BASSAM ABDO MAJDOUB

Análises espaço-temporais de viagens por transporte público a partir de matriz Origem-Destino gerada por meio de bilhetagem eletrônica

São Paulo

2024

BASSAM ABDO MAJDOUB

Análises espaço-temporais de viagens por transporte público a partir de matriz Origem-Destino gerada por meio de bilhetagem eletrônica

Versão Corrigida

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Mestre em Ciências.

Área de concentração: Engenharia de Transportes

Orientadora: Prof.ª Dr.ª Mariana Abrantes Giannotti

São Paulo

2024

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, _____de _____de _____

Assinatura do autor: _____

Assinatura do orientador: _____

Catalogação-na-publicação

Majdoub, Bassam

Análises espaço-temporais de viagens por transporte público a partir de matriz Origem-Destino gerada por meio de bilhetagem eletrônica / B. Majdoub - versão corr. -- São Paulo, 2024.

124 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Transportes.

1.Engenharia de Transportes 2.Planejamento de Transportes 3.Transporte Público 4.Padrões de Viagem 5.Desigualdades Socioespaciais I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Transportes II.t.

AGRADECIMENTOS

Gostaria de iniciar agradecendo a Deus, à minha família e aos meus pais, pelo amor, pelo apoio incondicional e por serem a base que me sustentou durante todo o percurso.

Expresso minha profunda gratidão à Universidade de São Paulo e à Escola Politécnica pela oportunidade de realizar este estudo. Agradeço especialmente à professora e minha orientadora Mariana, cuja paciência, esforço e constante apoio foram fundamentais para eu superar desafios, não desistir e persistir até o fim. Um agradecimento especial ao pessoal do laboratório de geoprocessamento da Escola Politécnica, com destaque para o Leonardo, pela ajuda indispensável ao longo dessa jornada.

Cada um de vocês desempenhou um papel crucial no meu sucesso acadêmico, e sou imensamente grato por todo o suporte e contribuições. Este trabalho não teria sido possível sem a presença e influência positiva de cada um de vocês. Muito obrigado.

RESUMO

Título: Análises espaço-temporais de viagens por transporte público a partir de matriz Origem-Destino gerada por meio de bilhetagem eletrônica.

Os sistemas de cobrança de tarifa automática estão sendo amplamente utilizados no transporte urbano. Suas principais vantagens incluem a fácil utilização dos passageiros, o aumento na eficiência do gerenciamento da receita, facilidade na integração com outros modos de transporte e coleta de dados, que contribuem no processo de planejamento de transportes. Este estudo realiza análises espaço-temporais das viagens por transporte público a partir de matriz Origem-Destino gerada por meio de bilhetagem eletrônica para a cidade de São Paulo. Para isso, estuda-se a aplicação do método de encadeamento de viagens *(trip-chaining method)* para estimar os locais de desembarque e inferir as origens e os destinos dos usuários. Após a construção da matriz OD, faz-se avalia os padrões de comportamento espaço-temporais das viagens, na identificação dos grupos homogêneos de usuários e na atribuição de informação semântica aos dados de bilhetagem com a perspectiva das desigualdades espaciais e dos grupos sociais, baseado nos dados de uso e ocupação do solo, do censo demográfico e da pesquisa OD domiciliar. Alguns conjuntos principais foram encontrados: casa-trabalho (com diferenças por classe social e por raça, menos transferências e distâncias percorridas para classes mais altas com predominância de raça/cor branca) e educação/saúde/assuntos pessoais (desigualdade no acesso à saúde e falta de acessibilidade da periferia). A variabilidade dentro dos clusters é alta, com perfis socioeconômicos distintos para usuários fora dos padrões. O estudo revela a complexa realidade da mobilidade em São Paulo e a necessidade de políticas públicas que considerem a diversidade dos padrões de viagem.

Palavras-chave: Planejamento de transportes; Dados de bilhetagem; Método de encadeamento de viagens; Padrões de comportamento de viagem; Desigualdades Socioespaciais.

ABSTRACT

Automatic fare collection (AFC) systems are being widely used in urban transportation. Its main advantages include easy usability by passengers, improvement of the revenue management efficiency, ease of integration with other transportation modes and data collection, which contributes to the transportation planning process. This study explores the application of Smart Card Data in transportation planning, focusing on the behavior pattern of public transportation passengers, including origin-destination (OD) inference for the city of São Paulo. This work starts with the application of the trip-chaining method (TCM) to estimate alighting and infer users' origins and destinations. After the construction of the OD matrix, this study evaluates the spatial-temporal patterns of travel, the identification of the homogeneous groups of users and the assignment of semantic information to the smart card data with the perspective of spatial and social groups inequalities based on the land use, the demographic census, and the household survey. Several main clusters have been identified: home-work (with differences by social class and race, fewer transfers and distances traveled for higher classes with a predominance of white race/color) and education/health/personal affairs (inequality in access to healthcare and lack of accessibility from the periphery). The variability within the clusters is high, with distinct socioeconomic profiles for users outside the norms. The study reveals the complex reality of mobility in São Paulo and the need for public policies that consider the diversity of travel patterns.

Keywords: Transport planning; Smart Card Data; Trip-chaining method; Travel behavior patterns; Socio-spatial inequalities.

SUMÁRIO

1. Introdução

O aumento na utilização de sistemas automáticos de cobrança de tarifa (bilhetes eletrônicos) nos sistemas de transporte tem proporcionado um volume massivo de dados de viagem dos usuários. Estes conjuntos de dados com resolução espacial e temporal, coletados durante um longo período de tempo, proporcionam a oportunidade de investigar melhor os padrões e comportamentos de viagem dos usuários de transporte público, além de subsidiar o planejamento de transportes (Nassir; Hickman; Ma, 2015; Kujala et al., 2018).

A identificação de grupos de comportamento de viagem é útil em uma série de aplicações. Pode auxiliar os planejadores de transporte público a entender o comportamento dos usuários e assim, oferecer um serviço mais adequado de acordo com sua demanda (Zhao et al., 2017). Do ponto de vista social, avaliar os diferentes grupos pode ajudar a mensurar e caracterizar as desigualdades de mobilidade enfrentadas pelos grupos menos privilegiados. As desigualdades socioespaciais se concretizam no ambiente urbano, onde os bens e serviços (saúde, educação) não são universalizados e a segregação residencial reduz o acesso às oportunidades (Bittencourt; Giannotti; Marques, 2021).

Historicamente, a obtenção de dados para o planejamento de transportes é realizada a partir de pesquisas de campo tradicionais, que fornecem uma quantidade razoável de informações sobre viagens, atividades realizadas e características socioeconômicas, em um período específico de tempo, o que as limita, pela curta validade dos dados, já que estes são estáticos e o padrão de viagens se modifica com certa velocidade (Pelletier; Trépanier; Morency, 2011). Comparativamente os dados de bilhetagem eletrônica permitem a obtenção de uma amostra maior captada pelo sistema automático (muitas vezes próxima ao universo), sendo que não são necessárias infraestruturas adicionais neste sistema, já que as informações são geradas a partir do uso do cartão inteligente no pagamento da tarifa, apesar de não captarem os atributos socioeconômicos e demográficos dos usuários (Devillaine; Munizaga; Trépanier, 2012).

A extração dos padrões de viagem dos passageiros de transporte público a partir de dados de cartões inteligentes pode ser particularmente desafiadora

dado que o sistema de coleta automática de tarifa (*AFC – Automatic Fare Collection*) não foi originalmente projetado para suportar o planejamento e as medidas de desempenho do transporte público (Pelletier; Trépanier; Morency, 2011). Como resultado, os dados do cartão inteligente coletados pelo sistema AFC impactam a demanda por processamento de dados.

Para alcançar o nível de detalhe desejado a partir das informações disponíveis, isto é, transformando o grande volume de dados coletado em informações valiosas, é necessário entender a melhor maneira de processá-los, com técnicas adequadas de mineração de dados. Existem desafios no uso dessa grande quantidade de dados, o seu tratamento depende de computadores com grande capacidade de processamento e é necessário ter um espaço para o armazenamento desses dados (Devillaine; Munizaga; Trépanier, 2012).

Em resumo, explorar os padrões de viagem dos dados de bilhetagem para superar as restrições mencionadas acima, proporcionando maior entendimento ao adicionar informações semânticas para caracterizar o usuário do sistema de transporte público é um grande desafio (Ma et al., 2013).

Este estudo explora os dados de bilhetagem eletrônica e a aplicação de métodos para a inferência e a avaliação dos padrões espaço-temporais de viagem dos usuários, adicionando informações socioeconômicas e motivacionais às viagens de transporte público produzidas em toda a cidade de São Paulo.

O primeiro desafio foi a construção da matriz Origem-Destino (OD) de viagens a partir dos dados de bilhetagem eletrônica e, para isso, foi utilizada a metodologia de encadeamento de viagens (*trip-chaining*), desenvolvida originalmente por Barry et al. (2002) e aplicada para estimar os locais de destino das viagens, já que, no caso do sistema de transporte público de São Paulo, registram-se apenas o registro em um momento da viagem.

O encadeamento de viagens é utilizado para completar a sequência de viagem do passageiro, conectando as pernas da viagem de cada usuário, com base no pressuposto de que é inconveniente para um passageiro percorrer uma longa distância para embarcar em um ponto de parada diferente do desembarque prévio. (Alsger et al., 2015; Alsger et al., 2016; Hora et al., 2017).

Para conectar essas pernas de viagens, é necessário distinguir novos embarques, entre o início de uma nova viagem e as transferências dentro da

mesma viagem. Um dos desafios do uso dos dados de bilhetagem eletrônica é diferenciar as transferências dos locais de atividade (Nassir; Hickman; Ma, 2015). Para isso, é utilizado um critério de diferença de horário entre o desembarque e o próximo embarque: se o próximo embarque ocorrer após esse critério de diferença de horário, infere-se que uma atividade foi realizada naquele período e uma nova viagem deve ser considerada. Se o tempo entre o desembarque e o próximo embarque for menor que esse critério de diferença de horário, será atribuída uma etapa de transferência a esse embarque (Alsger et al., 2016).

Além do encadeamento de viagens, a inferência do motivo de viagem é um componente fundamental na análise dos padrões de mobilidade dos passageiros e um dos atributos mais importantes ausentes do conjunto de dados de bilhetagem eletrônica. Enriquecer esses conjuntos de dados com informações sobre o motivo das viagens amplia as possibilidades desses dados para fins de planejamento e pesquisa. Os dados de bilhetagem eletrônica podem, por exemplo, ser complementados por outras fontes de dados para reduzir a necessidade de métodos caros de coleta de dados e inferir o motivo das viagens (Alsger et al., 2018).

Neste contexto, este estudo infere os motivos das viagens dos usuários com base em características temporais, preenchendo esse atributo ausente nos dados de bilhetagem eletrônica. Essa informação será usada em conjunto com outros dados provenientes da bilhetagem eletrônica para identificar grupos homogêneos de passageiros, permitindo uma análise aprofundada da mobilidade urbana e do planejamento de transporte.

1.1 Motivação

Os dados de bilhetagem contêm informações espaço-temporais inerentes, fornecendo uma oportunidade única para compreender os padrões de viagem e desigualdades socioespaciais enfrentadas pelos usuários de transporte público. Adicionalmente, observa-se um gap científico, principalmente na América Latina e no contexto do sul global, destacando a importância de usar *big data* nessas avaliações para captar a variabilidade de viagens e padrões de forma contínua, complementarmente às pesquisas tradicionais de Origem-Destino (OD). O uso de dados massivos, como os provenientes de cartões inteligentes, emerge como uma alternativa valiosa para gerar matrizes de viagem a um custo menor, com uma maior frequência e abrangência de usuários do sistema de transporte público.

1.2 Objetivos

Este trabalho tem como objetivo principal a análise dos padrões de deslocamento espaço-temporais dos usuários de transporte público no município de São Paulo, utilizando dados de bilhetagem eletrônica.

Adicionalmente, o trabalho estabelece os seguintes objetivos específicos:

- Explorar método para a inferência de motivos de viagens a partir de enriquecimento semântico e sequenciamento de viagens.
- Comparar análises dos dados da Pesquisa OD, baseada em método tradicional de coleta, com os resultantes dos processamentos de dados de bilhetagem eletrônica.
- Analisar as características espaço-temporais dos deslocamentos e sua variabilidade, tanto da população em geral como de forma particular na população de baixa renda;

Com esses objetivos, este estudo se propõe a aprofundar a compreensão dos padrões de mobilidade na cidade de São Paulo, contribuindo para a identificação de melhorias e políticas que possam beneficiar os usuários do

sistema de transporte público, com atenção especial àqueles em situação de vulnerabilidade econômica.

1.3 Estrutura do Texto

Esta pesquisa está organizada em cinco capítulos. O capítulo 1 consiste nesta introdução, o capítulo 2 apresenta a revisão da literatura dos assuntos relacionados à pesquisa em desenvolvimento. O capítulo 3 descreve as características do conjunto de dados utilizado e explica os métodos aplicados e adaptados. O capítulo 4 mostra a estruturação e o processamento do banco de dados ao longo do processo. Já o capítulo 5 apresenta os resultados das avaliações dos padrões de viagem. O capítulo 6 mostra as conclusões e as potenciais direções que podem seguir esta pesquisa. Ao final são apresentadas as referências bibliográficas utilizadas.

2. Revisão Bibliográfica

Neste capítulo serão apresentadas as abordagens desenvolvidas a partir de dados de bilhetagem eletrônica. Primeiramente, serão destacados aspectos gerais sobre o uso de dados de cartão inteligente, depois se tratará da metodologia de encadeamento de viagens e da caracterização dos padrões de viagem espaço-temporais. Ao final são destacados alguns estudos que conectaram as desigualdades sociais ao acesso por transporte público.

2.1 Uso de bilhetagem eletrônica

Sistemas de cobrança automática a partir de cartões estão sendo cada vez mais utilizados pelas agências de transporte público. Enquanto o principal motivo é cobrar de forma mais fácil o usuário, este sistema produz grandes quantidades de dados que podem ser úteis para os planejadores de transporte (Pelletier; Trépanier; Morency, 2011).

