex e Cunha

(2017), que utilizaram sistemas da informação para estimação de matrizes O/D por paradas, utilizando-se de muitas premissas simplificadoras e uma confiabilidade no pendularismo dessas viagens.

Pfitscher *et. al*. (2020) utilizaram a linguagem de programação *python* para manipulação e construção de uma matriz O/D para o munícipio de Guaíba-RS, com uso de dados de bilhetagem e ADC do sistema de transportes público da cidade. Neste estudo, os autores procuravam registros de viagens nos dados de bilhetagem, demarcavam seu identificador e procuravam o mesmo em outras linhas, para que pudessem inferir o local de desembarque. Caso apenas um registro por dia fosse encontrado, aquela viagem era desconsiderada. Vale destacar também, que registros acima de um raio de 250m da parada de ônibus mais próxima, também foram desconsiderados. Estes fatos embora facilitem a análise, descartam muitas possíveis viagens, que poderiam ser estimadas por outros métodos.

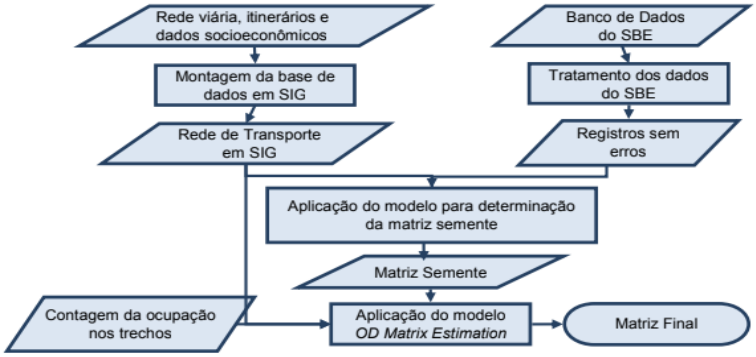
Munizaga e Palma (2012) utilizaram dados de *smartcard* para geração de matrizes desagregadas para a cidade de Santiago do Chile. Um fato interessante é que esta cidade em específico, tem grande foco no transporte público e investimento em tecnologias, tendo taxa de penetração nesse sistema de 97%. Sem contar no sistema de "pulsos", que emitem uma área de influência de 500m no início, meio e fim das rotas, para verificação da adequação da linha com o GPS. Dessa forma os autores levantam uma série de considerações para definir a diferença entre início e fim da viagem, com embarque e desembarque, sendo a primeira demarcada pelos pontos de início e fim da rota, e a segunda onde o passageiro validou para entrar no veículo e onde ele validou para descer do veículo.

Posteriormente foram considerados 3 casos necessários para tratamento: (i) viagens com origem conhecida e destino desconhecido, (ii) viagens detectadas, mas cuja origem e destino não são desconhecidos e (iii) viagens que ocorrem, mas não foram validadas pelo sistema. No primeiro caso, pode-se supor que o destino da viagem desconhecida segue alguma distribuição das viagens conhecidas, podendo ser determinado um fator de expansão. No segundo, é possível utilizar o mesmo fator de expansão do primeiro caso, mas com uma desagregação temporal. No terceiro caso, deve-se ter esses dados contabilizados por outras fontes.

Guerra *et al*., 2014 apresentam uma diferenciação entre métodos indiretos e diretos de estimação, sendo o primeiro de previsão da demanda e o segundo de pesquisa de campo. No qual o foco do trabalho, foi baseado nos métodos indiretos junto a montagem da rede em um

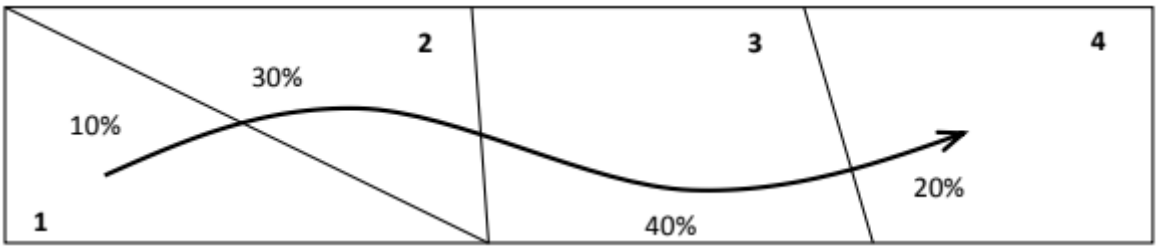
sistema SIG, além da caracterização e tratamento dos dados de bilhetagem. Desse modo o foco maior do trabalho se deu na proposição metodológica (Figura 3). Partindo-se da extração de dados de várias bases, como a rede viária do município de Maceió, itinerário e dados socioeconômicos. Seguindo da montagem da rede de transportes e tratamento dos dados do Sistema de Bilhetagem Eletrônica, averiguando registros sem erro. Para determinação da origem de cada usuário, foi necessário delimitar o tempo de início e término de cada linha, além da velocidade do veículo que percorre a mesma. Com o tempo inicial e final, a velocidade e os dados georreferenciados é possível saber o tempo estimado que cada veículo fica dentro de uma zona, tempo esse representado como a porcentagem do tempo total do percurso (Figura 4).

Figura 3 - Fluxograma da metodologia




Fonte: Guerra et. al (2014)

Figura 4 - Localização das linhas segundo o zoneamento




Fonte: Guerra et. al (2014)

### *2.2.2 Estimativa do local de embarque*

Na literatura a maioria dos modelos desenvolvidos são para os sistemas de entrada, enquanto os estudos realizados nos sistemas de entrada e saída, servem apenas para validar os

modelos desenvolvidos no sistema de entrada ou para avaliar outras questões que não sejam fluxos OD. No caso de trem/metrô, os passageiros devem acessar a entrada por uma estação. Se mais de um trem ou metrô serve a mesma estação, o trem de embarque e a direção não pode ser inferida pelos métodos tradicionais, requerendo suposições apropriadas. O mesmo cenário se assemelha no caso de terminais de transporte público, onde as pessoas validam dentro do veículo antes de entrar no terminal, ou na própria entrada do terminal, não tendo como saber, de forma direta, para quais linhas cada passageiro se deslocou dentro do terminal e efetuou o embarque.

Com base na disponibilidade dos dados registrados pelo sistema de entrada, a estimativa das paradas pode ser dividida em 3 categorias: (i) quando a parada de embarque e o tempo são registrados nos dados, sendo possível utilizar diretamente esses dados; (ii) quando o tempo de embarque é registrado, mas faltam informações sobre a parada de embarque, nesse caso pode-se obter a informação da parada integrando a base de bilhetagem com outras bases; (iii) Quando nem o tempo, nem informação das paradas estão disponíveis. Sendo este último incomum, mas existente. O caso I ocorre sempre para sistemas que são tap-on/tap-off. O caso II é comum em sistemas tap-on, pelo motivo do SBE e GPS não serem integrados (não armazenam a informação em conjunto). Para o caso III, não é possível identificar o local de embarque, nem mesmo com integração do GPS, porém esses dados podem ser distribuídos como proporções a partir de dados conhecidos (Hussain *et al*. 2021).

Mesmo sendo possível realizar essa integração, devido aos erros do GPS não é possível encontrar 100% das coordenadas das validações. É possível também, definir um raio para cada parada baseado em algum critério topológico da rede ou conhecimento do pesquisador, e atribuir as validações dentro desse raio para uma mesma parada (Munizaga; Palma, 2012). O problema desse método é que dependendo da linha, uma validação pode estar dentro de mais de um raio estabelecido para as paradas, podendo causar inconsistências na hora de atribuir a uma das paradas. Outro método engloba o uso de dados programados, quando os dados de AVL não estão disponíveis. Porém isso excluirá muitas transações das análises pela diferença entre os valores programados e o reais de operação.

Arbex e Cunha (2020) foram os pioneiros no Brasil a citar a problemática da validação dos sistemas *tap-on* não ser exatamente no local de embarque, devido a distância entre a porta de embarque e o validador. Embora os mesmos definam a problemática no escopo de análise

da variação da demanda, não avançam na discussão ou proposição de um método que englobe essa problemática.

### *2.2.3 Cadeia de Viagens*

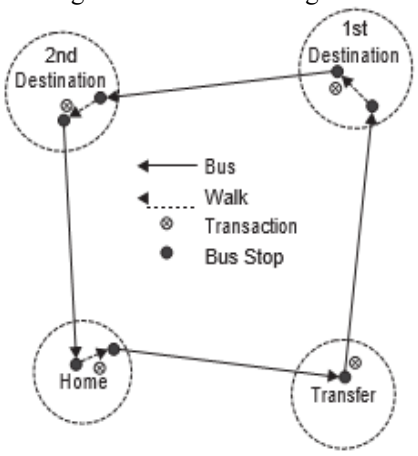
O toque em um Smartcard geralmente é suficiente para identificar o local de validação e o horário de início do trecho da viagem. Se o destino de um trecho de viagem e a hora de chegada forem desejados, será necessária uma derivação adicional do cartão inteligente ou um meio de inferir esse destino. Devido à prevalência de sistemas de validação única, muitos pesquisadores investigaram o problema de inferir destinos. LI *et al.* (2018) fez uma revisão da literatura sobre estimativas de destinos em sistemas apenas de entrada e classificou esses modelos como de probabilidade, encadeamento de viagem e de aprendizado profundo. A mescla desses modelos em um modelo híbrido não é bem explorada na literatura.

A técnica mais comum de inferir o destino e a hora de chegada usa a noção de uma "cadeia de viagem", descrevendo a cadeia de trechos de viagem que uma pessoa fará em um único dia. A cadeia pressupõe que o destino de um trecho de viagem esteja próximo à origem do próximo trecho de viagem e que o destino do último trecho de viagem na cadeia seja próximo à origem do primeiro trecho de viagem (Kurauchi; Schomocker, 2016). Também se parte da premissa de que nenhuma jornada é feita por um modo diferente, logicamente sendo necessário pelo menos duas viagens para constituir uma cadeia (ida e volta). Um exemplo de cadeia de viagens é demonstrado na Figura 5

O problema, então, é inferir os locais de transferência ou destino. Como uma técnica, pode-se escolher a parada mais próxima na rota anterior. Na Figura 5, o ponto de desembarque no primeiro ônibus pode ser inferido como a parada nessa rota mais próxima da segunda transação. Da mesma forma, o ponto de desembarque do segundo ônibus pode ser inferido como a parada na segunda rota mais próxima do terceiro local de transação. Também é necessário estimar o tempo de desembarque do passageiro, através do tempo que o ônibus demora para chegar a esse local. Portanto dentre as suposições mais comuns para inferir uma cadeia de viagem, incluem: (i) O destino da última etapa de uma viagem é idêntico à origem da primeira etapa da viagem; (ii) Os passageiros geralmente percorrerão as trilhas a pé mais diretas entre os serviços, conforme medido pelo tempo, pela distância ou algum tempo ou

custo generalizado; e (iii) Os passageiros farão o próximo serviço disponível após chegar a uma estação/parada.

Figura 5 - Cadeia de viagens




Fonte: Kurauchi e Schumocker, 2016.

Barry *et al.* (2002) usaram esse algoritmo para inferir destinos no metrô de Nova York. Para certificar o pressuposto da cadeia de viagens, eles empregaram uma amostra de 100 passageiros que fizeram apenas 2 viagens e 150 passageiros que fizeram cadeias de 3 ou mais viagens em um único dia. Em ambas as amostras, 90% dos destinos puderam ser inferidos com sucesso. Trépanier *et al.* (2007) sugeriu algumas melhorias para esse algoritmo. Em primeiro lugar, nos casos em que vários dias de dados de cartão inteligente estão disponíveis, o último local de desembarque em um determinado dia é dado como o local de embarque inicial da primeira viagem no dia seguinte. Outro aspecto levantado refere-se, aos os trechos da viagem em que um local de desembarque não pode ser inferido de outra forma, o destino pode ser inferido como um ponto de desembarque para o mesmo passageiro, se ele tiver usado historicamente a mesma rota e ponto de embarque. Através dessas duas melhorias, embora 21% dos dados tenham sido errôneos e 13% sem sucesso na inferência, teve-se 66% sucesso na inferência dos desembarques. Li *et al.* (2018) propuseram um método melhorado para viagens desvinculadas cujo desembarque não pode ser estimado baseado na cadeia de viagens. Nesse tipo de inferência se utiliza um método probabilístico de densidade de kernel para encontrar a parada de desembarque, observando os dados históricos do cartão inteligente.

Métodos de inferência podem ser utilizados para outros fins, não apenas para encontrar o destino de uma viagem. Jang (2010) por exemplo, utilizou um algoritmo de classificação em árvore escrito em linguagem R (o C503, uma derivação do C50 ambos os

modelos de inferência em árvore) para prever os propósitos das viagens. Além desses algoritmos baseados em regras, recentemente, os pesquisadores estão adotando modelos de aprendizado de máquina para fazer inferências. O principal objetivo é diminuir as premissas do encadeamento de viagem devido a complexa natureza do comportamento dos passageiros. Jung e Sohn (2017) usaram aprendizado de máquina supervisionado, empregando unidade linear retificada, com duas camadas ocultas na rede entre as camadas de entrada e saída para estimar o local de desembarque. Eles utilizaram 27 variáveis relacionadas ao *smartcard* e uso do solo. Assemi *et al.* (2020) propôs uma metodologia envolvendo rede neural para inferir informações que o encadeamento de viagem, feito previamente, não conseguiu realizar. Na realidade a proposta de modelos de aprendizado para uma mesma abordagem é aumentar a precisão do algoritmo. Esse modelo, apresentou 79,5% de precisão em comparação a um modelo unicamente baseado em regras com 72,2%.

## 2.3 Identificação de padrões de uso e dos locais de embarque

Um aspecto essencial do sistema de transportes público de passageiros é a previsão dos deslocamentos que é uma questão significativa na área do planejamento de transportes, devido à sua importância operacional (Thiagara e Prakashkumar, 2021). São muitos avanços e aplicações inovadoras que têm sido introduzidos para um ambiente mais seguro e eficiente. Para tanto as implicações destas alterações na oferta, apenas são possíveis com a compressão de como a demanda de usuários por transporte público se desloca na rede. A fim de reconhecer tendências úteis para melhorar o plano de programação, é necessário implementar abordagens adequadas de aprendizagem automática. Estudos recentes adotaram uma abordagem utilizando agrupamentos espaciais baseados na densidade de aplicações com ruído (DBSCAN) com o algoritmo de média móvel integrada autoregressiva sazonal (SARIMA) para analisar como os passageiros utilizavam o sistema (Thiagara e Prakashkumar, 2021).

Cats *et. al* (2015) examinaram a distribuição espaço-temporal do fluxo de passageiros dos transportes públicos de Estolcomo (Suécia) para identificar e classificar centros de atividades baseados nestes dados de mobilidade. Conforme apresentado a maioria dos estudos que se dispõem a analisar padrões no sistema de transporte público está interessado unicamente na questão da mobilidade do usuário, sem se a unificação com aspectos metodológicos de como se deu o movimento (localização dos embarque e desembarques). Lin

*et. al* (2020) propuseram um estudo que avaliou como os usuários pendulares se deslocam na rede. A compreensão dos padrões de deslocamento pendular pode fornecer um apoio eficaz ao planeamento e à operação dos sistemas de transportes públicos. Estes usuários podem ser divididos em grupos menores baseado em características como hora de partida da primeira viagem e tipo de usuário. Ainda não foi verificado a relação destas variáveis no processo de deslocamento e na formação dos grupos, bem como outras variáveis relacionadas a operação ainda não foram testadas. Diferentemente dos trabalhos citados anteriormente que se dispuseram de técnicas de mineração de dados para avaliar padrões, Chakirov e Erath (2012) investigaram as oportunidades de detecção das atividades primárias, como atividades domésticas e profissionais, e suas localizações com base nos registros do sistema de pagamento de tarifas por cartão inteligente para os transportes públicos. O método utilizado foi uma abordagem de escolha discreta, e avaliado a inclusão de variáveis de uso do solo nestas identificações de padrão.

Quanto a identificação do local de embarque de usuários de transporte público é um problema importante na otimização do sistema de transporte e no aprimoramento da experiência do usuário. Vários estudos têm sido conduzidos para desenvolver modelos e técnicas que possam ser utilizados para identificar o local de embarque com base em dados disponíveis, como a localização do local de validação e informações históricas de viagens. Alsrehin *et al.* (2019) propuseram uma abordagem de agrupamento de trajetórias baseada em *clustering* para a identificação de passageiros em ônibus. Eles utilizaram dados de GPS para obter informações de trajetória e, em seguida, aplicaram um algoritmo de *clustering* para agrupar os passageiros com base em suas trajetórias.

# 3 INTELIGÊNCIA ARTIFICIAL PARA TRANSPORTE PÚBLICO

## *3.3 Inteligência Artificial para identificar padrões em sistemas de TP*

Antes da discussão sobre mineração dos padrões de uso é necessário primeiramente deixar claro do que se trata esta etapa, sendo também um aprendizado de máquina. A mineração de dados é uma abordagem exploratória que visa descobrir padrões e informações úteis em grandes conjuntos de dados. Ela envolve o uso de técnicas estatísticas e computacionais para identificar padrões e tendências em dados brutos, assim como o modo correto de extraí-los e armazená-los. Portanto, é necessário deixar claro algumas vantagens e desvantagens do aprendizado de máquina. De forma geral, estes algoritmos podem ser utilizados sem restrições matemáticas ou de distribuição populacional (ou mesmo para dados com colinearidade). Também são úteis quando se têm grandes contingentes de dados em comparação a outros tipos de modelos. Porém, mesmo podendo ser utilizados para classificação de observações, perdem bastante explicações em relação a alguns modelos tradicionais. Não se pode de fato quantificar relações com todo o embasamento de intervalos de confiança, por exemplo (Shalit *et al.*, 2022; Tang *et al.*, 2023).

## *3.2 Aprendizado de Máquina*

Esse assunto despertou atenção novamente com a evolução da capacidade de processamento dos computadores, não sendo mais um processamento apenas serial nas Unidades de Processamento Central (CPU), mas paralelo nas chamadas Unidade de Processamento Gráficas (GPUs), além da necessidade de técnicas de análises e previsões a partir de grandes massas de dados (Davis e Goadrich, 2006; Power, 2011; Géron, 2019)

Dessa forma, assim como seres humanos conseguem aprender, os computadores são submetidos a técnicas de aprendizado para construção de modelos que conseguem “aprender” com os dados e fazem com que sejam capazes de resolver problemas. Aprender nesse caso pode significar classificar objetos, agrupar itens ou responder perguntas sobre imagens. São várias as tarefas que podem ser realizadas através de aprendizado de máquina. Por exemplo:

- **Tomada de decisão:** O computador é capaz de analisar uma base histórica de dados, entender os padrões e decidir se vai aceitar ou rejeitar um parâmetro.

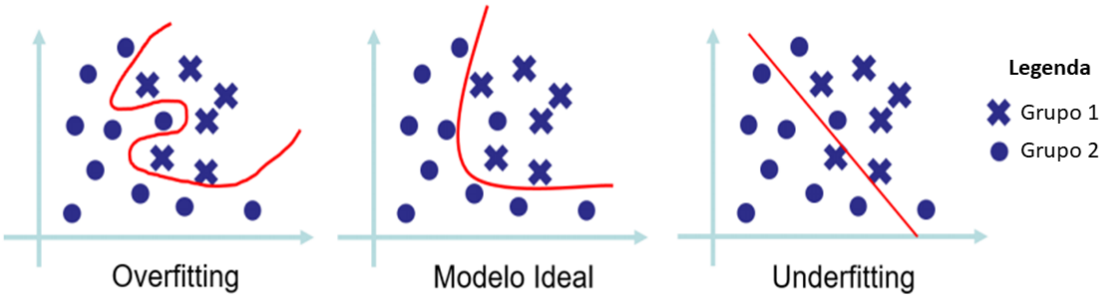
- **Regressão:** A partir de um conjunto de características de um conjunto de treinamento, é possível prever padrões em grupos semelhantes.
- **Classificação:** A partir de um conjunto de dados para treinamento do modelo, é possível identificar se um novo dado pertence a uma dada categoria estabelecida previamente.
- **Clustering (ou agrupamento):** É possível também utilizar as técnicas de aprendizado de máquina para identificar padrões nos dados e criar grupos similares, ou seja, uma base de usuários de transporte público pode ser subdividida de acordo com características de sexo, idade, renda familiar, etc.

Para que o computador consiga aprender a resolver uma determinada tarefa é preciso instruí-lo da melhor forma possível. Isso pode ser feito por exemplo a partir de uma base de dados históricos. Porém existem casos em que não é possível coletar os dados históricos, por questões de tempo, esforço ou recursos. Para esses casos, existem técnicas de aprendizado específica:

- **Aprendizagem supervisionada:** Para esse tipo de aprendizagem existe previamente um conjunto de dados onde é possível conhecer qual é o resultado esperado para cada entrada do modelo. Com isso o sistema é capaz de avaliar se o resultado encontrado está próximo ao resultado correto. Esse tipo de análise pode-se utilizar a relação de 75-80% dos dados disponíveis para o grupo de treinamento e 20-25% para o grupo de teste, sendo cabível sempre verificar a distribuição dos mesmos. Se necessário o autor ainda afirma que se pode retirar 10% do conjunto total de dados para se fazer uma validação.
- **Aprendizagem não-supervisionada:** nesse caso não existe um conjunto de treino. Consequentemente não existe um resultado específico esperado e não é possível prever os resultados do cruzamento das informações. Por isso é um tipo de aprendizado comumente utilizado para descobrir padrões entre os dados e resolver problemas de agrupamento.
- **Aprendizagem por reforço:** nesse tipo de aprendizagem existe um ambiente e um agente que interagem entre si através de percepções e ações. A cada iteração o agente recebe uma indicação do estado atual do ambiente e escolhe uma nova ação. A ação altera o estado do ambiente e o agente recebe um sinal de reforço. Ao final, o agente terá uma política de comportamento.

É comum em sistemas de aprendizado de máquina que um modelo construído seja específico demais ou que não consiga generalizar para outros casos. O ideal é que seu aprendizado seja bastante equilibrado. Quando um modelo é treinado pode acontecer de, na fase de treinamento, alcançar uma taxa de erro muito baixa, dando a impressão de que é um ótimo modelo (Figura 6).

Figura 6 - Overfitting e Underfitting




Fonte: Géron (2020).

Porém o desempenho é péssimo quando aplicado a um conjunto de teste. Isso provavelmente significa que é um caso de *overfitting*, ou seja, o modelo não tem a capacidade de generalização. Nesse caso o modelo memoriza os dados em vez de aprender. Quando um modelo é treinado e na fase de treinamento a taxa de erro é relativamente alta e na fase de testes é mais alta ainda, provavelmente é um caso de *underfitting*, ou seja, o modelo é simples demais e acaba subestimando a realidade. Modelos ideais tem taxas de erros moderadas, possibilitando a generalização para outros casos (Davis e Goadrich, 2006; Power, 2011; Géron, 2019).

### *3.3 Algoritmos supervisionados e não-supervisionados*

No aprendizado supervisionado, os dados de treinamento fornecidos ao algoritmo incluem as soluções desejadas, chamadas de rótulos. A *classificação* é uma tarefa típica de aprendizado supervisionado. Sendo utilizado também para prever um alvo de valor numérico, através de um modelo de regressão. Alguns algoritmos de regressão também podem ser usados para classificação e vice-versa. Os principais algoritmos supervisionados são (Géron, 2019; Cats e Ferranti, 2022; Zhao et. al, 2023): (i) K-Nearest Neightbours; (ii) Máquinas de Vetores de Suporte (SVM); (iii) Árvore de Decisão e Florestas Aleatórias; e (iv) Redes Neurais.

No aprendizado não-supervisionado, utilizado para identificar padrões os dados de treinamento não são rotulados. O Sistema tenta aprender sem um “professor”. A seguir estão alguns dos principais algoritmos de aprendizado não-supervisionado (Géron, 2019; Cats e Ferranti, 2022; Zhao et. al, 2023): (i) Clustering (K-Means, Clustering Hieráquico, Maximização de Expectativa); (ii) Visualização e redução de dimensionalidade (Análise de Componentes Principais, Locally Linear Embedding, *t-distributed Stochastic Neighbor Embedding*); e (iii) Aprendizado da regra de associação (A priori, ECLAT)

A dimensionalidade é uma tarefa relacionada na qual o objetivo é simplificar os dados sem perder muita informação. Uma maneira de fazer isso é mesclar várias características correlacionadas em uma, também conhecido como extração de características. Outro importante tarefa não supervisionada é a *detecção de anomalias,* como detectar transações incomuns nos dados de *smart card* ou remover *outliers* de um conjunto de dados antes de fornecê-lo a outro algoritmo de aprendizado.

Outro critério utilizado para classificar os sistemas de Aprendizado de Máquina é se o sistema pode ou não aprender de forma incremental a partir de um fluxo de dados recebido. No aprendizado em lote, o sistema é incapaz de aprender de forma incremental, devendo ser treinado com a utilização de todos os dados disponíveis. Esse tipo de sistema aprende uma única vez e depois é colocado em produção, sem aprender mais nada. Caso necessite treinar com novos dados é necessário treinar uma nova versão do modelo. Caso envolva grandes quantidades de dados, o sistema deve ser automatizado, podendo ficar impossível utilizar esse tipo de algoritmo caso contrário. Uma segunda opção é o aprendizado *online*, onde se treina o sistema de forma incremental, alimentando sequencialmente as instâncias de dados individualmente ou em pequenos grupos, chamados *minilotes*. Esse tipo de modelo requer menor poder computacional, se configurando como um processo mais barato. Um parâmetro importante dos sistemas de aprendizado online é a rapidez com que eles devem se adaptar às mudanças dos dados, chamada de *taxa de aprendizagem*. Taxas de aprendizado alta fazem com que o sistema se adapte rapidamente a novos dados, mas também se esqueça rapidamente dos dados antigos. O inverso faz com que o sistema se mantenha em inércia, sendo menos sensível ao apontar novos dados ou sequências de pontos de dados não representativos.

Dessa forma, é necessário definir uma medida de desempenho, que pode ser uma função de utilidade (*função fitness*) que mede o quão bom é o modelo, ou uma função de custo, que mede o quão ruim é esse modelo. Encontrar os melhores parâmetros de um modelo

com vários dados de treinamento é o que se referência como *treinar o modelo* e a etapa de aplicação do modelo com dados de teste é chamado de *Inferência*. Essas soluções são conhecidas como *regularização* do modelo.

Por fim, a única maneira de saber o quão bem um modelo generalizará em novos casos é de fato testá-lo em novos casos. Um modo de testar é pôr o modelo em produção e monitorar a qualidade do seu desempenho. Outra opção é dividir os dados em dois conjuntos: o *conjunto de treinamento* e o *conjunto de teste*. A taxa de erro em novos casos é chamada de *erro de generalização* (ou erro fora da amostra), e deve ser avaliado ao testar o modelo. Se o erro de treinamento for baixo, mas o erro de generalização é alto, isso significa que o modelo está sobreajustado aos dados de treinamento. Para evitar o "desperdício" de dados é comum utilizar técnicas de *validação cruzada,* onde o conjunto de treinamento é dividido em subconjuntos complementares e validado em relação às partes restantes (Géron, 2019).

De acordo com Habib & Weiss (2014), a maioria das viagens que ocorrem no Transporte Público é realizada por viajantes pendulares. As viagens de diferentes passageiros de ônibus, podem não ser consistentes, o que tem um impacto na precisão de suas previsões de demanda por viagens. As viagens estáveis, detém a prioridade temporal e características espacialmente dinâmicas, como frequência de viagem, rotas de viagem e valor de transação, enquanto captura o movimento de longo prazo dos viajantes devido à relação entre estabilidade e mobilidade (Cui e Long, 2015; Cui et al., 2014). Assim, infere-se que a estabilidade das viagens reflete a longo prazo, movimentos dinâmicos característicos dos viajantes no cotidiano influenciados pelas áreas de atividades, ambiente construído e políticas de viagens.

Cui e Long (2015) primeiro classificaram os viajantes em duas categorias e, em seguida, as interdependências da mobilidade e estabilidade temporal dos residentes foram analisadas combinando a matriz de transição dos padrões de viagem temporal de passageiros e dados socioeconômicos para Pequim. Enquanto a maioria dos estudos existentes usam apenas alguns indicadores para descrever a estabilidade de tempo de viagem do passageiro, os métodos de pesquisa e os resultados são relativamente simples para refletir o significado inerente de estabilidade de viagem.

Cats e Ferranti (2022) propuseram um estudo para avaliar os padrões temporais de mobilidade usando dados de *smart card* no sistema de TP em Estolcomo, Suécia. Se ampararam em técnicas de classificação (*k-means*) padrão e classificação hierárquica. Dessa

forma foi possível encontrar 10 padrões de deslocamento desde viajantes regulares em horário pico, até viajantes regulares durante a madrugada. Apenas 70% dos dados foram utilizados por conta do grande número de inconsistências. Ainda sobre a classificação dos usuários em relação aos padrões, Zhao *et. al*. (2023) definem outras duas categorias, os pendulares, que são um tipo típico de passageiros, que possuem fortes repetições ao viajar entre lugares em tempo fixo. E os não-pendulares que exibem maior variabilidade no espaço e tempo de trânsito para atividades não-obrigatórias, tornando-os difíceis de serem detectados por abordagens baseadas em regularidade.

Os estudos anteriores analisaram apenas parcialmente as viagens através do horário de partida, modo de viagem e localização do emprego, detendo abordagens limitadas de coleta de dados e tecnologias. No entanto, o rápido progresso da inteligência de transportes forneceu uma fonte de dados favorável para o estudo de identificação de comportamento de viagem e análise das características dos passageiros durante um longo período. Em geral, estes estudos visam identificar as diferentes formas de uso do sistema para caracterizar os usuários em diferentes níveis de similaridade tanto de espaço quanto de tempo. Porém ainda existem lacunas quanto a aplicação dos métodos levantados para compreensão dos padrões de forma que possibilite reconstruir atributos das viagens baseado nos padrões levantados.

Os principais indicadores encontrados na literatura para avaliar padrões se referem ao uso pela intensidade e regularidade, havendo variação desses conceitos entre os estudos (Morency *et al*.,2007; Ma *et al*., 2013; Huang *et al*., 2015). Ma *et al*. (2013) utilizaram dados de Pequim, para gerar dois tipos de agrupamentos, um individual para verificar a constância de uso no tempo e espaço, e em grupos para agrupar os usuários em relação à aspectos de regularidade. Os algoritmos de mineração podem ser classificados de acordo com sua tarefa, ou propósito particular, variando suas implementações e adequando-se a novas finalidades. 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, 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, 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 "descendente" (Ex.: AGS, Genitor, 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*, FCMdd).

Dessa forma, o conjunto de modelos utilizados neste trabalho configuram modelos de mineração de dados com foco supervisionado e não-supervisionado. O avanço tecnológico tem possibilitado o desenvolvimento e a aplicação de diversos modelos de aprendizagem de máquina em diferentes áreas, incluindo o transporte público. Dentre os principais algoritmos nesse novo contexto o algoritmo de *k-means* tem sido utilizado em diversas áreas para agrupar dados e identificar padrões em grandes conjuntos de informações. Esse modelo pode ser aplicado para segmentar usuários de transporte público baseado em suas preferências de rotas. Outra aplicabilidade de algoritmos se refere ao *Random Forest (*Cutler *et. al*, 2008*).* Lin *et. al* (2020) previu a demanda de passageiros em diferentes horários do dia, o que permitiu às empresas de transporte planejarem a oferta de serviços de forma mais eficiente. Já para modelagem de classificação é comum a utilização da modelagem com *naive bayes.* Por fim, tem-se exemplos de redes neurais para previsão de atraso nos horários de chegada dos ônibus, permitindo que os passageiros fossem informados em tempo real sobre eventuais alterações no serviço.

### *3.4 Redes Neurais*

Redes neurais artificiais (RNA) são sistemas de processamento de informações que foram desenvolvidos com base em modelos matemáticos inspirados no funcionamento do cérebro humano. As redes são capazes de aprender a partir de exemplos e de tomar decisões autônomas com base em padrões complexos e informações não lineares. Segundo Haykin (2008), uma rede neural é composta por muitos processadores simples, altamente interconectados, que operam em paralelo para resolver problemas complexos de processamento de informação.

Uma das principais aplicações das redes neurais é em visão computacional, onde elas são usadas para reconhecimento de padrões e identificação de objetos em imagens. Rumelhart *et. al* (1986) introduziram o algoritmo de retropropagação, que é a base para o treinamento das redes neurais. Esse algoritmo permite que a rede ajuste seus pesos sinápticos para minimizar o erro entre a saída da rede e a saída desejada para um determinado conjunto de dados. Segundo Kocadaglı (2015) as redes neurais podem ser utilizadas no processo de planejamento para identificar padrões nos dados históricos e fazer previsões com base nesses

padrões. Em particular, a rede neural de retropropagação tem sido aplicada com sucesso na previsão de demanda por ônibus e metrô (Haykin, 2008) .

### *3.5 Considerações finais*

Ainda existem muitos desafios para os gestores e analistas, alguns específicos para uso de dados AFC, relacionado a proteção da privacidade dos usuários e acesso ao AFC. Outros desafios afetam todas as fontes de dados, como: falta de recursos internos; conhecimento técnico para geração, extração, transformação e carregamento inadequados; dados conflitantes; gerenciamento dos dados por parte dos gestores de forma pouco eficaz, voltados apenas para o volume massivo, esquecendo-se da veracidade, velocidade de obtenção e valor. Dessa forma, se faz mais presente no cotidiano o uso de termos como *Business Intelligence* e Tomada de Decisão orientada a dados. Pesquisadores de todo mundo, nas mais abrangentes áreas utilizam esses conceitos, além de ferramentas analíticas como suporte à resolução de problemas. E embora tenha tido um crescente "aumento na curva" de exploração do *Data Science*, a maioria dos sistemas de trânsito não optou, ou não foi capaz de seguir por esse caminho, tendo seus esforços voltados para melhora da operação em tempo real. O que pode ser um entrave para a evolução do sistema, já que a demanda é fator determinante para se ter uma operação condizente com a necessidade.

Mais recentemente, a discussão aumentou no domínio público em geral, bem como na indústria de transportes, em tópicos relacionados, incluindo propriedade de dados, acesso a dados e aplicativos de dados abertos. O setor de transporte público recentemente começou a explorar como se relaciona com essas questões e desenvolvimentos e está cada vez mais abrindo seus dados, a passos curtos, para pesquisadores e desenvolvedores de aplicativos. Por fim, com o intuito de contribuir com a temática de estimação de padrões de deslocamento a partir de dados automáticos do transporte público e para que futuros pesquisadores e profissionais que vierem a ler este trabalho tenham um guia que ajude a nortear suas pesquisas, apresenta-se um resumo dos principais trabalhos (Tabela 1) citados nestes capítulos, bem como suas áreas temáticas de atuação (ordenados por essa categoria) e respectivas contribuições. Vale ressaltar, que cada trabalho pode haver mais de uma contribuição em áreas diferentes, portanto foi definido apenas uma área de maior contribuição nestes casos.

Tabela 1 - Resumo dos principais trabalhos científicos citados

| Temática | Autor | Ano | Contribuição |
|---|---|---|---|
| Sistemas da informação para Transporte Público | Hwang et. al. | 2006 | Definição conceitual do Sistemas Avançados de Transporte Público (APTS) |
| Sistemas da informação para Transporte Público | Chapleu et al | 2011 | Derivação do APTS para Sistemas da informação: AVL, APC, AFC e GIS. |
| Sistemas da informação para Transporte Público | Pelletier et. al. | 2011 | Revisão da literatura sobre os usos de sistemas de Smartcard para transporte público |
| Sistemas da informação para Transporte Público | Kurauchi e Schmocker | 2016 | Discursão e resumo de trabalhos que utilizaram sistemas da informação nas últimas décadas, e principalmente sistemas de bilhetagem em várias cidades pelo mundo |
| Sistemas da informação para Transporte Público | Rabay | 2017 | Utilizou Sistemas de Bilhetagem e geolocalização de linhas para gerar indicadores do TP |
| Sistemas da informação para Transporte Público | Braga | 2019 | Análise da Variabilidade de indicadores da acessibilidade às oportunidades de trabalho e educação, por intermédio de dados de GTFS, GPS e Bilhetagem. |
| Sistemas AVL para transporte público | Rossetti | 1996 | Sistema de monitoramento de trânsito baseado em identificação por radiofrequência |
| Sistemas AVL para transporte público | Okunieff | 1997 | Revisão sobre alguns sistemas automáticos de localização (AVL) que incluíam postos de controle, pontos de rádio e leitura de hodômetros |
| Sistemas AVL para transporte público | Chakroborty e Kikuchi | 2004 | Previsão dos tempos de viagem em corredores urbanos |
| Sistemas AVL para transporte público | Córtes et al. | 2011 | Delimitação de 3 linhas de pesquisa com uso de GPS: Para controle da operação; Estimação da velocidade e Planejamento de horários. |
| Sistemas AVL para transporte público | Córtes et al. | 2011 | Delimitação de 3 linhas de pesquisa com uso de GPS: Para controle da operação; Estimação da velocidade e Planejamento de horários. |

Fonte: Autor.