Com a base de dados das transações e os dados provenientes dos aparelhos GPS dos veículos, é possível coletar as seguintes informações: o número identificador do cartão (ID), as coordenadas geográficas no momento da transação, o modo de transporte, custo da transação, entre outros. Com base no ID de cada cartão, é possível reconstruir a cadeia de diária de viagens de um usuário. Com a informação de transações e registros de localização por GPS pode-se estimar os tempos de viagem e as transferências realizadas pelo usuário (Amaya; Munizaga, 2013).

Os estudos baseados em dados de cobrança eletrônica podem ser divididos em três grandes categorias: de nível estratégico, que está relacionado ao planejamento das redes de transporte público a longo prazo, análises de comportamento do usuário e previsão de demanda; de nível tático, que se relaciona ao ajuste de horários e padrões de viagem; e de nível operacional, que estuda os indicadores de oferta e demanda além da operação do sistema de bilhetagem (Pelletier; Trépanier; Morency, 2011).

Inicialmente as pesquisas anteriores baseadas em bilhetagem eletrônica sobre comportamentos dos usuários de transporte público se concentraram na estimativa de pontos de transferência e na inferência de origem e destino

(Munizaga; Palma, 2012). Pelletier et al. (2011) revisaram estudos anteriores relacionados ao planejamento de transporte de longo prazo e expôs que, a partir de dados de cobrança eletrônica como data/horário/local da transação, frequência do uso e dados históricos, é possível analisar os hábitos dos usuários para entender melhor seu comportamento, criar uma matriz origem-destino para ajuste da rede de transporte e comparação com a pesquisa origem-destino domiciliar para melhorar a precisão de ambas.

Morency, Trépanier (2007) utilizaram dados de bilhetagem eletrônica de vários dias para analisar a variabilidade das viagens dos passageiros de transporte público e apontaram que um melhor entendimento da variabilidade das viagens pode auxiliar na redução de custos operacionais e no gerenciamento a demanda. Lee e Hickman (2011) descobriram que os padrões de viagem de usuários regulares típicos, que fazem duas ou mais viagens durante os dias da semana, variavam de acordo com o tipo de cartão.

Também apresentando uma visão geral das pesquisas realizadas a partir de dados de bilhetagem eletrônica, Anda, Fourie e Erath (2016) deram exemplos em que os dados dos cartões inteligentes podem ser explorados para reconstruir jornadas individuais e apontam para o potencial de melhorar a caracterização dessas viagens com o aumento do significado semântico pela inferência do motivo da viagem. Focando nesta atribuição de uma informação semântica às viagens extraídas dos dados de bilhetagem eletrônica, Devillaine, Munizaga e Trépanier (2012) desenvolveram um método para alocar motivos nas viagens, categorizados em estudo, trabalho, casa e outros. Os critérios utilizados para a esta inferência foram: tipo de cartão (adulto, estudante, menor de idade), tempo na atividade, uso do solo no destino da viagem e posição da transação no dia.

De outra forma, alguns estudos foram baseados tanto em dados de pesquisa de cartões inteligentes quanto em dados de pesquisa domiciliar para explorar as viagens pendulares em Pequim na China (Zhou; Murphy; Long, 2014; Long; Thill, 2015). Zhou, Murphy e Long (2014) avaliaram e destacaram o potencial dos dados do cartão inteligente para rastrear os padrões de viagem no transporte público e concluíram que quanto mais difícil o acesso do usuário ao trabalho, maior a tendência de uso do automóvel (quando possível).

2.2 Padrões de Viagem – Bilhetagem eletrônica

Algoritmos de agrupamento (clusterização) como o k-means têm sido amplamente utilizados em análises de agrupamento de padrões de viagem na área de transporte público devido à sua capacidade de calcular conjuntos de dados relativamente grandes e exigir poucos parâmetros a serem utilizados (MA et al., 2013). Agard et al. (2013) usaram a clusterização por k-means para entender as diferentes categorias de usuários, especialmente aqueles que têm um padrão regular de viagem. No caso de Ortega-Tong (2013), o algoritmo de agrupamentos identificou 8 grupos que representaram usuários regulares, compostos por trabalhadores e estudantes que fazem viagens pendulares durante a semana, e alguns deles fazem viagens de lazer durante os finais de semana e usuários ocasionais, a partir de análise detalhada das características dos padrões de viagens espaciais e mudanças temporais.

Na avaliação dos padrões de viagem, Zhao et al. (2017) conduzem, a partir de dados de bilhetagens de Shenzhen (China), dois tipos de análise de padrões de viagens: uma baseada em estatísticas e outra em agrupamentos (clusterização). Na análise estatística, uma análise de regularidade é feita através dos três tipos de padrões separadamente (espacial, temporal e espaço-temporal). Na análise baseada em agrupamentos, apenas os padrões espaciais e temporais são avaliados separadamente usando o algoritmo K-means. Seus resultados são então correlacionados através de uma matriz de probabilidade condicional. Pelos resultados do estudo, Zhao et al. (2017) concluem que se um usuário é temporalmente regular, é muito provável que o usuário também seja espacialmente regular.

Yu e He (2016) propõem utilizar outro algoritmo de agrupamento, o DBSCAN, para extrair os padrões de viagem dos passageiros de transporte público fornecidos pela agência de trânsito de Guangzhou, identificando a Origem, o Destino e o horário regular que o usuário normalmente viaja. Ma et al. (2016) também usaram o DBSCAN, mas modificado para incorporar os atributos espaciais e temporais dos deslocamentos dos usuários extraídos de um encadeamento de viagens realizado, além do algoritmo de clusterização espacial ISODATA (iterative self-organizing data analysis technique) utilizado para categorizar passageiros em 3 grupos diferentes com base em características

espaciais e temporais. A partir disso, a metodologia permitiu identificar a residência e o local de trabalho de cada usuário frequente com uma taxa de acerto de até 94,1%, oferecendo informações para a gestão eficiente do transporte público e a compreensão dos padrões de deslocamento.

Outro método de agrupamento utilizado para analisar padrões de viagem é o dos mapas auto-organizáveis, denominado em inglês como SOM – *Self-Organizing Map* (Himanen; Järvi-Nykänen; Raitio, 2019; Pieroni et al., 2021), que foram extensivamente aplicados como um classificador não supervisionado de dados multiespectrais de sensoriamento remoto (Yan; Thill, 2009). Himanen et al. (1998) são os primeiros a explorar a aplicabilidade dos SOMs na identificação de padrões de viagens diárias. Pieroni et al. (2021) utilizaram SOM como um dos métodos de agrupamento para avaliar os padrões de viagem dos residentes de baixa renda com empregos de baixa remuneração e comparar estes usuários com residentes de outras áreas da cidade, analisando suas diferenças e semelhanças de deslocamento em toda a cidade.

Já Zhong (2015) abordou o assunto de padrão de viagem medindo a variabilidade em um nível individual e agregado utilizando dados de bilhetagem eletrônica de uma semana de Singapura. Os dados de cartão inteligente registraram o local de embarque e desembarque; portanto, o desembarque neste caso não foi inferido. Para a avaliação temporal, construiu-se um perfil do horário de início da viagem durante uma semana e uma matriz de correlação dos padrões temporais de cada dia. Para a análise espacial, uma rede espacial foi construída a partir de uma matriz Origem-Destino de viagens diárias. Zhong (2015) concluiu que, no caso de Cingapura, embora exista variabilidade espacial de padrões de viagem em escala individual e agregada ao longo da semana analisada, a estrutura espacial do movimento urbano permanece praticamente a mesma no período.

Sob outro enfoque, Lee e Hickman (2013) exploram o uso de dados de bilhetagem eletrônica para inferir o motivo das viagens e também revelar padrões de deslocamento em áreas urbanas. A metodologia envolveu a análise das informações temporais e espaciais dos usuários, além do desenvolvimento de regras para criar um conjunto de treinamento para derivar os motivos de viagem. Em seguida, foi conduzida uma técnica de classificação baseada em árvore de decisão com um conjunto de teste para determinar o desempenho do modelo, e

os resultados foram comparados com dados de pesquisa embarcada. Os resultados mostraram que a técnica de classificação baseada em árvore de decisão pode inferir com precisão o motivo da viagem usando dados de bilhetagem.

Alsger et al. (2018) buscam explorar o uso de dados bilhetagem eletrônica para inferir o propósito de viagens dos passageiros, com o objetivo de aprimorar a eficiência e custo-benefício em relação às pesquisas domiciliares de viagens. O estudo adota uma abordagem metodológica que integra várias fontes de dados, incluindo pesquisas domiciliares, informações de uso do solo e GTFS (General Transit Feed System), utilizando atributos espaciais e temporais para classificar o comportamento de viagem dos passageiros. Os resultados demonstram um aumento na precisão da inferência de motivo de viagem com a inclusão de atributos temporais, enquanto a clusterização dos usuários é empregada como um método de classificação adicional. A análise destaca a complexidade dos padrões de viagem e a importância de compreender a motivação dos passageiros para viajar, com a aplicação de regras de modelagem baseadas em atributos espaciais, temporais e de frequência, culminando na discussão dos resultados da inferência do propósito das viagens.

Já Faroqi e Mesbah (2021) propõem um método para inferir o propósito da viagem a partir de atributos temporais registros de cartões inteligentes e da pesquisa Origem-Destino domiciliar de Queensland, Austrália. O método proposto adota uma perspectiva baseada em indivíduos e leva em consideração a relação entre as viagens anteriores e subsequentes de um indivíduo durante o dia, a definindo um sequenciamento de viagens. Os resultados da avaliação mostram que o método proposto supera os métodos existentes em termos de precisão e sensibilidade, oferecendo uma abordagem promissora para inferir o propósito da viagem a partir de registros de cartões inteligente.

.

2.3 Bilhetagem eletrônica – Matriz OD

Para a criação de uma matriz Origem-Destino primeiramente é necessário saber o tipo do sistema de cobrança automática, que podem ser categorizados em 2 classes: sistema somente de entrada e sistema de entrada e saída.

No sistema somente de entrada, os usuários são obrigados a utilizar o cartão apenas ao embarcar no serviço de transporte público (ou ao entrar na estação no caso de sistemas ferroviários). No sistema de entrada e saída, os passageiros devem validar o cartão tanto no embarque quanto no desembarque do veículo de transporte público. A maioria dos estudos revisados possuem sistema somente de entrada, caso dos dados utilizados nesse estudo e, requer a inferência do local de desembarque dos passageiros.

Munizaga e Palma (2012) propõem e implementam uma metodologia para obtenção da matriz de transporte público por meio dos dados de bilhetagem baseado na estimativa do local de desembarque, para a cidade de Santiago, no Chile. A metodologia utilizada é aplicada a sistemas de transporte público multimodais em larga escala. O destino foi inferido como a mesma zona de origem da próxima viagem e, consideraram que a última viagem do dia terminava na zona de origem da primeira viagem. Neste caso, a máxima distância de caminhada permitida foi de 1000m.

No Brasil, Arbex e Cunha (2017) apresentam uma metodologia para analisar dados de bilhetagem eletrônica em sistemas de transporte sobre trilhos, com o objetivo de obter informações sobre a lotação e a matriz origem-destino dos usuários. A metodologia proposta inclui a utilização de dados de embarque em ônibus para auxiliar na inferência da estação de destino dos usuários. Os resultados obtidos mostraram que é possível obter informações importantes sobre o padrão espaço-temporal dos níveis de lotação e os volumes de transferências nas estações, além de permitir um mapeamento diário da demanda por transporte público.

Outro aspecto a ser considerado, para criação de uma matriz OD, é distinção entre transferências e atividades de curta duração. Os critérios utilizados variam desde limites de tempo simples entre transações sucessivas até critérios mais elaborados.

Devillaine 2012, Munizaga e Palma, 2012 adotam alguns critérios para detectar locais de atividade: quando o usuário utiliza a mesma linha de ônibus em duas pernas consecutivas, independente do sentido, é razoável assumir que não há razão para desembarcar e embarcar em uma mesma linha, a não ser a realização de alguma atividade. Critérios de tempo também são adotados para detectar atividade, diferentes limites de tempo são utilizados para o intervalo de tempo entre duas transações seguidas (abordado mais a frente).

Devido à limitação dos sistemas automatizados de cobrança eletrônica em registrar as informações de desembarque dos passageiros, os resultados da estimativa da matriz OD precisam ser reavaliados antes de seu uso e análise do comportamento de viagem dos indivíduos (Munizaga et al., 2014).

O método de encadeamento de viagens (*trip-chaining method*), descrito mais adiante na metodologia, é normalmente utilizado para construir a sequência de viagens de um passageiro, conectando as pernas de viagem realizadas pelo uso de seu bilhete eletrônico (Alsger et al., 2015). Alguns estudos tentaram avaliar este método e suas suposições. Farzin (2008) validou os resultados da matriz OD estimada para 2006, obtida a partir de 5% de todas as viagens de transporte público em São Paulo, com os resultados da pesquisa domiciliar de OD de 1997.

Hora et al. (2017) mostram a implementação da metodologia de encadeamento de viagens para estimar os locais de desembarque dos vários estágios de uma viagem em um estudo de caso da cidade de Porto, em Portugal. Para este estudo foram consideradas quatro premissas:

(i) Usuários iniciam sua próxima perna de viagem no local ou próximo ao local de desembarque anterior;
(ii) Usuários terminam sua última viagem do dia no local de embarque da primeira viagem;
(iii) Usuários desembarcam em algum ponto de parada ainda não percorrido por sua rota e;
(iv) Usuários possuem um tempo de transferência e uma distância de caminhada limitados entre duas pernas de viagem.

Ma et al. (2013) propõem um procedimento robusto de mineração dos dados de bilhetagem para extrair padrões de viagens e regularidade dos usuários. Duas questões principais são examinadas: primeiro, os padrões de

viagem espaciais e temporais, depois, se determina a regularidade deste padrão de viagem. Para realizar estas análises, se fez necessária a construção do encadeamento de viagens para construção da matriz OD. Neste caso, se utilizou um tempo de transferência máximo de 60min entre 2 transações para definir se pertencem à mesma viagem.

Alsger et al. (2015) avaliaram as premissas comuns do método de encadeamento de viagens usando um conjunto de dados de cartão inteligente obtido da autoridade de transporte público do sudeste de Queensland (SEQ), Austrália. A vantagem importante desse conjunto de dados para a avaliação das premissas do método de encadeamento de viagens é que ele inclui horários e locais de embarque e desembarque para cada passageiro dos serviços de transporte público que incluem ônibus, trens e balsas. O estudo concentrou-se em premissas individuais (tempo de transferência permitido, distância permitida a pé e último destino de um determinado dia igual à primeira origem daquele dia) do método de encadeamento de viagens, em uma situação em que as informações reais de embarque e desembarque eram conhecidas.