(Continuação)

| Temática | Autor | Ano | Contribuição |
|---|---|---|---|
| Sistema de Bilhetagem de Transporte Público | Pelletier; Trépanier; Morency | 2011 | Estimativa de padrões de viagem e geração de matrizes origem-destino com dados de bilhetagem |
| Estudos específicos de Smartcards | Barry et al | 2002 | Obtenção de Indicadores de TP com dados de Smartcard |
| Estudos específicos de Smartcards | Morency et al. | 2007 | Mediu o desempenho da rede de trânsito para resoluções espaciais e temporais |
| Estudos específicos de Smartcards | Trépanier et. al. | 2007 | Utilizaram técnicas de mineração de dados, para geração de clusters dos usuários em grupos diferentes |
| Pesquisas Domiciliares | Henrique | 2004 | Discute as premissas, vantagens e desvantagens na utilização de pesquisa domiciliar para compreensão da demanda |
| Geração de Matrizes OD | Barry | 2009 | Pioneiro em incluir viagens por todos os modos por meio da inferência dos desembarques e estimação dos embarques. |
| Geração de Matrizes OD | Guerra | 2011 | Defini conceitualmente os parâmetros de uma matriz OD |
| Geração de Matrizes OD | Muniziga e Palma | 2012 | Determinação de indicadores de nível de serviço por meio da reconstrução de matrizes OD para Santiago do Chile |
| Geração de Matrizes OD | Arbex e Cunha | 2018 | Utilizaram sistemas da informação para estimação de matrizes O/D por paradas |
| Expansão de matrizes semente | Cui | 2006 | Geração de matrizes OD Semente e Expansão por meio do método de Furness, estimativa de verossimilhança e ajuste proporcional |
| Expansão de matrizes semente | Li et al | 2011 | Geração de matrizes OD Semente e Expansão por meio do método de Furness, estimativa de verossimilhança e ajuste proporcional |
| Expansão de matrizes semente | Gordon et al | 2013 | Métodos de expansão baseados em caminhos O-D individuais, em vez dos fluxos O-D agregados, e regras para diferenciar transferências de atividades curtas |
| Comportamento do usuário na escolha de rotas | Lee e Hickman | 2013 | Estudaram as variações no comportamento de diferentes grupos de usuários (Estudantes, funcionários, etc.) |
| Geração de Matriz OD a nível de ZAT | Farzin | 2008 | Utilizou o GPS para encontrar a localização de ônibus e assim conseguir atribuir suas respectivas zonas de origem para a cidade de São Paulo |

Fonte: Autor.

(Continuação)

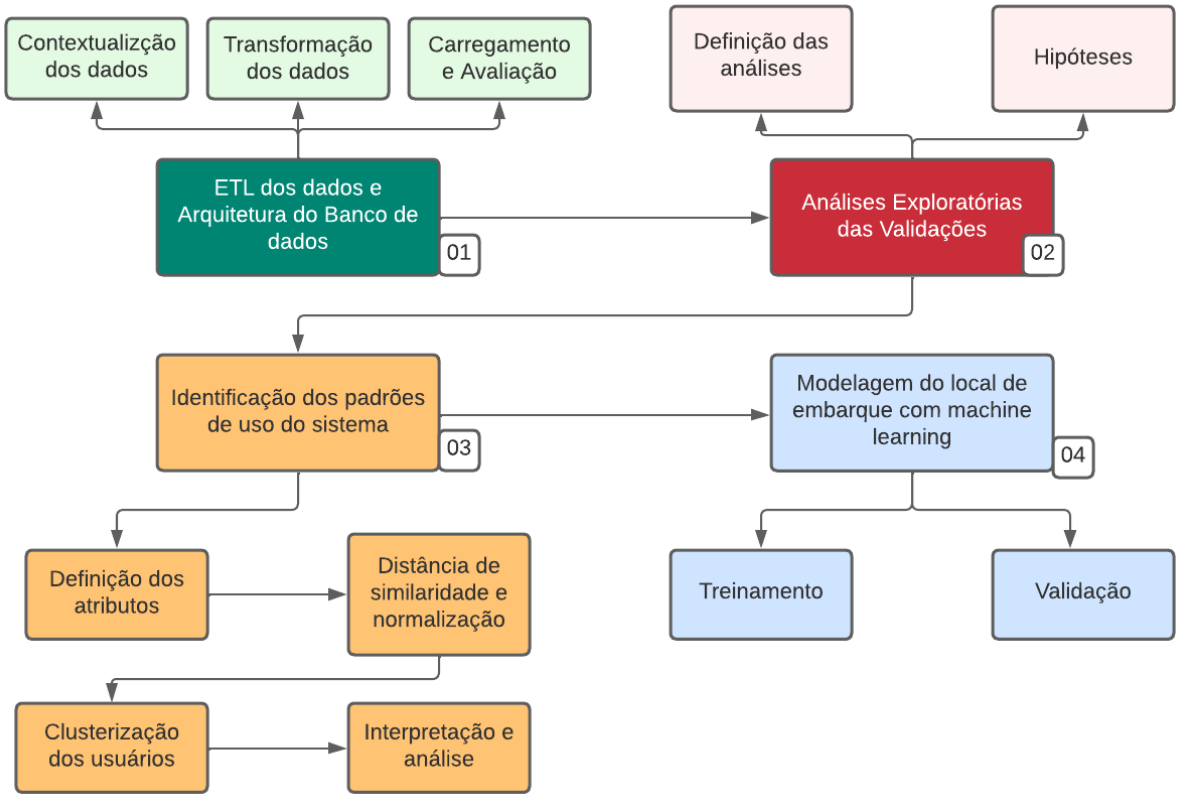
| Temática | Autor | Ano | Contribuição |
|---|---|---|---|
| Inferência sobre rotas | Agard et al. | 2006 | Investigou a variabilidade temporal e espacial de viajantes com validações através de SmartCard e agrupamento pelo método k-means |
| Inferência sobre rotas | Nassir *et Al* | 2018 | Propõem um modelo de escolha de caminho recursivo baseado em dados de cartão inteligente |
| Estudos sobre pesquisa de campo / Reconstrução de Matrizes OD | Liu et. al | 2019 | Discute as limitações sobre a pesquisa de campo sobre a qualidade dos dados, além de replicar o sistema de TP de Singapura com dados de smartcard |
| Reconstrução de Matriz OD | Pfitscher et. al | 2020 | Utilizaram a linguagem Python, dados de bilhetagem e ADC para reconstruir a matriz por paradas da cidade de Guaíba-RS |
| Reconstrução de Matriz OD | Zhao | 2004 | Utilizou dados GIS, bilhetagem e AVL para desenvolvimento de uma matriz OD para o sistema de transporte público por trilho de Chicago |
| Reconstrução de Matriz OD | Arbex e Torres | 2007 | Inferiram premissas baseado no trabalho de Zhao (2004) para viabilizar a reconstrução de matrizes OD com falta de informações |
| Inferência de Origens e/ou Destinos | Zhao et al | 2007 | Estimativa do local de validação a partir de dados da oferta de TP |
| Inferência de Origens e/ou Destinos | Li e Trépanier | 2015 | Método probabilístico de densidade de kernel para encontrar a parada de desembarque |
| Inferência de Origens e/ou Destinos | Jang | 2010 | utilizou um algoritmo de classificação em árvore escrito em linguagem R (o C503) para prever os propósitos das viagens |
| Transferência de viagens e atividades | Nassir et al. | 2015 | Avaliou diversos critérios para inferir uma transferência ou uma atividade |
| Mineração de dados | Habib; Weiss | 2014 | Mineração de padrões de viajantes pendulares |
| Mineração de dados | Cui e Long | 2015 | Características dos padrões temporais para viagens estáveis |
| Mineração de dados | Cats; Ferranti | 2022 | Padrões temporais de mobilidade utilizando técnicas de k-means e mistura gaussiana |

Fonte: Autor

# 4. MÉTODO

O método deste trabalho está representado na Figura 7 e pode ser simplificado em 4 macro etapas: (i) Extração, transformação, carregamento dos dados e delimitação da arquitetura do banco de dados; (ii) Análises exploratórias das validações; (iii) Identificação dos padrões de uso do sistema; e (iv) Modelagem da probabilidade de validar ao embarcar com aprendizado de máquina.

Figura 7 - Método global para reconstrução das viagens por intermédio dos padrões espaço-temporais




Fonte: Autor.

A primeira etapa consiste na construção do modelo de entidade-relacionamento do banco de dados, bem como a contextualização dos dados (***etapa 1.1***), meios de transformação aplicados (***etapa 1.2***), carregamento no Sistema Gerenciador de Banco de Dados Relacional (SGBDR) e avaliação do processo com indicadores de rendimento do banco (***etapa 1.3***). Posteriormente foram realizadas as análises exploratórias (***etapa 2.1***) com foco na compreensão dos padrões espaciais, temporais e por frequência de uso. Estas

análises estão segregadas entre análises espaciais e temporais agregadas, não-espaciais agregadas e a nível de indivíduo. Ainda nesta etapa são definidas hipóteses consideradas neste estudo (***etapa 2.2***), em relação aos padrões de validação e de como estes podem auxiliar na reconstrução dos atributos de deslocamento para o caso do sistema de Fortaleza. A partir dos levantamentos realizados são definidos os atributos que serão utilizados para identificação dos padrões de forma a representar os hábitos de uso do sistema (***etapa 3***), sendo essas categorias delimitadas por *clusterização* através do método *k-means*. No caso do sistema de transportes público de Fortaleza, como não existe a informação da origem (apenas local de validação) e nem do local de desembarque, neste trabalho verifica-se a principal hipótese de que *os locais de validação para usuários regulares seguem um padrão espacial e/ou temporal que podem ser identificados e utilizados para ajustar o real local de embarque*. Dessa forma, pretende-se identificar os locais de embarque, já que os métodos propostos na literatura de encadeamento não são suficientes para o SIT-FOR. A identificação será realizada analisando distâncias de validação (distância entre o local de embarque e de validação) nas linhas regulares de uso do sistema para cada grupo de usuário. Na ***etapa 4***, são treinados e aplicados os modelos supervisionados com uma rede neural, um modelo de *naive bayes* e o algoritmo de *Random Forest* para modelagem da distância mínima de validação de uma amostra de usuários do cadastro de bilhetagem. Essa distância foi utilizada para encontrar o real local de embarque. Como não se sabe como estes dados se comportam entre si, espera-se que modelos que absorvam essa complexidade não-linear tenham melhor rendimento, e por isso serão testados e avaliados comparativamente. Dessa forma, espera-se reconstruir os atributos e por consequência os deslocamentos dos usuários que podem variar conforme os hábitos de utilização do sistema, podendo modificar-se conforme o dia da semana ou mês, assim como por tipo de usuário (Vale Transporte e Estudante) ou tipologia das linhas (Alimentadora, Troncal, Convencional e Complementar). A Tabela 2 apresenta o resumo das bibliotecas utilizada em cada etapa da proposta metodológica.

Dessa forma, com este método pretende-se reconstruir as viagens dos usuários, partindo da hipótese de que o padrão de regularidade de uso (seja espacial ou temporal) de cada usuário, permite identificar os locais de atividades, residência, transferências e de embarque. Algumas questões que cabem maior discussão na literatura e que motivaram a

estruturação desse método estão dispostas a seguir (Munizaga, Palma; 2012; Mesquita *et. al*, 2017; Braga, 2019; Mesquita; Neto, 2021; Cats; Ferranti, 2022):

- Muitos estudos não explicitam a forma de tratamento dos dados de bilhetagem, GPS e GTFS, embora seja uma etapa que requer um considerável esforço computacional, e que influencia diretamente nos resultados;
- Sistemas *tap-on* não detém a informação do local de embarque e desembarque da viagem, apenas o local de validação.
- No tipo de sistema de Fortaleza não há diferenciação entre transbordos e atividades.
- Os atributos dos deslocamentos não são representados para identificação do local de embarque.

Tabela 2 - Resumo das bibliotecas python utilizadas em cada etapa metodológica

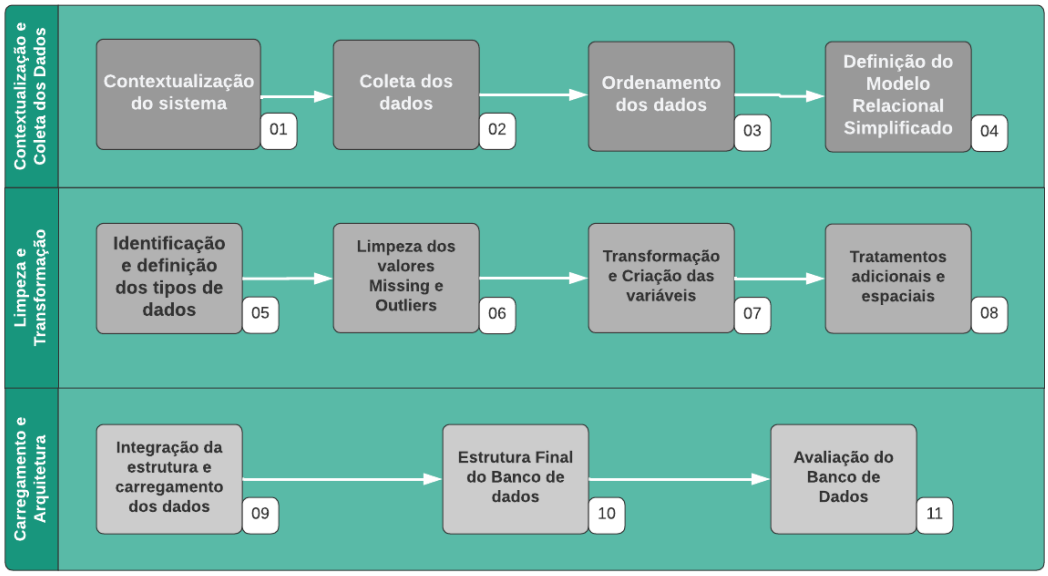
| **Etapa Metodológica** | **Bibliotecas** | **Funcionalidade** |
|---|---|---|
| Construção do Banco de Dados | mysql | Conexão e manipulação de banco de dados relacional |
| Construção do Banco de Dados / Análises Exploratórias | pandas | Análise de dados e tratamento |
| Construção do Banco de Dados / Análises Exploratórias | Numpy | Objeto de matriz multidimensional para análises matemáticas em grandes conjuntos de dados |
| Construção do Banco de Dados / Análises Exploratórias / Modelagem | Matplotlib | Visualização gráfica dos dados |
| Análises Exploratórias / Modelagem | Seaborn | Visualização gráfica dos dados |
| Construção do Banco de Dados / Análises Exploratórias | GeoPandas | Tratamento e análise de dados espaciais |
| Análises Exploratórias / Modelagem | Shapely | Análises geométricas planas para operações espaciais |
| Modelagem | Scikit-Learn | Aprendizado de máquina (k-means, PCA, Random Forest e Naive Bayes) |
| Modelagem | Tensor-Flow | Aprendizado de máquina e redes neurais |
| Modelagem | Keras | API para redes neurais utilizada em conjunto com o TensorFlow |
| Modelaegem | PyTorch | Treinamento e aplicação de redes neurais |

Fonte: Autor.

## 4.1 Método de Estruturação do Big Data de Transporte Público

Como primeira etapa do método de construção do banco de dados, é necessário conhecer bem os dados e armazená-los em uma estrutura que possibilite fácil acesso e alto desempenho. Para isso, esses dados devem estar devidamente tratados, respeitando a granulometria dos seus tipos e a normalização do banco de dados relacional. A Figura 8 representa o *sub-método* correspondente ao tratamento dos dados e estruturação do banco, sendo dividido em 3 etapas: (i) A contextualização da região de estudo, junto a extração de dados dessa região; (ii) Limpeza e Transformação dos dados; (iii) Carregamento dos dados e definição da arquitetura final do banco de dados.

Figura 8 - Método de Estruturação do Banco de Dados para Big Data de Transporte Público




Fonte: Autor.

As etapas anteriores correspondem a um processo chamado de ETL (*Extract, Transform and Load*), ou melhor: Extração, Transformação e Carregamento. O processo de extração é responsável por captar os dados das fontes, como uma espécie de recuperação apenas de dados necessários. Partindo-se para segunda etapa, tem-se a que demanda maior esforço computacional dentre as três, a de transformação. Como primeiro passo é necessário delimitar a integração dos dados, transformando os campos para um único padrão e deixando aptos a serem armazenados. Esta etapa busca transformar os dados brutos como são coletados de acordo com a arquitetura e tipagem especificada no projeto. É nesta etapa que são realizados pré-processamentos, nos quais são identificadas os dados duplicados,

integração, substituição de valores, limpeza de campos e quaisquer outras transformações necessárias.

A etapa de transformação também é responsável por resolver problemas oriundos das fontes de dados, como ausência de informação, valores inválidos, ausência de integridade referencial, cálculos inválidos, duplicação de informação, inconsistência de dados e falhas na modelagem das fontes de dados. É importante ressaltar que se as etapas de extração e transformação não forem bem planejadas e executadas, as análises feitas a partir desses dados serão duvidosas, podendo apresentar inconsistência com a realidade do fenômeno. Uma das principais transformações utilizadas neste trabalho foi a identificação das coordenadas das residências a partir do endereço salvo no cadastro. Foi utilizado linguagem *python* e a API do Google Maps.

Por fim, a etapa de carregamento é responsável por armazenar os dados no banco de dados. Nesta etapa será utilizado uma plataforma de dados estruturados, do tipo SGBD, onde serão alimentados, podendo ser atualizados à medida que novas tabelas sejam carregadas.

### *4.1.1 Limpeza e Transformação*

As principais bases de dados utilizadas nesta pesquisa foram: (i) Bilhetagem; (ii) GPS da Frota; (iii) GTFS; e (iv) Cadastro dos Usuários. Estas bases de dados serão explicadas detalhadamente no Capítulo 5 que apresenta o banco de dados desenvolvido. O ano escolhido para a análise foi 2018, por ser o ano que detinha dados disponíveis de todas as bases e menor número de arquivos faltantes pré-pandemia, utilizando-se 3 meses de coleta para as análises exploratórias e 6 meses para a modelagem, compreendidos entre junho e novembro. De forma sucinta: a Bilhetagem contém todos os registros diários de validação dos usuários; O GPS contém as coordenadas geográficas a cada 30 segundos dos veículos da frota; o GTFS é uma base com a programação operacional de todo o sistema, incluindo itinerários e as tabelas horárias das linhas ofertadas; e o Cadastro contém informações sobre os usuários cadastrados no sistema de bilhete único. A principal informação do cadastro para esta pesquisa é o endereço dos usuários.

Essa etapa foi responsável por 80% do tempo de processamento dos dados. O processo de tratamento dos dados, conhecido como limpeza e transformação, podem ser

divididos em tratamento de valores faltosos e tratamento de valores extremos. Valores faltosos podem ocorrer por falhas no sistema de coleta ou no próprio processo de extração. É necessário visualizar se existe um padrão, podendo remover o registro ou tentar inferir valores faltosos. Já valores extremos, distantes da média dos valores, podem ser simplesmente removidos ou tratados individualmente. O processo de tratamento também envolveu padronização de nomes e formatos de variáveis.

Dessa forma, o processo foi dividido entre o tratamento de dados faltantes e dados ruidosos. Na primeira etapa foram removidos dados com atributos nulos ou com valores com tipos errados na respectiva coluna de dados. Na segunda etapa para tratar os dados ruidosos (que não podem ser interpretados por máquinas) foi utilizado o método de *Binning*. Sendo um processo de suavização de dados, usado para minimizar os efeitos de pequenos erros de observação. Os valores dos dados originais são divididos em pequenos intervalos conhecidos como compartimentos e, em seguida, são substituídos por um valor geral calculado para esse compartimento. Pode-se substituir todos os dados em um segmento por seus valores médios ou limites.

### *4.1.2 Carregamento dos dados e Arquitetura*

Com os dados devidamente limpos e transformados, foi construído o modelo relacional final que compreende a integração de todas as bases levantadas com algum grau de relacionamento, que possibilite consultas de forma eficiente. Para tanto, adotou-se o sistema de Gerenciamento de Banco de Dados Relacional (SGBDR), sendo as tabelas carregadas por intermédio das linguagens *Python* e SQL. Ao final foi apresentado um resumo com indicadores que refletem o esforço despendido no processo. Foi calculado a taxa de tratamento, como sendo a relação entre os registros finais e os registros brutos, além da taxa de compactação dos arquivos, conforme segue:

$$Taxa\ de\ compactação = \frac{Tamanho\ final - tamanho\ inicial}{tamanho\ inicial} \quad (1)$$

## 4.2 Análises Exploratórias dos padrões de validação

Neste tópico serão apresentadas as análises exploratórias que têm como objetivo compreender e dar indícios de como ocorrem os padrões de deslocamento, reforçar as hipóteses sobre os padrões dos usuários e verificar se existe algum tipo de pendularismo. Assim, a partir destas análises, indicadores foram propostos para que permitam definir os atributos que influenciam os padrões, vistos por intermédio de *clusterização* em etapa posterior.

### *4.2.1 Definição das hipóteses de partida*

Inicialmente, foram definidas as hipóteses norteadoras que devem guiar a identificação dos padrões de uso do sistema. Dentre os padrões iniciais apontados na literatura, apresenta-se: (i) usuários frequentes que usam o sistema constantemente; (ii) usuários intermediários que apresentam constância em algum período específico da semana; e (iii) esporádicos, que em muitos estudos são considerados irregulares (Agard *et al*. 2006; Morency *et al*. 2007). Neste trabalho o aspecto de regularidade contribui na definição dos atributos que caracterizarão os padrões e como esses atributos auxiliam na identificação do real local de embarque. Algumas hipóteses prévias que serão verificadas estão dispostas a seguir:

1. A maioria dos usuários tende a validar assim que embarca, porém, uma proporção não tem essa tendência, por isso não se pode assumir que todos os usuários validem no momento do embarque. Hipótese relacionada ao tipo de método de tarifação que é realizado por meio de barreira física dentro do veículo;
2. As regiões prováveis de residência e atividades podem ser identificadas a partir dos padrões de regularidade (temporal e espacial) de uso do sistema, que podem ser identificados por técnicas de mineração de dados e utilizados para ajustar o real local de embarque;
3. Os tipos de atividades realizadas e itinerário das linhas influenciam os padrões temporais, e os locais das atividades e da residência influenciam os padrões espaciais relacionados às linhas/rotas escolhidas para realizar as viagens. Ambos os aspectos devem afetar os padrões de uso do sistema e os locais de validação.

4. Usuários regulares do sistema tendem a repetir certos comportamentos na rede que podem ser identificados através de técnicas de mineração, e que podem auxiliar na identificação de características dos deslocamentos.
5. Diferentes grupos de usuários tendem a ter diferentes padrões de uso do sistema.

Estas hipóteses partem da proposta deste trabalho em explicar como os padrões regulares obtidos da observação de vários dias apresentam características que melhor representam os deslocamentos do que a avaliação de um único dia. Essas hipóteses estão associadas ao hábito de utilização do sistema, ou seja, a análise de padrões se remete a análise de hábitos dos usuários que é influenciado pelo tipo, localização e horário das atividades, assim como pela oferta do sistema de transportes. Embora, espera-se que uma parcela de validações não ocorra no momento do embarque, acredita-se que o padrão regular obtido de uma série histórica de validações permita evidenciar os locais onde os usuários realizam atividades, o número de atividades, e como ele utiliza o sistema. Tal padrão permite identificar os locais de embarque e por conseguinte a cadeias de viagens.

### *4.2.2 Distância de validação e identificação do embarque para amostra do cadastro*

Para este estudo, definiu-se a distância de validação como a distância euclidiana entre a parada do embarque e o local de realização da primeira validação do dia. Estas distâncias foram determinadas a partir de endereços válidos dos usuários do cadastro e da localização das paradas de embarque mais próximas das residências, como será descrito. Como os endereços registrados no cadastro do bilhete único podem estar desatualizados, definiu-se os procedimentos a seguir para identificar registros com endereço considerado válido e as paradas de embarque.

Limitou-se inicialmente aos registros com endereços completos (nome logradouro, número e bairro) e com cadastro recente no ano considerado para análise, em 2018. Determinou-se então as coordenadas geográficas desses endereços, utilizando-se a API do *Google Maps*. Em seguida, com base nos dados de bilhetagem e GTFS, definiu-se a localização da parada mais próxima à residência de cada usuário da linha mais frequente utilizada na primeira viagem do dia, respeitando o sentido da linha. Está parada foi considerada como parada de embarque e salva no banco de dados. Adotou-se como endereço

válido aqueles cujas distâncias euclidianas à parada de embarque sejam menores do que 1000 m, assumindo-se como distância máxima que um usuário estaria disposto a caminhar.

A amostra de usuários também foi restringida conforme a frequência de uso do sistema, excluindo-se usuários que apresentam baixa utilização do sistema, em que a identificação de suas regularidades não seria possível. Estes usuários são os que dentro do período de 6 meses apresentaram frequência média de validação inferior a 1 em todos os dias da semana. Espera-se que os demais usuários apresentem características regulares passíveis de agrupamento.

Ao final, obteve-se uma amostra de 20,6 mil usuários com endereções válidos, dentre os 330 mil que usaram o sistema em 2018. Vale ressaltar que etapas adicionais para verificação da veracidade das coordenadas dos endereços foram realizadas, como verificação amostral das coordenadas em comparação ao endereço cadastrado.

### *4.2.3 Análises e determinação dos indicadores*

Foi realizado uma etapa de exploração dos dados para identificar padrões iniciais a partir do banco de dados consolidado para os meses setembro, outubro e novembro de 2018. As análises estão divididas entre ***análises espaciais e temporais agregadas (com todas as validações), e análises das primeiras validações (com a amostra do cadastro) e no nível do indivíduo***. Quanto ao primeiro grupo de análises, busca-se compreender espacialmente e temporalmente os padrões de frequência de validação (número de validações). Definiu-se as seguintes análises:

- *(i) Distribuição espacial das validações por faixa horária:* Nesta análise, busca-se analisar como o padrão espacial de validações varia ao longo de um dia típico, verificando as zonas de maior concentração e identificando fatores que podem contribuir para tal concentração;
- (*ii*) *Variação da frequência de validações ao longo do dia e entre dias:* Esta análise busca verificar variações na frequência de validações entre dias e semanas do período de análise;
- (*iii*) *Frequência de validação por faixa horária:* Nesta análise foi calculado a frequência de validação utilizando as primeiras e últimas validações, de modo a identificar picos de validação e padrão temporal de validação.

Quanto ao segundo grupo de proposição de análises, estão as análises que foram realizadas a partir da amostra de usuários do cadastro com endereços válidos definida anteriormente, buscando compreender o padrão de validações que ocorrem no momento do embarque, como segue:

- *(i) Perfis horários da proporção de primeiras validações do dia no momento do embarque:* Esta análise busca verificar a variabilidade temporal das primeiras validações que ocorrem no momento do embarque. Além disso, a análise também permite verificar se existe evidências de que a demanda ou ocupação dentro dos veículos pode influenciar o local de validação;
- *(ii) Proporção das primeiras validações do dia por tipo de linha:* As proporções foram tabuladas e comparadas por faixa de número de validações para cada categoria de linha (ALM – Alimentadora, TRC – Troncal, CNV – Convencional e CMP – Complementar). Suspeita-se que o tipo de linha e a configuração da rede tronco-alimentadora pode influenciar no local de validação. Por exemplo, usuários com linhas regulares do tipo alimentadora tem maior propensão a passar por terminais na sua cadeia. E usuários com linhas regulares do tipo troncal têm maior propensão ao destino ser no centro da cidade.
- *(iii) Distância entre a residência e as primeiras validações do dia:* Esta análise foi realizada a partir da amostra de usuários do cadastro, cujo local de residência é conhecido. A análise busca verificar se os usuários validam próximo das residências e, portanto, permitindo inferir o local mais provável de embarque.

Por fim, a última categoria busca analisar os padrões individuais de validação dos usuários. Também é nessa etapa que propõem análises que poderão não somente auxiliar a identificar os padrões, mas participar diretamente da etapa de identificação dos atributos do deslocamento, principalmente do local de embarque. Para essas análises também se optou por utilizar a mesma amostra de usuários do cadastro, e que posteriormente podem ser aplicados a todos os usuários do sistema. Dessa forma apresenta-se as análises e indicadores avaliados:

- *(i) Perfil de validação temporal dos usuários*: análise do perfil de variação horária de um dado usuário, buscando evidenciar se o usuário tende a validar em períodos específicos do dia, com padrão mais concentrado ou disperso, podendo evidenciar os

horários e número de atividades realizadas. Essa informação em conjunto com a frequência média diária de validações, pode dar indícios sobre a quantidade de atividades que o usuário costumeiramente realiza, e dessa forma indicar número de viagens realizadas por dia;

- *(ii) Distância temporal média entre validações:* média das diferenças (horas) entre as primeiras e últimas validações durante um certo período de meses de utilização. Esse indicador pode indicar o tipo de atividade realizada, conforme seja o tamanho do intervalo temporal e o tipo de usuário;
- *(iii) Perseguição espacial dos usuários:* Nesta etapa, buscou-se verificar como determinados tipos de usuários se locomovem no espaço e ao longo do período de análise, com a finalidade de verificar a ocorrência de validações em locais próximos e em mesmo horário, e se elas pertencem as mesmas linhas.

.

### 4.3.Identificação dos padrões de uso do sistema

Esta etapa consiste em identificar diferentes padrões de uso do sistema que possam auxiliar na identificação dos atributos do deslocamento. Para identificar os padrões de uso do sistema, foram considerados sete aspectos que podem influenciar esses padrões, conforme segue: (i) Localização da residência; (ii) Localização das atividades; (iii) Tipo de atividade; (iv) Duração da atividade; (v) Horário de realização das atividades; (vi) Número de atividades realizadas por dia e (vii) Oferta em termos de itinerário, conectividade da rede, e nível de serviço (frequência e lotação). Acredita-se, portanto, que estes aspectos afetam o padrão de deslocamento dos usuários do sistema, e por conseguinte nos padrões de validações.

Com base neste aspectos e em trabalhos apontados na literatura (Morency *et al*.,2007; Ma *et al*., 2013; Huang *et al*., 2015; Ma *et al*., 2013), suspeita-se que existem alguns padrões bem definidos de uso, tais como: (i) usuários com uma única atividade compulsória e, portanto, com um padrão pendular de uso do sistema; (ii) usuários que realizam atividades adicionais, além das compulsórias, apresentando diferentes padrões de uso ao longo da semana (em termos de linhas e horários); e (iii) usuários que usam o sistema para acessar atividades eventuais, não tendo um horário de validação bem definido ao longo

do dia. Nestes padrões, o local de residência e das atividades pode determinar se ocorrerá integrações (mudança de veículo em terminais ou em pontos da rede) ao longo do dia.

Para identificar os possíveis grupos, adotou-se o método *k-means* de clusterização de acordo com as seguintes etapas: (i) Definição dos atributos; (ii) Definição de uma distância de similaridade; (iii) Normalização dos dados aplicando o *z-score*; (iv) Definição do número de clusters pelo método do cotovelo e método da silhueta; (v) Aplicação do algoritmo de *clusterização*; (vi) Análise dos componentes principais sobre a influência dos atributos nos agrupamentos; e (vi) Interpretação dos agrupamentos de dados.

### *4.3.1 Definição dos Atributos*

Partiu-se de alguns indicadores individuais definidos nas análises exploratórias e baseados na literatura (Cats & Ferranti, 2022), conforme segue:

- ***Atributos I1-I5 – Frequência média de validações por dia da semana*** *(Segunda à Sexta)*, representando o aspecto (vi) Número de atividades. Estes atributos correspondem a frequência média de validações por dia útil da semana, sendo calculados separadamente para a amostra de usuários válidos do cadastro;
- ***Atributo I6 – Validações próximas aos terminais***, sendo influenciado pela (ii) Localização das atividades. Este atributo corresponde ao número de vezes em média que o usuário utiliza o terminal, ou seja, realiza transferências entre linhas, sendo um aspecto importante do deslocamento, possibilitando viagens mais longas dentro do sistema. Este atributo é calculado identificando as zonas dos terminais e as imediatamente vizinhas, obtendo-se a frequência média diária de validações nessas regiões;
- ***Atributos I7-I11 – Distância temporal média entre as primeiras e últimas validações por dia da semana*** *(Segunda à Sexta)*, tendo relação com o (iii) Tipo de atividade e (iv) Duração da atividade. Estes os atributos correspondem as distâncias temporais médias, medida em horas, entre as primeiras e últimas validações para cada um dos dias da semana;
- **Atributos *I12-I13 – Faixa Horária de maior frequência das primeiras e últimas validações*,** relacionadas aos aspectos (iii) Tipo de atividade e (v) Horário de realização das atividades**.** Para determiná-los, divide-se o dia em 24 faixas

horárias, sendo contabilizado em qual faixa horário o usuário apresenta maior número de validações para as primeiras e últimas validações, respectivamente, especificando sempre pelo limite superior da faixa horária;

- **Atributos *I14-I16 – Espaço de atividades e seus atributos espaciais*,** relacionados às características da: (i) Localização da residência, (ii) Localização das atividades e (vii) Aspectos operacionais do sistema. O atributo I14 corresponde a área ($km^2$) de atividades definida a partir do conjunto convexo formado pelos pontos de validação para o período de análise (6 meses). Já os atributos I15 e I16 correspondem as distâncias máximas (km) horizontal (leste-oeste) e vertical (norte-sul) da área de atividades. Estes atributos auxiliam na compreensão da extensão e cobertura dos deslocamentos dentro do município de Fortaleza;
- **Atributos *I17-I20 – Proporção de validações por tipo de linha*** *(Alimentadora, Troncal, Convencional e Complementar),* relacionados aos aspectos (vii) 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 (configuração espacial e função da linha) 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.

### *4.3.2 Definição da distância de similaridade e normalização*

A distância de similaridade utilizada foi a distância Euclidiana, se aplicando melhor a dados padronizados, e devido a isso o resultado é invisível a outliers (exceções, ou dados com uma diferença muito grande em relação à média). Uma desvantagem sobre essa medida de distância pode acontecer se houver diferença de escala entre as dimensões; por isso a importância de se normalizar os dados. Essa distância é calculada como a soma da raiz quadrada da diferença entre coordenadas de dois pontos, onde P é número de registros:

$$d(x,y) = \sqrt{\sum_{i=1}^{P}(x_i - y_i)^2} \quad (2)$$

Onde x e y são dois pontos que se está comparando.

Uma vez que a distância de similaridade é definida, o próximo passo é padronizar os dados para garantir que todas as dimensões (ou atributos) tenham a mesma importância na determinação da distância (Japkowicz e Shah, 2011; Géron, 2019). Neste caso, optou-se pela padronização por *z-score*, também conhecida como padronização, em que se transforma cada valor em um valor *z*, representando o número de desvios padrões acima ou abaixo da média dos dados. Isso é realizado subtraindo a média dos dados e dividindo pelo desvio padrão, conforme segue:

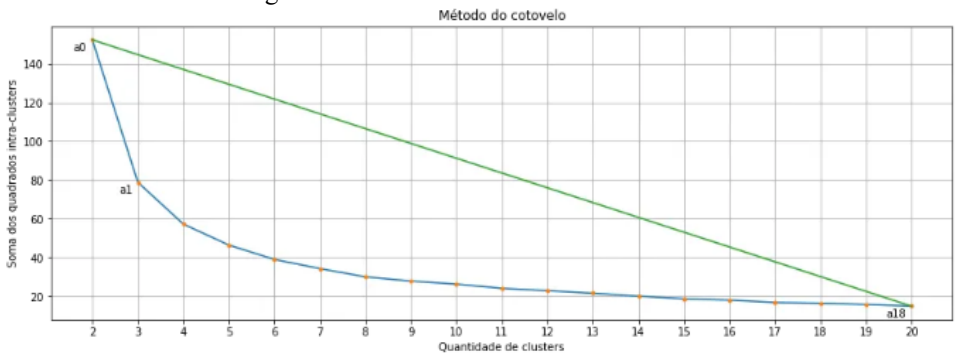
$$z = \frac{(x - média)}{Desvio\ Padrão} \tag{3}$$

Este tipo de padronização, além de garantir a mesma importância dos atributos, também evita que atributos com valores extremos dominem a distância de similaridade.

### *4.3.3 Clusterização dos usuáriosF*

Inicialmente, foi necessário definir o número de grupos, o qual é o parâmetro do método de *clusterização*. Optou-se pela utilização do método do cotovelo e uma validação pelo método da silhueta. Dessa forma foi executado o *k-means* variando o número de clusters, *k*, de 1 a 20 . Em cada rodada, calculou a variação dentro dos grupos pela soma dos desvios quadráticos em relação ao centróide de cada grupo (MacQueen, 1967; Jain *et. al*, 1999). A partir do gráfico relacionando as variações dentro dos grupos com os valores de *k*, identifica-se o "ponto de cotovelo" como sendo o ponto em que a variação começa a se estabilizar, ou seja, a curva começa a ficar mais suave. Assume-se que é neste ponto de cotovelo o número ideal de grupos. A figura a seguir apresenta visualização gráfica (Ketchen, 1996). O ponto de cotovelo, ou número ótimo de clusters seria 3, demarcado pela quebra da curva e maior distância a reta.

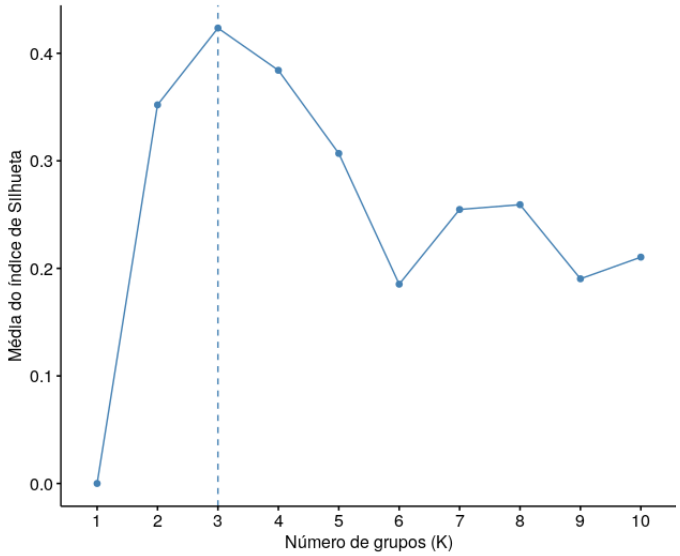
Figura 9 - Método do cotovelo de 2 a 20 clusters




Fonte: Adaptado de Géron (2019)

Como sua interpretação do método do cotovelo pode ser subjetiva, foi utilizado também o método da silhueta para validar o número de agrupamentos. Esse método determina o quão bem cada objeto está alocado em um grupo, ou seja, a homogeneidade de um grupo. A métrica da Silhueta varia de -1 a 1. Valores próximos de 1 indicam que o objeto possui semelhança com os objetos do seu grupo (coesão) e dessemelhança com os objetos de outros grupos (Rousseeuw, 1987; Kaufman e Rousseeuw, 2009). O número ideal de grupos é escolhido com base na maior média do índice de silhueta.

Figura 10 - Método da silhueta




Fonte: Adaptado de Géron (2019)

Além da utilização destas técnicas, cabe uma análise dos atributos após separação para validar os agrupamentos ou até testar novos valores à nível de comparação. Nesta dissertação, além do valor indicador pelos métodos foram testados os valores imediatamente inferiores e superiores de *clusters*.

A partir do número de grupos, utilizou-se o algoritmo do *k-means* (Géron, 2019; Cats e Ferrati, 2022) para agrupar os usuários por modo de uso do sistema. Outro aspecto importante delimitado foi o método de inicialização dos centroides, optando-se pelo ***k-means***++. O k-means++ é uma técnica de inicialização para o algoritmo *k-means* que visa selecionar centróides iniciais mais distantes entre si, com o objetivo de melhorar a qualidade do agrupamento resultante e reduzir o tempo de convergência do algoritmo.

O *k-means* começa selecionando *k* centróides aleatórios do conjunto de dados, que serão os primeiros clusters. Em seguida, o algoritmo atribui cada observação ao *cluster* mais próximo, calcula a média de cada *cluster* e redefine os centróides com as médias. Esse processo é repetido até que não haja mais mudanças nos clusters (MacQueen, 1967; Jain *et. al*, 1999). O problema com a seleção aleatória dos centróides iniciais é que isso pode levar a agrupamentos subótimos, dependendo de quais centróides são selecionados inicialmente. Por exemplo, se os centróides iniciais estiverem muito próximos um do outro, é mais provável que o algoritmo fique preso em um ótimo local, em vez de encontrar o ótimo global. O *k-means*++ resolve esse problema selecionando os centróides iniciais com base em sua distância. O algoritmo começa selecionando o primeiro centróide aleatoriamente de todo o conjunto de dados. Em seguida, ele seleciona os centróides subsequentes a partir de uma distribuição de probabilidade ponderada pela distância dos centróides já selecionados. Em outras palavras, os próximos centróides são selecionados em locais que são mais distantes dos centróides existentes.

Após a geração dos grupos, avaliou-se a importância de cada atributo na formação dos clusters. O indicador utilizado foi o score de silhueta que leva em conta distâncias entre pontos dentro dos clusters e distâncias entre pontos de clusters diferentes, tendo ao final uma silhueta média para cada variável. Para cada variável é calculado a soma dos desvios quadráticos entre grupos (SSB) e a soma dos desvios quadráticos dentro do grupo (SSW). O *score* é a razão entre o SSB e a SSW. Quanto maior o valor dessa razão (variação entre sendo maior do que a variação dentro dos grupos), maior é a importância da variável na separação dos grupos (MacQueen, 1967; Jain *et. al*, 1999; Japkowicz, 2011).

### *4.3.4 Interpretação e análise dos locais de embarque*

Inicialmente, para visualização dos grupos, foi utilizado a técnica de Análise dos Componentes Principais (PCA). Essa técnica é uma transformação linear ortogonal que transforma os dados para um novo sistema de coordenadas de forma que a maior variância por qualquer projeção dos dados fica ao longo da primeira coordenada, a segunda maior fica ao longo da segunda, e assim sucessivamente, sendo possível reduzir a dimensionalidade dos dados independentemente da quantidade de atributos (Jolliffe, 2002; Géron, 2019; Cats e Ferrati, 2022; Zhao *et. al*, 2023).

De forma mais detalhada, a análise de PCA é composta por 5 etapas, sendo elas: (i) Obtenção da média e centralização dos dados; (ii) Cálculo da matriz de covariância; (iii)

Decomposição da matriz de covariância em autovalores e autovetores; (iv) Montagem da matriz de projeção com os k autovetores correspondentes aos k maiores autovalores; e (v) Projeção dos dados originais no novo espaço de k dimensões (Jolliffe, 2002; Jackson, 2005).

Para cada atributo foram determinados 20 componentes principais. Porém apenas, uma parcela destes componentes é necessária para se compreender como cada atributo influencia na separação dos grupos. Portanto, será apresentado graficamente a relação do número de componentes (em ordem decrescente de variância) com a soma cumulada da variância explicada pelos componentes principais. A partir desta análise foi escolhido o grau de componentes necessários para se discutir os resultados, além de apresentá-los. Os dois componentes principais, de maior variância dos dados, foram também visualizados graficamente, permitindo avaliar como alguns atributos influenciam na formação dos grupos a partir da direção e o sentido dos autovetores de cada atributo. Esta última análise é conhecida como mapa perceptual, mostrando a relação entre os componentes e as variáveis.

Portanto, através da análise dos componentes principais integrada a análise do mapa perceptual, foi possível identificar quais atributos (espaciais, temporais e características operacionais) mais influenciaram na segregação dos grupos, auxiliando na compreensão de como cada grupo se desloca no sistema e os principais aspectos influentes nesses comportamentos.

Dessa forma, a análise de PCA auxilia na compreensão de quais atributos por componente dos grupos mais influenciaram na segregação deles, onde a direção dos autovetores indicado no mapa perceptual aponta para a ordem de grandeza da variância. Enquanto os autovalores representam a variância explicada pelos componentes principais. Também é válido destacar que não foi realizada rotação, pois não se conhece a relação entre os dados. Um dos principais ganhos em se realizar o PCA é relacionar um determinado grupo a uma direção apontada pelo vetor de um atributo, ou um conjunto de atributos que mais influenciam na formação dos grupos. Auxiliando na identificação dos padrões e tornando mais claro as características de cada grupo (Jolliffe, 2002; Jackson, 2005).

Com relação a análise dos atributos em cada grupo encontrado na etapa anterior, realizou-se uma análise descritiva, calculando-se medidas de tendência central e dispersão, permitindo interpretar que tipos de padrões de deslocamento podem ser evidenciados e as diferenças entre os grupos.

Avaliou-se também em cada grupo a variável de interesse para localização dos embarques, sendo ela a distância mínima de validação (entre a residência e a primeira validação do dia), considerando o período de análise de 6 meses. Esta variável foi determinada junto ao processo de obtenção dos atributos, porém não faz parte da *clusterização* uma vez que é a variável a base para os modelos supervisionados. Essa distância mínima pode indicar se um indivíduo valida no momento de embarque. Assim, comparou-se as distribuições e as medidas descritivas desta variável entre os grupos. Neste caso busca-se verificar a hipótese de que o padrão de validação é diferente para diferentes grupos, pois acredita-se que as diferentes localizações das residências e atividades, bem como os tipos de atividades e a própria configuração das linhas, segreguem esses padrões, fazendo com que aumente a similaridade entre usuários com aspectos de uso parecidos e aumente a dissimilaridade entre usuários com aspectos de uso diferentes.

## 4.4 Modelagem supervisionada do local de embarque

Nesta etapa, propõe-se um método para modelar a probabilidade de um usuário validar ao embarcar. Dessa forma, o modelo permitirá a aplicação em um novo conjunto de dados onde não se sabe a localização da residência. O método tem como base caracterizar o comportamento de validação para cada padrão identificado na etapa anterior. Portanto, a variável a ser modelada é a probabilidade de validar ao embarcar, sendo uma variável categórica, onde o valor 1 significa que o usuário valida ao embarcar e 0 para os usuários que não validam ao embarcar. Para tanto, utilizou-se a variável de distância mínima (euclidiana) entre o local de validação e a parada de embarque. Conforme, apresentado a distância entre paradas da rede está entre 400 e 550m, portanto assumiu-se um valor médio de 500m, para categorizar os usuários quanto a ação de validar ao embarcar. Caso o usuário tenha uma distância mínima inferior ou igual a 500m, a variável categórica de validação ao embarcar recebe o valor 1, e em caso contrário, recebe o valor 0. Portanto os modelos supervisionados nesta etapa são categóricos.

### *4.4.1 Modelos Supervisionados Categóricos*

Dado a complexidade dos dados, adotou-se os seguintes modelos supervisionados fundamentados em aprendizado de máquina: *Naive Bayes (NB), Random Forest (RF), e Rede Neural (RN)*.

O modelo de *NB* é um modelo de previsão comum para aplicações de classificação quando não se sabe ao certo as relações entre as variáveis (Rish, 2001). Sua escolha se deve à sua simplicidade e ao seu bom desempenho comparado à de vários outros modelos de classificação, necessitando de um pequeno conjunto de dados para obter resultados precisos. O modelo baseia-se na premissa "naive" de independência das variáveis, prevendo a classe que pertence os valores das variáveis de acordo com as suas distribuições de probabilidades, que são predefinidas baseada nas características da amostra. Neste trabalho, adotou-se distribuições empíricas para as variáveis quantitativas com base em estimação de densidade com base em função de suavização do tipo *Kernel* (Rish, 2001).

Nos modelos *RF* e *RN (*Breiman, 2001*)*, assume-se uma forma estruturada para as funções de previsão, que pode criar algum tipo de viés, mas reduzir variância nas estimativas de previsão. O modelo *RF* faz parte do conjunto de modelos *ensemble*, que se beneficiam da combinação de diferentes modelos de árvore de decisão para se obter um único resultado, necessitando de um maior custo computacional. O método *RF* gera várias árvores de decisão, em que a escolha das variáveis em cada nó da árvore é aleatória (Breiman, 2001). Para problemas de classificação, a previsão corresponde à estimativa com maior moda. A premissa principal é a de que não é possível definir uma relação linear entre variáveis e que a curva de separação entre os grupos não é linear. Esta premissa parece razoável para o caso de padrões de validação já que a decisão de validar no momento do embarque pode variar consideravelmente na rede, não sendo possível definir uma função de classificação global linear. As etapas do modelo consistem em: (i) Amostragem aleatória; (ii) Seleção aleatória de preditores; (iii) Crescimento da árvore através da divisão dos nós; e (iv) Agregação das previsões (Cutler, 2007; Géron 2019).

As redes neurais são modelos matemáticos inspirados no funcionamento do cérebro humano que podem ser utilizados para a identificação de padrões em dados. Elas são particularmente úteis para problemas que envolvem grandes quantidades de dados e relações complexas entre as variáveis, ou não lineares em que não se sabe ao certo como e em que grau ocorrem as interações entre elas. No modelo de *RN*, novas variáveis são construídas (nós de uma rede) a partir de combinações lineares dos atributos de entrada. A variável

resposta é então modelada como função destas novas variáveis (LeCun, 2015; Goodfellow, 2016).

### *4.4.2 Especificação e treinamento dos modelos*

Optou-se por uma modelagem baseada em *machine learning*, definindo um grupo de treinamento e outro para teste, seguindo um método de validação cruzada com a seguinte proporção 7:3 (LeCun, 2015). Foram definidos modelos para cada padrão de uso do sistema, assim como considerando todos os dados sem segmentação por grupo. Adotou-se como atributos para previsão os mesmos atributos utilizados na etapa de *clusterização*. Contudo, desconsiderou-se desta modelagem o atributo Faixa Horária da Última Validação, acreditando-se que este fator não deve influenciar a decisão de validação no início do dia. Considerou-se todas variáveis como sendo quantitativas, com exceção da varável relacionada a primeira faixa horária de validação, a qual foi considerada categórica (*dummy* com classes horárias de 0 – 23).

Em cada modelo de *RN* foi instanciado uma rede neural com uma camada de entrada com 19 neurônios (1 para cada atributo no conjunto de treinamento), não havendo ativação nessa camada. Em seguida foram instanciadas duas camadas ocultas intermediárias (*Hidden Layers*) cada uma com 64 neurônios e a função de ativação ReLU (*Rectified Linear Unit*), aplicada a cada neurônio. A função ReLU é uma função de ativação comumente utilizada em redes neurais para introduzir não linearidade. Ela é aplicada elemento por elemento em um tensor (ou vetor) de entradas e retorna o valor máximo entre zero e a entrada. Posteriormente, foi instanciado a camada de saída com 2 neurônios correspondendo as classes para classificação. A função de ativação foi a *softmax,* utilizada para gerar probabilidades para cada classe. Ela garante que a soma das probabilidades de todas as classes seja igual a 1. Por fim, a função de perda utilizada foi a *categorical crossentropy* sendo comumente utilizada em problemas de classificação. Ela é adequada para problemas de classificação *one-hot* (classificação binária), calculando a diferença entre as distribuições de probabilidade preditas pelo modelo e as classes reais dos dados (Bridle, 1990; Bishop, 1995; Nair, 2010; Glorot, 2011; Nielsen, 2015; Goodfellow, 2016;).

### *4.4.3 Avaliação dos modelos*

Após a modelagem, os modelos em cada grupo foram comparados, avaliando-se a capacidade preditiva dos modelos de classificação com os dados de teste, verificando se os modelos realmente "aprenderam", ou seja, conseguiram expandir aceitavelmente seus resultados de predição para conjuntos de dados desconhecidos. Quanto ao desempenho de previsão, os modelos podem ser classificados em *underfitting*, quando não conseguem aprender o padrão dos dados para predição ou *overfitting*, quando aprendem o padrão apenas dos dados de treinamento, não conseguindo expandir para novos conjuntos (Davis,2006; Japkowicz, 2011).

Os indicadores utilizados para avaliar o nível de previsão dos modelos de classificação foram o *precision, recall, f1-score e a acurácia*. Na discussão a seguir, considera-se que um evento positivo pode corresponder a ação de validar ao embarcar ou o caso contrário (Jain, 1999; Sokolova, 2009; Powers, 2011).

O *recall* (Revocação) mede a capacidade do modelo de identificar corretamente os eventos positivos, ou seja, é a proporção de eventos positivos corretamente previstos em relação ao número total de eventos positivos no conjunto de dados. Valores mais altos indicam que o modelo é capaz de recuperar uma maior proporção de eventos positivos. O cálculo do *recall* é dado pela fórmula a seguir (Sokolova, 2009; Powers, 2011; Géron, 2019):

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \tag{4}$$

Onde:

- ***TP (True Positive)*** é o número de eventos positivos corretamente classificados.
- ***FN (False Negatives)*** é o número de eventos positivos erroneamente classificados como negativos.

A precisão mede a capacidade do modelo de classificar corretamente os eventos como positivos. Dessa forma, ele mede a proporção de eventos positivos corretamente previstos em relação ao número total de eventos positivos. A precisão fornece informações sobre a confiabilidade das previsões positivas do modelo. Valores mais altos indicam que o modelo possui uma proporção menor de falsos positivos. O cálculo da precisão é dado pela fórmula (Sokolova, 2009; Powers, 2011; Géron, 2019):

$$Precision = \frac{\text{TP}}{(\text{TP}+\text{FP})} \quad (5)$$

Onde:

- ***FP (False Positive)*** é o número de eventos negativos erroneamente classificados como positivos.

O *f1-score* é uma métrica que combina o recall e a precisão em um único valor, fornecendo uma medida geral do desempenho do modelo. O *f1-score* leva em consideração tanto a capacidade do modelo de recuperar eventos positivos quanto a precisão de suas previsões. É a média harmônica entre recall e precisão e é calculada pela equação a seguir (Sokolova, 2009; Powers, 2011; Géron, 2019):

$$f1\_score = \frac{2*(\text{Precision} * \text{Recall})}{(\text{Precision}+\text{Recall})} \quad (6)$$

A acurácia é a métrica mais objetiva dentre as quatro apresentadas. Ela mede a proporção de exemplos corretamente classificados em relação ao número total de exemplos, sendo calculada conforme segue (Sokolova, 2009; Powers, 2011; Géron, 2019):

$$Acuracia = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})} \quad (7)$$

Onde:

- ***TN (True Negative)*** é o número de exemplos negativos corretamente classificados.

Por fim, para verificar a hipótese central deste estudo de que a segmentação em padrões auxilia na identificação dos embarques, comparou-se o desempenho dos modelos por grupo em relação à modelos treinados sem considerar os dados do grupo de interesse. Para tanto, para um *Grupo i* são treinados todos os modelos e gerados todos os indicadores de previsão. Os dados do Grupo *i* são retirados da base, e novos modelos são treinados com os grupos restantes. Esses novos modelos são aplicados nos dados de teste do *Grupo i*. Dessa forma, espera-se que os modelos treinados e aplicados com os dados do grupo

específico apresentem melhores resultados do que os modelos generalizados dos outros grupos.

## 4.5. Restrições Metodológicas

Para tornar o estudo viável de ser executado é importante deixar claro as restrições metodológicas consideradas e que podem vir a ser aprimoradas em trabalhos posteriores. Dessa forma estão listadas as principais restrições que serão aprofundadas no decorrer do trabalho:

- As análises ditas espaciais se remetem ao deslocamento e a visualização em rede/zona e não a técnicas estatísticas espaciais;
- Foram utilizados apenas 3 meses como série histórica para as análises exploratórias e 6 meses para a amostra de usuários no processo de modelagem dos grupos;
- Os dados de cadastro apresentaram muitas inconsistências que dificultaram sua utilização, sendo reduzida a base de amostra para menos de 10% dos usuários no cadastro, sendo estes os considerados coerentes para se utilizar em todo o processo proposto.
- O Trabalho tem como foco caracterizar os principais padrões, assim como identificar os reais locais de embarque, sendo a reconstrução da cadeia, um resultado importante, mas não abordado neste trabalho.
- É assumido uma premissa de que o usuário embarca na parada mais próxima da sua residência durante a primeira validação do dia.
- Os dados de endereço dos usuários foram georreferenciados através de linguagem *python* e API do Google, sendo colocado como critério de validação, uma distância máxima de 1000 m entre a residência e o local de embarque.

# 5. CONSTRUÇÃO DO BANCO DE DADOS PARA O BIG DATA-TP

## 5.1.Coleta, identificação, definição dos tipos de dados e modelagem

Neste tópico serão apresentadas as principais bases de dados utilizadas na pesquisa, bem como suas características e tipos de dados. Vale ressaltar que a integração total constituiu 20 bases de dados (**APÊNDICE A – DESCRITIVO DAS BASES DE DADOS**) divididas em 4 grupos de coleta, determinados de acordo com o meio de disponibilização dos dados.

Embora se tenha tratado 20 tipos de dados, apenas 4 bases foram efetivamente utilizadas na dissertação, conforme segue: (i) Bilhetagem; (ii) GPS da frota; (iii) GTFS; e (iv) Cadastro dos usuários. Vale a ressalva de que a base de cadastros é sensível de caráter confidencial. Dados complementares como o dicionário dos identificadores dos veículos e o zoneamento também foram utilizadas. A totalidade das bases foi mantida para se conseguir chegar a um ponto com pouca ou quase nenhuma redundância de dados, além do fato que este modelo relacional poderá ser replicado para outras realidades de transportes público pelo mundo. Dessa forma, visando uma contribuição técnica, optou-se por manter o modelo com todos os dados disponíveis, mesmo apenas alguns sendo necessários para esta pesquisa.

Os dados de GTFS, embora tenham sido disponibilizados pelo órgão gestor do Transporte Público de Fortaleza, são repassados semanalmente para o Google, em uma formatação específica e por isso foi colocado em um grupo à parte. Vale ressaltar que todos os dados dessas bases correspondem ao ano de 2018, por ser o ano que detinha dados disponíveis de todas as bases e menor número de arquivos faltantes pré-pandemia, utilizando-se 3 meses de coleta para as análises exploratórias e 6 meses para a modelagem, compreendidos entre junho e novembro. Na Figura 13 está a representação do modelo conceitual relacional das bases de dados de Transporte Público de Fortaleza. Ainda neste modelo, tem-se os dados disponibilizados pelo PASFOR, referentes a coletas de *Embarque nos terminais, Transbordo nos terminais e Zoneamento*, além de dados de pesquisa de campo que foram integrados em etapas posteriores ao cadastro de usuários. Esses dados de pesquisa são de suma importância para complementar os dados do Sistema da Informação, auxiliando no processo de reconstrução das viagens.

Os dados de Bilhetagem e GPS em seu formato bruto correspondem a arquivos referentes a todos os dias do ano, sendo os maiores de toda a base, tornando-os os últimos no

processo de limpeza e transformação, e obtidos massivamente de forma passiva. O quadro de dicionário corresponde na realidade a um conjunto de arquivos CSV que interligam os códigos dos ônibus do arquivo de bilhetagem com os códigos dos ônibus dos arquivos de GPS. Por fim, o cadastro dos usuários, corresponde a todas as informações dos usuários no momento de solicitação do *Smartcard*. Portanto, para fins de pesquisa e para não transgredir a Lei Geral de Proteção de Dados (LGPD), nenhuma identificação pessoal foi utilizada nesta pesquisa.

Por fim, tem-se uma categoria de *shapes* que foram disponibilizados de outras pesquisas (Braga, 2019), correspondentes ao *shape de bairros de Fortaleza* e *shape dos terminais de integração*. Nesse caso como são muitos registros que se repetem, na etapa de limpeza dos dados teriam duas opções: Excluir as informações dos bairros nas bases, ou tentar interligar essa informação, normalizando os dados. Optou-se pela segunda opção, pois esses dados podem ser úteis para pesquisas futuras.

Posteriormente, foi apresentado os zoneamentos escolhidos para as análises do projeto utilizados no cálculo de alguns atributos dos usuários e das análises exploratórias. Dessa forma o zoneamento utilizado está em formato hexagonal com 1 km de lado, inspirado no projeto de acesso a oportunidades do Instituto de Pesquisa Econômica Aplicada (Pereira *et. al*, 2022). A delimitação de 1 km foi baseada na distância máxima de caminhada de um usuário para acessar o sistema, segundo o Desenvolvimento Orientado ao Transporte Sustentável (DOTs). O zoneamento foi utilizado por se adequar a necessidade de contextualização dos deslocamentos do sistema e ao reconhecimento dos padrões individuais, que fica mais evidente quando se minimiza as diferenças espaciais.

### *5.1.1 Descrição dos dados de GTFS*

Originalmente o GTFS é composto por 13 arquivos do tipo txt, porém sendo 7 opcionais, portanto, o feed de Fortaleza compõe apenas os 6 arquivos obrigatórios: Agência, Rotas, Viagens, Paradas, Horário das paradas e Calendário. E mais 4 arquivos opcionais: Regras de Tarifa, Atributos de Tarifa, Calendário e Shapes.

O arquivo de Paradas contém um identificador numérico para cada parada, sem repetições e que pode ser utilizado como identificador geral desse conjunto, bem como informações importantes de latitude e longitude, que servirão para identificar pontos de embarque/desembarque. O arquivo de Atributos da tarifa basicamente remete ao valor da

tarifa e a moeda da região. O arquivo da agência contém dados referentes a agência reguladora, no caso a ETUFOR. O arquivo de Rotas é o que contém maior grau de dependência dos outros arquivos, pois 4 bases necessitam de informações referentes as linhas de ônibus. O arquivo de shape contém as informações geográficas das viagens.

Por fim, os arquivos de horário das paradas e de viagens, representam respectivamente o cronograma dos veículos de cada linha para cada parada e a sequência de viagens por veículo em cada linha. No caso do primeiro conjunto, esse é o maior arquivo e que contém a maior quantidade de informação com 116 MB, tendo a última prioridade dos dados de GTFS. Vale ressaltar que os dados de GTFS são um agrupamento excelente de informações, porém não foram pensados para um armazenamento relacional seguindo os parâmetros de normalização, portanto muitas adaptações foram necessárias, principalmente quanto aos identificadores.

### *5.1.2 Descrição dos dados externos*

Os arquivos denominados como "externos" foram estipulados por órgãos gestores, mas foram adquiridos de outras pesquisas, e sua principal funcionalidade nessa pesquisa é manter a integridade do banco, uma vez que os dados de terminais e bairros se repetem constantemente em outras bases, como nos dados de cadastro dos usuários e transbordo. A base de zoneamento hexagonal, conforme apresentado no método foram gerados a partir do zoneamento de bairros de Fortaleza, disponibilizados pelo PASFOR. Por fim, os dados dos terminais são *shapes* de pontos com as coordenadas e identificação de cada terminal, podendo ser utilizado em conjunto com os dados de GPS para identificar o momento em que os ônibus estão dentro dos terminais.

### *5.1.3 Descrição dos dados do Órgão Gestor*

A última categoria e com prioridade mais baixa na hierarquia de relacionamento são os dados disponibilizados diretamente pela ETUFOR: Bilhetagem, GPS e Cadastros dos usuários. Escolheu-se o ano de 2018, por ter dados de todos os meses disponíveis e condizer com o mesmo espaço temporal dos dados de GTFS (pré-pandemia).

A Bilhetagem contém registros de validações diárias dos usuários com as seguintes informações: o identificador do cartão, nome da linha, número da linha, prefixo do carro, descrição do tipo de cartão, sentido da viagem e se houve integração durante a validação.

Os dados da base de GPS são os mais pesados de toda a base de dados, tendo cada arquivo uma média de tamanho de 7 GB, sendo necessário dividi-los por dia. Seus valores correspondem ao azimute da direção, latitude e longitude, data e hora, distância percorrida, identificador do dispositivo e do veículo. Uma ressalva é que a informação de latitude e longitude consta apenas nos dados de GPS e não nos de Bilhetagem sendo necessária uma integração entre as duas bases, conforme será descrito.

Por fim os dados de Cadastro dos Usuários, contém a identificação do usuário como nome, nome da mãe e CPF, porém essas informações não foram utilizadas na pesquisa. O Cadastro também contém o endereço do usuário, endereço do local de trabalho, nome da universidade/órgão solicitante e informações do *smart card*. A base de Cadastro pode ser relacionada a Bilhetagem por meio de um identificador comum nas duas bases.

As bases de Bilhetagem e de GPS são independentes. A relação entre as duas bases é feita a partir de um dicionário que relaciona o identificador do veículo nos dados de GPS com o prefixo do carro nos dados de Bilhetagem. Braga (2019) desenvolveu um método de integração das bases de Bilhetagem e GPS. Neste estudo, aplicou-se um método similar, mas sem a necessidade de duplicar os códigos dos veículos em diferentes tabelas, mantendo um único identificador dos veículos para toda a base. Outra alteração no método de integração foi a utilização dos dados de GTFS para atrelar a validação a uma parada.

## 5.2. Tratamento dos dados e criação de variáveis

Nesta etapa, os arquivos de cada base foram tratados e transformados num único arquivo. O tratamento dos dados de Bilhetagem e GPS será descrito nas subseções seguintes. Inicialmente, aplicou-se um tratamento nos dados para limpar valores faltosos ou errôneos. Dessa forma, garante-se que dados errôneos não sejam tratados posteriormente, afetando a eficiência computacional do processo. No processo de limpeza, considerou-se as seguintes correções mais comuns: valores de variáveis localizados em colunas erradas, dados nulos, paradas inexistentes, etc. Os dados também foram tratados quanto a padronização do formato dos campos. Assim, adotou-se um único formato padrão para data (AAAA/MM/DD) e formatos adequados para variáveis do tipo numérica.

### *5.2.1 Tratamento dos dados do GTFS*

Na base de GTFS foram realizados os seguintes tratamentos:

- No arquivo de Paradas do GTFS do ano de 2018, contendo 4969 linhas e 12 colunas, excluiu-se todas as colunas com dados vazios, mantendo-se os campos com dados de coordenadas (latitude e longitude) das paradas;
- Nos arquivos de Atributos de Tarifa, Agência, Calendário, Regras de Tarifa e Data do Calendário, ligados às linhas de ônibus, foram feitos os seguintes tratamentos: a) nos dados de Regras de Tarifa e Agência, excluiu-se colunas vazias; b) nos arquivos de Calendário e Data de Calendário, foi necessário inserir uma coluna com identificador numérico, já que o formato original era caractere, para garantir a normalização dos dados;
- Nos arquivos de Rotas, que originalmente continha 9 colunas e 319 linhas, excluiu-se registros duplicados e colunas vazias, totalizando apenas 4 colunas ao final. Os nomes das rotas foram também corrigidos e padronizados, permitindo uma integração destes dados com outras tabelas do banco. Assim, todas as tabelas que necessitarem da informação das linhas, conforme apresentado no modelo relacional se reportam diretamente a essa tabela de rotas;
- Nos arquivos geográficos (*Shape Files*) do GTFS, foi necessário inserir um identificador número para cada registro (sendo o formato original de caracteres), e um atributo identificando o sentido da linha. O sentido de cada linha (I – IDA ou V - VOLTA) foi obtido do próprio identificador original de cada linha e alocado em um campo específico do tipo caractere. Ao final, 96,96% dos dados foram devidamente limpos e transformados para serem carregados;
- No arquivo bruto de Viagens, excluiu-se 3 colunas com dados faltosos, totalizando ao final 6 colunas, sendo elas: ID_VIAGEM, ID_ROTAS, ID_SERVIÇO, TRIP_ID, SHAPE_ID, ACESSÍVEL_CADEIRANTE. Alterou-se também o identificador do calendário do tipo caractere para número, permitindo a associação com os arquivos de calendário modificados acima. Os dados de viagens continham 82615 registros, tendo após a sua limpeza e transformação uma taxa de tratamento de 99,96%.
- Por fim, no arquivo de Horário das Paradas, reduziu-se de 8 colunas para 6 colunas. Os identificadores dessa base foram também transformados de caractere para numérico a partir de sua associação com o arquivo de viagens. Este processo resultou em um elevado

tempo de processamento (18 horas). Vale ressaltar que o arquivo original tinha um tamanho de 116MB, e a taxa de aproveitamento dos dados após tratamento foi de 94,12%.

### *5.2.2 Tratamento dos dados Externos*

Para os arquivos referentes aos terminais, eram 5 arquivos *shapes* georreferenciados. Os dados foram devidamente tratados e normalizados para maiúsculo, não sendo necessário a limpeza de nenhuma informação nesses arquivos. Por se tratar de um arquivo pequeno de 6 KB, não houve grande influência na sua tabulação final. Conforme apresentado na caracterização da rede, existem 9 terminais, sendo 7 fechados e 2 abertos, cada um detendo um código identificador, conforme segue na Figura 11. Dentre os 9 terminais, foi inserido manualmente um décimo terminal chamado "NÃO INTEGRA", pois os dados de linhas em alguns arquivos estão ligados aos terminais de integração, porém nem todas as linhas integram, causando uma inconsistência na hora de armazenar os dados no banco. Portanto, para essas linhas, recebem o identificador 10 na coluna de terminais, mantendo a integridade do banco de dados.

Figura 11 - Tabulação dos dados referentes aos terminais

| ID_TERMINAL | NOME | TIPO | LATITUDE | LONGITUDE |
|---|---|---|---|---|
| 1 | ANTÔNIO BEZERRA | FECHADO | -3.737488 | -38.584713 |
| 2 | CONJUNTO CEARÁ | FECHADO | -3.772920 | -38.607601 |
| 3 | LAGOA | FECHADO | -3.771688 | -38.570013 |
| 4 | PARANGABA | FECHADO | -3.776106 | -38.563565 |
| 5 | SIQUEIRA | FECHADO | -3.789877 | -38.586774 |
| 6 | MESSEJANA | FECHADO | -3.831225 | -38.501589 |
| 7 | PAPICU | FECHADO | -3.738198 | -38.485087 |
| 8 | PRAÇA DA ESTAÇÃO | ABERTO | -3.722341 | -38.530191 |
| 9 | CORAÇÃO DE JESUS | ABERTO | -3.722341 | -38.530191 |
| 10 | NÃO INTEGRA | NÃO INTEGRA | 0.000000 | 0.000000 |

Fonte: Autor.

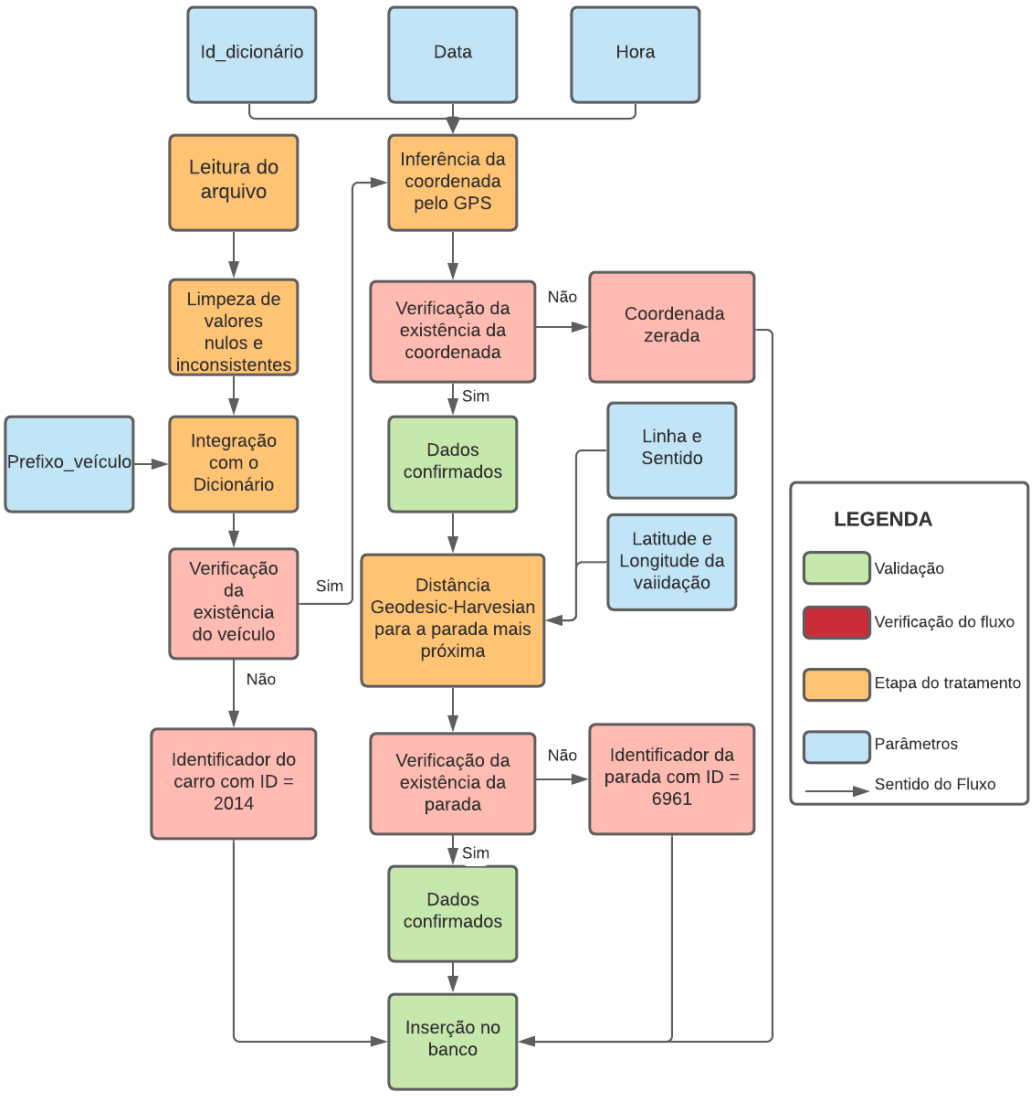
### *5.2.3. Tratamento dos dados do Órgão Gestor*

Foram realizados os seguintes tratamentos nos dados de Cadastro, Bilhetagem e GPS da frota.

- Todos os arquivos de Cadastro, originalmente separados por mês, foram agrupados em um único arquivo com pouco mais de 400 mil registros. Os arquivos em cada mês foram tratados para corrigir as seguintes inconsistências: (i) Número de colunas diferentes em cada arquivo; (ii) Informações de uma coluna no campo errado (Ex.: Telefone no campo do bairro); (iii) Nomes de bairros e cidades escritos em mais de uma forma (sendo substituído pelo correspondente identificador numérico); (iv) Campos de CEP e NUMERO_SIGOM (Identificador do cartão) vazios (sendo eliminados da base); e (v) Campos de endereço separados em mais de uma coluna. O identificador do cartão e o CEP são informações fundamentais para o método nesta pesquisa, já que serão necessárias para identificar padrões e gerar os modelos de identificação de embarques. A base inicial tinha 111MB e a base final carregada com 50MB, com uma taxa de tratamento de 47,60%.
- Os arquivos de Dicionários dos veículos foram tratados de modo a permitir uma consistência com os registros de veículos nos dados atuais de GPS e de Bilhetagem, eliminando-se também registros duplicados. Registros com código defasado de veículos da frota foram eliminados dos arquivos. No total foi possível identificar 2013 veículos, cada um com um identificador único referente a tabela no banco de dados. Um identificador adicional foi crido na tabela final para representar os veículos que aparecem em uma das duas bases, mas não estão presentes na tabela de dicionários. Isso é necessário, pois antecipa um possível erro de compilação na hora do tratamento das bases de Bilhetagem e GPS.
- O tratamento dos dados brutos de GPS, embora tenha demandado muito tempo (90 horas de processamento), consistiu na padronização dos dados de direção, latitude e longitude para o tipo decimal com 6 casas, do campo data e hora limpeza, e dos campos de odômetro e velocidade. O identificador do veículo foi padronizado conforme o formato estabelecido na tabela de dicionário dos veículos. Registros com coordenadas fora da região metropolitana de Fortaleza foram excluídos da base. O tratamento desta base demandou um elevado tempo de processamento, pois cada arquivo de um dia contém em média mais de 300.000 KB. Assim, devido a esta restrição de processamento computacional, restringiu-se neste trabalho a um período de três meses de dados. Os meses escolhidos foram meses com menos dias atípicos.

- A Figura 12 apresenta o processo de tratamento da base de Bilhetagem. Como pode ser visto, tratou-se inicialmente valores nulos/inconsistentes, incluindo a formatação de data, hora e nome da linha. Em seguida, padronizou-se o identificador dos veículos conforme a tabela de Dicionários do Veículos. Os registros de bilhetagem foram georreferenciados a partir de uma adaptação do método de Braga (2019), que consiste em associar a hora da validação aos instantes de registros de GPS para o mesmo veículo onde a validação foi realizada. As coordenadas encontradas são registradas na base de Bilhetagem como sendo o local de validação. Foi possível identificar as coordenadas de validação para 80% dos usuários. O restante não foi possível devido a existirem códigos de veículos que estão na base de GPS e não estão no dicionário, impossibilitando a integração.