Nunes et al. (2016) apresentam uma nova metodologia para estimar os destinos das viagens de passageiros usando dados de sistemas de coleta automatizada de tarifas e técnicas de validação espacial. A metodologia envolve um processo de duas etapas: primeiro, um modelo probabilístico é usado para estimar a probabilidade de um passageiro viajar para cada destino possível com base em seu comportamento de viagem observado. Em segundo lugar, técnicas de validação espacial são usadas para verificar a precisão das probabilidades de destino estimadas e identificar possíveis resultados incorretos. Os resultados da aplicação dessa metodologia a dados de serviços de ônibus em Porto, Portugal, mostram que ela é eficaz para estimar os destinos das viagens de passageiros em um nível detalhado e confiável na detecção de resultados incorretos. As características de validação espacial também sugerem que suposições-chave presentes na literatura anterior no campo são amplamente válidas para o caso de Porto.

A Tabela 3 apresenta as premissas que cada estudo revisado utilizou para a construção da matriz Origem-Destino.

Conforme observado, inferir se uma transação realizada por um usuário é uma transferência, é um dos principais pontos da metodologia de encadeamento

de viagens. A diferenciação entre uma transferência ou uma atividade realizada pelo usuário é realizada aplicando restrições espaciais e temporais e observando se a transação realizada é a última do dia.

A restrição temporal é conhecida como tempo máximo de transferência (TMT) e a restrição espacial é denominada distância máxima de transferência (DMT). TMT é definido como a diferença de tempo entre um desembarque e o embarque consecutivo de um usuário de bilhete eletrônico dentro de um dia. DMT é a distância máxima a pé entre um desembarque e o embarque consecutivo, durante os quais se presume que um passageiro tenha feito a transferência de um serviço para outro.

Observando-se a literatura, o TMT varia de 18 minutos a 90 minutos. Huang et al. (2020) adotaram o TMT como um tempo inferior à frequência de viagem das linhas de transporte coletivo. Nassir et al. (2011) propuseram o TMT como um mínimo de 90 minutos ou o tempo necessário para caminhar da estação de desembarque para a estação de embarque, mais o tempo necessário para uma atividade menor. Se o tempo entre desembarques sucessivos e embarques for inferior ao tempo de atividade menor (30 minutos), é rotulado como transferência sem a aplicação de nenhuma outra regra. O mesmo princípio também é utilizado por Kumar et al. (2018).

Em vez de usar um único valor de TMT, Chu e Chapleau (2008) converteram a distância euclidiana entre paradas em tempo, dividindo a distância pela velocidade de caminhada (considerada como 4,3 km/h). Em continuação a esse trabalho, Gordon et al. (2013) usaram o TMT correspondente à distância euclidiana máxima de 750 m (assumindo uma velocidade de caminhada de aproximadamente 3 km/h), mais uma tolerância de 3 minutos, com o tempo máximo de espera para o ônibus considerado como 45 minutos.

Seaborn et al. (2009) consideraram uma faixa de tempo de transferência para transferências entre vários modos de transporte público. O estudo utilizou TMT de 15 a 25 minutos, 30 a 50 minutos e 40 a 60 minutos para transferência de metrô para ônibus, ônibus para metrô e ônibus para ônibus, respectivamente.

Na Tabela 1, apresentam-se os tempos máximos de transferência utilizados por cada um dos autores citados.

Tabela 1 – Tempos máximos de transferência na literatura

Estudo	Sistema de bilhetagem	Tempo máximo de transferência (min)
Barry et al. (2009)	Entrada	18
Munizaga e Palma (2012)	Entrada	30
Munizaga et al. (2014)	Entrada	30
Alsger et al. (2015)	Entrada/Saída	60
Nassir et al. (2015)	Entrada/Saída	60
Alsger et al. (2016)	Entrada/Saída	60
Nassir et al. (2011)	Entrada	90
Kumar et al. (2018)	Entrada	90
Huang et al. (2020)	Entrada	< frequência dos ônibus
Seaborn et al. (2009)	Entrada	Variável
Gordon et al. (2013)	Entrada	Variável

Fonte: Elaboração própria

Com relação à distância máxima de transferência (DMT), Li et al. (2011) afirmaram que essa distância depende do status econômico da cidade, cobertura de transporte público, etc. Vários valores para DMT utilizados por pesquisadores são resumidos na Tabela 2.

Alsger et al. (2016) utilizaram dados de um sistema de entrada e saída do Sudeste de Queensland (SEQ), Austrália. O estudo estimou o número de viagens considerando uma distância de transferência de 400, 800, 1000 e 1100 metros, utilizando apenas dados de bilhetagem no embarque. O número de viagens estimado foi comparado com o número real de viagens a partir dos dados do sistema de entrada e saída. Os autores relataram que, ao usar um tempo de transferência máximo de 60 minutos, a porcentagem de correspondência foi de 86% para 800 metros de distância máxima de transferência.

A Tabela 2 mostra que o DMT varia entre o valor mínimo de 400 metros (Zhao et al., 2017; Nassir; Hickman; Ma, 2015) e um valor máximo de 1500 metros (Huang et al., 2020) na literatura. Seaborn et al. (2009) e Gordon et al. (2013) utilizaram tempo de transferência variável para transferência entre modos diferentes. No entanto, nenhum dos estudos usou distância de transferência variável para transferência entre modos. A análise de Munizaga et al. (2014) mostra que a distância a pé depende do uso do solo ao redor dos pontos de embarque e desembarque.

Tabela 2 – Distâncias máximas de transferência na literatura

Estudo	Sistema de bilhetagem	Distância máxima de transferência (m)
Zhao et al. (2007)	Entrada	400
Nassir et al. (2015)	Entrada/Saída	400
Alsger et al. (2016)	Entrada/Saída	530
Gordon et al. (2013)	Entrada	750
Alsger et al. (2015)	Entrada/Saída	800
Munizaga et al. (2014)	Entrada	1000
Yan et al. (2019)	Entrada	1000
Huang et al. (2020)	Entrada	1500

Fonte: Elaboração própria

Para este estudo, as sete primeiras premissas apresentadas na Tabela 3 serão utilizadas na construção da matriz Origem-Destino de transporte público da cidade de São Paulo.

.

Tabela 3 – Premissas utilizadas na construção da matriz OD – Revisão dos estudos

Premissas Construção Matriz OD	Barry (2002)	Farzin (2008)	Nassir et al. (2011)	Munizaga e Palma (2012)	Gordon et al. (2013)	Ma et al. (2013)	Hora et al. (2017)
(1). Passageiros iniciam a próxima viagem no local de desembarque da viagem anterior ou próximo a ele;	✓	✓	✓	✓	✓	✓	✓
(2). Passageiros terminam a última viagem do dia no local de embarque da primeira viagem do dia;	✓	✓	✓	✓	✓	✓	✓
(3). Os passageiros não usam outros modos de transporte em segmentos consecutivos de viagens em trânsito;		✓	✓				
(4). Limite de distância a pé de transferência: os passageiros não andam mais do que um limite para transferir;		✓	✓	✓	✓		✓
(5). O horário do local estimado de desembarque deve ser anterior ao horário da próxima transação;			✓				✓
(6). Limite de tempo de transferência (um passageiro tem um tempo máximo para fazer a transferência);			✓		✓	✓	✓
(7). Os passageiros que viajam para longe de seus próximos locais de embarque são descartados;					✓		
(8). Um local de desembarque só é inferido quando há um registro AVL compatível para aquela viagem naquele momento;					✓		

2.4 Desigualdades socioespaciais

As desigualdades estão relacionadas às diferenças sociais e espaciais entre indivíduos e principalmente grupos sociais. As desigualdades se concretizam no ambiente urbano, onde a segregação residencial reduz o acesso às oportunidades (Bittencourt; Giannotti; Marques, 2021).

O acesso às oportunidades é distribuído desigualmente no território e entre os grupos sociais. Famílias de baixa renda e minorias étnicas sofrem com desvantagens no transporte – a falta de opções no transporte público e privado e outras dificuldades para a realização das viagens (custos financeiros, falta de segurança, etc.) (Lucas et al., 2016).

Esses problemas são ainda mais visíveis nas áreas periféricas de baixa renda, onde o transporte escasso, combinado com a falta de oportunidades, tende a acentuar as desigualdades socioespaciais. Ao mesmo tempo que essas limitações às oportunidades revelam as desigualdades, elas também agem de forma a aprofundar e reforçar estas dificuldades (Pereira; Braga; Serra, 2019).

Alguns estudos recentes relacionaram as desigualdades sociais com o planejamento de transporte. Martens (2012) foca seu estudo nas teorias de justiça para avaliar as desigualdades a partir do transporte e da acessibilidade dos usuários. Também a partir de uma abordagem teórica, Schwanen et al. (2015) avaliam as ligações estre a exclusão social e as desvantagens no transporte por meio da lente do capital social. Currie (2010) utilizou dados GTFS de ônibus e trens em Melbourne, Austrália e avaliou parâmetros para quantificar as lacunas espaciais na oferta de transporte público com base nas necessidades sociais. Nieuwenhuis et al. (2019) utilizaram dados nacionais e de censos de alguns países da Europa para investigar a mobilidade socioespacial definida como o movimento de pessoas entre diferentes bairros residenciais com distintos níveis de privação social. Já Bittencourt et al. (2020), utilizando dados demográficos e socioeconômicos do Censo de 2010 do IBGE, dados GTFS de transporte público e tempos de viagem, exploraram como as diferenças de escala, geografia, raça e classe social estão relacionadas à segregação espacial, levando a diferentes níveis de acessibilidade em 4 cidades brasileiras – São Paulo, Rio de Janeiro, Curitiba e Fortaleza.

Zhao e Cao (2020) tiveram como objetivo analisar a desigualdade nos tempos de deslocamento de passageiros em Xangai, uma megacidade, usando dados de cartões inteligentes de transporte público. Focando principalmente nas viagens de metrô, que representaram a maioria das viagens, o estudo identificou que cerca de 20% dos passageiros em Xangai enfrentavam longos deslocamentos diários, gastando mais de 60 minutos em cada sentido. A análise geograficamente ponderada revelou que áreas desfavorecidas, caracterizadas por baixos aluguéis e baixa acessibilidade ao trabalho, abrigavam uma alta concentração de trabalhadores com deslocamentos longos. Surpreendentemente, áreas com grandes populações de migrantes tinham menos viagens longas, indicando a importância do desenvolvimento de áreas habitacionais e industriais nos subúrbios para melhorar a correspondência entre empregos e habitação para migrantes de baixa renda. Além disso, o estudo destaca a necessidade de políticas de transporte inclusivas em megacidades em crescimento, enfatizando o acesso a serviços de transporte público.

Basso et al. (2020) tiveram como objetivo avaliar a acessibilidade do sistema de transporte público de ônibus em Santiago, Chile, em relação a áreas de educação, saúde e emprego. Utilizando dados de GPS em tempo real de 6.681 ônibus, o estudo calcula os tempos de viagem com base nas operações reais dos ônibus, proporcionando uma avaliação precisa da acessibilidade às oportunidades. Os resultados revelam que o transporte público não consegue mitigar eficazmente as desigualdades resultantes da distribuição geográfica das oportunidades na cidade. Enquanto a acessibilidade a oportunidades públicas de emprego é relativamente homogênea, as oportunidades privadas mostram maior disparidade. O estudo enfatiza a importância de políticas destinadas a reduzir essas desigualdades, destacando a necessidade de melhorar a qualidade dos serviços de transporte público, expandir a infraestrutura dedicada e planejar o desenvolvimento urbano para garantir uma distribuição equitativa das oportunidades em toda a cidade.

Pieroni et al. (2021) abordam a análise de padrões de viagem de 8 regiões (4 áreas precárias e 4 áreas de classe média/alta) por meio da mineração de dados de cartões inteligentes. Utilizando dados de usuários de transporte público de São Paulo, a metodologia emprega a análise de agrupamento para identificar padrões de viagem. O estudo identifica um agrupamento com características de

trabalhadores manuais de baixa renda que trabalham em áreas residenciais de alta renda e que, em média, registram suas primeiras bilhetagens do dia duas horas mais cedo do que o grupo de média/alta renda.

3. Dados e Método

Neste capítulo serão apresentados os dados que serão utilizados no desenvolvimento deste estudo, assim como o método a ser explorado.

3.1 Dados

A cidade de São Paulo, área de estudo adotada neste trabalho, pode ser considerada o centro financeiro, corporativo e mercantil da América do Sul e é uma das cidades mais populosas das Américas, com 12.106.192 habitantes (Estimativa de 2017, IBGE).

Em 2017, segundo dados da Secretaria Municipal de Mobilidade e Transportes, foram registrados 2.864.266.074 de passageiros transportados, cerca de 7 milhões de pessoas por dia. Para melhor compreender a dimensão dos dados do cartão inteligente obtidos, de acordo com a SPTrans - empresa responsável pela gestão do sistema de transporte público de ônibus em São Paulo - a cidade opera com mais de 1.300 linhas de ônibus e mais de 14.000 veículos. Os registros de GPS dessa frota são obtidos a cada 40 segundos, resultando em aproximadamente 26 milhões de registros diários de GPS (Arbex; Alves; Giannotti, 2016).

Os dados dos cartões inteligentes e GPS utilizados nesta pesquisa foram fornecidos pela SPTrans. Os dados de bilhetagem eletrônica analisados foram de 1 dia útil (03 de agosto de 2016), dentro do intervalo de tempo de 30 de maio a 14 de agosto de 2016. Isso representou aproximadamente 3,58 milhões de cartões inteligentes diferentes no sistema, com mais de 12,6 milhões de transações no dia.