Figura 12 - Método de Tratamento dos dados de Bilhetagem




Fonte: Autor

- Durante o processo de tratamento da base de Bilhetagem também foram estimadas as paradas de embarque das primeiras validações do dia para os usuários da amostra de cadastro. Esta estimativa foi feita com base na informação das coordenadas dos endereços dos usuários, nas linhas usadas para primeira viagem (incluindo sentido e suas paradas correspondentes do GTFS). A localização das paradas de embarque foi então usada para determinação das distâncias de validação (distância euclidiana entre a parada de embarque e a validação). Caso nenhum registro de parada fosse encontrado, um valor padrão de identificador nulo de paradas era salvo, para garantir a integridade do banco.

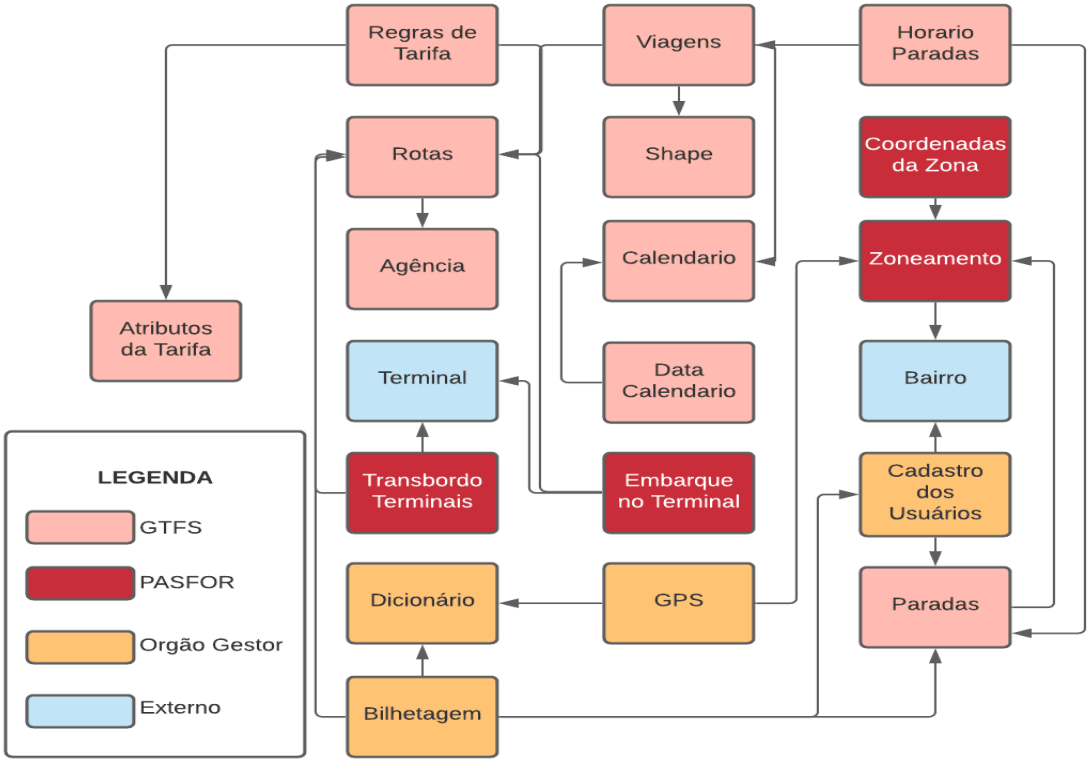
## 5.3. Carregamento dos dados e Estrutura do banco de dados

Após todo o processo de tratamento, os dados tratados foram carregados no banco. A Tabela 3 apresenta um resumo descritivo dos dados antes e depois do tratamento, e as taxas de tratamento (razão entre o número de registros nos dados tratados e os registros nos dados brutos) e compactação dos dados. Vale ressaltar que para algumas bases foi necessário inserir novos campos para carregamento, aumentando um pouco o tamanho da base. Cada linha da tabela representa uma base de dados utilizada na modelagem relacional, junto as respectivos resultados de tratamento e compactação das bases.

Todas as bases, exceto a de Cadastro dos Usuários, tiveram mais de 80% dos dados tratados e recuperados para utilização. Na base de Cadastro, obteve-se uma taxa de tratamento de 48%, provavelmente devido as inconsistências da base com endereços e números identificadores dos cartões vazios. Vale ressaltar, as bases de GPS e Bilhetagem tiveram uma excelente taxa de tratamento de 80,35% e 91,62%, respectivamente, dada a grande quantidade de dados, garantindo um grau de confiança para a próxima etapa do trabalho. Destaca-se também que, conforme a taxa e compactação, nota-se que somente 5 dentre as 20 bases tiveram redução do tamanho.

Por fim, como produto deste capítulo, além dos dados devidamente tratados e armazenados, apresenta-se a estrutura final do banco de dados relacional de *Big Data* de Transporte Público para Fortaleza (Figura 13), que norteará todas as consultas das próximas etapas e possibilitará maior otimização para busca de registros e atualização de dados, quando necessário. A ideia geral é que essa estrutura sirva também para outras pesquisas de modo a facilitar o uso de várias bases de dados, por intermédio da sua integração.

Figura 13 - Modelo Relacional simplificado para o Big Data de Transportes Público de Fortaleza.




Fonte: Autor.

Tabela 3 - Resumo dos resultados de tratamento do banco de dados

| ID | Bases de dados | Quantidade Dados Brutos | | Tamanho (KB) | Quantidade de Dados Após Tratamento | | Tamanho (KB) | Quantidade de dados no carregamento | | Tamanho (KB) | Taxa de Tratamento | Taxa de Compactação |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Linhas | Colunas | | Linhas | Colunas | | Linhas | Colunas | | | |
| 1 | Bairro | 119 | 5 | 489 | 119 | 4 | 6 | 119 | 4 | 8 | 100,00% | -98,36% |
| 2 | Zoneamento | 253 | 5 | 579 | 253 | 4 | 8 | 253 | 4 | 10 | 100,00% | -98,27% |
| 3 | Coordenadas zoneamento | 253 | * | 623 | 20716 | 3 | 780 | 20716 | 4 | 804 | 8188,14% | 29,05% |
| 4 | Paradas | 4969 | 12 | 312 | 4969 | 5 | 321 | 4696 | 5 | 343 | 100,00% | 9,94% |
| 5 | Cadastro Usuário | 475408 | 20 | 111777 | 345326 | 17 | 57585 | 226297 | 20 | 50940 | 47,60% | -54,43% |
| 6 | Atributos Tarifa | 2 | 6 | 1 | 2 | 6 | 1 | 2 | 4 | 2 | 100,00% | 100,00% |
| 7 | Agência | 2 | 7 | 1 | 2 | 6 | 1 | 2 | 6 | 2 | 100,00% | 100,00% |
| 8 | Rotas | 319 | 9 | 21 | 318 | 4 | 13 | 318 | 4 | 15 | 99,69% | -28,57% |
| 9 | Regras de Tarifa | 319 | 5 | 4 | 319 | 3 | 6 | 319 | 3 | 7 | 100,00% | 75,00% |
| 10 | Terminal | 9 | 5 | 6 | 9 | 5 | 1 | 9 | 5 | 3 | 100,00% | -50,00% |
| 11 | Embarque Terminal | 5359 | 14 | 474 | 5221 | 9 | 376 | 5221 | 9 | 387 | 97,42% | -18,35% |
| 12 | Transbordo Terminal | 88942 | 16 | 7290 | 88158 | 11 | 7469 | 88158 | 11 | 7328 | 99,12% | 0,52% |
| 13 | Calendário | 4 | 10 | 1 | 4 | 11 | 1 | 4 | 11 | 3 | 100,00% | 200,00% |
| 14 | Shapes GTFS | 111723 | 5 | 4113 | 108324 | 5 | 6187 | 108324 | 6 | 6530 | 96,96 | 58,76% |
| 15 | Data Calendário | 18 | 3 | 1 | 18 | 4 | 1 | 18 | 4 | 3 | 100,00% | 200,00% |
| 16 | Viagens | 82615 | 9 | 3308 | 82586 | 5 | 4407 | 82586 | 6 | 4338 | 99,96% | 31,14% |
| 17 | Hora Parada | 2665108 | 8 | 116201 | 2508282 | 5 | 135292 | 2508282 | 6 | 133057 | 94,12% | 14,51% |
| 18 | Dicionário | 2051 | 2 | 24 | 2013 | 2 | 46 | 2014 | 3 | 43 | 98,20% | 79,17% |
| 19 | GPS | 12594520 | 9 | 733000 | 100678235 | 6 | 7705000 | 100678235 | 7 | 7909000 | 80,35% | 7,9% |
| 20 | Bilhetagem | 20739540 | 9 | 1750000 | 1900120 | 8 | 2040000 | 19001200 | 12 | 2180000 | 91,62% | 24,57% |
| - | Média | - | - | - | - | - | - | - | - | - | 95,59% | 29,13% |
| - | Variância | - | - | - | - | - | - | - | - | - | 1,67% | 66,90% |
| - | Des. Padrão | - | - | - | - | - | - | - | - | - | 7,23% | 61,10% |

*Não se aplica

Fonte: Autor.

# 6. ANÁLISES EXPLORATÓRIAS

Neste capítulo é apresentada uma contextualização do Sistema de Transporte Público de Fortaleza e os resultados das análises exploratórias das validações.

## 6.2. Contextualização do Sistema de Transporte Público de Fortaleza

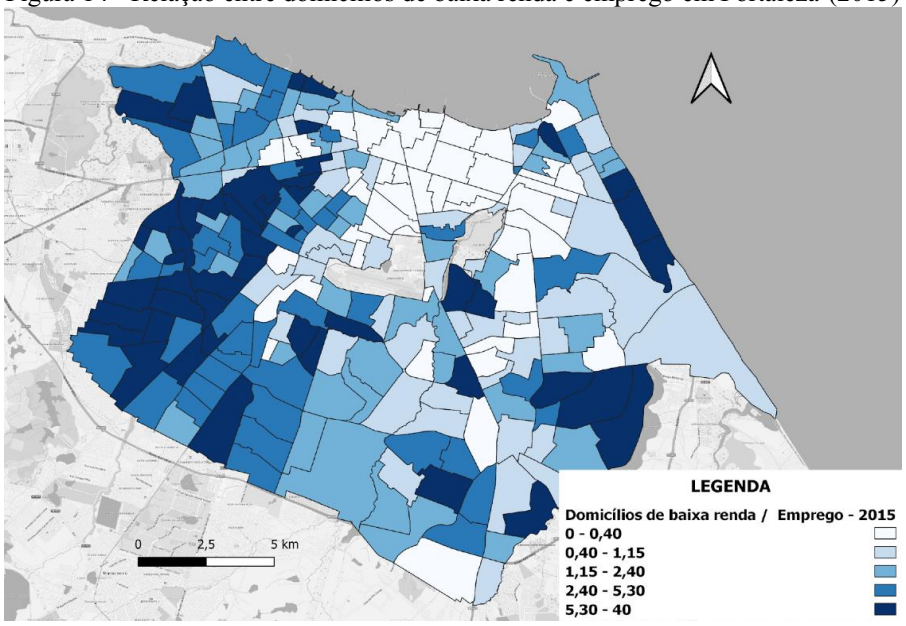
De acordo com dados do Censo, Fortaleza cresceu de 270 mil habitantes em 1950 para quase 2,452 milhões em 2010, e posteriormente para 2,703 milhões em 2021. Essa expansão trouxe grandes desafios aos planejadores da cidade no que se refere aos transportes dessa nova massa de habitantes (Braga, 2019). As análises de Lima (2017) apresentam Fortaleza com a perspectiva da representação do sistema de atividades e do Uso do Solo, assumindo uma inter-relação com o sistema de transportes. Nesta perspectiva, a cidade foi caracterizada pela distribuição espacial das atividades e dos domicílios, a partir de dados do Ministério do Trabalho (MTE) e do Censo do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2010. Essa caracterização pode ser observada na Figura 14, onde demonstra a relação de espacialização dos domicílios de baixa renda pela distribuição dos empregos (extrapolados para 2015) com concentração das atividades de Fortaleza nos bairros com maior quantidade de domicílios de alta renda. Conforme pode ser visto, a população de baixa renda reside nas regiões periféricas de Fortaleza, com baixos índices de mobilidade e acessibilidade, proporcionando uma maior necessidade de utilização do transporte público. Fato que é evidenciado também em análises posteriores da concentração de validações do TP em regiões periféricas do município.

A Autarquia Municipal de Trânsito, Serviços Públicos e Cidadania (AMC) é a responsável pela fiscalização do Código de Trânsito Brasileiro em Fortaleza e atua junto à Empresa de Transporte Urbano de Fortaleza (ETUFOR), que é a responsável pelo controle, regulação e fiscalização dos sistemas de transporte coletivos da cidade, sendo ônibus, táxi e mototáxi à exemplo. O sistema integrado de Transporte público da cidade de Fortaleza (SIT-FOR) foi implantado em 1992, deixando de ser um sistema radial e tornando-se tronco-alimentador (Fortaleza, 2015). O sistema é operado por ônibus e, desde 2012, estava dividido em cinco áreas de operação e uma área neutra, sendo que a operação segue regime

consorciado, porém em dezembro de 2020 os bairros da cidade foram reorganizados dentro de 12 regionais.

A Figura 15 demonstra a relação espacial das linhas de ônibus, os seus pontos de parada e os terminais de integração. É possível observar uma grande concentração de linhas e terminais na parte oeste da cidade, além de um ponto bastante importante, a sobreposição das linhas com o zoneamento, ou seja, existe forte predominância das linhas passarem pelas principais vias divisórias dos bairros.

Figura 14 - Relação entre domicílios de baixa renda e emprego em Fortaleza (2015)

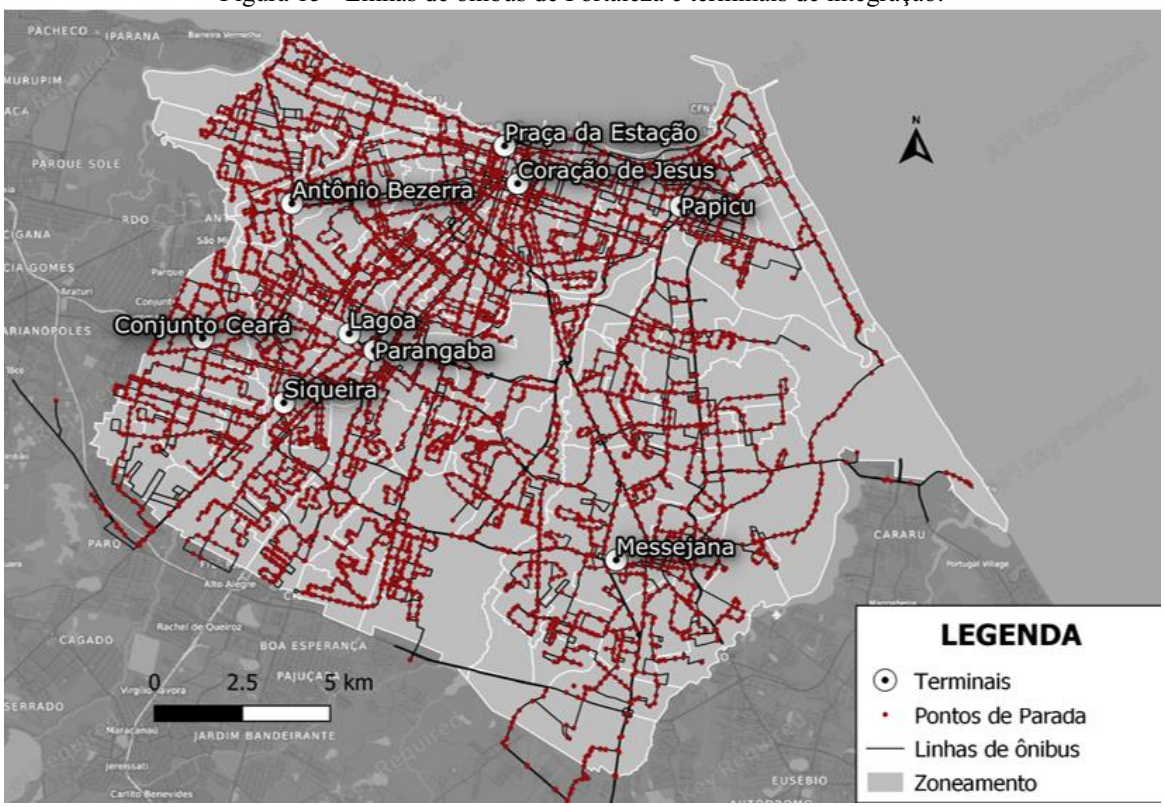



Fonte: Adaptado de Lima (2017)

O sistema, que registrava em 2018 com mais de 1 milhão de validações, contava com 279 linhas de ônibus regulares e 22 linhas complementares. De acordo com os dados de GTFS de 2021, a extensão total da rede era de 4969 km, com uma média de 11 km por linha. Ao todo, são 14 empresas de ônibus gerenciando as linhas regulares e 320 cooperados gerenciando a operação das linhas complementares (Fortaleza, 2015). A rede de linhas conta com pouco mais de 5000 pontos de parada distribuídos na rede e uma frota aproximada de 2700 veículos (BRAGA, 2019). O sistema conta atualmente também com 7 terminais fechados de integração física (Antônio Bezerra, Conjunto Ceará, Messejana, Papicu, Parangaba, Siqueira e Lagoa) e 2 terminais abertos de integração tarifaria (Praça da Estação e Coração de Jesus).

O SIT-FOR tem atualmente 12 tipos de linhas operando, sendo considerado nesse esforço de contextualização as 7 linhas descritas no anuário de Fortaleza de 2010, sendo elas: Alimentadora (ALM), Circular (CIR), Complementar (CMP), Troncal (TRC), Troncal Expressa (TRE), convencional (CNV) e especial (ESP). As Linhas Alimentadoras levam passageiros dos bairros até os terminais, enquanto as Linhas Troncais levam os passageiros dos terminais até a área central da cidade (BRAGA, 2019). As Linhas complementares atuam como as linhas alimentadoras, porém sua diferenciação está na extensão geográfica de atuação, levando a demanda dos bairros mais distantes (periféricos) aos terminais. As Linhas Convencionais ligam diretamente os bairros ao Centro, sem passar pelos terminais. As Linhas Circulares ligam diversos terminais (com percurso poligonal) passando pelos bairros, evitando o tráfego das regiões centrais. As Linhas Corujões realizam o transporte a partir das 00:00 quando as demais linhas cessam a operação. Em resumo, aproximadamente 36% das linhas em Fortaleza são do tipo Alimentadora, 22% do tipo Convencional, 20% do tipo Complementar, 6% do tipo Troncal e 16% dos tipos restantes (Circular, Especial e Corujão).

Figura 15 - Linhas de ônibus de Fortaleza e terminais de integração.




Fonte: Elaborado pelo autor.

Atualmente o valor da tarifa cobrada para se utilizar o TP de Fortaleza é de R$3,90 para passagens inteiras, e com passagem estudantil no valor de R$ 1,80. É válido ressaltar que durante a hora social (segunda a sábado, de 9h às 11h e de 14h às 16h) o valor da passagem é de R$ 3,30 a inteira e R$1,50 a estudantil, sendo uma tentativa da prefeitura de lidar com a queda da demanda por conta da descentralização das atividades nesses horários entre picos (Filho, 2002).

Desde sua criação, a rede de transportes público de Fortaleza é integrada operacional e fisicamente, pois os serviços são operados de forma coordenada com relação a horários, frequências e itinerários, além da reestruturação espacial, para que os usuários realizem caminhadas mínimas, quando houver necessidade de transferência, dentro de um espaço que proporcione segurança e comodidade. Desde 2013 foi incorporada a integração temporal, a partir de um bilhete eletrônico (*smart card*), possibilitando realização de transferências em qualquer ponto da rede, dentro de um período de 2 horas. O sistema de Bilhetagem Eletrônica é aberto (*tap-on*), ou seja, a validação (coleta da tarifa pelo cartão eletrônico) pode ser realizada em qualquer momento da viagem. Assim, para o caso de Fortaleza os locais de validação não são necessariamente os locais de embarque o que torna a estimativa de atributos das viagens (p. ex., locais de embarques, desembarques e transferências) desafiadora. Também não se tem informações do motivo da viagem e do destino.

Tabela 4 - Comparativo percentual do número de paradas por comprimento das linhas (2010 e 2021)

| **Intervalo (m)** | **Percentual - 2010** | **Percentual – 2020** |
|---|---|---|
| 0 – 100 | 3,5% | 5,6% |
| 100 – 200 | 18% | 15,4% |
| 200 - 300 | 35,9% | 30,4% |
| 300 – 400 | 21,1% | 21,8% |
| 400 – 500 | 9,7% | 11,4% |
| 500 - 600 | 4,7% | 6,5% |
| 600 – 700 | 2,6% | 2,8% |
| 700 – 800 | 1,3% | 2,2% |
| 800 – 900 | 1,2% | 1,4% |
| 900 - 1000 | 0,6% | 0,9% |
| 1000 > | 1,5% | 1,5% |

Fonte: Autor.

Um aspecto importante sobre o sistema e que será de grande valia para as modelagens seguintes é a influência da acessibilidade ao sistema, por intermédio da distância entre as paradas, com distância média de 550m, e detalhado na Tabela 4. Aparentemente em 10 anos, o sistema viário da cidade obteve mudanças significativas, crescendo junto a urbanização das

periferias da cidade, fato que é refletido no aumento percentual das distâncias entre paradas de intervalos maiores (acima de 400 m). Esse impacto tem forte influência sobre como os usuários se deslocam, para onde e a motivação desses deslocamentos.

## 6.2 Análises Espaciais e temporais agregadas

Neste tópico serão apresentadas as análises exploratórias para mineração dos padrões de validação dos usuários de Transporte Público de Fortaleza. Conforme apresentado no método, as análises estão divididas em 3 categorias para melhor avaliar-se os possíveis padrões invisíveis no grande contingente de dados, sendo elas: *(i) Análises espaciais e temporais agregadas*; *(ii) Análises das primeiras validações*; e *(iii) Análises a nível do indivíduo*.

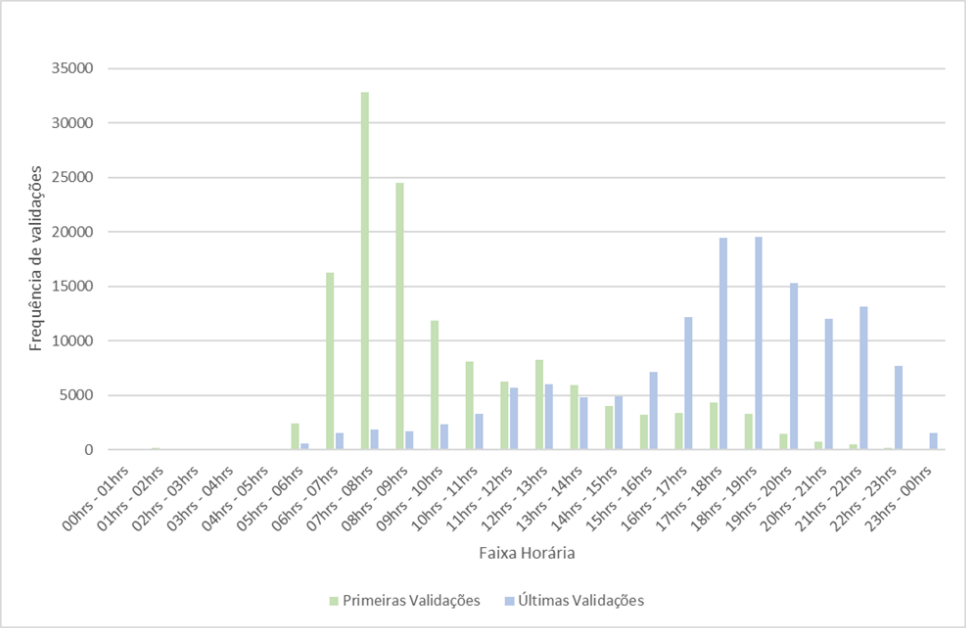
Nesta etapa foram utilizados os dados tratados do banco de dados. Para o ano de 2018, registrou-se em média 969.672 validações por dia, que correspondem a 328.260 usuários. Destes usuários, 156.682 estão na base de Cadastro, sendo que somente 20,6 mil foram classificados com residência válida (veja capítulo 4). Para análises agregadas (i), adotou-se uma amostra de 42,4 mil usuários do cadastro, com informações de 2018, e um período de três meses típicos de dados, devido a restrição computacional. Já para as análises (ii) e (iii), adotou-se somente a amostra de 20,6 mil usuários com endereços validados conforme a distância da residência a parada de embarque, e um período de seis meses, sendo apenas dias típicos e desconsiderando integrações.

### *6.2.1. Variação da frequência de validações ao longo do dia e entre dias*

A frequência de validações conforme apresentado no método é um importante indicador para categorizar os padrões, seja para definir grau de regularidade ou para representar os números de atividades acessadas pelos usuários.

Analisou-se inicialmente as médias de validações por faixa horária das primeiras e últimas validações, conforme exposto na Figura 16. Os picos de validações das primeiras ocorrem entre 07:00 e 09:00 horas da manhã, enquanto os picos da tarde ocorrem entre 17:00 e 19:00 horas. Dessa forma há indícios de que esta frequência possa contribuir como atributo na segregação dos padrões de uso.

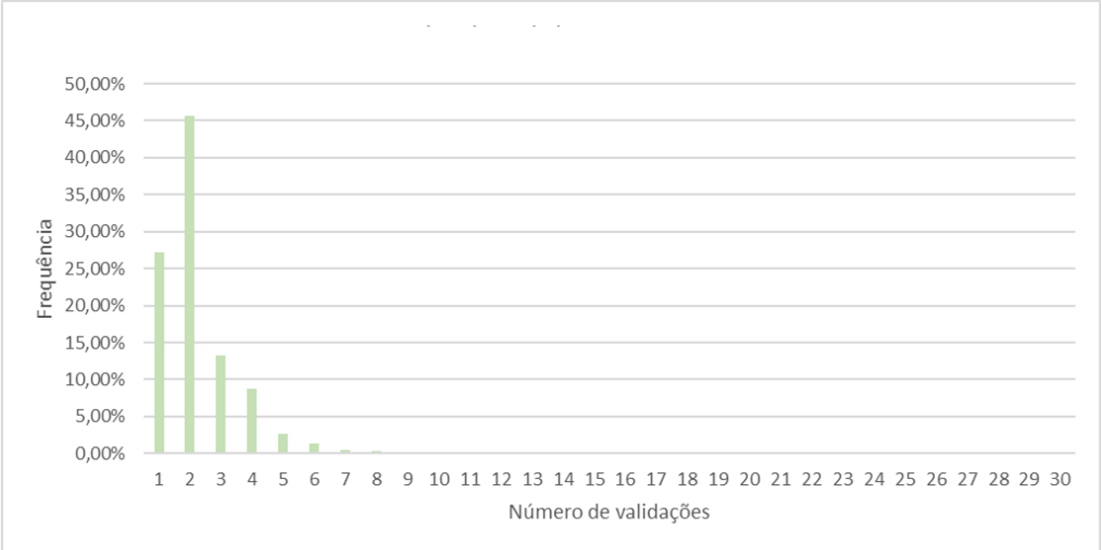
Figura 16 - Frequência média de validações por faixa horária




Fonte: Autor.

A figura a seguir apresenta a frequência de validações diárias, onde 59% das validações diárias ocorrem entre 2 ou 3 vezes e que na maioria das vezes (> 70%), ocorrem 1 ou 2 validações, indicando que a maioria dos usuários normalmente acessam apenas 1 atividade utilizando o Transporte Público. Vale destacar que 27% dos usuários apresentam validações intermediárias, sendo possíveis integrações ou atividades curtas.

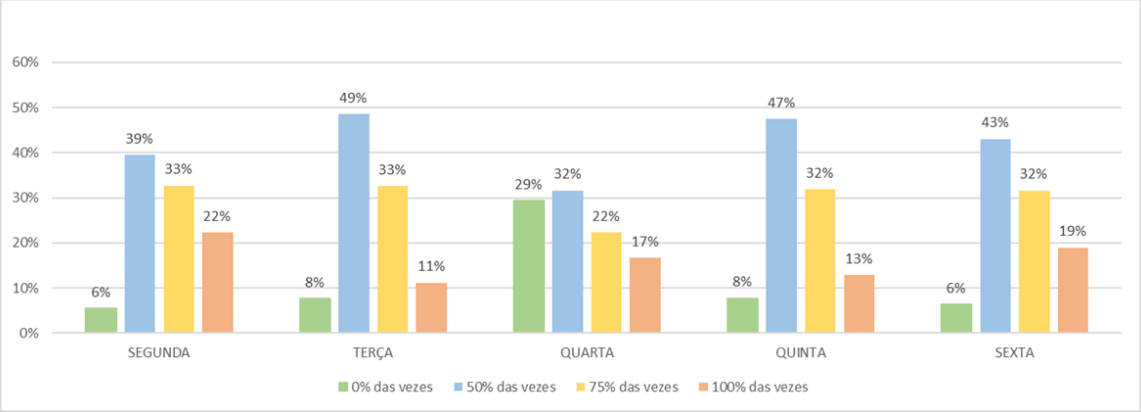
Figura 17 - Média de validações diárias




Fonte: Autor.

A figura a seguir apresenta uma comparação dos padrões de validações entre os dias da semana. Os grupos de frequência de validação foram segregados por dia útil (segunda a sexta) e foi verificado para cada usuário se ele validava 0%, 50%, 75% ou 100% das vezes com a mesma frequência em determinado dia, de modo a segregar em faixas de frequência de validação. Os resultados indicam um certo nível de regularidade de uso do sistema já que na maioria das vezes, algo em torno de 87%, os usuários repetem o mesmo número de validações do sistema para um dia específico da semana. Contudo, observe que há uma diferença entre dias, com a quarta-feira apresentando uma menor regularidade, provavelmente devido às diferentes atividades que são realizadas ao longo da semana. Fatores como hábitos de deslocamento e tipo de atividades podem auxiliar na compreensão desse comportamento.

Figura 18 - Análise dos padrões de frequência por dia da semana



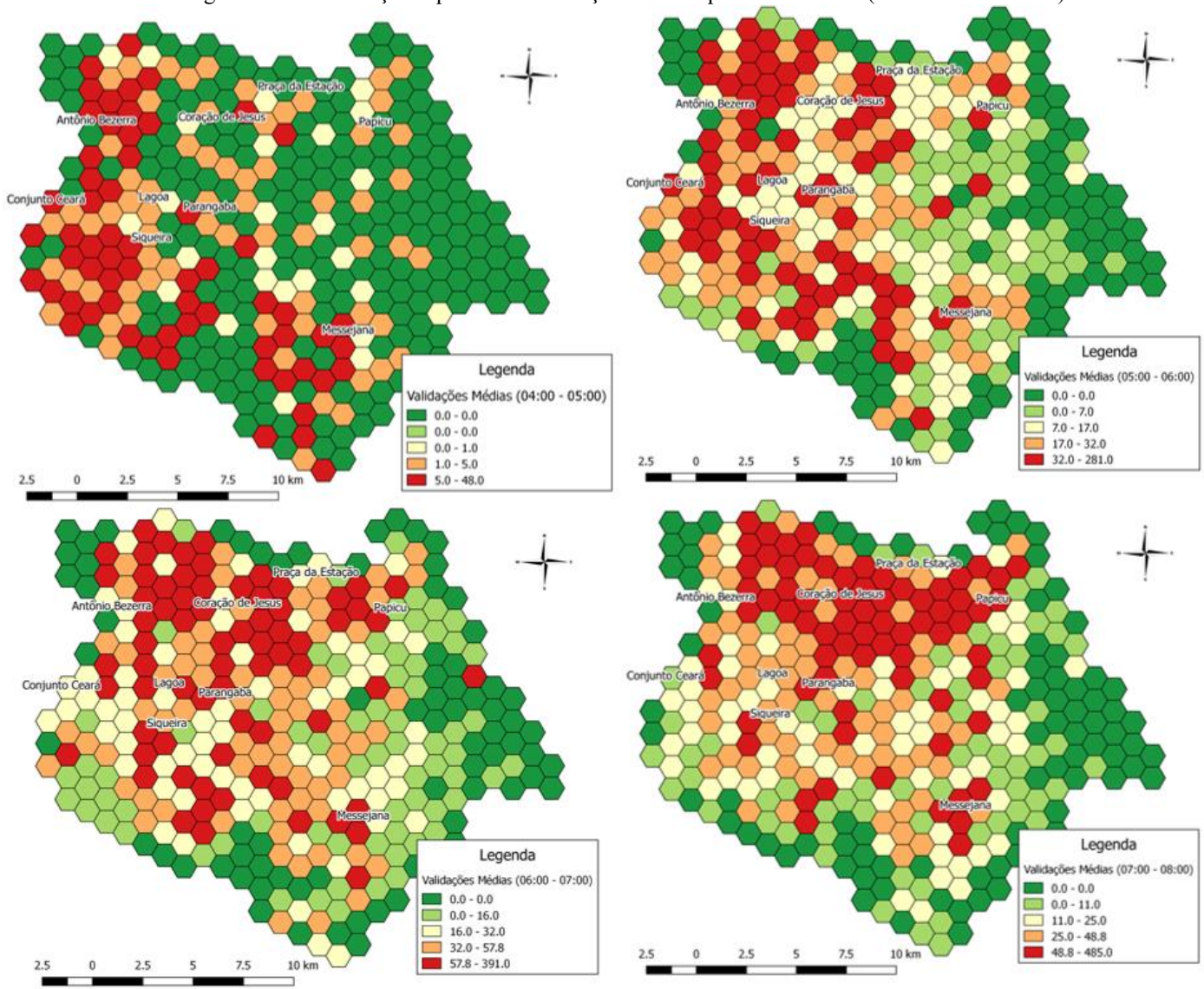

Fonte: Autor.

### *6.2.2. Padrão espacial das validações por faixa horária*

As Figura 19 a Figura 23 apresentam os mapas da frequência média de validação por faixa horária das primeiras e últimas validações dentro do período em análise em cada zona. Pode-se visualizar que no período da manhã, as primeiras validações se concentram nas regiões periféricas da cidade, longe do centro comercial (local onde está localizado os terminais da Praça da Estação e Coração de Jesus). Regiões adjacentes às regiões mais periféricas mantiveram-se no quarto quintil, enquanto as zonas no centro da cidade estão no primeiro quintil. No período da tarde, uma maior concentração de validações ocorre nas áreas centrais, dando um indício de que em períodos fora do pico da manhã os usuários tendem a validar próximo ao destino das viagens. Vale destacar também que durante todo o período em análise, os hexágonos próximos aos terminais se mantiveram nas faixas mais altas de

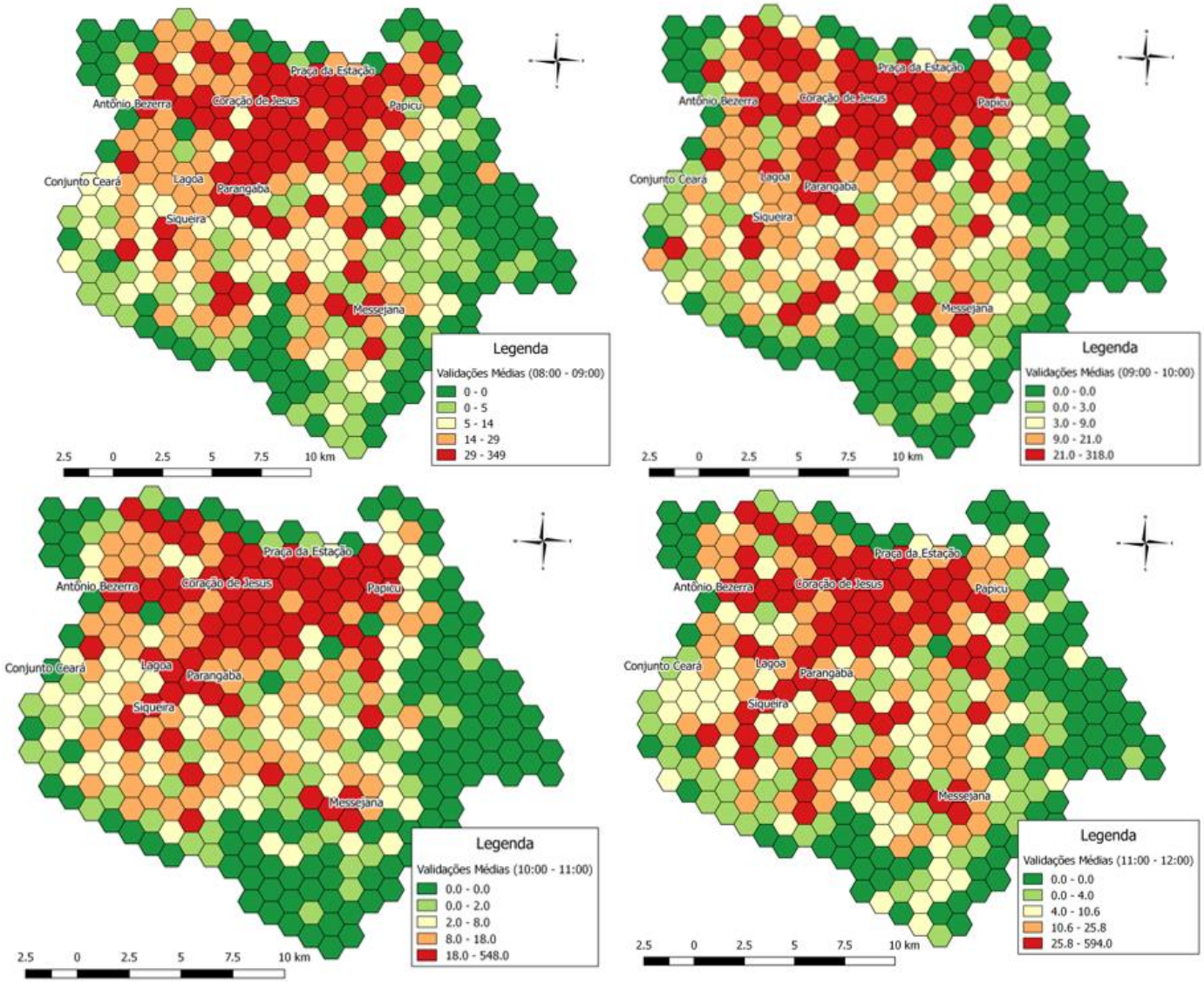
frequência de validação, como por exemplo os terminais da Messejana, Antônio Bezerra e Siqueira. Esses padrões agregados, se comportam de forma similar, porém não igual em diferentes dias da semana, sendo utilizado como atributo para modelagem dos grupos de usuários e do local de embarque, uma vez que existe diferentes padrões espaciais apontados nesta análise, mas que ainda não estão claros como se distribuem, papel este da etapa de clusterização e de identificação dos padrões, conforme será abordado adiante.

Figura 19 – Distribuição espacial das validações médias por faixa horária (04:00hrs - 08:00hrs)



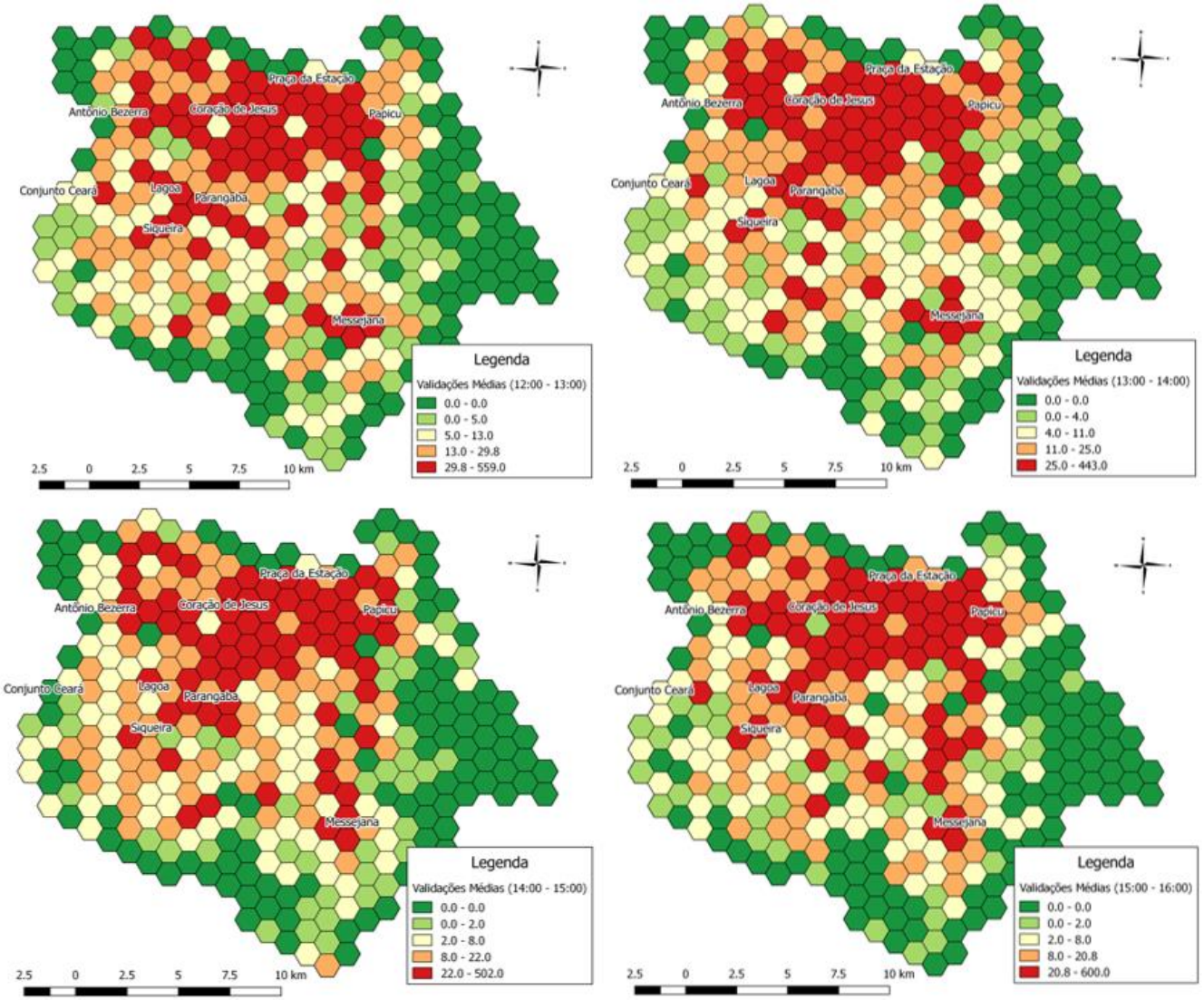

Fonte: Autor.

Figura 20 - Distribuição espacial das validações médias por faixa horária (08:00hrs - 12:00hrs)



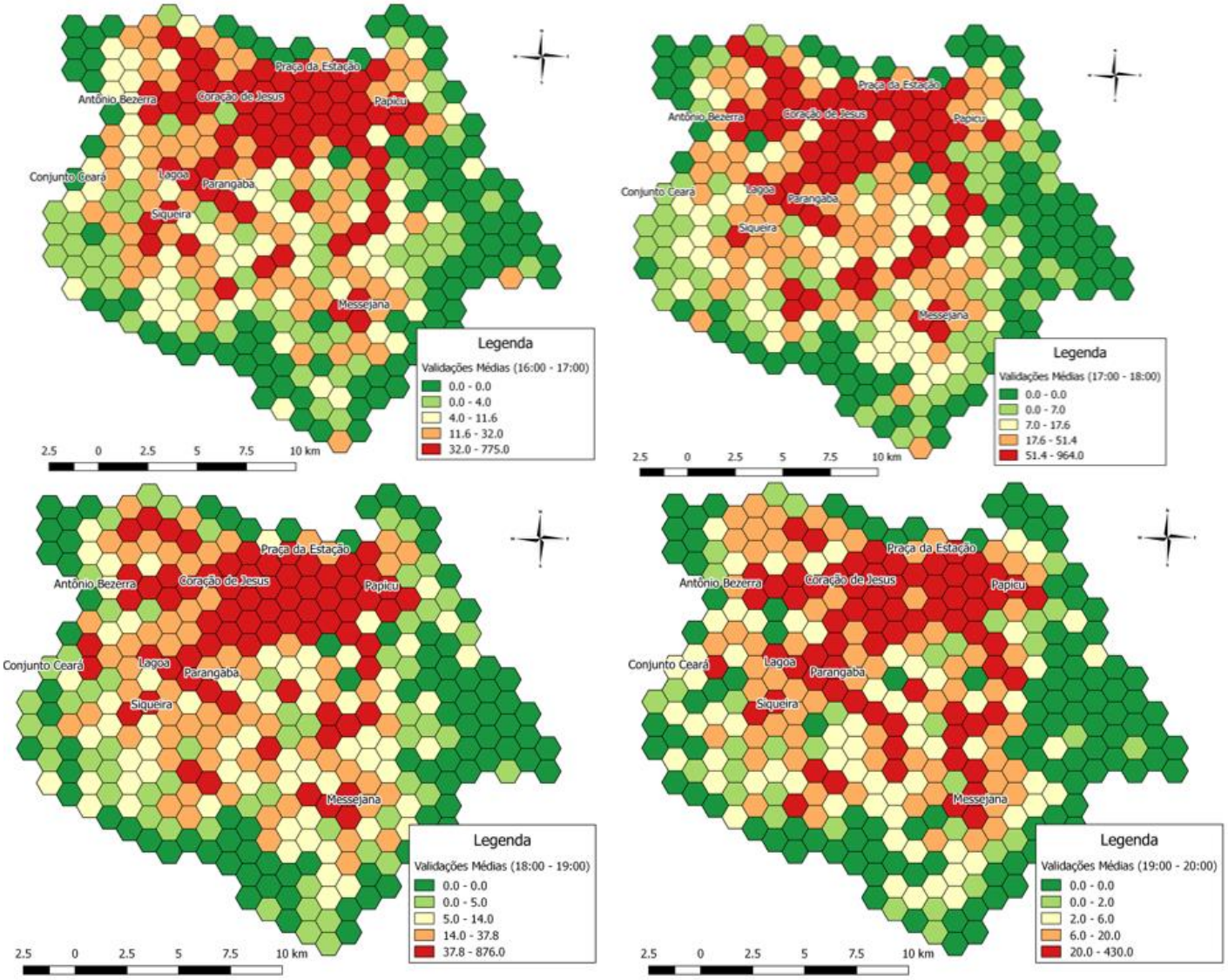

Fonte: Autor.

Figura 21 - Distribuição espacial das validações médias por faixa horária (12:00hrs - 16:00hrs)



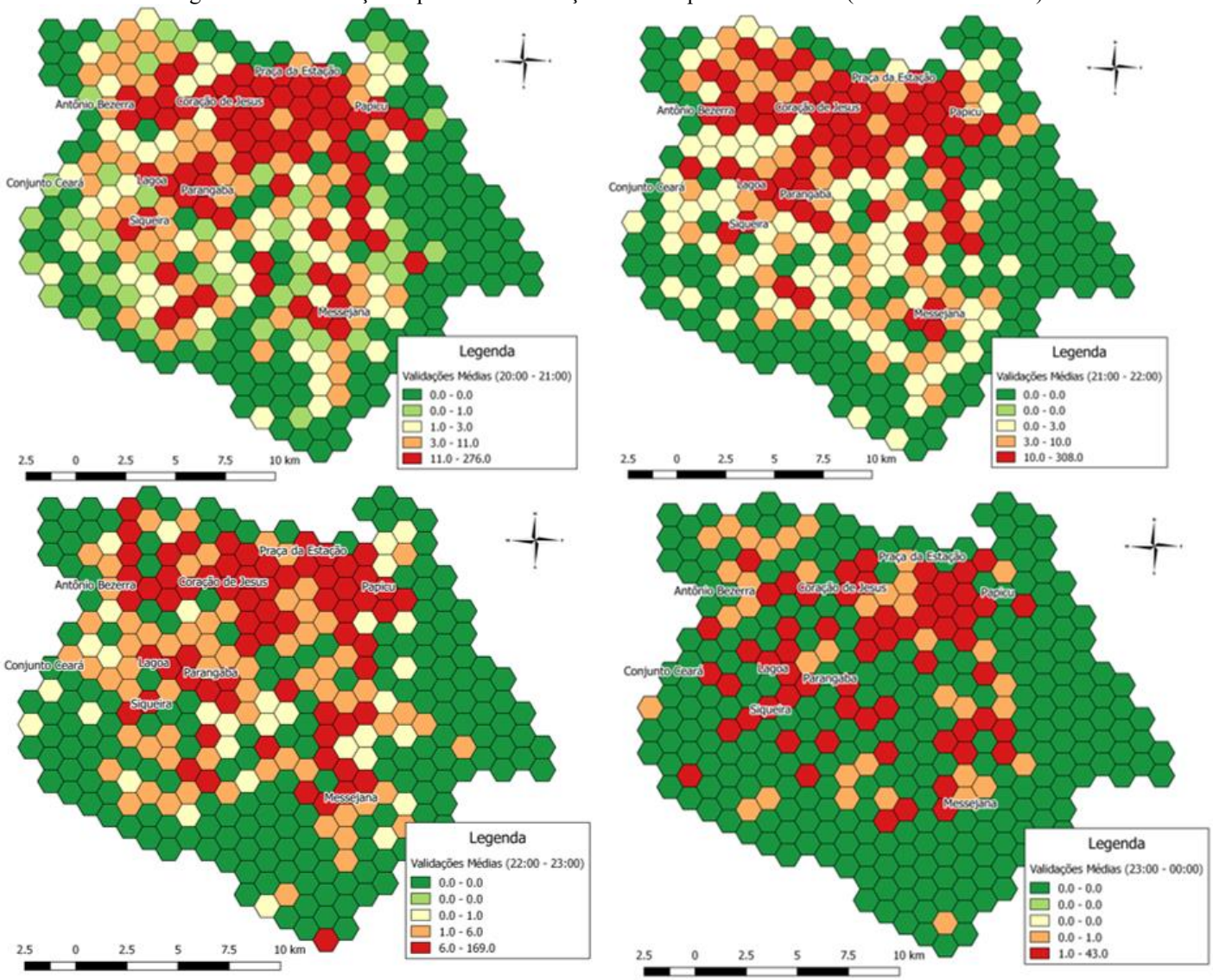

Fonte: Autor.

Figura 22 - Distribuição espacial das validações médias por faixa horária (16:00hrs - 20:00hrs)




Fonte: Autor.

Figura 23 - Distribuição espacial das validações médias por faixa horária (20:00hrs - 00:00hrs)




Fonte: Autor.

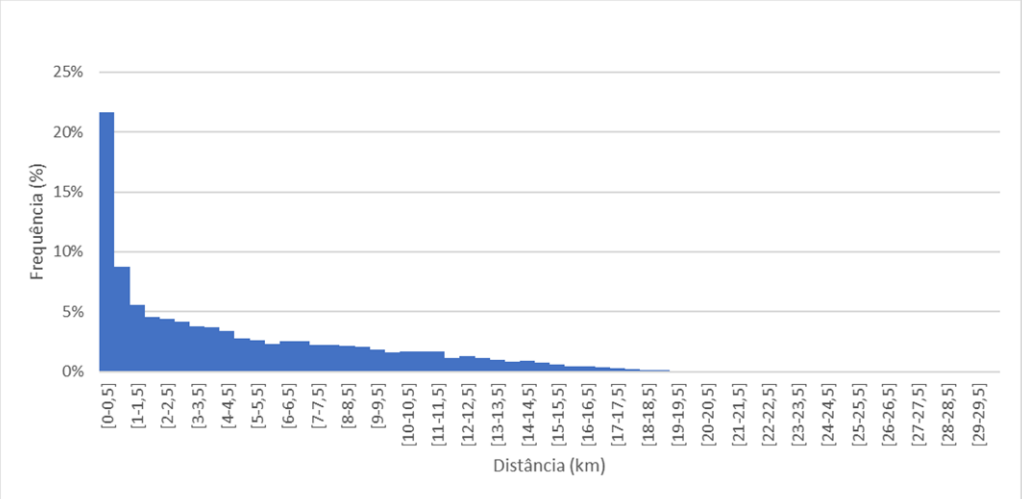
## 6.3. Análises das primeiras validações

Nestas análises, utilizou-se a amostra de 20,6 mil usuários do cadastro com endereços válidos, juntamente com as estimativas de locais de embarque para as primeiras validações e de distâncias de validação em relação ao local de embarque. Vale ressaltar que nesta amostra foram considerados os usuários com distâncias da residência a parada de embarque de até 1000 m.

### *6.3.1. Distribuição da distância de validação e de caminhada*

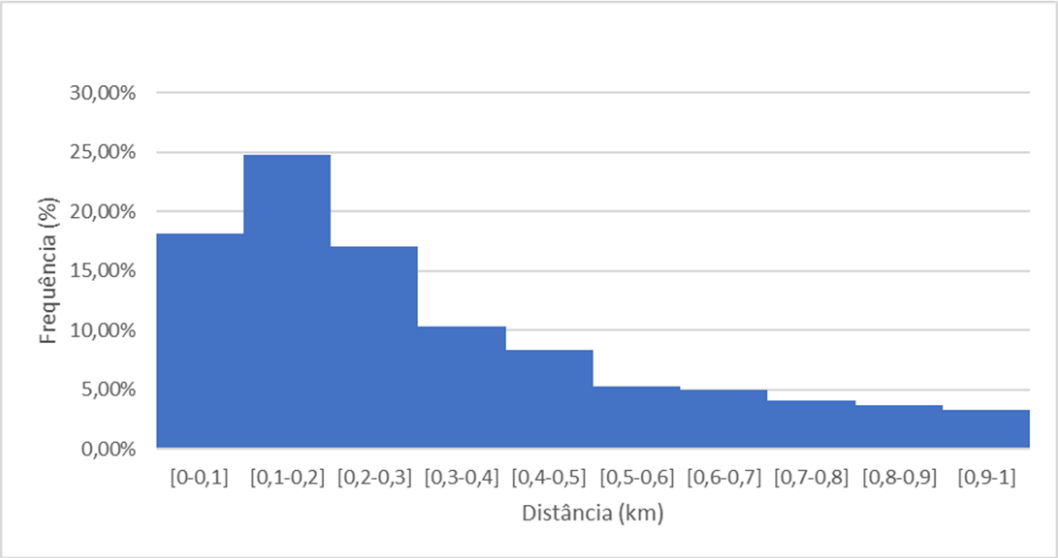
Na Figura 24 está apresentado a distribuição das distâncias entre as residências e as paradas de embarque, considerando toda a base de Cadastro com endereço completo. Observa-se uma parcela considerável (35% dos usuários) com distância longas, acima de 5 km, e apenas 22% com distância de até 500m. Estes resultados indicam que os endereços na Base de Cadastro podem estar incorretos ou incompletos gerando erros nas estimativas de distâncias. Assim, conforme proposto, usuários com distância à parada de embarque maior do que 1000 m foram excluídos da análise. A Figura 25 apresenta apenas os usuários considerados válidos para a pesquisa (20.632 usuários), onde 78% apresentam distâncias de até 500 m.

Figura 24 – Distância de caminhada (m) pela frequência de usuários – Base Completa




Fonte: Autor.

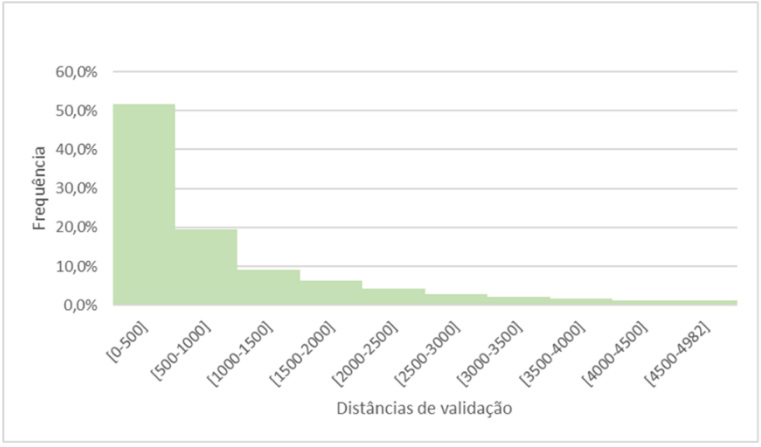
Figura 25 – Distância de caminhada (m) pela frequência de usuários – Base Amostral




Fonte: Autor.

A Figura 26 apresenta o histograma da variável distância de validação para os usuários válidos. Observa-se que mais de 50% validações ocorrem a uma distância de até 500 m da parada de embarque. A média e desvio padrão encontrados foram de 1,31 e 1,95 km, respectivamente. Existe, portanto uma elevada variabilidade das distâncias de validação, provavelmente devido a diferentes comportamentos e a diversidade de características das viagens. Embora, na maioria das vezes ocorre uma tendência de se validar no momento do embarque, percebe-se uma grande parcela de validações distantes, acima de 1 km, do local de embarque. Portanto, a identificação dos locais de origem das viagens é uma etapa essencial para que os dados de Bilhetagem possam ser usados para qualquer análise.

Figura 26 - Distribuição das distâncias (m) entre embarques e validações dos usuários

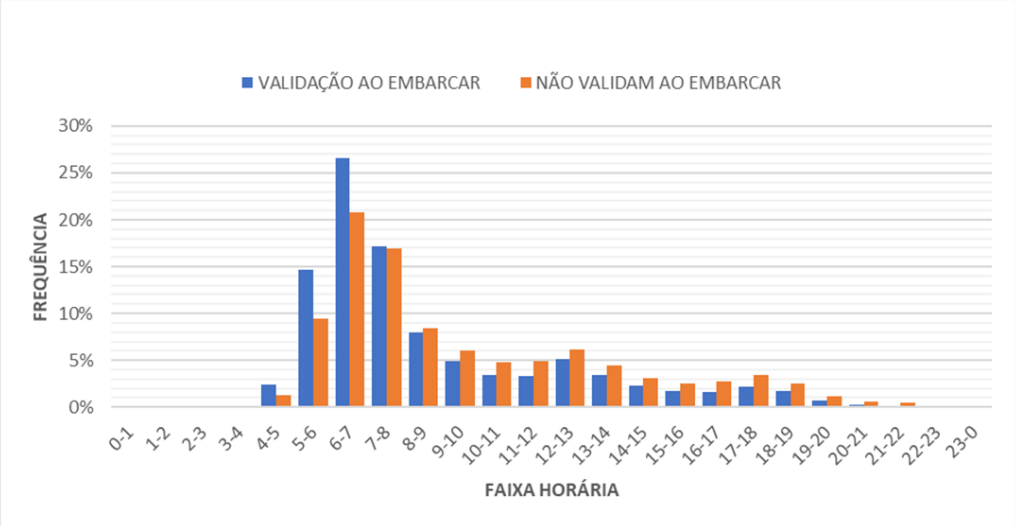



Fonte: Autor.

### *6.3.2. Proporção de validações por hora*

Dentre os usuários válidos do cadastro, 76% detêm a linha mais frequente de utilização do dia como a mesma da primeira validação do dia, e 27% têm a primeira validação do dia próximo aos terminais. A Figura 27 apresenta o perfil de variação horária da proporção de primeiras validações que ocorrem e que não ocorrem no momento do embarque. Assumiu-se que uma validação ocorre no momento do embarque se a distância de validação for inferior a 500 m. Os picos da primeira validação ocorrem no período da manhã, principalmente entre 6h e 7h. Observa-se uma maior tendência de validar ao embarcar nos primeiros horários do dia, entre 4h e 8h. Essa tendência de validar assim que embarca aparenta desaparecer após o pico da manhã, dando um indício que as validações no pico da tarde e período noturno apresentam maiores distâncias entre o local de embarque e de validação.

Figura 27-Frequência de usuários por faixa horária




Fonte: Autor.

### *6.3.3. Proporção de validações por tipo de linha*

A Tabela 5 apresenta as proporções de validação ao embarcar conforme o tipo de linha e volume horário de passageiros da linha, tendo como finalidade demonstrar a influência da lotação nas validações. Foram consideradas 4 tipologia de linhas: Alimentadora (ALM), Complementar (CMP), Convencional (CNV) e Troncal (TRC). Para todas as linhas, as proporções descrescem a medida que aumenta o volume horário de passageiros da linha. Apenas a linha complementar apresenta uma menor proporção na primeira classe. Este tipo de

linha contém mais da metade (65%) de suas rotas do tipo integração entre terminais, o que indica que a ocupação dentro dos veículos tem menor influência neste tipo de linha. Dessa forma, essa análise indica que existe uma relação entre a lotação do veículo e o comportamento de validar ao embarcar, indicando que uma maior ocupação dos veículos pode resultar em menos validações no embarque. Além disso, a integração no terminal pode levar os usuários a validarem distantes do local de embarque.

Tabela 5 - Proporções de usuários que validam ao embarcar por linha e validação média.

| Classes (N° de Validações) | ALM | CMP | CNV | TRC |
|---|---|---|---|---|
| 0 -\|200 | 0,67 | 0,37 | 0,58 | 0,64 |
| 200 -\| 400 | 0,24 | 0,34 | 0,29 | 0,35 |
| 400 -\| 600 | 0,08 | 0,18 | 0,09 | 0,01 |
| 600 -\| 800 | 0,01 | 0,10 | 0,02 | - |
| 800 -\| 1000 | - | 0,02 | 0,02 | - |

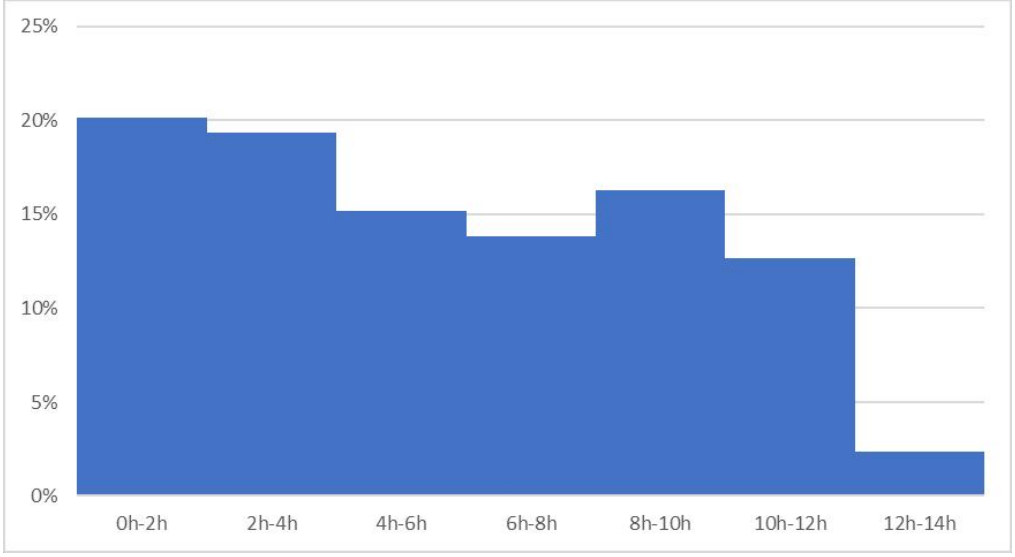
Fonte: Autor.

## 6.4. Análises à nível do indivíduo

### *6.4.1. Distância Temporal entre as primeiras e últimas validações*

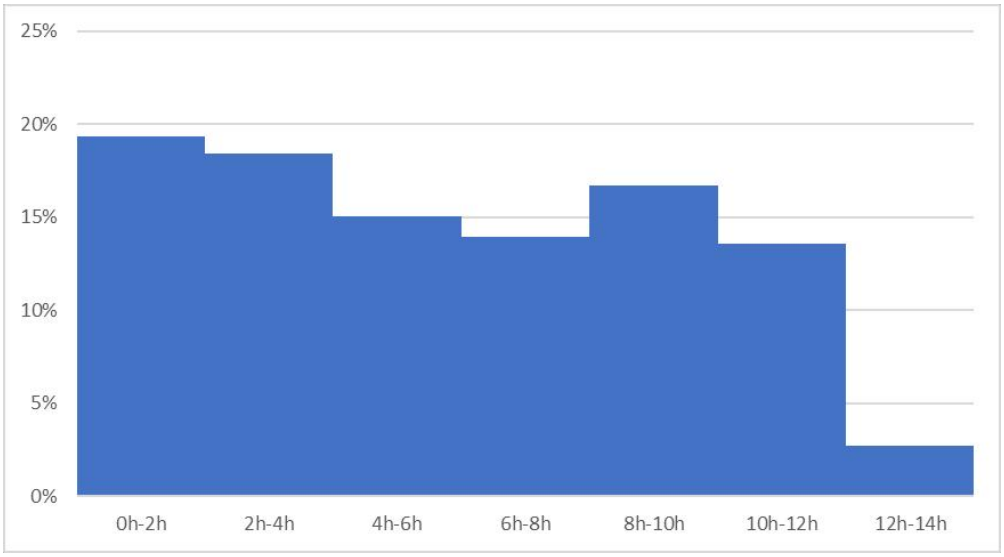
Foram realizadas análises das distâncias temporais entre as primeiras e a ultimas validação dos usuários do cadastro (Figura 28 à Figura 32). Este indicador está associado a natureza das atividades realizadas. Observa-se uma elevada heterogeneidade nestes tempos indicando uma diversidade de atividades que são acessadas usando o sistema. Nos 5 dias da semana as maiores frequências de distâncias temporais estão em até 4 horas, indicando a existência de usuários com atividades curtas (em relação ao horário diário de 8 horas estimado no mercado de trabalho). Porém, em média para todos os dias da semana, 44% dos usuários apresentam distâncias temporais entre 6 e 12 horas, indicando atividades mais longas. Nota-se ainda um padrão similar de distribuição comparando os dias da semana.

Figura 28 - Distância Temporal - Segunda-feira



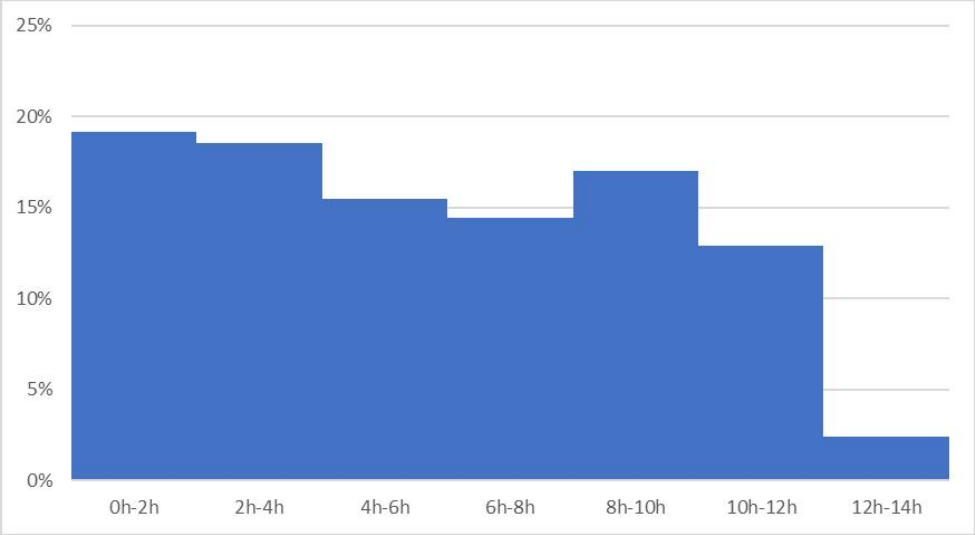

Fonte: Autor.

Figura 29 - Distância Temporal - Terça-feira



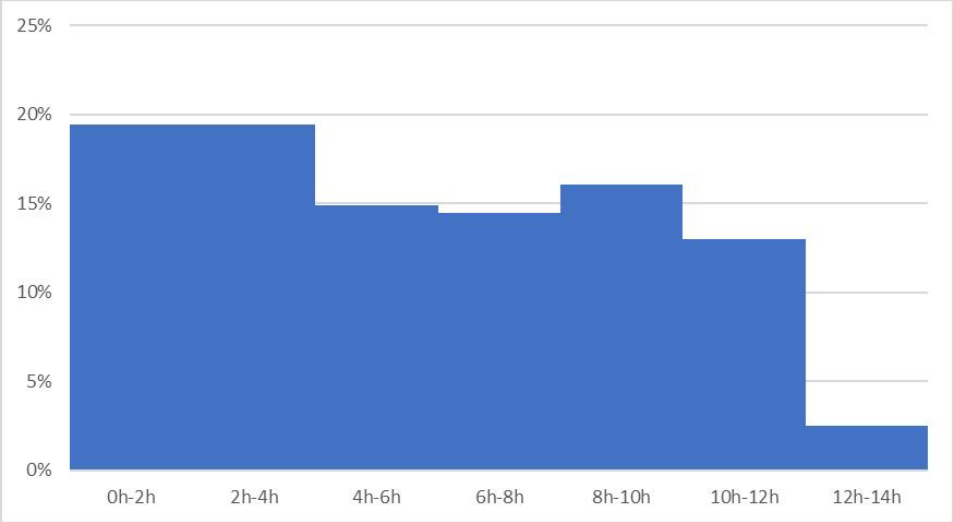

Fonte: Autor.

Figura 30 - Distância Temporal -Quarta-feira



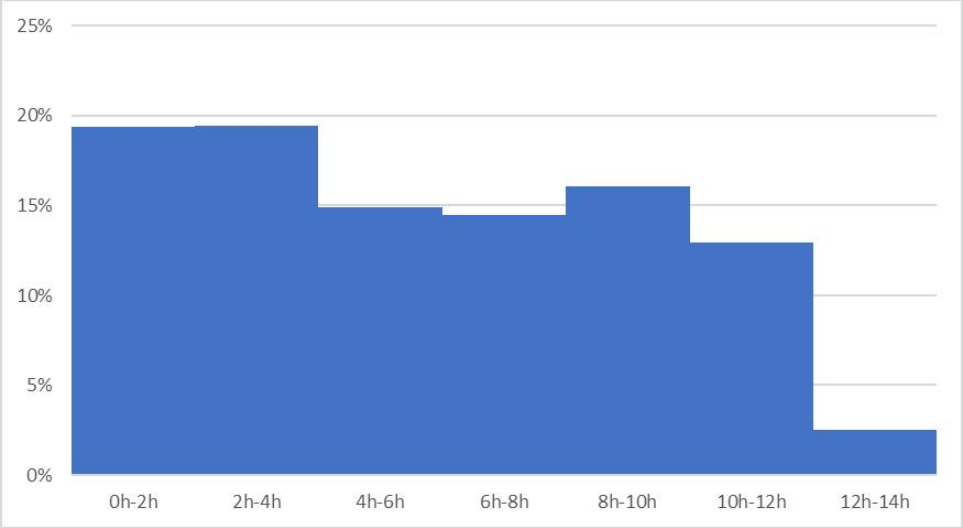

Fonte: Autor.

Figura 31 - Distância Temporal - Quinta-feira




Fonte: Autor.

Figura 32 - Distância Temporal - Sexta-feira



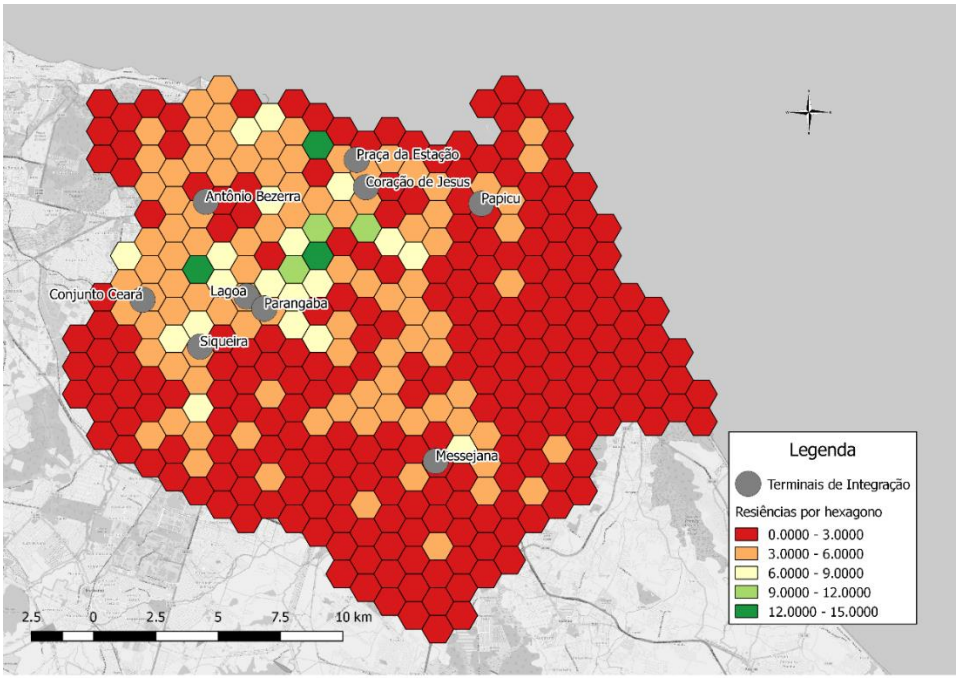

Fonte: Autor.

### *6.4.2. Perseguição Espaço-Temporal*

Nesta seção será apresentado de forma amostral o processo de "perseguição" de um usuário com *smartcard* de estudante de modo a evidenciar um padrão de uso. A Figura 33 apresenta a distribuição das residências dos usuários válidos no cadastro. A maioria das residências dessa amostra se concentra nos hexágonos próximo ao terminal da Parangaba (Oeste) e no centro comercial da cidade (norte).

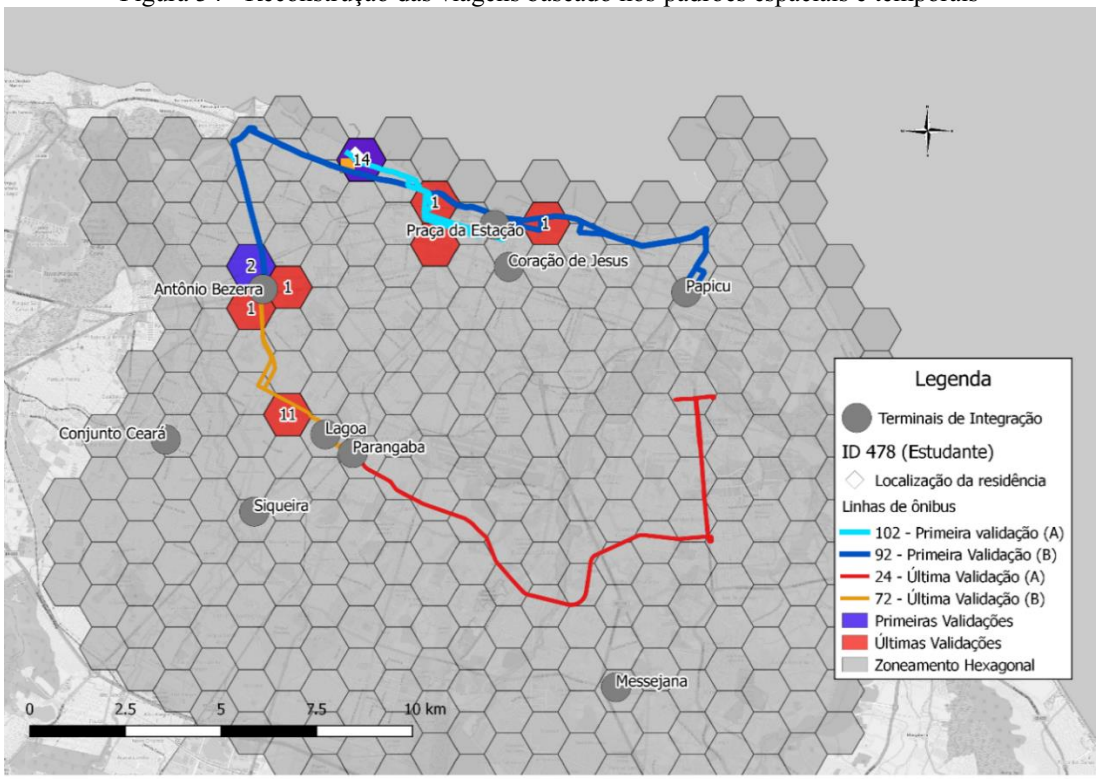
Na Figura 34 está representada a localização das validações no espaço temporal de um mês e a linhas mais frequentes, onde a letra "A" representa a linha de maior frequência e a letra "B" a subsequente. Existe uma alta concentração de validações do usuário no hexágono onde está localizado sua residência, dando fortes indícios de que esta seja sua origem. Avaliando as últimas validações existe uma forte concentração próximo aos terminais do Lagoa e Parangaba, e que seu embarque da última viagem provavelmente é no terminal ou em regiões próximas ao terminal. Avaliando separadamente as linhas por tipo de validação, as linhas 102 e 92 (linhas mais frequentes das primeiras validações) apresentam mais de 80% de similaridade, embora claramente suas funcionalidades são distintas, onde a primeira é do tipo convencional e a segunda do tipo complementar. Enquanto as linhas 24 e 72 (linhas mais frequentes das últimas validações) apresentam 59% de similaridade (no mapa estão sobrepostos parte da extensão). Dessa forma, baseado nesse indicador, o usuário tende a utilizar mais de uma linha, porém para exercer uma única atividade.