Em São Paulo, o usuário de transporte público utiliza o cartão somente depois que embarca no transporte público. Os dados obtidos incluem:

- Dados do cartão inteligente em todo o sistema de transporte público, incluindo estações de metrô, trem e ônibus da cidade (2016);
- Dados do Sistema de Posicionamento Global (GPS) contendo a localização dos veículos do serviço de ônibus urbano (2016);
- Estrutura da rede de transporte público da região no formato GTFS (General Transit Feed System), que contém informações como pontos de

ônibus e localização das estações, frequências e rotas de ônibus, trens e linhas de metrô (2016);

- Dados de uso do solo da cidade de São Paulo (2022);
- Dados do Censo Demográfico do IBGE para a cidade de São Paulo em relação à classe social e raça da população (2010);
- Dados da pesquisa Origem e Destino domiciliar do metrô de São Paulo (2017);

Para cada transação, os principais atributos a seguir estão disponíveis: Data e hora da transação de embarque; ID do cartão e tipo; Número e direção da linha de ônibus; Estação ou terminal; Número do veículo.

3.2 Método

Este capítulo apresenta o método utilizado para a construção da matriz Origem-Destino de viagens dos usuários de transporte público e posterior avaliação dos padrões de viagens espaço-temporais na cidade de São Paulo. Essa sequência da metodologia foi baseada na literatura revisada no capítulo 2 e no conjunto de dados disponíveis.

Alguns conceitos e definições que serão utilizados ao longo do texto estão definidos a seguir, para facilitar a compreensão da metodologia.

Define-se uma viagem como um deslocamento de um ponto de origem a um ponto de destino (De Ortuzar; Willumsen, 2011). Cada viagem pode ter um ou mais segmentos, que são deslocamentos em um serviço específico (ônibus, trem ou metrô).

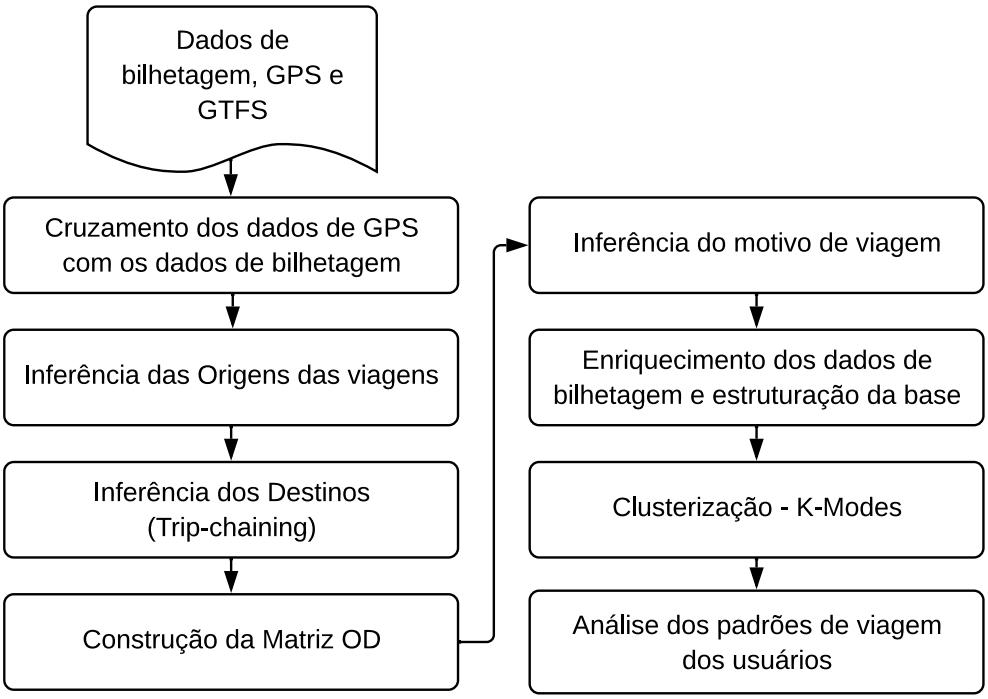
Origem e destino são os locais onde a viagem começa e termina, respectivamente. Pontos de embarque e desembarque são os locais onde os segmentos de viagem começam e terminam.

A transferência refere-se ao ato em que um passageiro muda de um veículo ou linha para outro durante uma única viagem. Esse processo envolve a transição entre diferentes modos de transporte (por exemplo, ônibus para metrô) ou a troca entre diferentes linhas dentro do mesmo modo (por exemplo, trocar de ônibus em ponto de parada).

A Figura 1 apresenta o fluxograma da metodologia prevista para este trabalho. O método para a avaliação dos padrões de viagem dos usuários possui

as seguintes etapas: detecção de transferência ou atividade para cada transação, construção da matriz OD a partir do encadeamento de viagens (*trip-chaining*), clusterização e avaliação dos padrões de comportamento. Todos os tratamentos de dados e desenvolvimentos deste trabalho foram realizados utilizando o software QGIS 3.6, Tableau Desktop, PTV-Visum e códigos em Python e R.

Figura 1 – Fluxograma da metodologia do estudo



Fonte: Elaboração própria

3.2.1 Construção da Matriz OD

Com base na revisão bibliográfica apresentada foi realizada a construção de uma matriz Origem-Destino (OD) de viagens de transporte público a partir dos dados de bilhetagem eletrônica. Para isso, uma série de processamentos foram adotados conforme descritos a seguir.

A primeira fase para o processamento da base de dados é a limpeza de dados. O processo envolve várias etapas, como validação, filtragem e transformação de dados. A validação de dados verifica a consistência e a completude dos dados, enquanto a filtragem de dados remove dados irrelevantes ou incorretos. Já a transformação de dados converte dados brutos em um formato mais utilizável e significativo. No caso dos dados de bilhetagem eletrônica, a limpeza de dados deve ser feita devido a falhas em equipamentos, leitores de cartões com defeito, relógios dessincronizados nos dispositivos de coleta de dados, falhas no aparelho de GPS e também a erros humanos, incluindo esquecimento de tocar o cartão, tocar o cartão errado, etc.

Zhao et al (2017) filtraram passageiros com menos de oito dias de uso do metrô ao longo de um mês de análise. Ma et al. (2013) filtraram sua base utilizando o critério de no mínimo 2 transações por cartão na semana analisada.

A base de bilhetagem eletrônica para o dia escolhido para este estudo contém 12.622.863 transações, considerando todo o sistema de bilhetagem de São Paulo (município e rede ferroviária que atende a grande São Paulo). O primeiro filtro realizado nessa base foi o de tipo de cartão utilizado. No sistema de transporte coletivo de São Paulo existem cartões do tipo ‘OPERAÇÃO’, esses cartões não são de usuários comuns e, portanto, as transações realizadas com esse tipo de cartão foram retiradas da base de dados, totalizando cerca de 4,75% da base.

Os dados de bilhetagem eletrônica obtidos para este trabalho não possuem informação de localização. Portanto, após a limpeza da base, é preciso estimar a posição de cada transação registrada de cartão utilizando dados de GPS. Os critérios da estimativa de embarque para os usuários de ônibus foram realizados com base em critérios delineados no trabalho de Arbex e Cunha (2017).

O modo da estimativa de localização dos embarques nos ônibus é realizado através da filtragem de todas as transações de cartão inteligente (dados de bilhetagem) registradas para cada veículo no banco de dados, para um dia.

Os horários destas bilhetagens são cruzados com os horários de todos os registros de GPS deste veículo neste dia, obtendo, para cada bilhetagem, o registro de GPS daquele veículo correspondente à menor diferença de tempo entre o horário que ocorreu a bilhetagem e o momento que o veículo emitiu a informação.

Para o sistema ferroviário (metrô e trem) e terminais de ônibus, uma associação de todos os embarques às respectivas estações/terminais é feita anteriormente por meio de uma tabela com os códigos correspondentes. Associa-se previamente todos os registros de embarque nos validadores às suas respectivas estações por meio de uma tabela auxiliar (validador – prefixo veículo), conforme o exemplo nas Tabela 4 e Tabela 5. Neste exemplo, o usuário de número de bilhete 3337074 fez sua transação das 13:35 do dia 30/05 no terminal Vila Prudente do Expresso Tiradentes.

Tabela 4 – Exemplo da base de bilhetagem eletrônica

data_hora	hash_bilhete	linha_validador	prefixo_veiculo	tipo_cartao
30/05/2016 09:06:03	2554789	933	45127	100-8
30/05/2016 08:18:28	4337191	933	45127	100-10
30/05/2016 08:17:48	3161653	933	45127	100-10
30/05/2016 09:06:19	2752530	933	45127	100-9
30/05/2016 09:49:01	4525585	933	45127	100-10
30/05/2016 08:18:03	3577784	933	45127	100-10
30/05/2016 09:04:10	3332493	933	45127	100-9
30/05/2016 08:18:18	3568738	933	45127	100-10
30/05/2016 13:35:47	3337074	1837	953982	100-10

Fonte: Elaboração própria

Tabela 5 – Exemplo da tabela auxiliar – estações e terminais

Validador	Estação	Modo	Lon	Lat
33201	9926PR-4310/10 - TER. DOM PEDRO II	Pré Embarque	-46.6302	-23.5467
33202	9926PR-4310/10 - TER. DOM PEDRO II	Pré Embarque	-46.6302	-23.5467
33993	9925PR - 4310/10 EST. TRANSF. ITAQUERA	Pré Embarque	-46.4541	-23.5352
500308	JUD- São Judas	Metrô 1	-46.6409	-23.626
500401	SAU - Saúde	Metrô 1	-46.6392	-23.6185
500606	SCZ- Santa Cruz	Metrô 1	-46.6367	-23.5991
500607	SCZ - Santa Cruz	Metrô 1	-46.6367	-23.5991
953982	EXP. TIRADENTES-TER. V. PRUDENTE	Pré Embarque	-46.5854	-23.5841

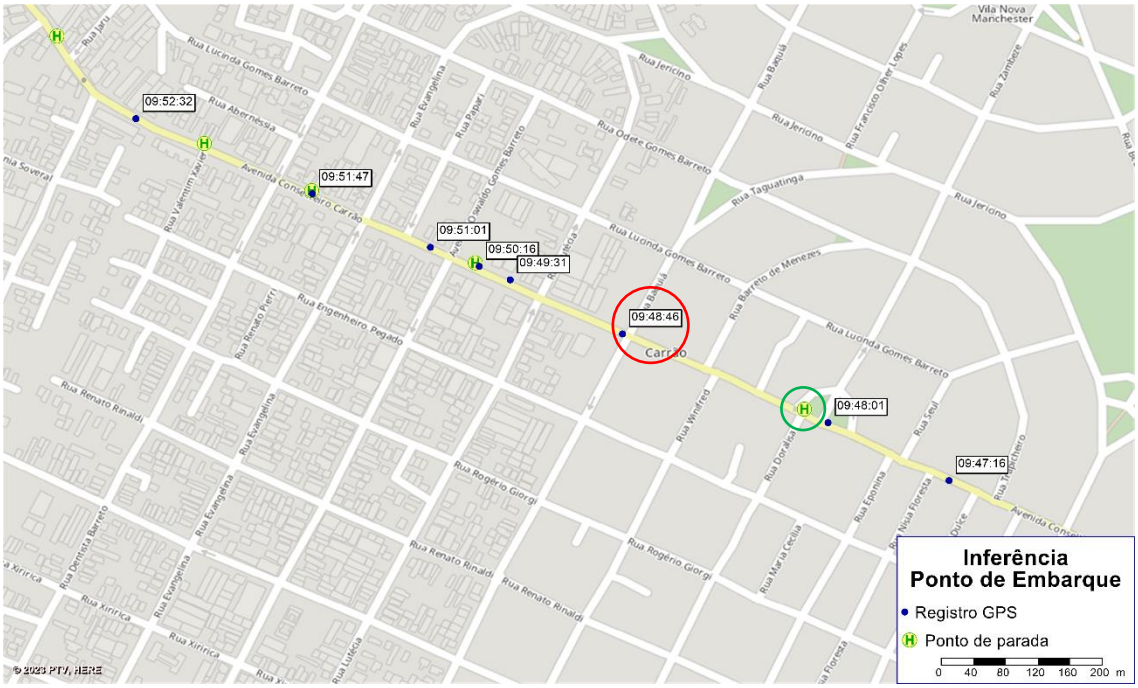
Fonte: Elaboração própria

Para os ônibus, esse processo é executado por cada veículo no banco de dados utilizado no dia de análise. Isso assegura que apenas as transações efetuadas em um veículo específico em um dia particular da análise sejam consideradas para a estimativa. Então, o horário exato dessas transações de cartões inteligentes (Tabela 4 acima) é então correlacionado com os registros de GPS de daquele veículo naquele dia, resultando no registro de GPS correspondente ao menor intervalo de tempo para cada transação de cartão inteligente.

Em seguida à aferição da localização dos embarques, faz-se a estimativa do ponto de parada dos embarques. Para isso, para cada linha de transporte coletivo, se utiliza os dados de GTFS, mais precisamente o arquivo stop.txt, que contém os pontos de parada para cada linha e sentido. A partir da menor distância entre a localização inferida dos embarques e o ponto de parada anterior à localização da transação, estima-se o ponto de parada em cada embarque ocorreu.

Um exemplo da inferência da localização da transação e do ponto de embarque está na Figura 2. Observa-se os registros de GPS ao longo de uma linha de transporte coletivo, os seus respectivos horários e também a localização dos pontos de parada dessa linha e sentido ao longo desse trecho. Portanto, se uma transação realizada por um usuário nessa linha aconteceu às 09:49:00, o registro de GPS associado à transação será o circulado em vermelho (às 09:48:46). Já o ponto de embarque inferido para essa transação será o circulado em verde, pois é o ponto de embarque anterior mais próximo.

Figura 2 – Exemplo de inferência do ponto de embarque



Fonte: Elaboração própria

Para a inferência dos destinos das viagens, é aplicado o método de encadeamento de viagens (trip-chaining method) que consiste em conectar as pernas de viagem de cada usuário, baseado na hipótese que o passageiro não caminha longas distâncias para embarcar em um novo ponto de parada.

No método de encadeamento de viagens, onde apenas as transações de embarque são registradas e conhecidas (caso deste estudo), as localizações de desembarque são inferidas a partir da localização de embarque da viagem subsequente. Nesse contexto, se encadeamos os embarques do usuário ao longo de todo o dia de maneira cíclica, a localização de desembarque de uma viagem deve ser espacialmente restrita à localização de embarque da viagem sucessiva.