Figura 33 - Zoneamento hexagonal com a distribuição das residências dos usuários válidos




Fonte: Autor.

Figura 34 - Reconstrução das viagens baseado nos padrões espaciais e temporais



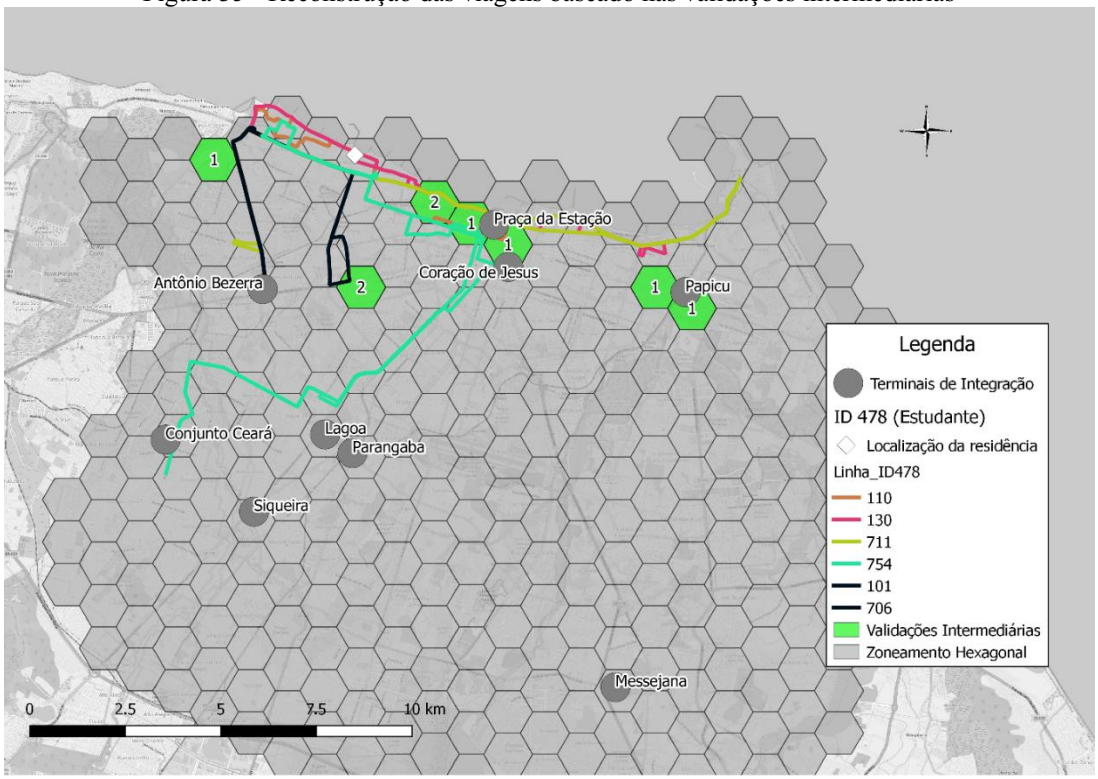

Fonte: Autor.

Para se reconstruir a cadeia completa ainda é necessário avaliar a "perna" intermediária desses trajetos. A Figura 35 apresenta as linhas utilizadas em validações

intermediárias. Na grande maioria foi possível identificar apenas uma validação, sendo a de maior frequência de utilização a linha 754 que liga os terminais da Praça da Estação e Coração de Jesus ao Conjunto Ceará. Estas validações intermediárias indicam alguma atividade esporádica realizada no centro da cidade. Além disso, indicam que em algum dia o usuário vai ao Centro, realiza alguma atividade, e depois segue para o destino da atividade principal que está localizada na zona vermelha de frequência 11, próxima ao Terminal da Lagoa.

Conforme estes resultados, é possível inferir as possíveis zonas onde ele realiza atividades, como as próximas ao local de embarque da última validação, e que ele realiza atividades esporádicas em zonas do centro da cidade. Estas atividades esporádicas são evidenciadas por validações intermediárias e últimas validações que ocorrem próximas a Praças da Estação e Coração de Jesus, no centro da cidade. O mapa também indica que provavelmente o usuário realiza transbordos no Terminal do Antônio Bezerra devido a concentração de validações próximas ao terminal, e devido à sequência de linhas (passando pelo terminal) utilizadas por ele.

Figura 35 - Reconstrução das viagens baseado nas validações intermediárias

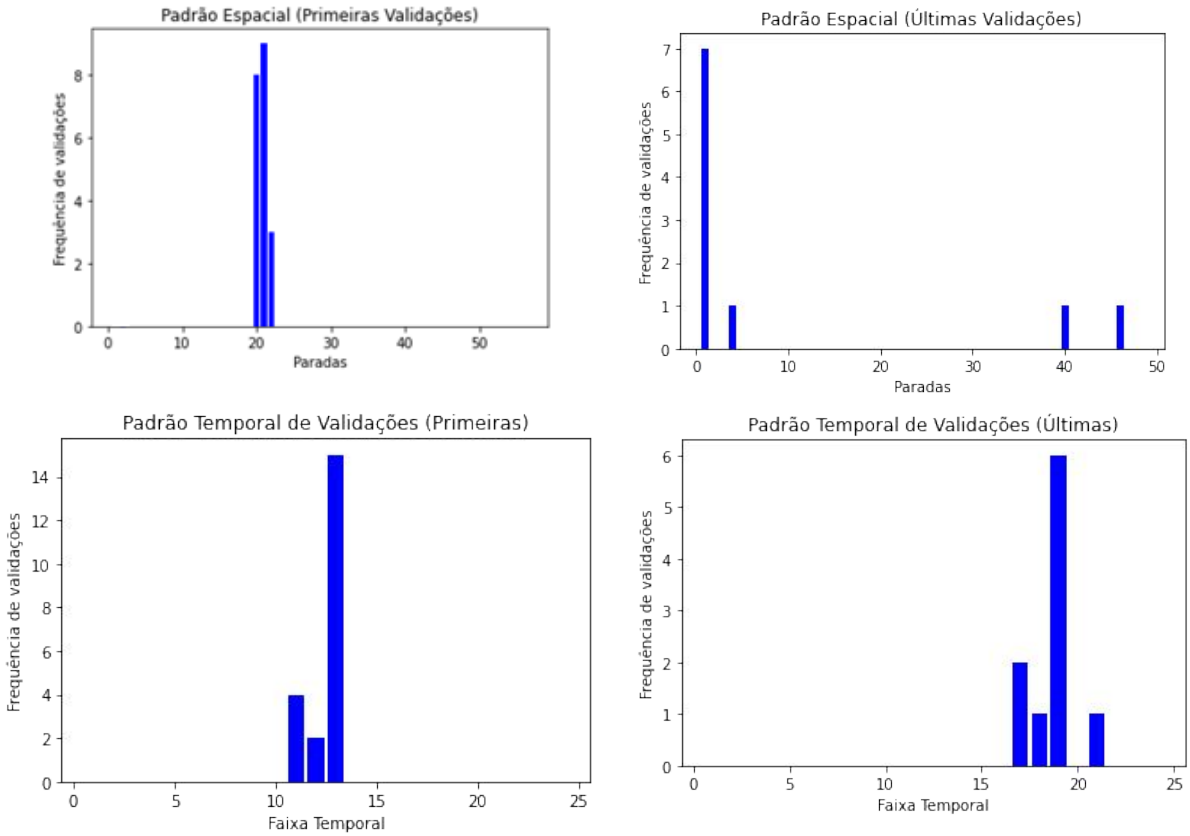



Fonte: Autor.

Embora com as informações iniciais já se tenha uma noção da origem e destino, é necessário identificar e validar o número de atividades e os locais de embarque das linhas

regulares. A Figura 36 apresenta os perfis de validação temporal e espacial para o usuário. Com base no perfil temporal, é visualizado que ele tem apenas um pico significativo de validações temporais, levantando-se a hipótese de que este usuário realiza uma atividade predominante, com uma origem e destino bem delimitadas, porém com viagens esporádicas pouco frequentes. Também é possível verificar que as viagens iniciam em torno de 13h00min e com validação final em torno das 19hh00min, com um indicador de diferença temporal entre a primeira e última validação de 6 horas.

Figura 36 - Distribuições espaciais e temporais das primeiras e últimas validações




Fonte: Autor.

Já com base no perfil espacial, as primeiras validações ao longo do sentido de “ida” indicam que o local de embarque da origem é na parada mais próxima à montante do primeiro local de concentração de validações, indicando então que o usuário tem uma tendência de validar ao embarcar. Portando, a origem da viagem desse usuário é o hexágono onde está localizado a residência. Para as últimas validações o usuário apresenta uma tendência de validar no início da linha logo nas primeiras paradas, sendo considerado uma validação ao

embarcar. Estes resultados indicam que local de embarque é numa zona próxima ao Terminal da Lagoa no início da linha. O usuário, então, parece realizar transbordos no Terminal do Antônio Bezerra, onde embarca em uma das linhas intermediárias (Linhas 130 – Vila do Mar / Náutico / Antônio Bezerra II, 711 – Barra do Ceará / Cais do Porto) para fechar a cadeia de viagem em direção a sua residência.

Diante do exposto neste capítulo, as análises espaciais e temporais agregadas, das primeiras validações e à nível do indivíduo foram as percussoras para definição dos atributos utilizados para segregação dos padrões de uso, conforme será apresentado adiante. As análises de frequência de validação por dia da semana, por faixa horária e por tipos de linhas foram decisivas para definição dos atributos da modelagem. As análises segregadas dos indivíduos ajudaram a compreender inicialmente os padrões de deslocamentos, que ficarão mais claros após as análises individuais dos atributos de cada grupo identificado.

# 7. IDENTIFICAÇÃO DOS PADRÕES DE USO E MODELAGEM

Neste capítulo serão apresentados os resultados para identificação dos padrões, interpretação dos padrões de uso e posteriormente modelagem do local de embarque para cada grupo. Considerou-se nas análises o período de 6 meses de validação da amostra de cadastro válida.

## 7.1 Identificação dos padrões de uso do sistema

Para cada usuário da amostra válida (20,6 mil usuários) foram calculados os atributos na qual acredita-se que apresentam características sobre o padrão de uso do sistema desses usuários. A Tabela 6 apresenta o descritivo de cada atributo para elucidar quais aspectos foram considerados como relevantes para a categorização dos usuários. A Figura 37 ilustra a base de dados dos usuários utilizada para *clusterização*, onde cada coluna representa um atributo. Vale destacar mais uma vez que foi utilizado uma série histórica de 6 meses para cálculo das variáveis.

Tabela 6 - Descrição dos atributos

| Código do Atributo | Descrição | Unidade |
|---|---|---|
| **FREQ_DIA_SEGUNDA** | Frequência média de validações nas segundas-feiras | Número de validações por dia |
| **FREQ_DIA_TERCA** | Frequência média de validações nas terças-feiras | Número de validações por dia |
| **FREQ_DIA_QUARTA** | Frequência média de validações nas quartas-feiras | Número de validações por dia |
| **FREQ_DIA_QUINTA** | Frequência média de validações nas quintas-feiras | Número de validações por dia |
| **FREQ_DIA_SEXTA** | Frequência média de validações nas sextas-feiras | Número de validações por dia |
| **FRQ_VALIDACAO_TERMINAL** | Validações próximas aos terminais | Número de validações por dia |
| **DIST_TEMPO_DIA_SEGUNDA** | Distância temporal media entre as primeiras e últimas validações nas segundas-feiras | Horas |
| **DIST_TEMPO_DIA_TERCA** | Distância temporal media entre as primeiras e últimas validações nas terças-feiras | Horas |
| **DIST_TEMPO_DIA_QUARTA** | Distância temporal media entre as primeiras e últimas validações nas quartas-feiras | Horas |
| **DIST_TEMPO_DIA_QUINTA** | Distância temporal media entre as primeiras e últimas validações nas quintas-feiras | Horas |
| **Código do Atributo** | **Descrição** | **Unidade** |

| | | |
|---|---|---|
| **DIST_TEMPO_DIA_SEXTA** | Distância temporal media entre as primeiras e últimas validações nas sextas-feiras | Horas |
| **FAIXA_HORARIA_FRQ_PRIMEIRA** | Faixa horária de maior frequência média das primeiras validações | Faixa Horária |
| **FAIXA_HORARIA_FRQ_ULTIMA** | Faixa horária de maior frequência média das últimas validações | Faixa Horária |
| **AREA_VALIDACOES** | Área das validações de um usuário durante o período de 6 meses | $Km^2$ |
| **DIST_HORIZONTAL_MAX** | Distância Horizontal Máxima média diária das validações | Km |
| **DIST_VERTICAL_MAX** | Distância Vertical Máxima média diária das validações | Km |
| **PROP_DIARIA_ALIMENTADORA** | Proporção de validações medias diárias por tipo de linha - alimentadora | Proporção de validações por dia |
| **PROP_DIARIA_TRONCAL** | Proporção de validações medias diárias por tipo de linha – Troncal | Proporção de validações por dia |
| **PROP_DIARIA_CONVENCIONAL** | Proporção de validações medias diárias por tipo de linha – Convencional | Proporção de validações por dia |
| **PROP_DIARIA_COMPLEMENTAR** | Proporção de validações medias diárias por tipo de linha - Complementar | Proporção de validações por dia |

Fonte: Autor.

Figura 37 - Dados importados para clusterização

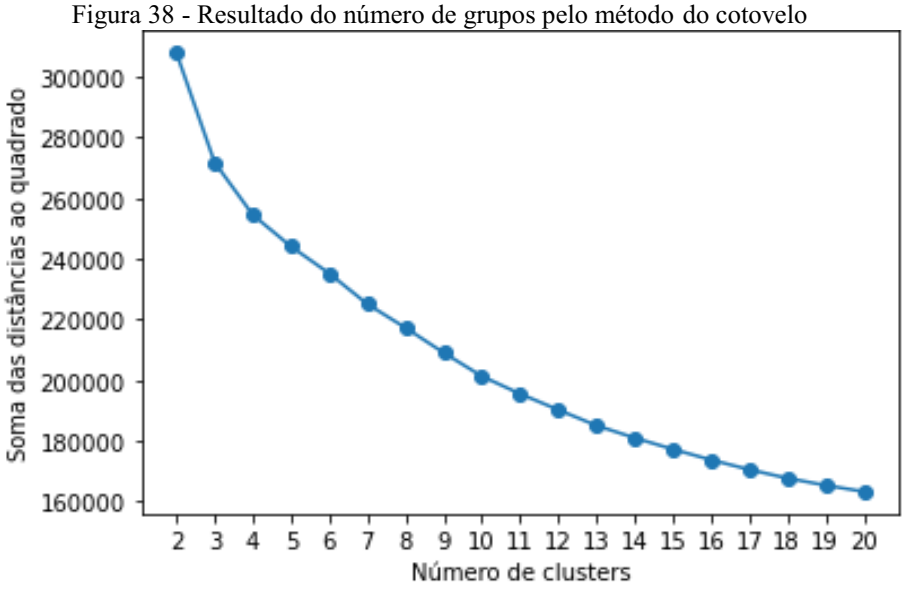
| | FREQ_DIA_SEGUNDA | FREQ_DIA_TERCA | FREQ_DIA_QUARTA | FREQ_DIA_QUINTA | FREQ_DIA_SEXTA | FRQ_VALIDACAO_TERMINAL |
|---|---|---|---|---|---|---|
| 0 | 1.25 | 1.000000 | 0.25 | 0.50 | 0.50 | 0.095238 |
| 1 | 0.50 | 0.000000 | 0.00 | 0.75 | 0.50 | 0.000000 |
| 2 | 0.00 | 0.000000 | 0.25 | 0.25 | 0.00 | 0.023810 |
| 3 | 1.75 | 0.000000 | 0.50 | 0.50 | 0.50 | 0.214286 |
| 4 | 0.00 | 0.000000 | 0.00 | 0.75 | 1.00 | 0.285714 |
| ... | ... | ... | ... | ... | ... | ... |

Fonte: Autor.

O método do cotovelo (Figura 38) e o score de silhueta (Figura 39) indicaram que existência de 3 ou 4 padrões nos dados. Após uma análise dos atributos aplicando o método *k-means* para com 3 e 4 grupos, foi possível visualizar com 4 grupos usuários mais regulares no

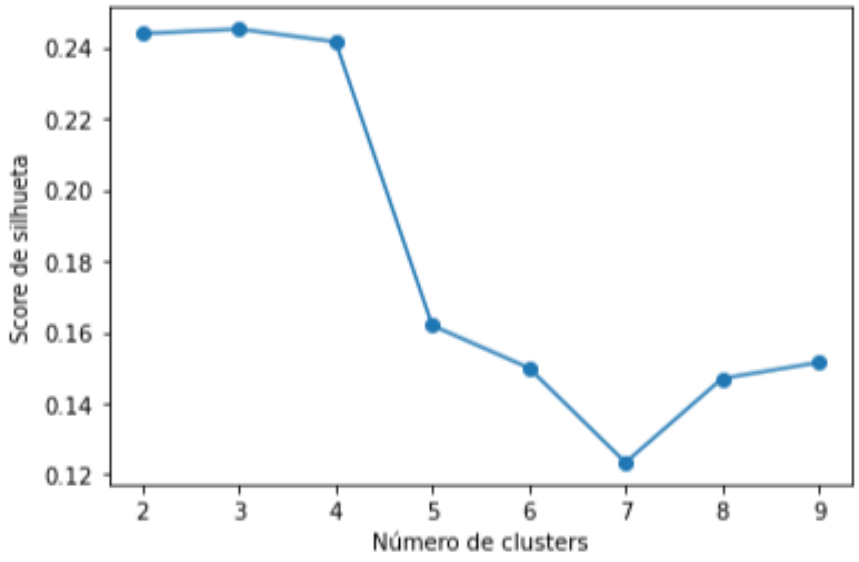
sistema e com maior distância temporal, e grupos de usuários menos frequentes com distâncias temporais reduzidas, assim como apresentado de forma agregada. Dessa forma, o agrupamento dos usuários foi realizado em 4 clusters. O método *k-means*++ foi então aplicado para gerar 4 grupos.

Figura 38 - Resultado do número de grupos pelo método do cotovelo




Fonte: Autor.

Figura 39 - Resultado do número de grupos pelo método da silhueta.




Fonte: Autor.

A Tabela 7 apresenta o score de importância de cada atributo na formação dos clusters. Nota-se que 36% da importância na formação dos grupos é referente a aspectos da tipologia das linhas. Assim, as diferentes funcionalidades das linhas na rede do SIT-FOR estão bem associadas aos diferentes padrões gerados. Nota-se também uma considerável relevância de aspectos espaciais (relacionados ao espaço de atividades) e aspectos relacionados aos horários das atividades (faixas horárias das primeira e últimas validações) na formação dos padrões de validação. Considerando os atributos de frequência diária e a área de validações, a importância geral não ultrapassa 4%. Os resultados indicam então que os horários das atividades, os tipos de atividades, e as suas distâncias impactaram mais na separação dos grupos do que o número de atividades.

Tabela 7 - Score de importância das variáveis para o agrupamento

| ID | Atributos | Score de Importância das Variáveis | Peso da Importância (%) |
|---|---|---|---|
| 20 | PROP_DIARIA_COMPLEMENTAR | 1203,7969 | 14,13% |
| 19 | PROP_DIARIA_CONVENCIONAL | 1063,9781 | 12,49% |
| 17 | PROP_DIARIA_ALIMENTADORA | 968,7161 | 11,37% |
| 16 | DIST_VERTICAL_MAX | 795,4570 | 9,34% |
| 13 | FAIXA_HORARIA_FRQ_ULTIMA | 640,0862 | 7,51% |
| 12 | FAIXA_HORARIA_FRQ_PRIMEIRA | 541,7824 | 6,36% |
| 18 | PROP_DIARIA_TRONCAL | 523,0156 | 6,14% |
| 15 | DIST_HORIZONTAL_MAX | 503,4942 | 5,91% |
| 7 | DIST_TEMPO_DIA_SEGUNDA | 418,2722 | 4,91% |
| 9 | DIST_TEMPO_DIA_QUARTA | 379,2931 | 4,45% |
| 8 | DIST_TEMPO_DIA_TERCA | 371,9453 | 4,37% |
| 11 | DIST_TEMPO_DIA_SEXTA | 345,0226 | 4,05% |
| 10 | DIST_TEMPO_DIA_QUINTA | 318,1097 | 3,73% |
| 6 | FRQ_VALIDACAO_TERMINAL | 113,5409 | 1,33% |
| 5 | FREQ_DIA_SEXTA | 70,6752 | 0,83% |
| 3 | FREQ_DIA_QUARTA | 64,9941 | 0,76% |
| 1 | FREQ_DIA_SEGUNDA | 63,8241 | 0,75% |
| 2 | FREQ_DIA_TERCA | 58,4298 | 0,69% |
| 4 | FREQ_DIA_QUINTA | 58,4298 | 0,69% |
| 14 | AREA_VALIDACOES | 16,9233 | 0,20% |

Fonte: Autor.

## 7.2. Interpretação dos grupos

Os quatro grupos gerados foram numerados de 0 a 3. As proporções de usuários em cada grupo foram de 15,8%, 26,2%, 20,7% e 37,3%, respectivamente.

### *7.2.1. Análise dos componentes principais*

Para visualização dos dados foi realizado uma redução de dimensionalidade com o método do PCA. A Figura 40 apresenta o gráfico relacionando o primeiro e o segundo componente principal dos dados, assim como as direções de maior variação positiva dos atributos relacionados aos componentes principais 1 (CP 1) e 2 (CP 2), respectivamente. Atributos com vetores na mesma direção são perfeitamente correlacionados, afetando a separação dos grupos da mesma forma. Pode-se destacar o efeito de alguns atributos em cada grupo. O Atributo 19 (proporção de validação em linha convencional), por exemplo, apresenta maior efeito no Grupo 0. O Atributo 10 (relacionado a distância temporal), apresenta maior contribuição no Grupo 1. Da mesma forma, os atributos 20 (proporção de validação por linha complementar) e 12 (faixa horária da primeira validação) tem uma maior intensidade nos grupos 2 e 3, respectivamente.

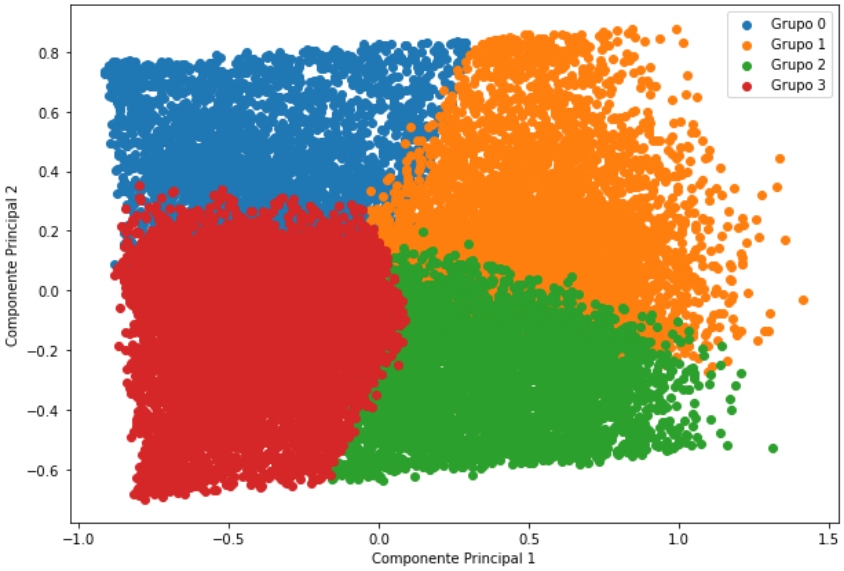
O principal aspecto evidenciado pelo componente 1 nas características dos grupos são aspectos temporais de uso do sistema, dado pelos maiores vetores (atributos 10 e 12), enquanto a segunda maior influência no aspecto de uso destes usuários são características operacionais das linhas, como a sua tipologia, que configura em como o usuário poderá se deslocar na rede. Estes resultados também são ilustrados nas Tabela 8 e Tabela 9 que apresentam os autovetores para os 6 primeiros componentes principais que agregam 87% da variância nos dados, conforme apresentado na Figura 41. Conforme a Tabela 8, a variância do CP 1 é realmente mais influenciada pelos atributos 10 e 12, enquanto a o CP 2 é mais afetado pelos atributos 19 e 20.

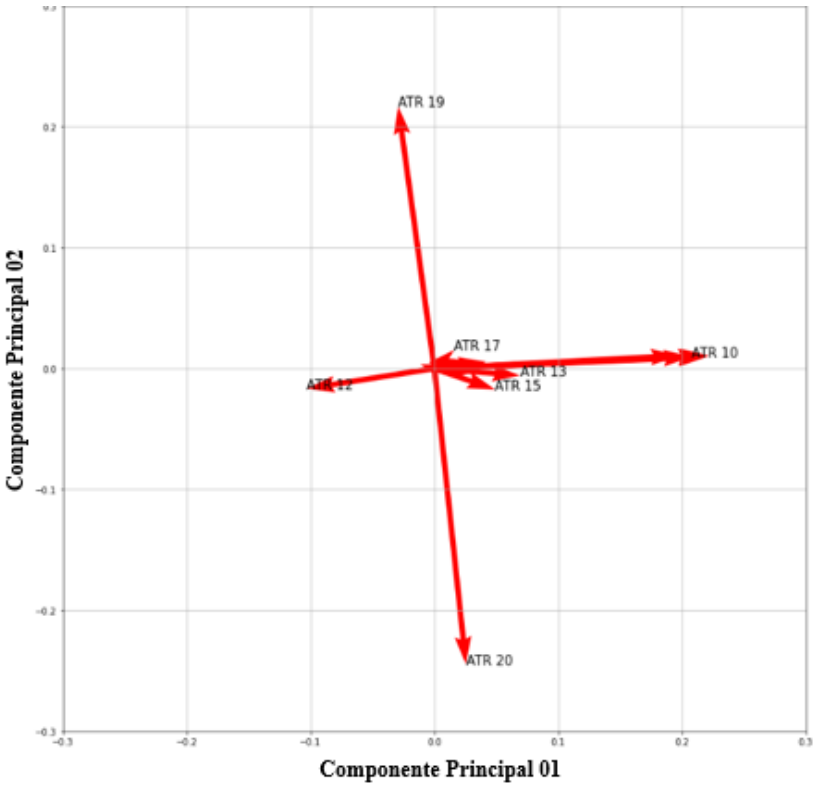
O Avaliando o terceiro componente principal (CP3), que já explica 68% da variância junto aos outros 2 componentes, apresenta uma maior influência de linhas alimentadoras. O CP 4 é mais afetado por atributos espaciais de distância dos deslocamentos. O CP 5 tem sua variância mais afetada pelos atributos relacionados aos horários das atividades,

principalmente aos horários de término (faixa horária da última validação). Por fim, o CP 6 é mais afetado pelo uso de linhas troncais nos deslocamentos.

Todos estes resultados indicam que os atributos adotados influenciam de forma diferente na variação dos dados e que, portanto, podem apoia a geração de grupos distintos.

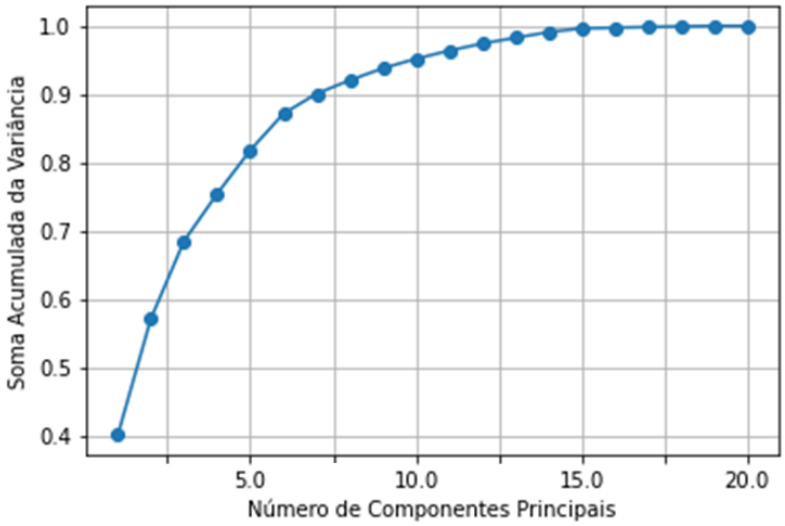
Figura 40 – K-means com PCA e mapa perceptual dos atributos




Fonte: Autor.

Figura 41 - Soma Acumulada da Variância dos componentes principais




Fonte: Autor.

Tabela 8 - Componentes Principais dos Atributos (1-3)

| ID | Atributos | Componente Principal 1 | Componente Principal 2 | Componente Principal 3 |
|---|---|---|---|---|
| 1 | FREQ_DIA_SEGUNDA | 0,030 | 0,006 | -0,006 |
| 2 | FREQ_DIA_TERCA | 0,028 | 0,006 | -0,006 |
| 3 | FREQ_DIA_QUARTA | 0,030 | 0,006 | -0,006 |
| 4 | FREQ_DIA_QUINTA | 0,028 | 0,006 | -0,006 |
| 5 | FREQ_DIA_SEXTA | 0,031 | 0,006 | -0,006 |
| 6 | FRQ_VALIDACAO_TERMINAL | 0,042 | 0,006 | -0,005 |
| 7 | DIST_TEMPO_DIA_SEGUNDA | **0,222** | 0,011 | -0,007 |
| 8 | DIST_TEMPO_DIA_TERCA | **0,219** | 0,011 | -0,007 |
| 9 | DIST_TEMPO_DIA_QUARTA | **0,220** | 0,010 | -0,005 |
| 10 | DIST_TEMPO_DIA_QUINTA | **0,208** | 0,010 | -0,004 |
| 11 | DIST_TEMPO_DIA_SEXTA | **0,197** | 0,011 | -0,004 |
| 12 | FAIXA_HORARIA_FRQ_PRIMEIRA | **-0,103** | -0,017 | -0,003 |
| 13 | FAIXA_HORARIA_FRQ_ULTIMA | 0,069 | -0,006 | -0,008 |
| 14 | AREA_VALIDACOES | 0,007 | 0,000 | 0,003 |
| 15 | DIST_HORIZONTAL_MAX | 0,049 | -0,017 | 0,024 |
| 16 | DIST_VERTICAL_MAX | 0,038 | -0,002 | 0,045 |
| 17 | PROP_DIARIA_ALIMENTADORA | 0,016 | 0,015 | **0,208** |
| 18 | PROP_DIARIA_TRONCAL | -0,011 | 0,003 | 0,015 |
| 19 | PROP_DIARIA_CONVENCIONAL | -0,029 | **0,217** | -0,124 |
| 20 | PROP_DIARIA_COMPLEMENTAR | 0,026 | **-0,245** | -0,101 |

Fonte: Autor.

Tabela 9 - Componentes Principais dos Atributos (4-6)

| ID | Atributos | Componente Principal 4 | Componente Principal 5 | Componente Principal 6 |
|---|---|---|---|---|
| 1 | FREQ_DIA_SEGUNDA | 0,014 | 0,003 | 0,007 |
| 2 | FREQ_DIA_TERCA | 0,014 | 0,002 | 0,007 |
| 3 | FREQ_DIA_QUARTA | 0,014 | 0,003 | 0,007 |
| 4 | FREQ_DIA_QUINTA | 0,014 | 0,002 | 0,007 |
| 5 | FREQ_DIA_SEXTA | 0,014 | 0,003 | 0,007 |
| 6 | FRQ_VALIDACAO_TERMINAL | 0,003 | -0,002 | 0,011 |
| 7 | DIST_TEMPO_DIA_SEGUNDA | -0,010 | -0,007 | 0,000 |
| 8 | DIST_TEMPO_DIA_TERCA | -0,011 | -0,007 | 0,000 |
| 9 | DIST_TEMPO_DIA_QUARTA | -0,008 | -0,006 | 0,000 |
| 10 | DIST_TEMPO_DIA_QUINTA | -0,008 | -0,007 | 0,000 |
| 11 | DIST_TEMPO_DIA_SEXTA | -0,007 | -0,005 | -0,001 |
| 12 | FAIXA_HORARIA_FRQ_PRIMEIRA | 0,069 | **-0,108** | -0,012 |
| 13 | FAIXA_HORARIA_FRQ_ULTIMA | 0,071 | **-0,139** | -0,016 |
| 14 | AREA_VALIDACOES | 0,011 | 0,005 | 0,001 |
| 15 | DIST_HORIZONTAL_MAX | **0,082** | 0,037 | 0,000 |
| 16 | DIST_VERTICAL_MAX | **0,157** | 0,079 | 0,030 |
| 17 | PROP_DIARIA_ALIMENTADORA | -0,019 | -0,006 | -0,070 |
| 18 | PROP_DIARIA_TRONCAL | -0,021 | -0,030 | **0,151** |
| 19 | PROP_DIARIA_CONVENCIONAL | 0,025 | 0,017 | -0,052 |
| 20 | PROP_DIARIA_COMPLEMENTAR | 0,008 | 0,021 | -0,046 |

Fonte: Autor.

### *7.2.2. Análise dos atributos em cada grupo*

As Tabela 10 a Tabela 13 apresentam os resumos dos atributos para cada grupo de usuários identificados. A Figura 42 apresenta uma representação espacial do espaço médio de validações de cada grupo. A partir estes resultados, pode-se interpretar os grupos, como será discutido a seguir.

O **Grupo 0** é formado por usuários que não tem um padrão regular de validação ao longo dia, com usuários realizando atividade em diferentes horários (faixas de maior frequência 10-11hrs e 13-14hrs, um elevado desvios padrão de 4 horas). As atividades realizadas por estes usuários podem incluir atividades de media duração, como atividade de estudo, considerando a média em torno de 3 horas de distância temporal, até atividade bem

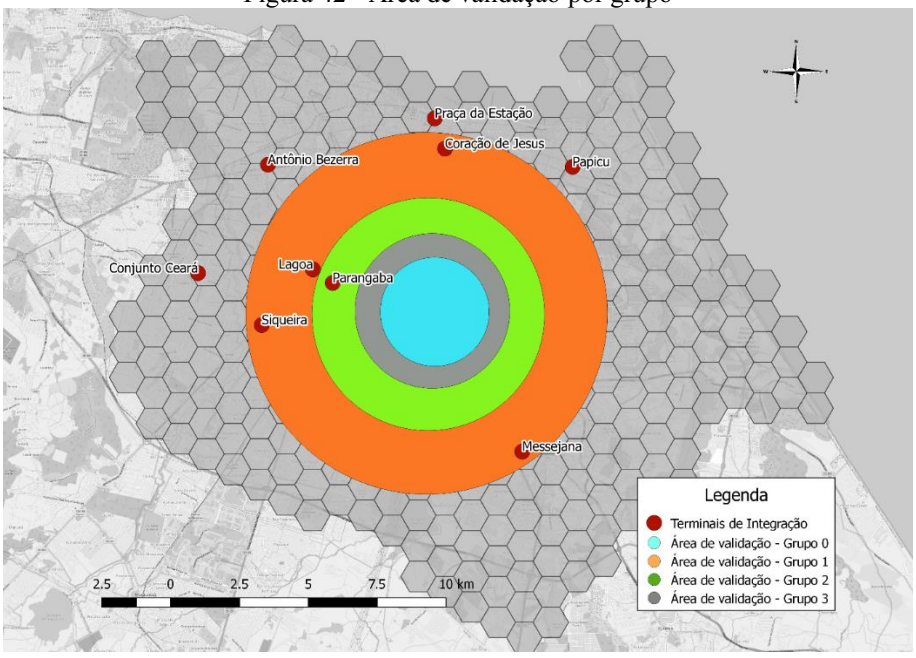
curtas com menos de duas horas. Por fim, avaliando a frequência de uso dos tipos de linha, o tipo mais utilizado são as linhas convencionais que interligam bairros periféricos diretamente ao centro comercial. Este grupo aparenta ser de usuários não muito regulares que têm atividades de curta e média duração, e com deslocamentos mais curtos, comparado aos outros grupos (Figura 42).