A maioria dos pesquisadores utilizou o método de encadeamento de viagens para estimar o local ou tempo de desembarque de uma viagem realizada em um sistema de entrada única, como discutido anteriormente. O método de encadeamento de viagens aplicado neste estudo foi utilizado para estimar o local de desembarque e usa as suposições citadas na revisão bibliográfica para associar viagens feitas por um passageiro usando um cartão específico. As suposições são as seguintes:

1. Passageiros iniciam a próxima viagem no local de desembarque da viagem anterior ou próximo a ele;
2. Passageiros terminam a última viagem do dia no local de embarque da primeira viagem do dia ou próximo a ele;
3. O horário do local estimado de desembarque deve ser anterior ao horário da próxima transação;
4. Um local de desembarque é inferido quando há um registro AVL compatível para aquela viagem naquele momento;

A primeira suposição também é conhecida como a suposição de continuidade. Isso implica que um passageiro não optaria por outro modo durante suas jornadas diárias. Em outras palavras, para um passageiro, o ponto de destino da viagem anterior pode ser encontrado a uma curta distância a pé do próximo ponto de embarque, neste caso, uma distância máxima de 1000m foi utilizada. A suposição de continuidade reduz o número de paradas candidatas para a estimativa da parada anterior de desembarque. O número de paradas candidatas é ainda filtrado selecionando apenas aquelas paradas que atendem à última linha embarcada. Portanto, facilitando o processo de identificação da parada anterior de desembarque.

A segunda suposição também é conhecida como a suposição de simetria diária da viagem. Pode ser compreendida pelo fato de que o ponto de embarque da primeira viagem e o ponto de desembarque da última viagem de um dia geralmente estão próximas (Trépanier; Tranchant; Chapleau, 2007). Essa suposição é utilizada para inferir os locais de desembarque da última transação de cada usuário.

Após obter as estimativas de localização dos pontos de desembarque, é necessário agora determinar os horários em que esses desembarques ocorreram. Para as linhas de ônibus, esse processo envolve associar cada ponto de desembarque ao registro GPS mais próximo da linha de transporte coletivo que o usuário está embarcado, minimizando a diferença de distância. O carimbo de data/hora desse registro AVL é então atribuído ao momento do desembarque. Importante destacar que o momento do desembarque deve estar após o embarque na mesma linha e anterior ao próximo embarque, caso este ocorra. Nos casos em que não há um registro AVL correspondente, a velocidade média da linha de ônibus é utilizada para estimar o horário de desembarque.

No caso do sistema ferroviário, o ponto de desembarque será a estação mais próxima do novo ponto de embarque (Arbex; Da Cunha; Speicys, 2021). Para tal, os tempos de viagem são calculados com base em uma velocidade média dos trens e um tempo médio de parada nas estações. Com esse propósito, foi construída uma matriz de tempos de viagem entre todas as 189 estações de trem e metrô e do expresso Tiradentes, registradas nos dados de bilhetagem. A matriz tem um tamanho total de 35.721 registros (189 x 189).

Com relação a distância máxima de caminhada para a inferência do destino, encontra-se na literatura uma variação desde 530m (Alsger et al., 2016) a 2000m (Trépanier; Tranchant; Chapleau, 2007). A Tabela 6 apresenta a relação dos estudos com a distância máxima de caminhada utilizada por cada um. Essa diferença pode ocorrer dependendo do terreno ou restrições geográficas, dependendo da cidade. Neste estudo, se utilizará a distância máxima de 1000m, também utilizada por Munizaga et al., 2014, Munizaga e Palma, 2012, Wang et al., 2011 e Yan et al., 2019.

Tabela 6 – Distância de caminhada para inferência de destino em vários estudos.

Estudo	Sistema de bilhetagem	Cidade de estudo	Distância de caminhada (m) (inferência de destino)
Trépanier et al. (2007)	Entrada	Quebec, Canadá	2000
Nassir et al. (2011)	Entrada	Queensland, Austrália	800
Wang et al. (2011)	Entrada	Shenzhen, China	1000
Munizaga e Palma (2012)	Entrada	Santiago, Chile	1000
Gordon et al. (2013)	Entrada	Londres, Reino Unido	750
Munizaga et al. (2014)	Entrada	Santiago, Chile	1000
Nunes et al. (2016)	Entrada	Porto, Portugal	640
Yan et al. (2019)	Entrada	Shenzhen, China	1000
Huang et al. (2020)	Entrada	Suzhou, China	500 a 1500
Alsger et al. (2015)	Entrada/Saída	Queensland, Austrália	800
Alsger et al. (2016)	Entrada/Saída	Queensland, Austrália	530

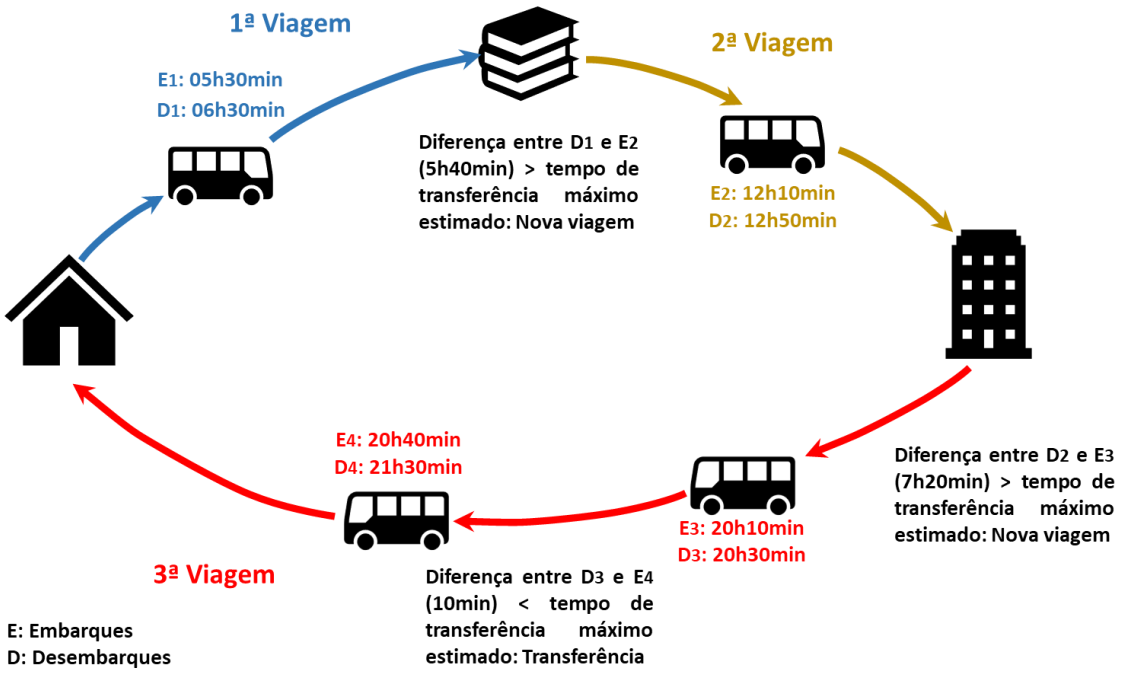
Fonte: Elaboração própria

A partir da estimativa de todos os locais de embarque e desembarque dos usuários e seus respectivos horários, é necessário construir as viagens e, para isso, novos embarques devem ser distinguidos entre o início de uma nova viagem e as transferências dentro da mesma viagem. Como já dito anteriormente, um dos desafios do uso de conjuntos de dados de cartão inteligente é distinguir os locais de transferência dos locais de atividade (Alsger et al., 2016; Nassir; Hickman; Ma, 2015). Para isso, é utilizado um critério de

diferença de horário entre o desembarque e o próximo embarque e também uma distância máxima de caminhada: se o próximo embarque ocorrer após o critério de diferença de horário ou a distância de caminhada do usuário entre o ponto de desembarque e o próximo ponto de embarque for maior que o critério adotado, infere-se que uma atividade foi realizada naquele período e uma nova viagem deve ser considerada. Se o tempo entre o desembarque e o próximo embarque for menor que esse critério de diferença de horário e a distância de caminhada for menor do que distância máxima de transferência, será atribuída uma etapa de transferência a esse embarque.

Este critério de diferença de horário (tempo de transferência) será estimado a partir dos tempos de caminhada entre o local de desembarque e próximo embarque e o tempo médio de espera dos usuários da próxima linha (intervalo de saída dos ônibus), explicados adiante. A Figura 3 apresenta o um exemplo do método a ser utilizado. Já a Figura 4 apresenta um fluxograma com a metodologia completa para o encadeamento de viagens.

Figura 3 – Exemplo fictício de encadeamento de viagens

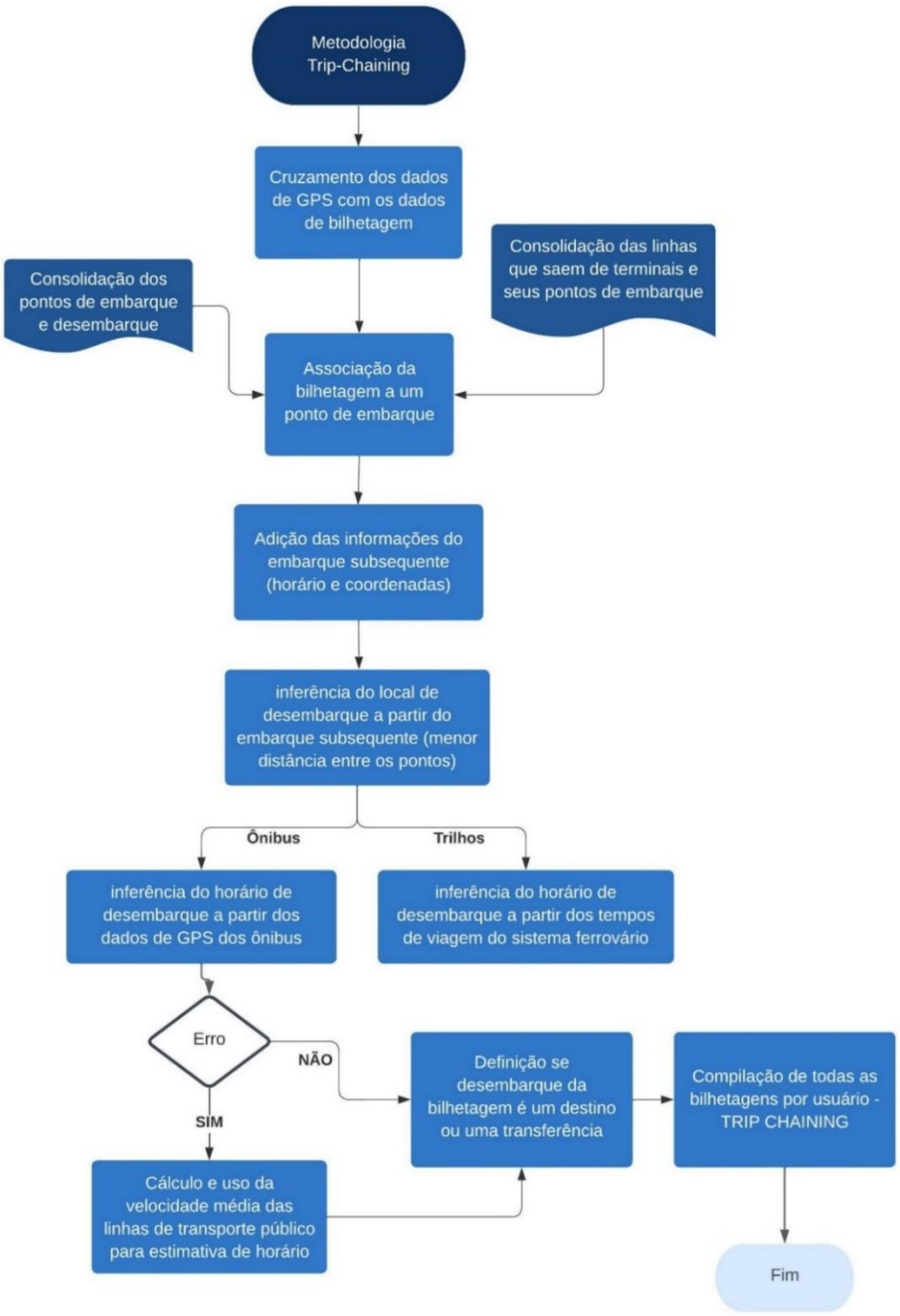


Fonte: Elaboração própria

Neste exemplo, um usuário faz seu primeiro embarque em E_1 e desembarca primeiro em D_1. Como o seu segundo embarque (E_2) acontece após

5h40min do primeiro desembarque, considera-se que este tempo é maior que o tempo máximo de transferência estimado e, portanto, o usuário faz uma atividade neste local e inicia uma nova viagem em E_2. Já no intervalo entre o D_3 e E_4, o tempo calculado foi de 10 minutos, tempo menor que o estabelecido como máximo e, portanto, o usuário realiza uma transferência neste local, continuando sua viagem.

Figura 4 – Método de encadeamento de viagens



Fonte: Elaboração própria

Durante esse processo, algumas filtragens foram realizadas, incluindo aproximadamente 1,44% das transações que não possuíam uma coordenada atribuída e, consequentemente, foram excluídas da base de dados. Foram ainda removidos da base de bilhetagem os usuários que apresentavam apenas uma transação no dia, correspondendo a aproximadamente 3,5% da base. Em seguida, uma filtragem foi aplicada para excluir erros na inferência do local de desembarque e nos horários associados a esses eventos. Transações que não apresentavam coordenadas atribuídas adequadamente ou que possuíam discrepâncias significativas entre o horário de embarque e desembarque foram excluídas, representando uma parcela de 21,0% do total de transações. Adicionalmente, procedeu-se à exclusão de todos os registros dos usuários que possuem um embarque ou desembarque fora do município de São Paulo, em algum momento do dia. Essa filtragem representou 4,0% das transações.

A seguir, serão definidos os parâmetros para que seja possível classificar um local de destino como uma transferência ou uma atividade.

3.2.1.1 Detecção de transferências

Como previamente abordado na revisão bibliográfica, um ponto importante da metodologia de encadeamento de viagens é a capacidade de determinar se uma transação realizada por um usuário corresponde a uma transferência. Essa distinção entre transferência e atividade é alcançada por meio da aplicação de restrições temporais e espaciais, no caso, o tempo máximo de transferência (TMT) e a distância máxima de transferência (DMT).

Com relação ao tempo máximo de transferência, neste estudo, se utiliza um tempo variável, utilizado por Gordon (2013), Seaborn (2009) e Huang (2020), de acordo com a intervalo de viagem da linha de transporte coletivo utilizada pelo usuário e o tempo de caminhada entre o ponto de desembarque anterior e o novo ponto de embarque.

Primeiro, calculou-se o intervalo de viagem médio para todas as linhas de ônibus presentes nos dados GTFS, hora a hora. Esse dado foi unido ao banco de dados de bilhetagem para cada transação, a partir da linha subsequente utilizada pelo usuário. Além desse tempo, converteu-se a distância de

Haversine[1] entre o ponto de desembarque e o ponto de embarque posterior em tempo, dividindo a distância pela velocidade de caminhada (considerada como 3,6 km/h ou 1m/s, valor intermediário entre os utilizados por Chu e Chapleau, 2008 e Gordon et al., 2013). Como exemplo, se um usuário desembarcou em um local às 12:30, andou 360 metros até o ponto de embarque seguinte, embarcou às 12:50 e a linha de ônibus que ele utilizou em seguida tem intervalo de 20 minutos, ou seja, uma frequência de 3 viagens por hora, o tempo máximo de transferência (TMT) dado a esta transação é de:

$$TMT = Tempo_{intervalo} + Tempo_{caminhada} \quad (1)$$

$$TMT = 20min + 6\,\mathrm{min} = 26min \quad (2)$$

Já para a distância máxima de transferência (DMT) utilizada neste estudo foi definido um valor fixo de 1000m, baseado na literatura (Munizaga et al., 2014; Yan; Yang; Ukkusuri, 2019).

Portanto, para o exemplo descrito acima, a distância de caminhada do usuário é menor que DMT e, o tempo entre o embarque e o desembarque anterior, 20 minutos, é menor que o TMT calculado, essa transação é considerada uma transferência.

A última restrição para detectar uma transferência é aplicada ao comparar a linha de transporte público utilizada no segmento de viagem anterior e no segmento de viagem atual. Se a linha de transporte público for a mesma para segmentos de viagem sucessivas, o desembarque é considerado uma atividade, sem considerar o TMT e o DMT. Vale ressaltar que, a última transação de cada usuário também tem como uma atividade o seu desembarque.

Uma vez que as transferências são detectadas, o próximo passo é identificar as atividades que ocorrem entre as transferências, isto é, identificar que o local de desembarque daquela transação realizada pelo usuário é o destino de sua viagem.

Em seguida, as atividades são vinculadas para formar viagens, o que é conhecido como encadeamento de viagens. Após a formação das viagens, o

[1] A distância mais curta entre dois pontos (ou seja, a 'distância em círculo máximo' ou 'em linha reta'), de acordo com o 'método de haversine'. Este método pressupõe uma Terra esférica, ignorando os efeitos elipsoidais (Sinnott, 1984).

próximo passo é a estimativa de Origem-Destino (OD) para o sistema de transporte público. A matriz OD representa o número de passageiros viajando entre cada par de zonas de origem e destino na rede de transporte público.

Após realizar o encadeamento de viagens para todos os usuários dentro de um dia de validações, se verifica a validade das cadeias. Para isso, alguns parâmetros serão verificados:

- Distância do local da transação ao ponto de parada mais próximo da linha de ônibus deve ser menor que 400m (Munizaga e Palma, 2012);
- O tempo total de viagem deve ser menor que 6 horas (tempo entre o primeiro embarque e o último desembarque de uma viagem);
- A distância total percorrida deve ser maior que zero;
- A distância mínima entre a primeira parada (de embarque) e a última parada (de desembarque) deve ser de 300 metros (Essa condição ocorre sempre que os usuários validam seu cartão antes de desembarcar, o que acontece com usuários em linhas lotadas);

Concluindo o processo de encadeamento de viagens, após todos os embarques e desembarques dos usuários apresentarem suas localizações com os pontos de parada e seus respectivos horários, é necessário detectar para cada transação se, seu local de desembarque é um local de transferência de viagem ou se é um destino de viagem, onde o usuário exerce uma atividade do seu dia.

Portanto, para todos os desembarques e embarques posteriores de todos os usuários da base de um dia de bilhetagem eletrônica, foi realizada essa comparação e se:

$$Inferência\ de\ Atividade\ e\ Transferência = \begin{cases} tempo \leq TMT \\ distancia \leq DMT = 1000m \\ transaçao \neq ultima \\ linha_{anterior} \neq linha_{atual} \end{cases}, \quad transferência\ detectada \quad (3)$$

Utiliza-se um exemplo real da base de dados para exibir o que foi realizado para etapa. A Tabela 7 apresenta os atributos de cada transação

realizada por um usuário, considerando os atributos estimados anteriormente. A sequência de viagens do usuário se inicia na estação Itaquera do Metrô. Como sua segunda transação ocorre na estação Pedro II do Expresso Tiradentes, inferiu-se que o desembarque relacionado à primeira transação ocorreu na estação Pedro II da linha de metrô. A terceira transação do dia ocorre na estação Ypiranga do Expresso Tiradentes, portanto, o desembarque relacionado à segunda transação foi inferido como esse mesmo local. A quarta transação ocorre na estação Pedro II do metrô e, inferiu-se que o desembarque anterior foi realizado na estação de mesmo nome do Expresso Tiradentes. Por fim, a última transação do dia foi realizada na linha de ônibus 2727-10 e, inferiu-se que o desembarque anterior foi realizado na estação de metrô mais próxima do ponto de parada (estação Artur Alvim) e o desembarque final desse usuário é em um ponto de parada da linha 2727-10 mais próximo do primeiro embarque do dia.

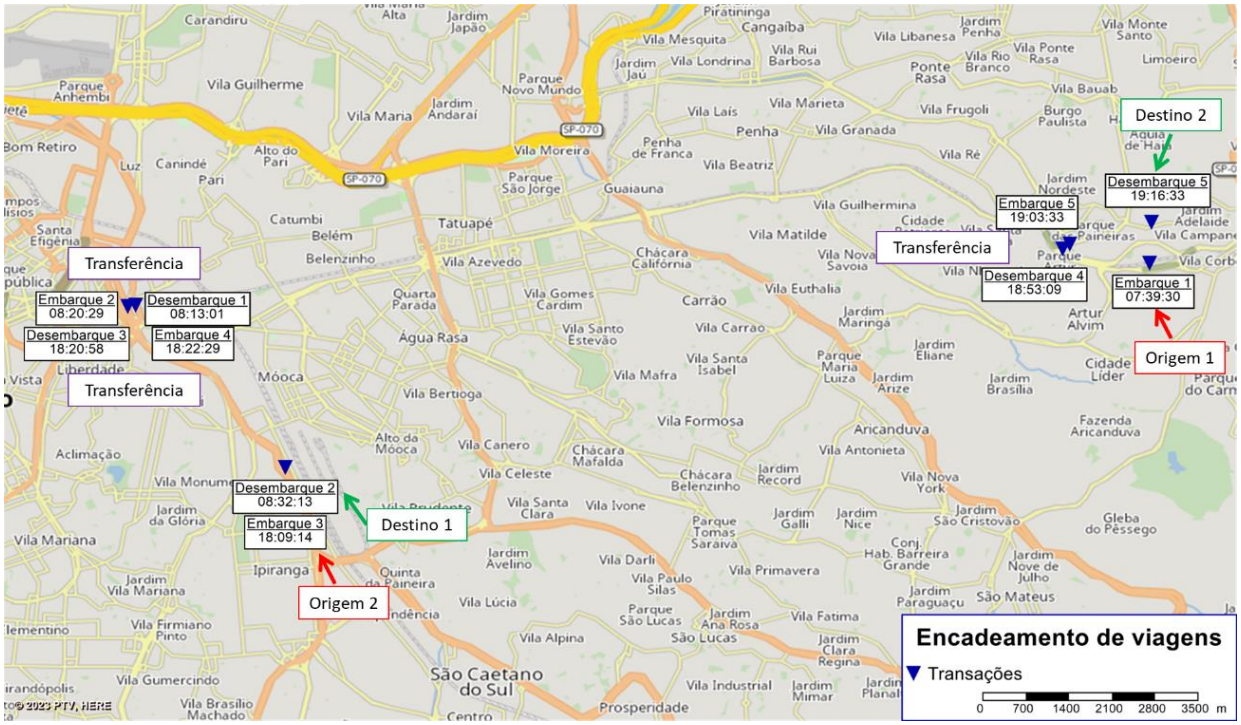
Após a inferência dos locais de desembarque e, a estimativa dos horários de desembarque de cada um dos segmentos de viagem com base no que foi apresentado na seção anterior, infere-se se o local de desembarque de cada segmento de viagem é uma atividade ou uma transferência.

Pelo critério de tempo entre o desembarque e o embarque posterior, observa-se que apenas o tempo entre o segundo desembarque e o terceiro embarque é maior do que o tempo máximo calculado, inferindo-se uma atividade neste local. A outra atividade inferida é no local de desembarque da última transação. Já os outros locais de desembarque foram classificados como transferência, pois respeitam todas as regras de inferência de transferência. Na Figura 5, apresenta-se um mapa com a localização dessas transações.

Tabela 7 – Exemplo dos atributos para cada transação de um usuário

Ponto Embarque	Ponto Desembarque	Linha utilizada	Hora Embarque	Hora Desembarque	Distância Desemb. e Emb. Posterior	Tempo máximo Transf. Calculado	Tempo Desemb. e Emb. Posterior	Local de Desembarque
ITQ - Corinthians Itaquera	PDS - Pedro II	Linha 03 - Metrô	07:39:30	08:13:01	280 m	00:12:40	00:07:28	Transferência
Exp. Tiradentes - Pedro II	Exp. Tiradentes - Ypiranga	Exp. Tiradentes	08:20:29	08:32:13	0 m	00:08:00	09:37:01	Atividade
Exp. Tiradentes - Ypiranga	Exp. Tiradentes - Pedro II	Exp. Tiradentes	18:09:14	18:20:58	100 m	00:05:40	00:01:31	Transferência
PDS - Pedro II	ART - Arthur Alvim	Linha 03 - Metrô	18:22:29	18:53:09	200 m	00:13:20	00:10:24	Transferência
50004439	360004866	2727-10	19:03:33	19:16:33	-	-	-	Atividade

Figura 5 – Exemplo real de encadeamento de viagens



Fonte: Elaboração própria

Com base nesse encadeamento de viagem realizado, é possível saber as origens e destinos de cada um dos usuários ao longo dia, todas as atividades são consideradas como destinos de viagem dos usuários. Neste exemplo, pela manhã, o usuário tem origem na região de Itaquera e destino na região do Ipiranga, realizando uma transferência no Centro. À noite, o usuário tem origem na região do Ipiranga, destino na região de Itaquera, realizando duas transferências, uma no Centro e uma no Artur Alvim.

Esse processamento foi realizado para todos os usuários da base de dados, estimando-se assim, todas as transferências realizadas pelos usuários e suas origens e destinos, construindo a matriz Origem-Destino a partir dos dados de bilhetagem eletrônica.

3.2.2 Inferência do motivo de viagem

Compreender por que os passageiros utilizam sistemas de transporte público e qual é o motivo por trás de suas viagens é de fundamental importância para o aprimoramento do transporte e o desenvolvimento de estratégias eficazes.

A metodologia aplicada tem como objetivo fundamental a inferência do atributo de motivo das viagens com base em transações de cartões inteligentes, adotando uma abordagem de aprendizado não supervisionado, especificamente o agrupamento. Essa estratégia parte da premissa de que ao agrupar passageiros com base em seus atributos temporais, como o horário do dia em que realizam suas viagens, é possível aplicar uma abordagem eficaz para inferir o propósito das viagens a partir dos registros do conjunto de dados da Pesquisa Origem e Destino Domiciliar (Faroqi; Mesbah, 2021).

O processo envolve algumas etapas listadas adiante. Inicialmente, os dados brutos de bilhetagem e da Pesquisa OD são submetidos a um pré-processamento. Durante essa etapa, características relevantes são extraídas, incluindo horários de início e término das viagens, número de viagens por usuário e duração das pernas de viagens. Essas características são utilizadas para comparar as duas bases de dados e também para a inferência dos motivos de viagem. Em sequência, esses atributos temporais são classificados e ordenados na criação de uma sequência temporal para cada usuário.

Após a limpeza dos conjuntos de dados, as sequências temporais de viagem para os usuários foram criadas. Para fazer isso, primeiro determinou-se um conjunto de seis intervalos de tempo para o horário de início das viagens (Tabela 8) e sete intervalos de tempo para o intervalo entre viagens (Tabela 9), da seguinte forma:

Início da viagem:

Tabela 8 – Atributo temporal – Início da viagem

A	Antes das 6h
B	Entre 6h e 7h
C	Entre 7h e 10h
D	Entre 10h e 16h
E	Entre 16h e 18h
F	Após as 18h

- Intervalo entre viagens (horas):

Tabela 9 – Atributo temporal – Intervalo entre viagens

a < 3h
3h ≤ b < 5h
5h ≤ c < 7h
7h ≤ d < 9h
9h ≤ e < 12h
12h ≤ f < 22h
g: última viagem do dia

Esses intervalos de tempo foram usados para modelar as viagens dos indivíduos nos conjuntos de dados. Por exemplo, se um usuário faz duas transações no dia, sendo a primeira transação do dia às 06h30 e a segunda transação às 17h, a sequência temporal desse usuário é: BeEg.

O passo seguinte é o agrupamento dos passageiros da Pesquisa OD domiciliar. A metodologia emprega o algoritmo K-means para essa finalidade, o que possibilita a formação de clusters com base em atributos temporais, citados no parágrafo anterior e nos motivos de viagem dos usuários entrevistados na Pesquisa Origem e Destino. Esses clusters representam grupos de passageiros

com padrões de viagem semelhantes. A similaridade nos padrões de viagem entre os passageiros agrupados será a base para a inferência subsequente do motivo das viagens nas transações coletadas com a bilhetagem eletrônica.

Uma vez que os clusters são formados, segue-se a etapa de rotulagem. Cada cluster é atribuído a um rótulo com base no motivo de viagem mais comum presente nesse grupo. Por exemplo, um cluster composto principalmente por viagens de ida e volta da residência para o trabalho é rotulado como "Residência-Trabalho-Residência". Essa rotulagem é fundamental, pois associa um motivo de viagem a cada cluster, o que possibilita a inferência do propósito das transações dos cartões inteligentes.