Tabela 10 - Resumo dos indicadores - Grupo 0

| Atributos | Média | Mediana | Desvio Padrão |
|---|---|---|---|
| FREQ_DIA_SEGUNDA | 1,94 | 1,90 | 0,81 |
| FREQ_DIA_TERCA | 1,98 | 1,93 | 0,75 |
| FREQ_DIA_QUARTA | 1,95 | 1,93 | 0,80 |
| FREQ_DIA_QUINTA | 1,98 | 1,93 | 0,75 |
| FREQ_DIA_SEXTA | 1,92 | 1,89 | 0,81 |
| FRQ_VALIDACAO_TERMINAL | 0,20 | 0,17 | 0,17 |
| DIST_TEMPO_DIA_SEGUNDA | 2,99 | 2,69 | 2,22 |
| DIST_TEMPO_DIA_TERCA | 3,09 | 2,76 | 2,26 |
| DIST_TEMPO_DIA_QUARTA | 3,06 | 2,80 | 2,23 |
| DIST_TEMPO_DIA_QUINTA | 3,12 | 2,84 | 2,20 |
| DIST_TEMPO_DIA_SEXTA | 2,97 | 2,64 | 2,23 |
| FAIXA_HORARIA_FRQ_PRIMEIRA | 10,45 | 9,00 | 4,20 |
| FAIXA_HORARIA_FRQ_ULTIMA | 13,73 | 14,00 | 4,42 |
| AREA_VALIDACOES | 3,55 | 1,05 | 6,53 |
| DIST_HORIZONTAL_MAX | 4,97 | 4,64 | 3,13 |
| DIST_VERTICAL_MAX | 3,69 | 2,89 | 2,95 |
| PROP_DIARIA_ALIMENTADORA | 0,05 | 0,01 | 0,10 |
| PROP_DIARIA_TRONCAL | 0,04 | 0,00 | 0,10 |
| PROP_DIARIA_CONVENCIONAL | 0,65 | 0,62 | 0,20 |
| PROP_DIARIA_COMPLEMENTAR | 0,16 | 0,14 | 0,13 |
|  | 1,53 | 0,24 | 2,80 |

Fonte: Autor.

Figura 42 - Área de validação por grupo




Fonte: Autor.

O **Grupo 1** parece ser formado por usuários com uso regular do sistema (maiores frequências de validação diária, sendo maior do que 2,6) e que usam o sistema para acessar diferentes atividades, incluindo trabalho (distância temporal média maior do que 9 horas e desvio padrão maior do que 2 horas). Outro indício de que seja um grupo regular de uso do sistema é que os horários de maior validação (faixas horárias de maior frequência estão entre 6-7hrs e 16-17hrs) correspondem aos horários de pico do dia. Muitos destes usuários parecem estar distantes das atividades, na região periférica, conforme indicam os atributos dos espaço de validações, e parecem realizar integrações em terminais físicos, conforme a proporção de validações em linhas alimentadoras e complementares neste grupo.

Tabela 11 - Resumo dos indicadores - Grupo 1

| Atributos | Média | Mediana | Desvio Padrão |
|---|---|---|---|
| FREQ_DIA_SEGUNDA | 2,63 | 2,29 | 0,88 |
| FREQ_DIA_TERCA | 2,64 | 2,32 | 0,88 |
| FREQ_DIA_QUARTA | 2,65 | 2,32 | 0,88 |
| FREQ_DIA_QUINTA | 2,64 | 2,32 | 0,88 |
| FREQ_DIA_SEXTA | 2,62 | 2,29 | 0,86 |
| FRQ_VALIDACAO_TERMINAL | 0,41 | 0,38 | 0,25 |
| DIST_TEMPO_DIA_SEGUNDA | 9,11 | 9,42 | 2,21 |
| DIST_TEMPO_DIA_TERCA | 9,37 | 9,59 | 2,10 |
| DIST_TEMPO_DIA_QUARTA | 9,26 | 9,46 | 2,04 |
| DIST_TEMPO_DIA_QUINTA | 9,29 | 9,47 | 2,03 |
| DIST_TEMPO_DIA_SEXTA | 8,86 | 9,05 | 2,14 |
| FAIXA_HORARIA_FRQ_PRIMEIRA | 6,74 | 6,00 | 2,39 |
| FAIXA_HORARIA_FRQ_ULTIMA | 16,80 | 17,00 | 3,25 |
| AREA_VALIDACOES | 11,08 | 3,90 | 22,03 |
| DIST_HORIZONTAL_MAX | 7,31 | 7,11 | 3,56 |
| DIST_VERTICAL_MAX | 5,14 | 4,12 | 3,41 |
| PROP_DIARIA_ALIMENTADORA | 0,25 | 0,15 | 0,26 |
| PROP_DIARIA_TRONCAL | 0,10 | 0,01 | 0,19 |
| PROP_DIARIA_CONVENCIONAL | 0,28 | 0,21 | 0,27 |
| PROP_DIARIA_COMPLEMENTAR | 0,25 | 0,26 | 0,16 |

Fonte: Autor.

O **Grupo 2** apresenta características similares **Grupo 1**, contudo com uma tendência de mais pendular (com frequência diária de validação pouco acima de 2,0 e com menor variação, distância temporal com média em torno de 8 hores). Possivelmente é formado por usuários que usam o sistema predominantemente para acessar atividade de trabalho. Muitos usuários deste grupo estão também mais próximos das atividades do que os do **Grupo 1**, como mostra a Figura 42, e assim acessam o sistema do SIT-FOR um pouco mais tarde do que os do Grupo 1 (faixa horária da primeira validação em torno de 7-8hrs). Estes usuários

apresentam um maior de linhas complementares, indicando um maior deslocamento aos terminais, principalmente de regiões mais distantes.

Tabela 12 - Resumo dos indicadores - Grupo 2

| Atributos | Média | Mediana | Desvio Padrão |
|---|---|---|---|
| FREQ_DIA_SEGUNDA | 2,34 | 2,13 | 0,70 |
| FREQ_DIA_TERCA | 2,37 | 2,14 | 0,71 |
| FREQ_DIA_QUARTA | 2,37 | 2,13 | 0,70 |
| FREQ_DIA_QUINTA | 2,37 | 2,14 | 0,71 |
| FREQ_DIA_SEXTA | 2,32 | 2,12 | 0,69 |
| FRQ_VALIDACAO_TERMINAL | 0,34 | 0,31 | 0,21 |
| DIST_TEMPO_DIA_SEGUNDA | 7,88 | 7,98 | 2,44 |
| DIST_TEMPO_DIA_TERCA | 8,20 | 8,24 | 2,32 |
| DIST_TEMPO_DIA_QUARTA | 8,08 | 8,13 | 2,26 |
| DIST_TEMPO_DIA_QUINTA | 8,11 | 8,11 | 2,27 |
| DIST_TEMPO_DIA_SEXTA | 7,61 | 7,67 | 2,35 |
| FAIXA_HORARIA_FRQ_PRIMEIRA | 7,62 | 7,00 | 3,08 |
| FAIXA_HORARIA_FRQ_ULTIMA | 16,51 | 17,00 | 3,72 |
| AREA_VALIDACOES | 7,59 | 2,98 | 13,76 |
| DIST_HORIZONTAL_MAX | 7,11 | 6,95 | 3,42 |
| DIST_VERTICAL_MAX | 4,34 | 3,75 | 2,74 |
| PROP_DIARIA_ALIMENTADORA | 0,07 | 0,01 | 0,11 |
| PROP_DIARIA_TRONCAL | 0,05 | 0,01 | 0,09 |
| PROP_DIARIA_CONVENCIONAL | 0,09 | 0,05 | 0,11 |
| PROP_DIARIA_COMPLEMENTAR | 0,71 | 0,70 | 0,17 |

Fonte: Autor.

Por fim, o **Grupo 3** é provavelmente formado por usuários que acessam o sistema para realizar atividades esporádicas (grupo com menor frequência de validações, abaixo de 2,0, e menores distâncias temporais, com média abaixo de 3 horas). Assim como os usuários do **Grupo 0**, é também constituído por usuários que utilizam o sistema em diferentes horários do dia. Contudo, a área de validações do **Grupo 3** é superior em 42% a do **Grupo 0**, e as linhas mais utilizadas no **Grupo 3** são do tipo complementar, indicando que neste grupo, diferente do **Grupo 0**, existe maior tendência de realizar integrações em terminais físicos.

A Figura 42 indica, portanto, que os grupos com maior regularidade de uso do sistema apresentam uma maior área de validação, indicando uma relação entre regularidade de uso e a localização espacial das atividades e residências. A Figura 43 à Figura 45 e Tabela 14, resumem comparativamente os aspectos discutidos sobre os grupos. Conforme apresentado, os grupos com maiores frequências de validação e distâncias temporais são os grupos 1 e 2, sendo os mais regulares. O grupo 0 apresentou uma maior predisposição em utilizar as linhas convencionais que interligam diretamente os bairros aos centros, enquanto o grupo 2

apresentou uma maior predisposição em utilizar linhas complementares que passam por terminais.

Tabela 13 - Resumo dos indicadores - Grupo 3

| Atributos | Média | Mediana | Desvio Padrão |
|---|---|---|---|
| FREQ_DIA_SEGUNDA | 1,89 | 1,85 | 0,78 |
| FREQ_DIA_TERCA | 1,96 | 1,89 | 0,72 |
| FREQ_DIA_QUARTA | 1,93 | 1,88 | 0,78 |
| FREQ_DIA_QUINTA | 1,96 | 1,89 | 0,72 |
| FREQ_DIA_SEXTA | 1,91 | 1,86 | 0,78 |
| FRQ_VALIDACAO_TERMINAL | 0,19 | 0,17 | 0,15 |
| DIST_TEMPO_DIA_SEGUNDA | 2,34 | 2,01 | 1,87 |
| DIST_TEMPO_DIA_TERCA | 2,42 | 2,11 | 1,88 |
| DIST_TEMPO_DIA_QUARTA | 2,49 | 2,20 | 1,90 |
| DIST_TEMPO_DIA_QUINTA | 2,48 | 2,22 | 1,83 |
| DIST_TEMPO_DIA_SEXTA | 2,41 | 2,06 | 1,94 |
| FAIXA_HORARIA_FRQ_PRIMEIRA | 11,53 | 11,00 | 4,43 |
| FAIXA_HORARIA_FRQ_ULTIMA | 13,85 | 15,00 | 4,50 |
| AREA_VALIDACOES | 5,05 | 1,78 | 8,80 |
| DIST_HORIZONTAL_MAX | 6,09 | 5,89 | 3,58 |
| DIST_VERTICAL_MAX | 4,36 | 3,51 | 3,26 |
| PROP_DIARIA_ALIMENTADORA | 0,18 | 0,06 | 0,24 |
| PROP_DIARIA_TRONCAL | 0,11 | 0,03 | 0,18 |
| PROP_DIARIA_CONVENCIONAL | 0,13 | 0,10 | 0,12 |
| PROP_DIARIA_COMPLEMENTAR | 0,45 | 0,45 | 0,25 |

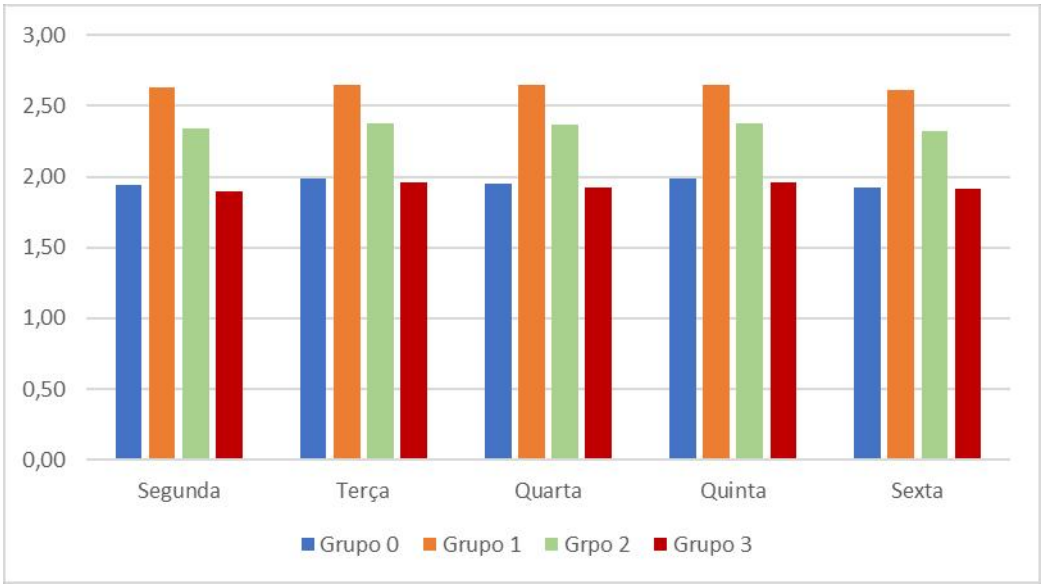
Fonte: Autor.

Tabela 14 - Resumo comparativo dos indicadores para os grupos de uso do sistema de transporte público

| Atributos | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 |
|---|---|---|---|---|
| FREQ_DIA_SEGUNDA | 1,94 | 2,63 | 2,34 | 1,89 |
| FREQ_DIA_TERCA | 1,98 | 2,64 | 2,37 | 1,96 |
| FREQ_DIA_QUARTA | 1,95 | 2,65 | 2,37 | 1,93 |
| FREQ_DIA_QUINTA | 1,98 | 2,64 | 2,37 | 1,96 |
| FREQ_DIA_SEXTA | 1,92 | 2,62 | 2,32 | 1,91 |
| FRQ_VALIDACAO_TERMINAL | 0,20 | 0,41 | 0,34 | 0,19 |
| DIST_TEMPO_DIA_SEGUNDA | 2,99 | 9,11 | 7,88 | 2,34 |
| DIST_TEMPO_DIA_TERCA | 3,09 | 9,37 | 8,20 | 2,42 |
| DIST_TEMPO_DIA_QUARTA | 3,06 | 9,26 | 8,08 | 2,49 |
| DIST_TEMPO_DIA_QUINTA | 3,12 | 9,29 | 8,11 | 2,48 |
| DIST_TEMPO_DIA_SEXTA | 2,97 | 8,86 | 7,61 | 2,41 |
| FAIXA_HORARIA_FRQ_PRIMEIRA | 10,45 | 6,74 | 7,62 | 11,53 |
| FAIXA_HORARIA_FRQ_ULTIMA | 13,73 | 16,80 | 16,51 | 13,85 |
| AREA_VALIDACOES | 3,55 | 11,08 | 7,59 | 5,05 |
| DIST_HORIZONTAL_MAX | 4,97 | 7,31 | 7,11 | 6,09 |
| DIST_VERTICAL_MAX | 3,69 | 5,14 | 4,34 | 4,36 |
| PROP_DIARIA_ALIMENTADORA | 0,05 | 0,25 | 0,07 | 0,18 |
| PROP_DIARIA_TRONCAL | 0,04 | 0,10 | 0,05 | 0,11 |
| PROP_DIARIA_CONVENCIONAL | 0,65 | 0,28 | 0,09 | 0,13 |
| PROP_DIARIA_COMPLEMENTAR | 0,16 | 0,25 | 0,71 | 0,45 |

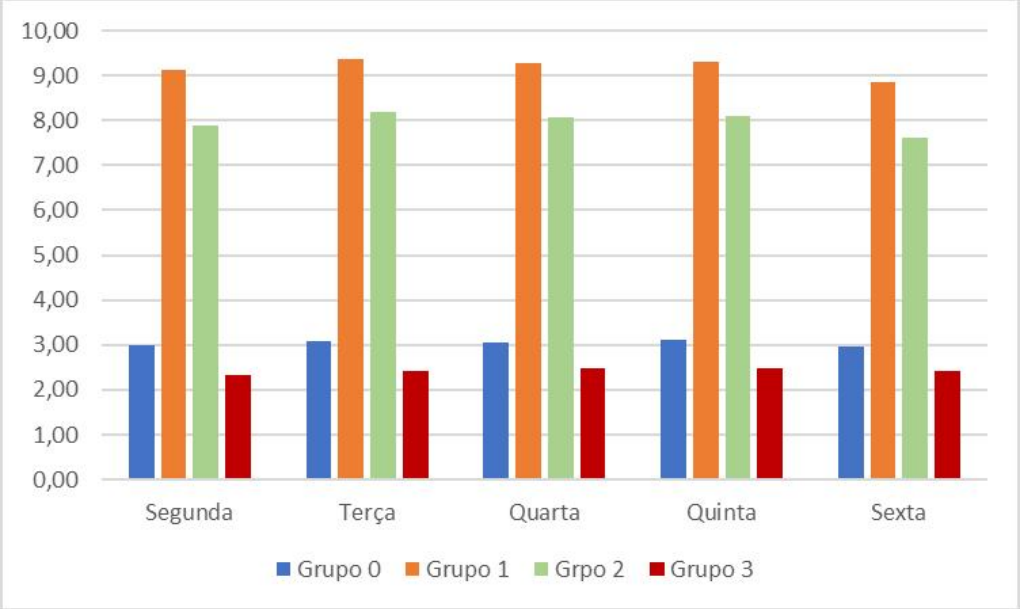
Fonte: Autor

Figura 43 - Frequência de validações por dia



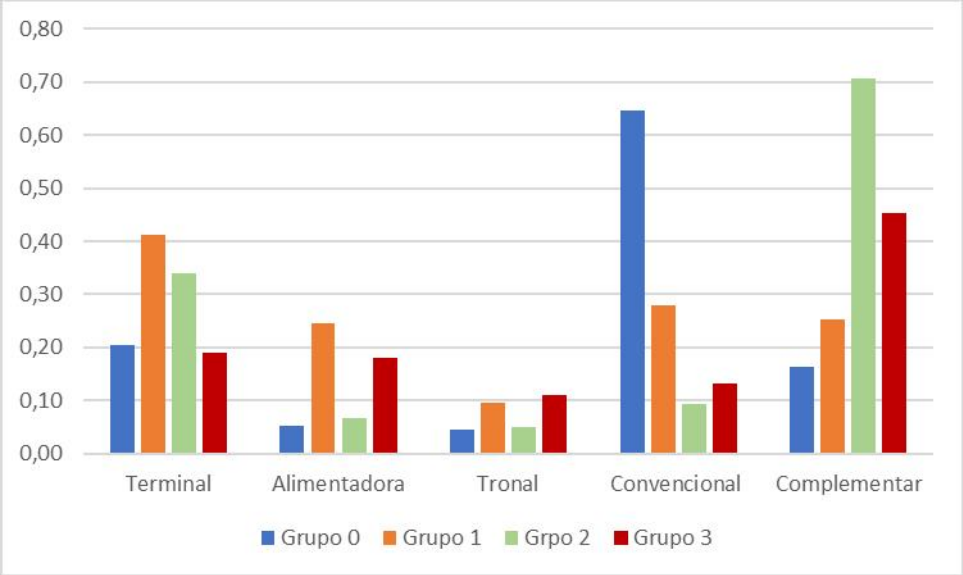

Fonte: Autor.

Figura 44 - Distância temporal de validações por dia




Fonte: Autor.

Figura 45 - Frequência de validações por tipo de linha

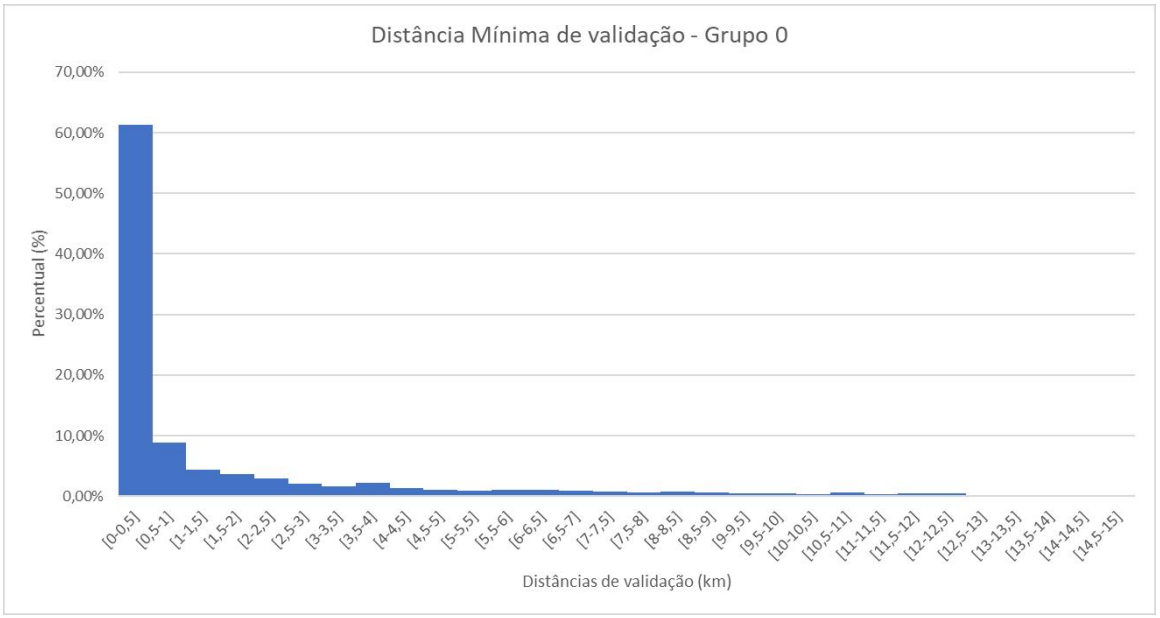



Fonte: Autor.

### *7.2.3. Análise das distâncias de validação em cada grupo*

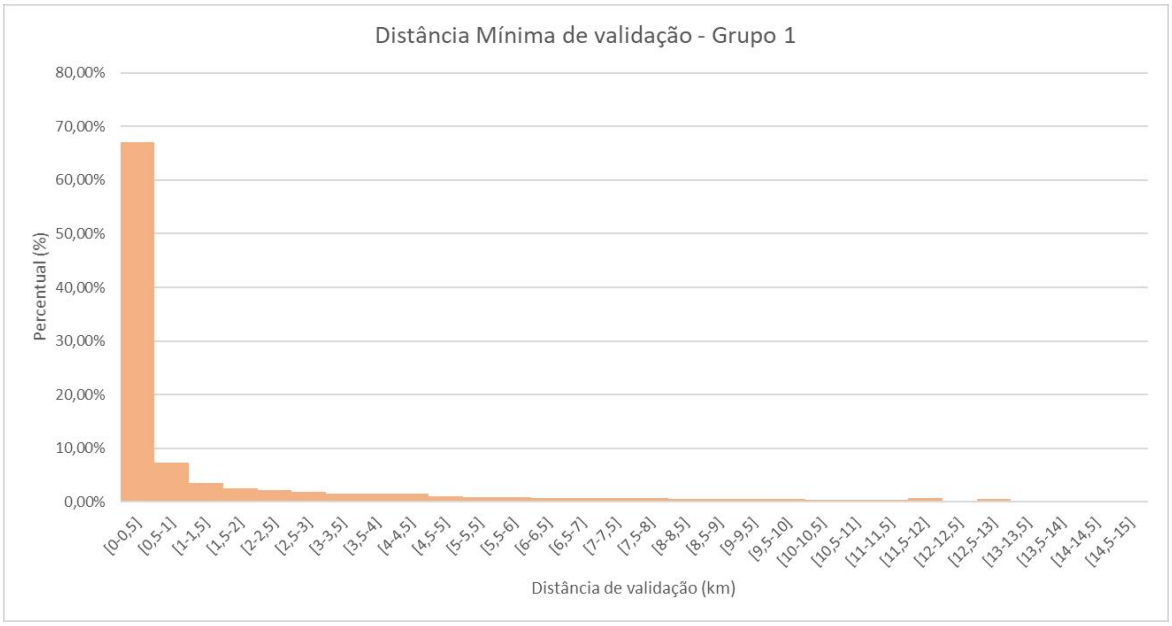
Para cada grupo, analisou-se as distâncias mínimas de validação (distância mínima entre o local de embarque e o local de validação) para o período de análise. As Figura 46 à Figura 49 apresentam os histogramas desta variável para cada grupo e a Tabela 15 apresenta o resumo da média, mediana e desvio padrão de cada grupo. Conforme os resultados, tem-se as seguintes proporções de validações com distâncias mínimas abaixo de 1000 m: 70,1%, 74,4%, 68,5% e 73,6%, para os grupos 0 a 3, respectivamente. Estes resultados indicam que os usuários do Grupo 1 tem uma maior tendência de validar ao embarcar, seguido pelo grupo 3. O Grupo 2 parece ser um grupo com uma tendência de validar em locais mais distantes do local de embarque. No geral, algo em torno de 58% dos usuários tendem a validar no momento do embarque, considerando uma distância mínima de até 400 m. Nota-se também que uma proporção expressiva (20%) de validações com distância mínima acima de 3 km, indicando validações ocorrendo no local do destino. Este padrão heterogêneo dos locais de validação impõe assim um desafio para identificar os reais locais de embarque dos usuários na rede. Contudo, acredita-se ser possível diferenciar entre usuários que validam ao embarcar ou não, e que os padrões de uso do sistema podem apoiar na modelagem dessa probabilidade de validar ao embarcar, conforme será visto na próxima seção.

Figura 46 - Histograma da distância de validação - grupo 0



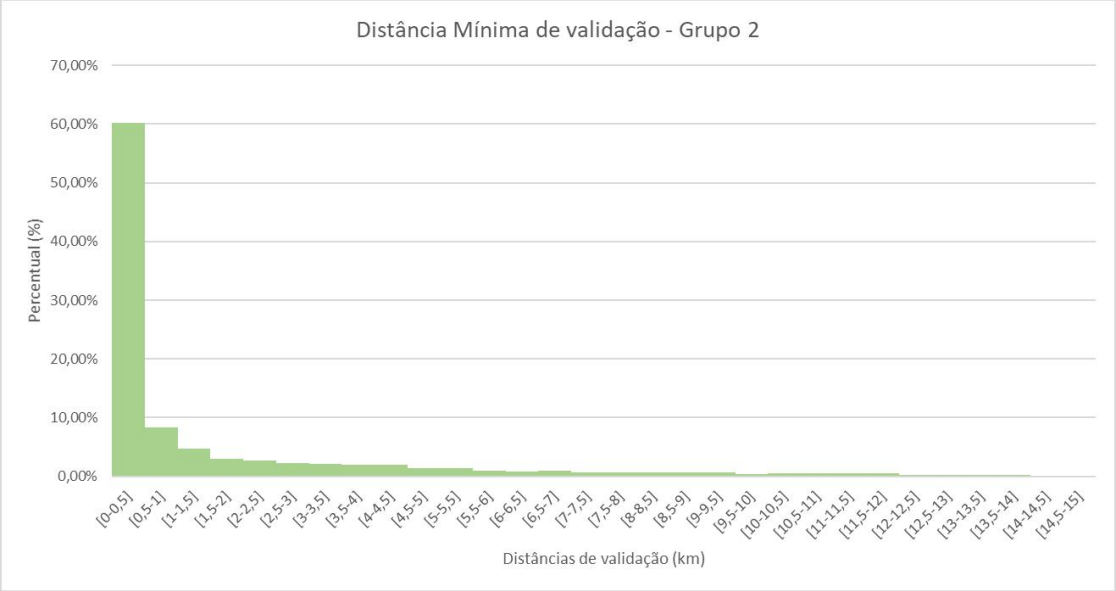

Fonte: Autor.

Figura 47 - Histograma da distância de validação - grupo 1



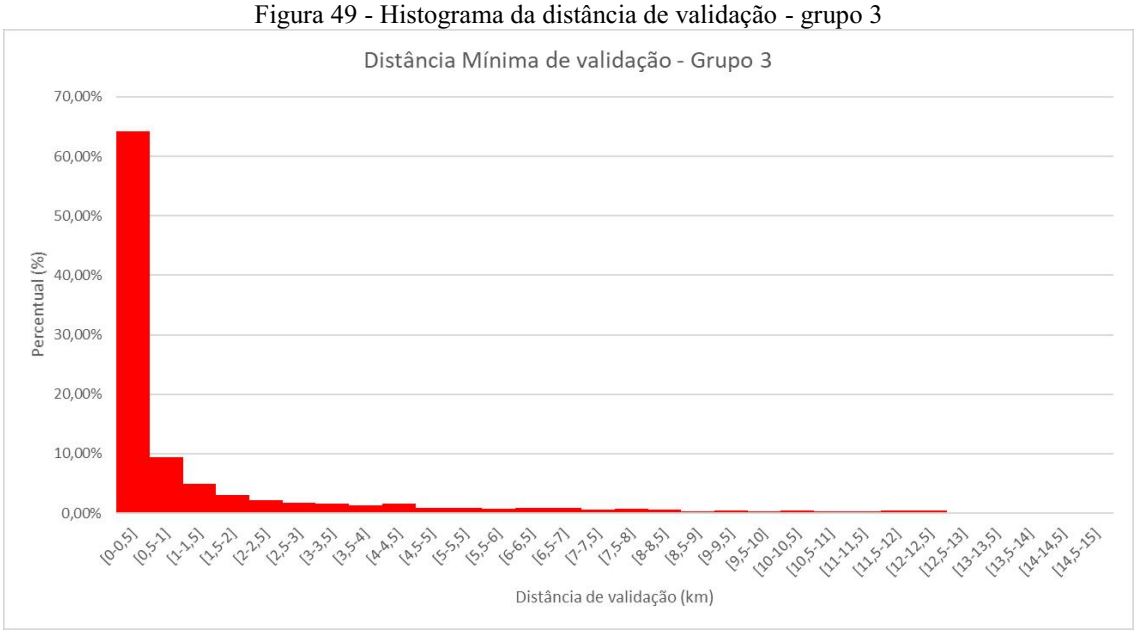

Fonte: Autor.

Figura 48 - Histograma da distância de validação - grupo 2




Fonte: Autor.

Figura 49 - Histograma da distância de validação - grupo 3




Fonte: Autor.

Tabela 15 - Resumo da média, mediana e desvio padrão da distância mínima (km)

| | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 |
|---|---|---|---|---|
| Média | 1,533 | 1,387 | 1,593 | 1,387 |
| Mediana | 0,241 | 0,097 | 0,207 | 0,216 |
| Desvio Padrão | 2,799 | 2,807 | 2,816 | 2,630 |

Fonte: Autor.

## 7.3. Modelagem categórica supervisionada e análise do desempenho

Conforme os padrões identificados, tem-se que as proporções de usuários que validam no momento do embarque são 61,3%, 67,1%, 60,2% e 62,8% para os grupos 0, 1, 2 e 3, respectivamente. Estes resultamos indicam que, em geral, a maioria dos usuários tende a validar ao embarcar. Estes resultados também indicam que os usuários do Grupo 1 (grupo regular e com várias atividades) são os que apresentam maior tendência de validar ao embarcar, seguidos pelo grupo 3 (grupo esporádico que utiliza o sistema para atividades curtas). Contudo, uma parcela considerável de usuários, em torno de 40%, não valida no momento do embarque e o interesse está em prever este comportamento, como será discutido nesta seção.

Um ponto importante sobre a escolha e utilização dos algoritmos é necessária neste ponto. Primeiramente é necessário deixar claro que não está claro a relação entre os atributos selecionados, logo não se pode afirmar que um modelo avaliado separadamente com um bom resultado foi gerado por uma boa adaptação do modelo aos dados. Outro ponto importante é que o foco não está intimamente ligado a avaliação dos modelos, mas sim, na verificação da hipótese de que a identificação dos diferentes grupos pode auxiliar nos locais de embarque, pois caso contrário, não seria necessária toda esta análise para identificação dos grupos. Por fim, a avaliação de um único modelo não é suficiente para validar ou não que a separação dos dados em grupos foi primordial para os melhores resultados da modelagem supervisionada, pois os hiper parâmetros dos modelos poderia influenciar nos resultados. Desta forma, foi selecionado 3 modelos com a finalidade de verificar esta hipótese, uma vez que com os 3 modelos obtendo melhores resultados no cenário de treinamento exclusivo com os dados do respectivo grupo, haverá evidência de que essa segregação dos usuários possibilita melhores resultados.

A nomenclatura **Grupo X** é usada neste tópico apenas para facilitar a identificação dos modelos treinados sem considerar os dados de um dado grupo específico, permitindo verificar se os modelos treinados para cada grupo apresentam um melhor desempenho do que o modelo generalizado sem considerar separação por grupos. As tabelas a seguir apresentam o resumo das métricas de desempenho para cada modelo e cada grupo.

Tabela 16 - Modelagem da probabilidade de validar ao embarcar - Grupo 0

| Treinamento: Grupo 0 Teste: Grupo 0 | Naive Bayes | | | Random Forest | | | Rede Neural | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0 | Acurácia | 1 | 0 | Acurácia | 1 | 0 | Acurácia |
| *Precision* | 0,60 | 0,50 | - | 0,60 | 0,45 | - | 0,56 | 0,40 | - |
| *Recall* | 0,99 | 0,01 | - | 0,89 | 0,13 | - | 0,04 | 0,95 | - |
| *fi- Score* | 0,75 | 0,02 | 0,60 | 0,72 | 0,20 | 0,58 | 0,08 | 0,57 | 0,41 |

Fonte: Autor.

Tabela 17 - Modelagem da probabilidade de validar ao embarcar - Grupo X-0

| Treinamento: Grupos 1, 2 e 3 Teste: Grupo 0 | Naive Bayes | | | Random Forest | | | Rede Neural | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0 | Acurácia | 1 | 0 | Acurácia | 1 | 0 | Acurácia |
| *Precision* | 0,60 | 0,00 | - | 0,60 | 0,42 | - | 0,60 | 0,44 | - |
| *Recall* | 1,00 | 0,00 | - | 0,81 | 0,21 | - | 0,95 | 0,06 | - |
| *fi- Score* | 0,75 | 0,00 | 0,60 | 0,69 | 0,28 | 0,57 | 0,73 | 0,10 | 0,59 |

Fonte: Autor.

Tabela 18 - Modelagem da probabilidade de validar ao embarcar - Grupo 1

| Treinamento: Grupo 1 Teste: Grupo 1 | Naive Bayes | | | Random Forest | | | Rede Neural | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0 | Acurácia | 1 | 0 | Acurácia | 1 | 0 | Acurácia |
| *Precision* | 0,67 | 0,23 | - | 0,69 | 0,52 | - | 0,70 | 0,41 | - |
| *Recall* | 0,99 | 0,01 | - | 0,92 | 0,19 | - | 0,74 | 0,36 | - |
| *fi- Score* | 0,80 | 0,01 | 0,66 | 0,79 | 0,28 | 0,67 | 0,72 | 0,38 | 0,61 |

Fonte: Autor.

Tabela 19 - Modelagem da probabilidade de validar ao embarcar - Grupo X-1

| Treinamento: Grupos 0, 2 e 3 Teste: Grupo 1 | Naive Bayes | | | Random Forest | | | Rede Neural | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0 | Acurácia | 1 | 0 | Acurácia | 1 | 0 | Acurácia |
| *Precision* | 0,67 | 0,00 | - | 0,71 | 0,45 | - | 0,67 | 0,44 | - |
| *Recall* | 1,00 | 0,00 | - | 0,78 | 0,36 | - | 1,00 | 0,01 | - |
| *fi- Score* | 0,80 | 0,00 | 0,67 | 0,74 | 0,40 | 0,64 | 0,8 | 0,01 | 0,67 |

Fonte: Autor.

Tabela 20 - Modelagem da probabilidade de validar ao embarcar - Grupo 2

| Treinamento: Grupo 2 Teste: Grupo 2 | Naive Bayes | | | Random Forest | | | Rede Neural | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0 | Acurácia | 1 | 0 | Acurácia | 1 | 0 | Acurácia |
| *Precision* | 0,59 | 0,49 | - | 0,63 | 0,57 | - | 0,58 | 0,48 | - |
| *Recall* | 0,87 | 0,17 | - | 0,81 | 0,34 | - | 0,97 | 0,04 | - |
| *fi- Score* | 0,71 | 0,26 | 0,58 | 0,71 | 0,43 | 0,62 | 0,73 | 0,08 | 0,58 |

Fonte: Autor.

Tabela 21 - Modelagem da probabilidade de validar ao embarcar - Grupo X-2

| Treinamento: Grupos 0,1, e 3 Teste: Grupo 2 | Naive Bayes | | | Random Forest | | | Rede Neural | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0 | Acurácia | 1 | 0 | Acurácia | 1 | 0 | Acurácia |
| *Precision* | 0,58 | 0,00 | - | 0,62 | 0,53 | - | 0,58 | 0,00 | - |
| *Recall* | 1,00 | 0,00 | - | 0,81 | 0,30 | - | 1,00 | 0,00 | - |
| *fi- Score* | 0,74 | 0,00 | 0,58 | 0,70 | 0,38 | 0,60 | 0,74 | 0,00 | 0,58 |

Fonte: Autor.

Tabela 22 - Modelagem da probabilidade de validar ao embarcar - Grupo 3

| Treinamento: Grupo 3 Teste: Grupo 3 | Naive Bayes | | | Random Forest | | | Rede Neural | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0 | Acurácia | 1 | 0 | Acurácia | 1 | 0 | Acurácia |
| *Precision* | 0,61 | 0,42 | - | 0,62 | 0,48 | - | 0,62 | 0,43 | - |
| *Recall* | 0,97 | 0,03 | - | 0,91 | 0,13 | - | 0,86 | 0,17 | - |
| *fi- Score* | 0,75 | 0,06 | 0,61 | 0,74 | 0,21 | 0,61 | 0,72 | 0,24 | 0,59 |

Fonte: Autor.

Tabela 23 - Modelagem da probabilidade de validar ao embarcar - Grupo X-3

| Treinamento: Grupos 0,1, e 2 Teste: Grupo 3 | Naive Bayes | | | Random Forest | | | Rede Neural | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0 | Acurácia | 1 | 0 | Acurácia | 1 | 0 | Acurácia |
| *Precision* | 0,61 | 0,39 | - | 0,62 | 0,42 | - | 0,61 | 0,42 | - |
| *Recall* | 0,76 | 0,24 | - | 0,84 | 0,18 | - | 0,98 | 0,02 | - |
| *fi- Score* | 0,68 | 0,30 | 0,56 | 0,71 | 0,25 | 0,59 | 0,75 | 0,05 | 0,61 |

Fonte: Autor.

Comparando os resultados dos grupos, notas que o modelo RF apresentou um leve melhor desempenho no nível de acurácia, com valores variando entre 0,58 e 0,67. O RF apresentou pior acurácia justamente em um dos grupos com menor regularidade de uso do sistema (Grupo 0) e maior acurácia no grupo com maior regularidade de uso (Grupo 1). No geral, o RF também apresentou melhor capacidade de separar entre usuários que validam e que não validam ao embarcar, principalmente no grupo com maior padrão pendular (Grupo 2). Logo há evidências de alguma relação da regularidade dos usuários com a melhor adaptação dos modelos para estes grupos de usuários.