A fase de inferência envolve a associação de cada usuário da base de bilhetagem ao cluster mais semelhante. Isso é determinado com base no índice de similaridade de Jaccard, que compara as sequências de viagens dos passageiros com as sequências de viagens de cada cluster. O motivo da viagem inferido para uma transação corresponde ao rótulo do cluster ao qual ela foi associada. Dessa forma, a metodologia permite identificar o propósito das viagens registradas nos cartões inteligentes (Faroqi; Mesbah, 2021).

O índice de similaridade de Jaccard é uma métrica estatística usada para medir a similaridade entre dois conjuntos. Essa métrica é amplamente aplicada em diversos campos, incluindo estatística, mineração de dados, aprendizado de máquina e, no contexto da metodologia de inferência de motivo de viagem, na comparação de padrões de viagem de passageiros.

A fórmula do índice de Jaccard é definida como a soma da interseção dos conjuntos dividida pela soma da união dos conjuntos. Matematicamente, é expressa da seguinte forma:

$$Índice\ de\ Jaccard\ (J) = |A \cap B| / |A \cup B| \qquad (4)$$

Onde:

$|A \cap B|$ representa o número de elementos comuns nos conjuntos A e B (ou seja, a interseção).

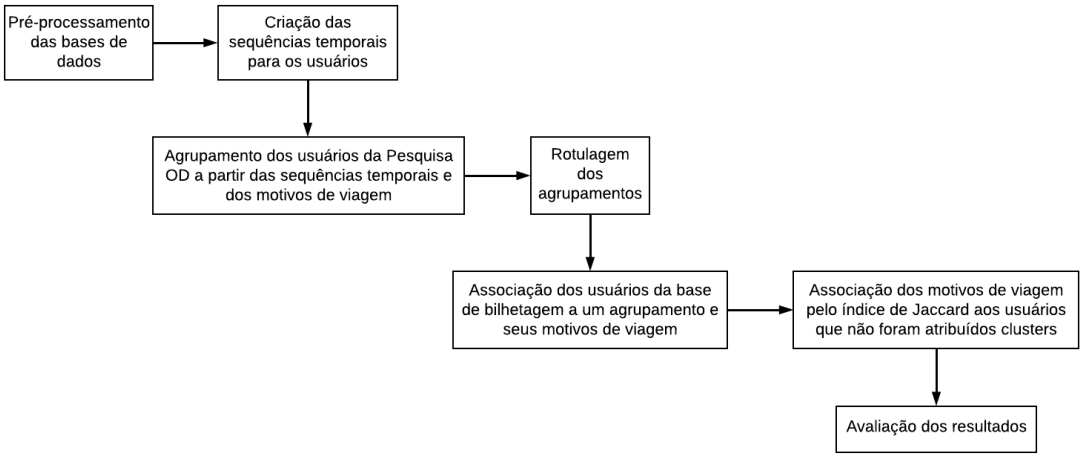
$|A \cup B|$ representa o número total de elementos nos conjuntos A e B (ou seja, a união).

Nessa tentativa de classificação por clusters enfrenta-se desafios ao lidar com a heterogeneidade nos dados de pesquisa Origem Destino domiciliar e,

principalmente, nos dados de bilhetagem eletrônica. A diversidade nas características sociodemográficas e comportamentais dos usuários dificulta uma classificação única para toda a base de dados.

Portanto, além de ser aplicado na fase de inferência para associar transações de cartão inteligente aos clusters, o índice de Jaccard foi utilizado para examinar a similaridade entre os padrões de viagem que não foram inicialmente associados. Essa abordagem permite uma análise abrangente, revelando padrões individuais que podem não ter sido capturados pelos clusters predefinidos.

Figura 6 – Método de inferência de motivo de viagem



Fonte: Baseado em Faroqi e Mesbah, 2021

Em um estágio final, após a inferência de atividade ou transferência para cada transação, realizou-se a exclusão de usuários que incorreram nessa inferência ou em algum dos erros mencionados anteriormente. Essa medida adicionou uma camada de refinamento, garantindo que a base de dados final tivesse todos os atributos para análise dos padrões de viagem do sistema de transporte público. A base de dados final, utilizada para a construção da matriz Origem-Destino e análises subsequentes, consistiu em um total de 4.560.409 transações, 3.231.254 viagens e abrangeu 1.446.058 usuários. A Tabela 10 apresenta cada uma das etapas com o resumo das transações filtradas ao longo do processamento.

Tabela 10 – Etapas do processamento e quantidade de transações filtradas

Etapa	Transações	% excluída em cada etapa
Dados brutos originais	12,622,863	-
Transações com coordenadas de embarque	12,440,553	1.4%
Dados com cartão diferente de 'Operação'	11,841,779	4.7%
Usuários com mais de 1 transação no dia	11,396,281	3.5%
Transações com coordenadas de desembarque	11,017,316	3.0%
Usuários sem inconsistências no embarque e desembarque	8,362,921	21.0%
Usuários com transações apenas no município de São Paulo	7,858,305	4.0%
Usuários sem erros de transferência/outros erros	4,560,409	26.1%

Fonte: Elaboração própria

Como comparação, no estudo de Trépanier et al. (2007), a estimativa dos destinos de viagem a partir da base de bilhetagem eletrônica também envolveu uma filtragem dos dados. Durante essa análise, foi possível atribuir destinos válidos para aproximadamente 66% do conjunto de dados. Vale destacar que 13% das transações, representando casos de viagens únicas, foram excluídos da análise, resultando no uso de 53% da base para a condução do estudo.

3.2.3 Avaliação dos padrões de viagem

Para descrever o processo de extração dos padrões de viagem de um indivíduo, esclarece-se primeiro o conceito de viagem. Uma viagem significa uma jornada de mão única de uma atividade para outra com um propósito específico, como viajar de casa para o espaço de trabalho. Uma viagem pode consistir em alguns segmentos de viagem definidos por uma transferência entre um mesmo modo (por exemplo, de um ônibus para outro) ou modos diferentes (por exemplo, de ônibus para o metrô).

A compreensão dos usuários de transporte público em termos de seus padrões de viagem pode apoiar o planejamento e o design de melhores serviços. A classificação do usuário facilita o planejamento por meio de um melhor desenho de pesquisa, bem como uma avaliação mais detalhada, através da análise de impactos com base na caracterização dos usuários afetados. A classificação dos usuários de transporte público pode ser aprimorada com o uso de dados de cartões inteligentes.

Para ilustrar os padrões de viagem de um passageiro individual, dividem-se os padrões em 2 categorias: espaciais e temporais. Viagens com padrões

espaciais mostram que o par Origem-Destino do usuário não varia significativamente, ou seja, na maioria de suas viagens, o usuário frequenta os mesmos locais. Viagens com padrões temporais mostram que as viagens são repetidas em determinados períodos de tempo.

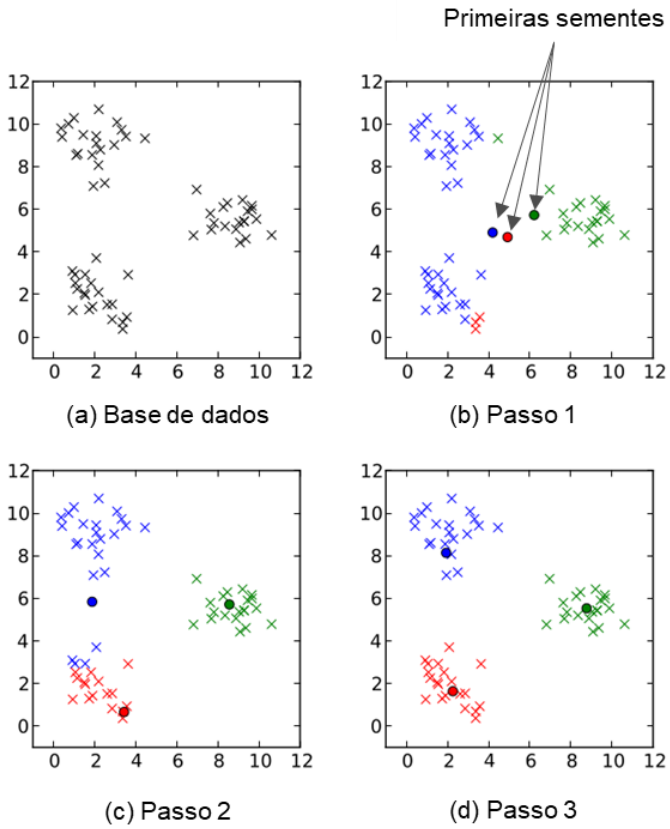
Várias características de viagem relacionadas à variabilidade temporal e espacial, padrões de atividade, características sociodemográficas e opções de modo são utilizadas para identificar grupos homogêneos.

Como a base de dados de cartões inteligentes não possui informações conhecidas sobre os padrões de viagem dos passageiros, é necessária a realização de um processo de agrupamento (*clustering*) para identificar diferentes grupos de usuários com padrões de viagem semelhantes aos dados de cartões inteligentes. Neste estudo, os dados utilizados serão de formato categórico, portanto escolheu-se o algoritmo de clusterização k-modes, descrito a seguir.

Proposto por Huang (1997) para lidar com o problema de agrupamento de grandes conjuntos de dados categóricos, o algoritmo k-modes estende o algoritmo k-means ao usar uma medida de dissimilaridade de correspondência simples para objetos categóricos, modas em vez de médias para clusters, e um método baseado em frequência para atualizar modas no processo de agrupamento, a fim de minimizar a função de custo do agrupamento.

Da mesma maneira que o k-means, uma desvantagem do agrupamento k-modes é que ele pode convergir para um ótimo local (e não global). Uma solução para este problema poderia ser uma exaustiva escolha de pontos de partida, buscando melhorar a performance do modelo (Morency; Trépanier; Agard, 2007). A Figura 7 abaixo ilustra um exemplo simples do método de agrupamento k-modes.

Figura 7 – Exemplo método k-modes



Fonte: Elaboração própria

Ademais das análises citadas acima, observa-se uma outra maneira de adicionar dados semânticos aos dados de bilhetagem eletrônica, o uso da pesquisa de Origem-Destino domiciliar de São Paulo, realizada em 2017.

A fusão de dados é uma das abordagens para integrar várias fontes de dados, aplicadas em vários campos, como aplicações e sistemas de transporte inteligentes (El Faouzi; Leung; Kurian, 2011). Shen e Stopher (2013). desenvolveram um método de imputação de finalidade de viagem para dados do Sistema de Posicionamento Global (GPS) utilizando a Pesquisa Origem-Destino Domiciliar Nacional (*National Household Travel Survey* (NHTS)) nos Estados Unidos. Em seu método, os motivos de viagem que não foram observados diretamente pelos dados do GPS foram estimados usando regras obtidas a partir dos dados do NHTS. Já Kusakabe e Asakura (2014) empregaram a metodologia de fusão de dados para derivar relações entre atributos comportamentais que não podem ser obtidos apenas a partir de dados de cartões inteligentes ou dados baseados em pesquisas.

Os dados baseados em pesquisa observam diretamente informações detalhadas sobre o comportamento da viagem, mas não conseguem fazê-lo continuamente, por um longo período. Por outro lado, os dados de cartão inteligente fornecem apenas informações parciais sobre o comportamento dos viajantes, embora possam fornecer dados contínuos a longo prazo, difíceis de obter por meio de uma pesquisa de viagem individual. Se as vantagens dos dados do cartão inteligente forem combinadas com as dos dados da pesquisa de viagem pessoal, melhoraria os efeitos do monitoramento contínuo das demandas de transporte. Portanto, a fusão de dados permite obter um bom entendimento das mudanças no comportamento dos viajantes durante o longo período contínuo.

A pesquisa domiciliar de Origem-Destino da cidade de São Paulo é o maior levantamento urbano sobre mobilidade do Brasil, ela revela uma “fotografia” do momento sobre as viagens diárias em São Paulo e na Região Metropolitana. A Pesquisa OD é um levantamento sobre o padrão e as escolhas de transporte da população de uma região, investigando os deslocamentos diários que as pessoas fazem, suas origens e destinos, que meios de transporte usam e quais os motivos de seus deslocamentos. Esse conjunto de informações é visto sob a ótica de outras variáveis, como renda, idade, escolaridade, locais de residência, de trabalho, de estudo – os chamados dados socioeconômicos, que são também levantados.

Os dados e bilhetagem eletrônica podem ser cruzados com os dados da pesquisa Origem-Destino principalmente pelas localizações dos pontos de embarque e desembarque das viagens, por meio do zoneamento da cidade de São Paulo e pelos dados de uso do solo.

4. Exploração e estruturação dos dados

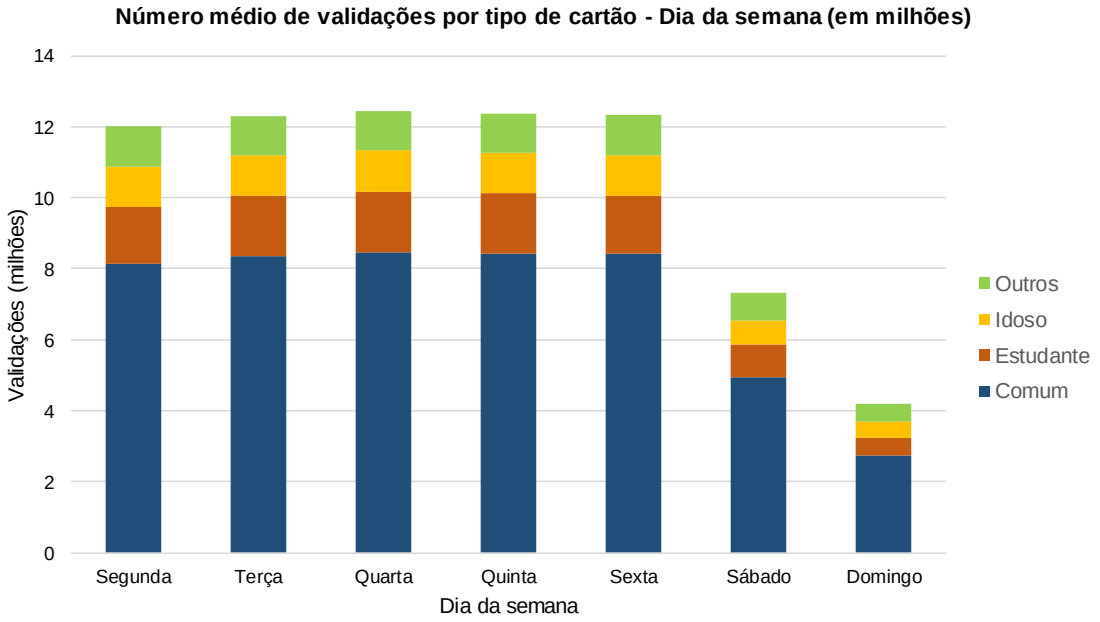
Esta seção apresenta a exploração prévia dos dados de bilhetagem eletrônica e da pesquisa Origem-Destino domiciliar antes de estruturá-los para o propósito de agrupamento e avaliação dos padrões de viagem, com o objetivo tanto de enriquecer as informações nos dados quanto de filtrar informações ou observações indesejadas que poderiam comprometer as avaliações a serem realizadas.