Ao se comparar os modelos em cada grupo com os modelos do Grupo X, nota-se que os modelos por grupo apresentam uma maior capacidade em prever os usuários que não

validam ao embarcar. Como pode ser visto nas Tabelas 17, 19, 21 e 23, os resultados dos indicadores de *recall, precision e f1-score* mostram que os modelos *NB* e *RN* não apresentaram desempenho satisfatório. Para os grupos 0, 1 e 2 estes modelos não foram nem mesmo capazes de classificar os usuários que não validam ao embarcar. Porém quando avaliados os mesmos indicadores para os mesmos grupos treinados com os dados do seu grupo de classificação original, foi possível identificar corretamente pelo menos 45% destes usuários. Sendo este resultado longe do aceitável para aplicação prática de um modelo, porém apresentando resultados melhores do que os modelos não segregados por grupo.

Excetuando-se o caso da rede neural para o Grupo 0, todos os outros modelos previram melhor os usuários que validam ao embarcar, podendo haver indícios de que o conhecimento do sistema possibilita o hábito de repetir o mesmo processo, enquanto para os usuários que não costumam ter um padrão bem delimitado, é mais complexo do modelo se adaptar aos dados, necessitando de outras abordagens e até mesmo uma nova proposta de variáveis preditoras para estes casos.

Para os modelos RF de cada grupo foram avaliadas o grau de influência dos atributos na modelagem da probabilidade de validar ao embarcar (Tabela 24). Será apresentado os resultados do RF, pois este obteve os melhores indicadores. Este indicador é uma medida que avalia o quanto cada variável contribui para a precisão das previsões realizadas pelo modelo. Essa importância é calculada com base na diminuição da impureza dos nós da árvore durante a construção do modelo. Essa medida é normalizada, de modo que a soma de todas as importâncias seja igual a 1.

Para o Grupo 0 os atributos mais influentes foram os relacionados a área, distâncias de validação e faixa horária da primeira validação. Dessa forma existe uma maior influência do horário de utilização do sistema, muito por conta da dependência temporal desse grupo com os horários das rotas. O Grupo 1 teve influência principalmente dos atributos de frequência de validação, dando mais um indício de que a regularidade influencia na modelagem da probabilidade de validar ao embarcar. Também teve maior influência dos atributos de distância temporal, corroborando com o aspecto da regularidade desse grupo. O Grupo 2 teve forte influência da frequência de uso e da área de validação, pois acredita-se que estes usuários tendem a se deslocar por maiores distâncias, também sendo usuários regulares. Por fim, o Grupo 3, teve influência da faixa horária da primeira validação e de aspectos da área e

distância de validação, conforme o Grupo 0, indicando que a probabilidade de validar nestes casos está diretamente ligada ao fator temporal de como funciona o sistema.

Tabela 24 - Grau de importância dos atributos para o modelo RF

| Atributos | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 |
|---|---|---|---|---|
| FREQ_DIA_SEGUNDA | 0,050 | 0,056 | 0,053 | 0,049 |
| FREQ_DIA_TERCA | 0,048 | 0,060 | 0,054 | 0,048 |
| FREQ_DIA_QUARTA | 0,053 | 0,059 | 0,054 | 0,049 |
| FREQ_DIA_QUINTA | 0,049 | 0,057 | 0,055 | 0,047 |
| FREQ_DIA_SEXTA | 0,047 | 0,056 | 0,061 | 0,049 |
| FRQ_VALIDACAO_TERMINAL | 0,044 | 0,046 | 0,043 | 0,041 |
| DIST_TEMPO_DIA_SEGUNDA | 0,053 | 0,053 | 0,056 | 0,051 |
| DIST_TEMPO_DIA_TERCA | 0,056 | 0,051 | 0,051 | 0,053 |
| DIST_TEMPO_DIA_QUARTA | 0,057 | 0,050 | 0,053 | 0,055 |
| DIST_TEMPO_DIA_QUINTA | 0,054 | 0,053 | 0,052 | 0,053 |
| DIST_TEMPO_DIA_SEXTA | 0,054 | 0,057 | 0,054 | 0,055 |
| AREA_VALIDACOES | 0,063 | 0,055 | 0,062 | 0,057 |
| DIST_HORIZONTAL_MAX | 0,059 | 0,058 | 0,056 | 0,061 |
| DIST_VERTICAL_MAX | 0,061 | 0,055 | 0,057 | 0,060 |
| PROP_DIARIA_ALIMENTADORA | 0,038 | 0,049 | 0,039 | 0,047 |
| PROP_DIARIA_TRONCAL | 0,032 | 0,040 | 0,045 | 0,043 |
| PROP_DIARIA_CONVENCIONAL | 0,060 | 0,053 | 0,052 | 0,053 |
| PROP_DIARIA_COMPLEMENTAR | 0,054 | 0,052 | 0,055 | 0,059 |
| FAIXA_HORÁRIA_PRIMEIRA | 0,068 | 0,042 | 0,048 | 0,072 |

Fonte: Autor.

Por fim, é importante destacar que para o problema central do trabalho da não compreensão do padrão de deslocamentos e a impossibilidade de identificação dos locais de embarque refletir negativamente na oferta do sistema, obteve-se uma contribuição. Contribuição está no sentido de propor uma metodologia que identifique os padrões de uso, bem como os atributos que mais impactam nestes padrões e a possibilidade de utilização de modelagem supervisionada para identificar o local de embarque e posteriormente até reconstruir as matrizes OD. Dentre as principais vantagens deste método está a possibilidade de analisar uma grande série temporal de dados de bilhetagem, a abrangência de uma região maior do que uma pesquisa OD cobriria, possibilidade de replicação em um curto espaço de tempo e a avaliação dos atributos de forma segregada analisando impacto deste na modelagem. Dentre as principais limitações, ainda não se sabe o motivo da viagem ou as

diferentes “pernas” desta. Outra desvantagem é a necessidade de um processo de tratamento adequado para não gerar resultados que não contribuam ao fenômeno estudado. Porém, mesmo com estas limitações é possível reproduzir este método para outros sistemas de transporte público, sendo necessário um esforço inicial para obtenção e modelagem dos dados, contribuindo assim em um nível operacional.

# 8. CONSIDERAÇÕES FINAIS

Diante do estudo proposto, é evidente que o sistema de transportes público apresenta um processo de análise e planejamento que varia de acordo com o tipo de rede e dados disponíveis. Portanto, as etapas de compreensão de fenômeno, análise dos padrões de viagem e dos fatores que influenciam esses padrões são de extrema importância. Neste estudo o foco se manteve na identificação, caracterização e aplicação desses padrões de modo a auxiliar na identificação dos locais de embarque, e que posteriormente podem apresentar uma noção macroscópica do sistema.

Tradicionalmente os padrões de viagem são identificados em Pesquisas Origem/Destino (O/D), ferramentas mais comuns para se obter informações sobre a mobilidade de uma cidade. Porém este método não incorpora a variação temporal das validações e tipos de deslocamento, ou seja, é uma afirmação errônea de que o usuário se comporta de forma fixa no tempo. Dessa forma, neste trabalho se evidenciou que a utilização de dados massivos possibilita a identificação dos diversos padrões de regularidade distribuídos no tempo. Ainda é necessário a utilização de ferramentas adequadas para manipulação e armazenamento dos dados, por isso, a estrutura integrada de banco de dados relacional proposta é um avanço nessa temática.

Conforme discutido, no Sistema Integrado de Transporte Público de Fortaleza as principais dificuldades para reconstrução das viagens é que não se sabe o local de embarque e desembarque, não existe verificação se a validação configura realmente uma transferência ou uma nova atividade, não se sabe o motivo da viagem, não se têm dados sobre os usuários não-rastreados e não se tem informações sobre os trasbordos nos terminais. Portanto, os métodos mais comuns geralmente utilizados para esse propósito, como o encadeamento de viagens, não podem ser aplicados diretamente, pela deficiência ocasionada dado o elevado número de premissas e suposições. Buscou-se, portanto, dar uma contribuição em relação a lacuna do ***sistema de Transporte Público de Fortaleza não possibilitar saber o real local de embarque e não apresentar informações sobre o aspecto de uso***, avançando sobre as técnicas de tratamento, mineração e modelagem de Big Data – TP.

A partir da etapa de tratamento e armazenamento dos dados finalizada, parte-se com maior segurança para as análises dos padrões intrínsecos das validações. As análises exploratórias auxiliaram na identificação desses padrões e na definição dos atributos, que

estão intimamente atrelados a regularidade de uso no sistema e aos hábitos de deslocamento. Essas análises foram divididas entre análises espaciais e temporais agregadas, análises das primeiras validações e a nível do indivíduo. Dessa forma, não apenas foi possível utilizar esses padrões para identificar os reais locais de embarque, mas identificar quando geralmente acontecem esses embarques, a frequência e a distância temporal das atividades baseado no padrão temporal de cada grupo. Diante disso, foi proposto um método baseado na mineração dos dados massivos do transporte público de Fortaleza. Um dos pontos principais da proposta é a identificação dos padrões de uso dos diferentes grupos, além de possibilitar a identificação dos locais de embarque, através das distâncias mínimas de validação.

## 8.1. Objetivos e Hipóteses

Diante da proposta e dos resultados obtidos, foi possível validar o objetivo geral do trabalho em identificar os reais locais de embarque das viagens do Big Data do Sistema Integrado de Transporte Público de Fortaleza (SIT-FOR). Porém, fica a ressalva que a identificação depende de um conjunto de atributos que apenas são passíveis de serem levantados, se tiverem dados suficientes de um mesmo usuário, que neste caso foi de 6 meses.

Para o primeiro objetivo específico foi construído o modelo de entidade-relacionamento (modelo conceitual) e posteriormente o modelo físico com 20 bases de dados, dentre elas as bases de bilhetagem, gps, cadastro de usuários e os arquivos obrigatórios do GTFS. Dentre todas as bases, a que apresentou maior número de inconsistências foi o cadastro de usuários, onde dentre os mais de 330 mil usuários identificados na bilhetagem, menos de 50% estavam na base e em torno de 20 mil estavam adequados para uso na modelagem dos padrões. As bases de bilhetagem e GPS tiveram taxa de tratamento aproximada de 91% e 80%, respectivamente, sendo essenciais no desenvolvimento das etapas subsequentes.

Através das análises exploratórias foram definidos 20 atributos espaciais e temporais que auxiliaram na identificação dos padrões por intermédio da mineração de dados com modelo não-supervisionado. No geral, foram encontrados e validados 4 grupos de uso através dos métodos do cotovelo e da silhueta. Dentre esses grupos 2 deles apresentam forte regularidade de uso espacial e com atividades bem delimitadas no tempo, enquanto os outros se distanciam dessa regularidade encontrada. No terceiro objetivo específico foi proposto

como estes padrões poderiam auxiliar na identificação do real local de embarque, desse modo averiguou-se que cada grupo apresentava uma tendência de validar (ou não) próximo ao local de embarque e que os atributos levantados poderiam auxiliar a explicar esse comportamento, dessa forma cada grupo individualmente através de seus padrões de uso foram segregados para modelagem da distância mínima de validação.

Por fim, para cada grupo e para cada conjunto remanescente de usuários do sistema, 3 modelos categóricos supervisionados foram propostos: o *Random forest*, o Naive Bayes (Gaussiano) e a rede neural. Os resultados indicaram uma aderência mediana (entre 0,58 e 0,67) dos modelos com os dados. Mas deram evidências de que segregar os usuários por grupo para modelar os aspectos de validação ao embarcar apresenta uma melhor acurácia do que apenas utilizar todos os usuários agregados. Este método, embora provavelmente apresente grupos diferentes para diferentes sistemas, poderá ser replicado para dar maior veracidade a identificação dos reais locais de embarque e consequentemente do encadeamento das viagens para construção de uma matriz O/D do sistema.

Durante o trabalho foram propostas 5 hipóteses que foram verificadas através das análises exploratórias e das modelagens. A primeira hipótese verificada foi a de que a maioria dos usuários tende a validar assim que embarca, e de forma geral, sem considerar as separações dos grupos identificados, 68% dos usuários validaram em distâncias de até 1km, dando indícios que realmente validam ao embarcar.

Acredita-se que os tipos de atividades influenciam os padrões temporais e os locais dessas atividades influenciam os padrões espaciais. Em relação à segunda hipótese, não houveram fortes evidências para dar indícios de que seja verdadeiro, ou não, necessitando ser melhor verificado. Porém avaliando o primeiro ponto, 7 dos 20 atributos mais influentes dos usuários mais frequentes são relacionados ao padrão temporal, ou seja, a demarcação temporal das atividades apresenta uma forte influência nas características desses grupos, por criar uma conexão entre a oferta do sistema e a necessidade de deslocamento dentro do prazo estabelecido em termos de contrato da atividade. A pouca independência financeira dos usuários causa uma dependência temporal do sistema de transporte público.

As duas últimas hipóteses avaliadas foram a de que os usuários regulares tendem a repetir certos comportamentos na rede que podem ser identificados através de técnicas de mineração e de que diferentes grupos tendem a ter diferentes padrões de uso. Ambas as hipóteses foram validadas na mineração e modelagem dos dados. Além do que, os grupos

com maior frequência de uso e maior regularidade temporal foram os usuários regulares na amostra e com melhores resultados na modelagem do local de embarque.

## 8.2. Propostas de Trabalhos Futuros

Embora a proposta metodológica desse trabalho tenha sido contemplada, é reconhecido que alguns processos/etapas poderiam ter sido mais detalhados e melhor aplicados. Além do que passos posteriores são importantes para ser possível utilizar estes dados no planejamento estratégico do sistema de transportes públicos de um município. Dessa forma, pontua-se algumas análises e propostas de trabalhos futuros:

- Como utilizar os dados de bilhetagem e GPS, num nível desagregado para identificar atributos dos deslocamentos, à exemplo: como diferentes séries temporais de atributos retirados dos dados podem ser usadas para identificar regularidade e atributos das viagens!?
- Realizar a *clusterização* para um espaço temporal maior que 6 meses e comparar com este método, averiguando se novos grupos serão identificados, ou se esse espaço temporal é suficiente para identificar a mesma quantidade de grupos.
- Avaliar novos modelos supervisionados e o impacto dos atributos na formulação desses modelos. Além de propor novos atributos, assim como uma análise mais detalhada da influência deles sobre a identificação dos grupos.

# BIBLIOGRAFIA

ALSREHIN, N. O.; KLAIB, A. F.; MAGABLEH, A. **Intelligent transportation and control systems using data mining and machine learning techniques: A comprehensive study**. IEEE Access, v. 7, p. 49830-49857, 2019.

ANTP: ASSOCIAÇÃO NACIONAL DE TRANSPORTES PÚBLICOS. **Sistemas Inteligentes de Transportes**. Série Cadernos Técnicos, vol. 8. São Paulo, 2012, 163 p.

ARBEX, R. O., & da CUNHA, C. B. (2017). **Estimação da matriz origem-destino e da distribuição espacial da lotação em um sistema de transporte sobre trilhos a partir de dados de bilhetagem eletrônica**. TRANSPORTES, 25(3), 166–177. https://doi.org/10.14295/transportes.v25i3.1347

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.

ASSEMI, B., Alsger, A., Moghaddam, M., Hickman, M., Mesbah, M., 2020. **Improving alighting stop inference accuracy in the trip chaining method using neural networks**. Public Transport 12, 89–121.

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), pp. 183-187.

BISHOP, C. M. (1995). **Neural Networks for Pattern Recognition**. Oxford University Press.

BREIMAN, L. (2001). **Random forests.** Machine learning, 45(1), 5-32. http://doi.org/10.1023/A:1010933404324

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, 2019.

BRIDLE, J. S. (1990). **Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition**. Neurocomputing: Algorithms, Architectures and Applications, 2(3), 227-236.

CATS, O.; FERRANTI, F. **Unravelling individual mobility temporal patterns using longitudinal smart card data**, Research in Transportation Business & Management, 2022,100816,ISSN 2210-5395, https://doi.org/10.1016/j.rtbm.2022.100816.

CATS, Oded; WANG, Qian; ZHAO, Yu. **Identification and classification of public transport activity centres in Stockholm using passenger flows data**. Journal of Transport Geography, v. 48, p. 10-22, 2015.

CHAKIROV, A., & ERATH, A. (2012). **Activity identification and primary location modelling based on smart card payment data for public transport**. In 13th International Conference on Travel Behaviour Research (IATBR 2012). IVT, ETH-Zürich.

CHAKROBORTY, P.; KIKUCHI, S. **Using Bus Travel Time Data to Estimate Travel Times on Urban Corridors**. Transportation Research Record: Journal of the Transportation Research Board, v. 1870, p. 18–25, 1 jan. 2004.

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.

CODD, E. F. (1970). **A relational model of data for large shared data banks**. Communications of the ACM, 13(6),377–387. doi:10.1145/362384.362685.

CUI, C. L., ZHAO, Y. L., DUAN, Z. Y. (2014). **Research on the stability of public transit passenger travel behavior based on smart card data**. 14th COTA International Conference of Transportation Professionals, 1318–1326.

CUI, Z. Y., & LONG, Y. (2015). **Perspective on stability and mobility of passenger's travel behavior through smart card data**. ACM SIGKDD Workshop on Urban Computing.

CUTLER, D. R.; EDWARDS JR, T. C.; BEARD, K. H.; CUTLER, A.; HESS, K. T.; GIBSON, J.; Lawler, J. J. (2007). **Random forests for classification in ecology.** Ecology, 88(11), 2783-2792.

DAVIS, J.; GOADRICH, M. (2006). **The relationship between Precision-Recall and ROC curves.** In Proceedings of the 23rd International Conference on Machine Learning, 233-240.

DESSOUKY, M.; HALL, R.; NOWROOZI, A.; MOURIKAS, K. **Bus dispatching at timed transfer transit stations using bus tracking technology**. Transportation Research Part C: Emerging Technologies, v. 7, n. 4, p. 187–208, 1 ago. 1999.

DoD, U. 2008. **Global positioning system standard positioning service performance standard**, 4th Ed. Assistant secretary of defense for command, control, communications, and intelligence.

DUMBILL, E. **What is big data? An introduction to the big data**. 2012. Disponível em: <http://radar.oreilly.com/2012/01/what-is-big-data.html> Acesso em: 2020-12-31.

ETUFOR. **Redes Radial e Troncal de Fortaleza**.2020. Disponível em:< https://www.fortaleza.ce.gov.br/ > Acesso em: 2020-04-09.

FORTALEZA, P.M. de. **Anuário de Transporte Público**. Fortaleza:[s.n.], 2010.

FORTALEZA, P. M. de. **Plano de Mobilidade de Fortaleza** PlanMob. [S.l.: s.n.], 2015.

FREITAS, A.T., **Metodologia de Caracterização da Problemática do Sistema de Transportes Públicos de Passageiros a partir dos Dados da Bilhetagem Eletrônica**. 2015. 100 f. Dissertação (Mestrado) - Curso de Engenharia Civil, Universidade Federal do Ceará, Fortaleza, 2015.

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.

GLOROT, X.; BORDES, A.; BENGIO, Y. (2011). **Deep sparse rectifier neural networks**. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 315-323). JMLR Workshop and Conference Proceedings.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. (2016). **Deep learning.** MIT press.

GORDON, J., KOUTSOPOULOS, H., WILSON, N. and ATTANUCCI, J. 2013. **Automated inference of linked transit journeys in London using fare-transaction and vehicle location data**. Transportation Research Record: Journal of the Transportation Research Board, (2343), pp. 17-24.

GUERRA, A. L.; BARBOSA, H. M.; OLIVEIRA, L. K. **Estimativa de Matriz Origem/Destino Utilizando Dados do Sistema de Bilhetagem Eletrônica: Proposta Metodológica**. Transportes, [s.l.], v 22, n. 3, p. 26-38, 2014.

HALEVY, A. Y.; NORVIG, P., PEREIRA, F. **The Unreasonable Effectiveness of Data.** IEEE Intelligent Systems 24 , no. 2 (2009): 8-12.

HAN, J., Pei, J., & Tong, H. (2022). **Data mining: concepts and techniques**. Morgan kaufmann.

HABIB, K. N., & Weiss, A. (2014). **Evolution of latent modal captivity and mode choice patterns for commuting trips: A longitudinal analysis using repeated cross-sectional datasets**. Transportation Research Part A: Policy and Practice, 66, 39–51.

HAYKIN, S. (2008). **Neural Networks and Learning Machines (3ª ed.).** Pearson Education.

HORA, J. (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.

HUANG, J., XU, L. e P. YE (2015). **Exploring transit use regularity using smart card data of students**. ICTE

HUSSAIN, E., BEHARA, K.N., BHASKAR, A., CHUNG, E., 2021. **A Framework for the Comparative Analysis of Multi-Modal Travel Demand: Case Study on Brisbane Network**. IEEE Transactions on Intelligent Transportation Systems, 23(7), pp.8126-8135.

HUSSAIN, E.; BHASKAR, A.; CHUNG, E.2021; **Transit OD Matrix Estimation UsingSmartcard Data: Recent Developments and Future Research Challenges**. 125 th. Transportat Research. doi: 10.1016/j.trc.2021.103044.

HWANG, M., KEMP, J., LERNER-LAM, E., NEUERBURG, N. and OKUNIEFF, P. (2006). **Advanced Public Transportation: State of the Art** Update 2006. Report FTA-NJ-26-7062–06.1, Federal Transit Administration, US Department of Transportation.

IPEA. **Projeto Acesso a Oportunidades.** Acesso a Oportunidades, 2022. Disponível em: < https://www.ipea.gov.br/acessooportunidades/sobre/> Acesso em: 06 de julho de 2022.

JACKSON, J.E., 2005. **A user's guide to principal components**. John Wiley & Sons.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. (1999). **Data clustering: a review.** ACM Computing Surveys, 31(3), 264-323.

JANG, W. 2010. **Travel time and transfer analysis using transit smart card data. Transportation Research Record: Journal of the Transportation Research Board**, (2144), pp. 142-149.

JAPKOWICZ, N.; SHAH, M. (2011). **Evaluating learning algorithms: a classification perspective.** Cambridge University Press.

JOLLIFFE, I. T. (2002). **Principal Component Analysis.** Springer.

JUNG, J., SOHN, K., 2017. **Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data**. IET Intelligent Transport Systems 11, 334–339.

KETCHEN JR, D. J.; SHOOK, C. L. (1996). **The application of cluster analysis in strategic management research: An analysis and critique**. Strategic Management Journal, 17(6), 441-458.

KOCADAGLI, O., 2015. **A novel hybrid learning algorithm for full Bayesian approach of artificial neural networks**. Applied Soft Computing, 35, pp.52-65.

KURAUCHI, F.; SCHOMOCKER, J. D. (2016) **Public transport planning with smartcard data**. 2016.

LECUN, Y.; BENGIO, Y.; HINTON, G. (2015). **Deep learning.** Nature, 521(7553), 436-444.

LEE, S.G. and HICKMAN, M. 2013. **Are transit trips symmetrical in time and space? Evidence from the Twin Cities. Transportation** Research Record: Journal of the Transportation Research Board, (2382), pp. 173-180.

LI, T., SUN, D., JING, P., YANG, K. (2018). **Smart card data mining of public transport destination**: A literature review. Inf.

LIN, P., WENG, J., ALIYANISTOS, D., MA, S., & YIN, B. (2020). Identifying and segmenting commuting behavior patterns based on smart card data and travel survey data. Sustainability, 12(12), 5010.

LIU, X.; ZHOU, Y.; RAU, A. **Smart card data-centric replication of the multi-modal public transport system in Singapore**. 2019. Journal Transport Geography. https://doi.org/10.1016/j.jtrangeo.2018.02.004.

LUO, D., Cats, O., van Lint, H., 2017. **Constructing Transit Origin-Destination Matrices with Spatial Clustering**. Transportation Research Record: Journal of the Transportation Research Board 2652, 39–49.

MA, X., WU, Y. J., WANG, Y., CHEN, F. and LIU, J. 2013. **Mining smart card data for transit riders' travel patterns**. Transportation Research Part C: Emerging Technologies, 36, pp. 1-12.

MACQUEEN, J. B. (1967). **Some methods for classification and analysis of multivariate observations.** Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1(14), 281-297.

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 com Big Data de Transporte Público**. 35° Congresso Nacional de Pesquisa em Transportes da ANPET, assíncrono, 2021

MORENCY, C., TRÉPANIER, M. and AGARD, B. 2007. **Measuring transit performance using smart card data**. Presented at World Conference on Transport Research, San Francisco, USA.

MUNIZAGA, M.A. 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, pp. 9-18.

NAGABHUSHANA, S. **Data Warehousing**, OLAP and Data Mining. New Delhi, Índia: New Age International, 2006.

NAIR, V. HITON, G.E., 2010. **Rectified linear units improve restricted boltzmann machines**. In Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 807-814).

NASSIR, N., HICKMAN, M., MA, Z.-L., 2015. **Activity detection and transfer identification for public transit fare card data**. Transportation 42, 683–705.

NASSIR, N.; HICKMAN, M.; MA, Z.L.A **Strategy-based recursive path choice model for public transit smart card data**. 2018. Journal Transport Geography. https://doi.org/10.1016/j.trb.2018.01.002.

NIELSEN, M. A. (2015). **Neural Networks and Deep Learning.** Determination Press.

OKUNIEFF, P. E. **AVL systems for bus transit: a synthesis of transit practice: Transit Cooperative Research Program (TCRP).** Washington, DC: Transportation Research Board, 1997.

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.

PFITSCHER, F.C.; MICHEL,F.D.; LADEIRA, M.C.M.;SANTOS, M.L. **Criação de matrizes Origem-destino para o sistema de ônibus de Guaíba com dados de bilhetagem eletrônica**. 34° Congresso Nacional de Pesquisa em Transportes da ANPET, assíncrono,2020.

POWER, D. M. (2011). **Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.** Journal of Machine Learning Technologies, 2(1), 37-63.

RABAY, P. F. B (2017). **Uso de dados secundários de rastreamento da frota na etapa de caracterização da problemática do sistema de transporte público**. (Dissertação de Mestrado) Universidade Federal do Ceará. 2017 .1 -63.

RISH, I. (2001). **An empirical study of the naive Bayes classifier.** IJCAI 2001 workshop on empirical methods in artificial intelligence, 3(22), 41-46.

ROSSETTI, M. D. (1996). **Automatic Data Collection on Transit Users Via Radio Frequency Identification,** Report of investigation, Transit-IDEA program, No. 10.

RUMELHART, D. E., HINTON, G. E., & WILLIAMS, R. J. (1986). **Learning representations by back-propagating errors**. Nature, 323(6088), 533.

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

SOKOLOVA, M.; LAPALME, G. (2009). **A systematic analysis of performance measures for classification tasks.** Information Processing & Management, 45(4), 427-437.

SOUSA, F.F.L.M.; MESQUITA, K.G.A.; LOUREIRO, C.F.G. **Caracterização da Evolução do Padrão de Mobilidade de Fortaleza a partir da calibração do Tranus**. 33° Congresso Nacional de Pesquisa em Transportes da ANPET, Balneário Camboriú, SC,2019.

TANG, T.; LIU, R.; CHOUDHURY, C.; FONZONE, A.; WANG, Y. (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.

THIAGARAJAN, R.; PRAKASHKUMAR, S. **Identification of passenger demand in public transport using machine learning**. Webology, v. 18, n. Special Issue on Information Retrieval and Web Search, p. 223-236, 2021.

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), pp. 1-14.

WONG, J. C. Use of the general transit feed specification (GTFS) in transit performance measurement. 2013. (Ph. D. thesis) Georgia Institute of Technology, 2013.

ZHAO, J. (2004) **The Planning and Analysis Implications of Automated Data Collection System: Rail Transit OD Matrix Inference and Path Choice Modeling Examples**. Thesis. Massachusetts Institute of Technology, Boston.

ZHAO (2007), J.; RAHBEE, A.; WILSON, N. H. **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, 2007. ISSN 10939687.

ZHAO, X., CUI, M., & LEVINSON, D. (2023). **Exploring temporal variability in travel patterns on public transit using big smart card data**. Environment and Planning B: Urban Analytics and City Science, 50(1), 198–217. https://doi.org/10.1177/23998083221089662

# APÊNDICE A – DESCRITIVO DAS BASES DE DADOS

Tabela 25 - Descrição da base de dados das paradas

| **Paradas** | |
|---|---|
| **Nome da variável** | **Descrição** |
| stop_id | Identificador da parada |
| stop_code | *Inexistente* |
| stop_name | Endereço da parada |
| stop_desc | *Inexistente* |
| stop_lat | Latitude |
| stop_lon | Longitude |
| zone_id | Inexistente |
| stop_url | Inexistente |
| location_type | Tipo de Localização |
| parent_station | *Inexistente* |
| stop_timezone | *Inexistente* |
| wheelchair_boarding | *Inexistente* |

Fonte: Autor.

Tabela 26 - Descrição da base de dados dos Atributos de tarifas

| **Atributos de Tarifa** | |
|---|---|
| **Nome da variável** | **Descrição** |
| fare_id | Identificador da tarifa |
| price | Valor da tarifa |
| currency_type | Moeda |
| payment_method | Método de pagamento |
| transfers | *Inexistente* |
| transfer_duration | *Inexistente* |

Fonte: Autor.

Tabela 27- Descrição da base de dados da Agência

| **Agência** | |
|---|---|
| **Nome da variável** | **Descrição** |
| agency_id | Identificador da Agência |
| agency_name | Nome da agência reguladora |
| agency_url | Endereço web |
| agency_timezone | Timezona |
| agency_lang | Idioma |
| agency_phone | Telefone |
| agency_fare_url | *Inexistente* |

Fonte: Autor.

Tabela 28 - Descrição da base de dados das rotas

| **Rotas** | |
|---|---|
| **Nome da variável** | **Descrição** |
| route_id | Identificador da rota |
| agency_id | Identificador da agência |
| route_short_name | Nome reduzido da rota |
| route_long_name | Nome completo da rota |
| route_desc | *Inexistente* |
| route_type | Tipo de rota |
| route_url | *Inexistente* |
| route_color | *Inexistente* |
| route_text_color | *Inexistente* |

Fonte: Autor

Tabela 29 - Descrição da base de dados das Regras de Tarifa

| **Regras de Tarifa** | |
|---|---|
| **Nome da variável** | **Descrição** |
| fare_id | Identificador da regra de tarifa |
| route_id | Identificador da rota |
| origin_id | *Inexistente* |
| destination_id | *Inexistente* |
| contains_id | *Inexistente* |
| route_type | *Inexistente* |
| route_url | *Inexistente* |
| route_color | *Inexistente* |
| route_text_color | *Inexistente* |

Fonte: Autor.

Tabela 30 - Descrição da base de dados do Calendario

| **Calendário** | |
|---|---|
| **Nome da variável** | **Descrição** |
| service_id | Identificador do tipo de serviço |
| monday | Identificador binário (1 – Sim, 0 - Não) |
| tuesday | Identificador binário (1 – Sim, 0 - Não) |
| wednesday | Identificador binário (1 – Sim, 0 - Não) |
| thursday | Identificador binário (1 – Sim, 0 - Não) |
| friday | Identificador binário (1 – Sim, 0 - Não) |
| saturday | Identificador binário (1 – Sim, 0 - Não) |
| sunday | Identificador binário (1 – Sim, 0 - Não) |
| start_date | Data de início do calendário |
| end_date | Data final do calendário |

Fonte: Autor.

Tabela 31 - Descrição da base de dados do shape

| **Shape GTFS** | |
|---|---|
| **Nome da variável** | **Descrição** |
| shape_id | Identificador do shape |
| shape_pt_lat | Latitude |
| shape_pt_lon | Longitude |
| shape_pt_sequence | Sequência da parada na rota |
| shape_dist_traveled | *Inexistente* |

Fonte: Autor.

Tabela 32 - Descrição da base de dados da data do calendário

| **Data do calendário** | |
|---|---|
| **Nome da variável** | **Descrição** |
| service_id | Identificador do serviço |
| date | Data |
| exception_type | Tipo de Exceção |

Fonte: Autor.

Tabela 33 - Descrição da base de dados das viagens

| Viagens | |
|---|---|
| **Nome da variável** | **Descrição** |
| route_id | Identificador da rota |
| service_id | Identificador do Serviço |
| trip_id | Identificador da viagem |
| trip_headsign | *Inexistente* |
| trip_short_name | *Inexistente* |
| direction_id | *Inexistente* |
| block_id | *Inexistente* |
| shape_id | Identificador do shape |
| wheelchair_accessible | Acessível a cadeira de rodas (1 – Não, 2 – SIM) |

Fonte: Autor.

Tabela 34 - Descrição da base de dados do Horário das Paradas

| Horário das Paradas | |
|---|---|
| **Nome da variável** | **Descrição** |
| trip_id | Identificador da viagem |
| arrival_time | Horário de chegada |
| departure_time | Horário de Partida |
| stop_id | Identificador da parada |
| stop_sequence | Sequência da parada na rota |
| stop_headsign | *Inexistente* |
| pickup_type | *Inexistente* |
| drop_off_type | *Inexistente* |
| shape_dist_traveled | *Inexistente* |

Fonte: Autor

Tabela 35-Descrição da base de dados de Embarques nos terminais

| Embarques nos Terminais | |
|---|---|
| **Nome da variável** | **Descrição** |
| Data | Data da pesquisa |
| Pesquisador | *Inexistente* |
| Supervisor | *Inexistente* |
| Cod_Linha | Número da Linha |
| Nome_Linha | Nome da Linha |
| Porta_Embarque | *Inexistente* |
| Prefixo | Identificador do veículo |
| Embarques | Número de embarques |
| Período | Turno do dia |
| Hora_Partida | Hora e Minutos |
| Hora | Hora |
| Minuto | Minuto |
| Sentido | Bairro de destino |

Fonte: Autor.

Tabela 36 - Descrição da base de dados das zonas

| Zonas | |
|---|---|
| **Nome da variável** | **Descrição** |
| ID | Identificador do registro |
| AREA | Área da zona |
| ZONA_NOVA | Código da zona |
| BAIRRO_MUN | Nome do bairro |
| MUNICPIO | Nome da cidade (região metropolitana) |

Fonte: Autor.

Tabela 37 - Descrição da base de dados do Transbordo nos Terminais

| Transbordo nos Terminais | |
|---|---|
| **Nome da variável** | **Descrição** |
| Local de Pesquisa | Nome do terminal |
| Formulário | Número do formulário |
| Data | Data da pesquisa |
| Pesquisador | Nome do pesquisador |
| Período | Turno do dia |
| Supervisor | Nome do supervisor |
| Cód_Linha | Código da linha |
| Nome_Linha | Nome da linha |
| Prefixo | Identificador do veículo |
| Sentido | Bairro de destino |
| Hora_Inicio | Hora e minuto de chegada |
| Hora_Partida | Hora e minuto da partida |
| Transferência | Código da linha de transferência |
| Descrição Linha | Nome da Linha de transferência |
| Total | Total de passageiros |
| Ocupação | *Inexistente* |

Fonte: Autor.

Tabela 38 - Descrição da base de dados dos bairros

| Bairros | |
|---|---|
| **Nome da variável** | **Descrição** |
| COD_BAIRRO | Identificador do Bairro |
| NOME | Nome do bairro |
| COD_BA_IBG | Código do IBGE para o bairro |
| REGIONAL | Número da divisão Regional (2018) |
| ZONA | Região da cidade |

Fonte: Autor.

Tabela 39 - Descrição da base de dados dos Terminais

| Terminais | |
|---|---|
| **Nome da variável** | **Descrição** |
| ID Terminal | Identificador geral do Terminal |
| Nome | Nome do terminal |
| Tipo | Fechado / Aberto |
| Lat | Latitude |
| Long | Longitude |

Fonte: Autor.

Tabela 40 - Descrição da base de dados do Cadastro dos Usuários

| **Cadastro dos Usuários** | |
|---|---|
| **Nome da variável** | **Descrição** |
| CIA | Identificador geral do cadastro |
| Nome | Nome do usuário |
| DataCadastro | Data do Cadastro |
| Endereco | Nome da rua e número do endereço do usuário |
| Bairro | Bairro do endereço do usuário |
| Cidade | Cidade do endereço do usuário |
| UF | Código do estado do endereço do usuário |
| Cep | CEP do endereço do usuário |
| TipoCartao | Identificador do tipo de cartão |
| NumeroSigom | Identificador do Smartcard |
| NumeroChip | Número do chip |
| Empresa | Nome da empresa/universidade solicitante |
| EnderecoEmpresa | Nome da rua e número do endereço da empresa/universidade |
| BairroEmpresa | Bairro do endereço da empresa/universidade |
| CidadeEmpresa | Cidade do endereço da empresa/universidade |
| UFEmpresa | Código do estado do endereço da empresa/universidade |

Fonte: Autor.

Tabela 41 - Descrição da base de dados do Dicionário

| **Dicionário** | |
|---|---|
| **Nome da variável** | **Descrição** |
| vehicleid | Identificador do veículo - GPS |
| numbus | Prefixo do carro - Bilhetagem |

Fonte: Autor.

Tabela 42 - Descrição da base de dados do GPS

| **GPS** | |
|---|---|
| **Nome da variável** | **Descrição** |
| direction | Azimute |
| latitude | Latitude |
| longitude | Longitude |
| metrictimestamp | Data e hora agregados |
| odometer | Distância percorrida |
| routecode | Inexistente |
| speed | Velocidade média |
| device_deviceid | Número do dispositivo |
| vehicle_vehicleid | Identificador do veículo |

Fonte: Autor.

Tabela 43 - Descrição da base de dados da Bilhetagem

| **Bilhetagem** | |
|---|---|
| **Nome da variável** | **Descrição** |
| id | Identificador do Smartcard |
| linha | Número da linha |
| nome_linha | Nome da Linha |
| prefixo_carro | Prefixo do carro |
| Dia | Data e hora da validação |
| tipo_cartao | Número do tipo de cartão |
| nome_cartao | Descrição do tipo de cartão |
| sentido_viagem | Sentido da viagem |
| integracao | Integração |

Fonte: Autor.

# APÊNDICE B – ESTRUTURA RELACIONAL DO BANCO DE DADOS