4.1 Exploração dos dados de bilhetagem e pesquisa OD

Alguns resultados da análise exploratória dos dados de bilhetagem eletrônica são apresentados no primeiro momento. Vale ressaltar que estes resultados foram calculados antes da filtragem de passageiros frequentes e do tipo de cartão utilizado na validação.

Também foram examinados os diversos tipos de cartões inteligentes inicialmente presentes na base de dados, sendo categorizados em quatro grupos: "Adulto", "Estudante", "Idoso" e "Outros". O cartão classificado como "Outros" inclui o cartão "Operação", que será excluído do processamento dos dados, pois não possui um usuário associado. A Figura 6 apresenta o número médio de validações por dia da semana, separadas por tipo de cartão. Observa-se que ao longo dos dias úteis, o volume de validações é próximo, varia de 12 a 12,4 milhões. Já aos sábados e domingos este volume é bem menor, cerca de 7,3 milhões de validações aos sábados e de 4,2 milhões aos domingos. O cartão do tipo "Outros" representa aproximadamente 9% nos dias úteis, sendo que o tipo "Operação" corresponde a quase 5% das transações, como será detalhado posteriormente.

Figura 8 – Número de validações por tipo de cartão

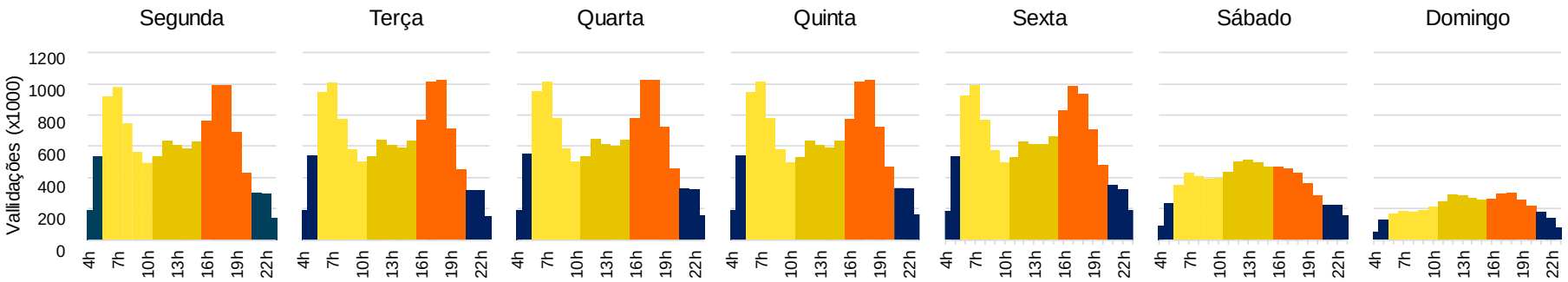


Fonte: Elaboração própria

A Figura 9 mostra um perfil do número médio de validações por hora, por dia da semana, levando em conta todas os dados das semanas disponibilizadas, com base em Zhong et al. (2015). O eixo y indica o número médio de validações (dividido por mil) e o eixo x representa a linha do tempo horária. Os horários de pico podem ser claramente identificados e a diferença entre dias úteis e finais de semana é significativa. Os picos da manhã e da tarde não se mostram claros nos fins de semana, onde as validações são mais distribuídas ao longo do dia. Outro fato interessante é que a soma das validações no período de pico da manhã é próxima da soma das validações no período de pico da tarde, ao longo dos dias úteis.

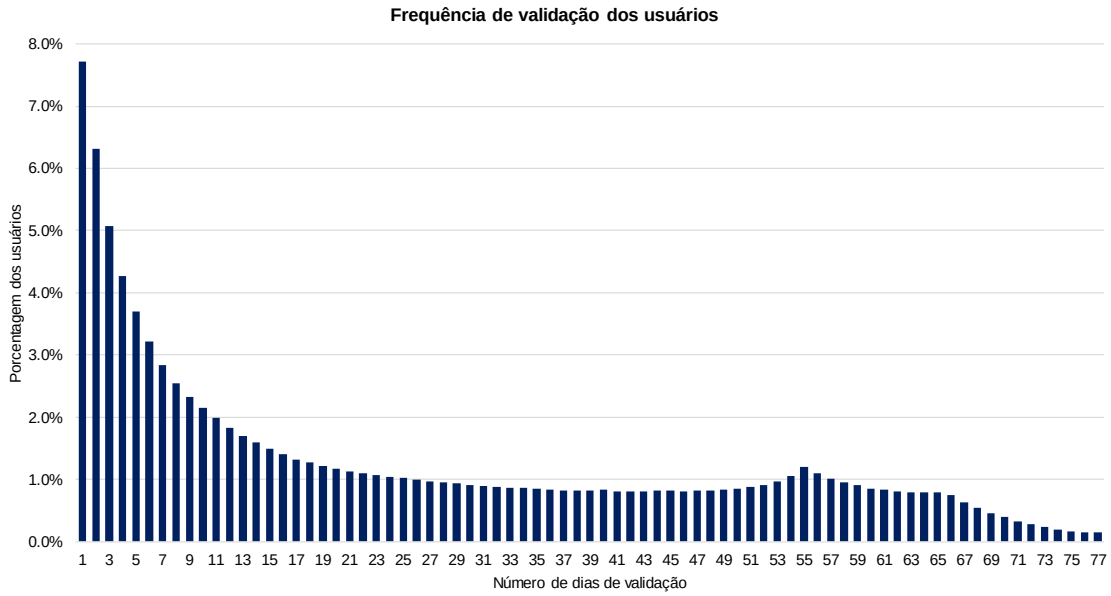
Construiu-se também um histograma da frequência das validações durante o período do conjunto de dados, ou seja, quantos dias cada usuário validou seu bilhete eletrônico. Na Figura 10 é apresentado este histograma. O número de passageiros ocasionais, com baixo número de dias de validações durante o período analisado, é alto.

Figura 9 – Perfil horário de validações



Fonte: Elaboração própria

Figura 10 – Histograma de frequência de validação

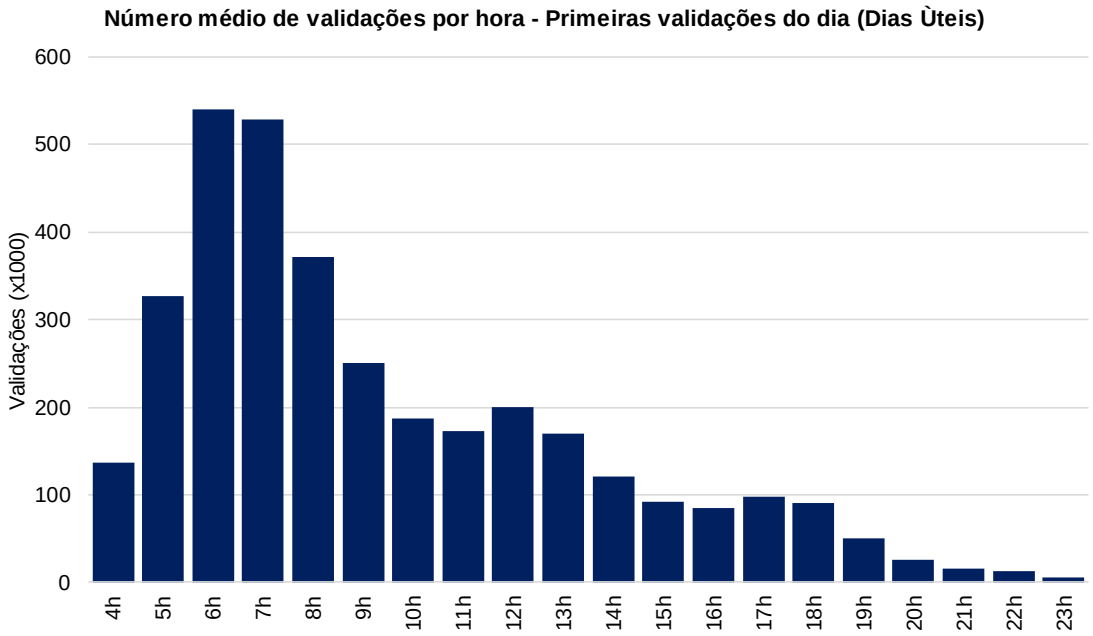


Fonte: Elaboração própria

Cerca de 40% dos usuários validaram seu bilhete 10 ou menos dias no período de 11 semanas. De aproximadamente 53 a 58 dias, há um ligeiro aumento na porcentagem de usuários, sugerindo que são passageiros regulares que usam o sistema durante dias úteis, cerca de 55 dias (pico entre os mais frequentes), que seriam 5 vezes por semana durante as 11 semanas.

Em se tratando das primeiras validações do dia de cada usuário, 2 análises foram realizadas. Primeiramente, plotou-se um histograma com o perfil horário dessas validações, apresentado na Figura 11. Como esperado, a grande maioria das primeiras validações é realizada no período da manhã, com o pico acontecendo entre 6 e 7 horas da manhã.

Figura 11 – Histograma de horário de validação – Primeira validação do dia

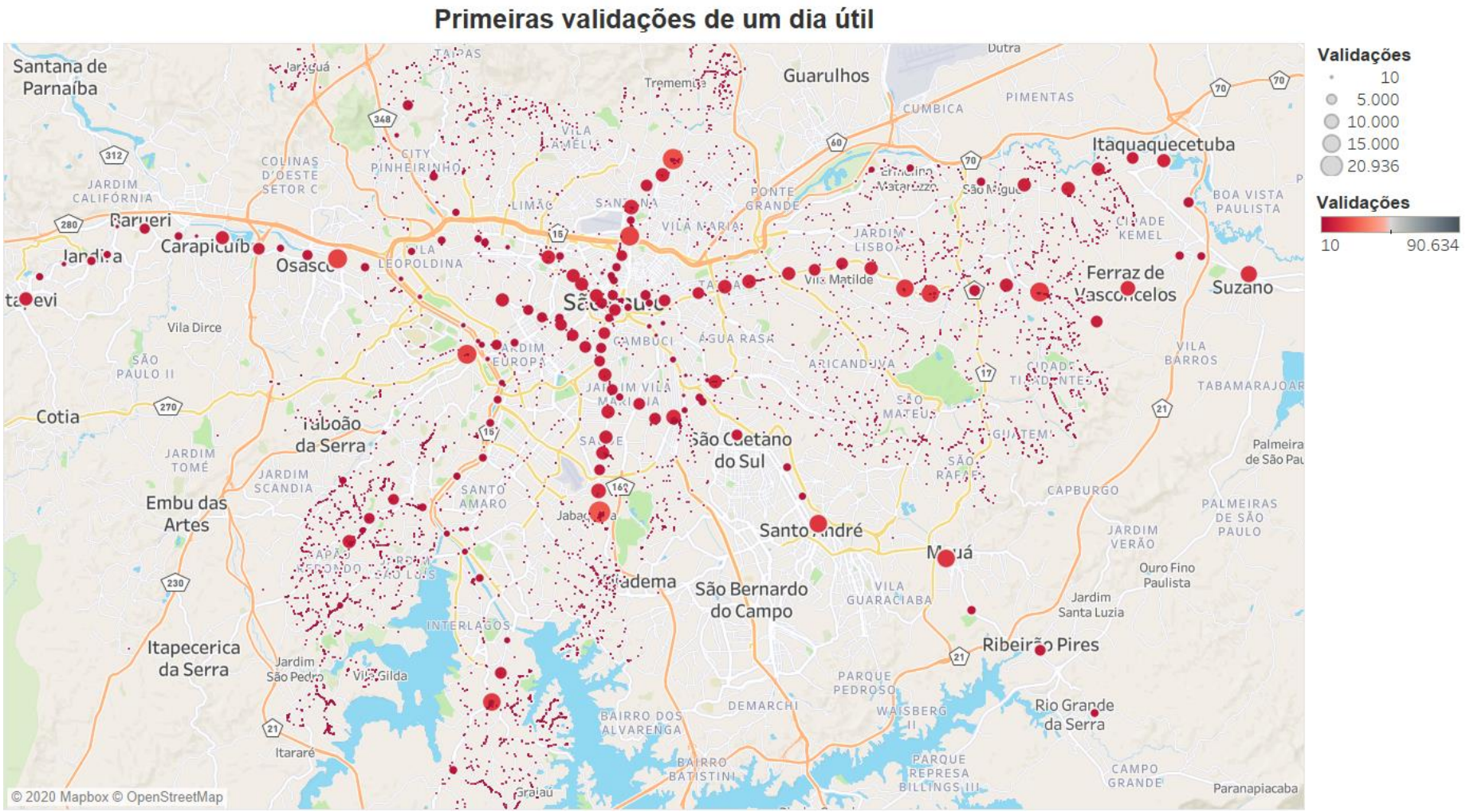


Fonte: Elaboração própria

Ainda avaliando as primeiras validações do usuário, agora examinamos a distribuição espacial da localização dessas validações de um dia útil de dados por meio de um mapa de círculos proporcionais, para visualizar como as validações se distribuem pela cidade (Figura 12). O tamanho do círculo é proporcional ao número de validações no local especificado. As primeiras validações estão localizadas principalmente nos eixos das linhas de metrô e trem da cidade de São Paulo, além dos terminais de ônibus, evidenciando círculos maiores de validações nas regiões mais periféricas ao centro.

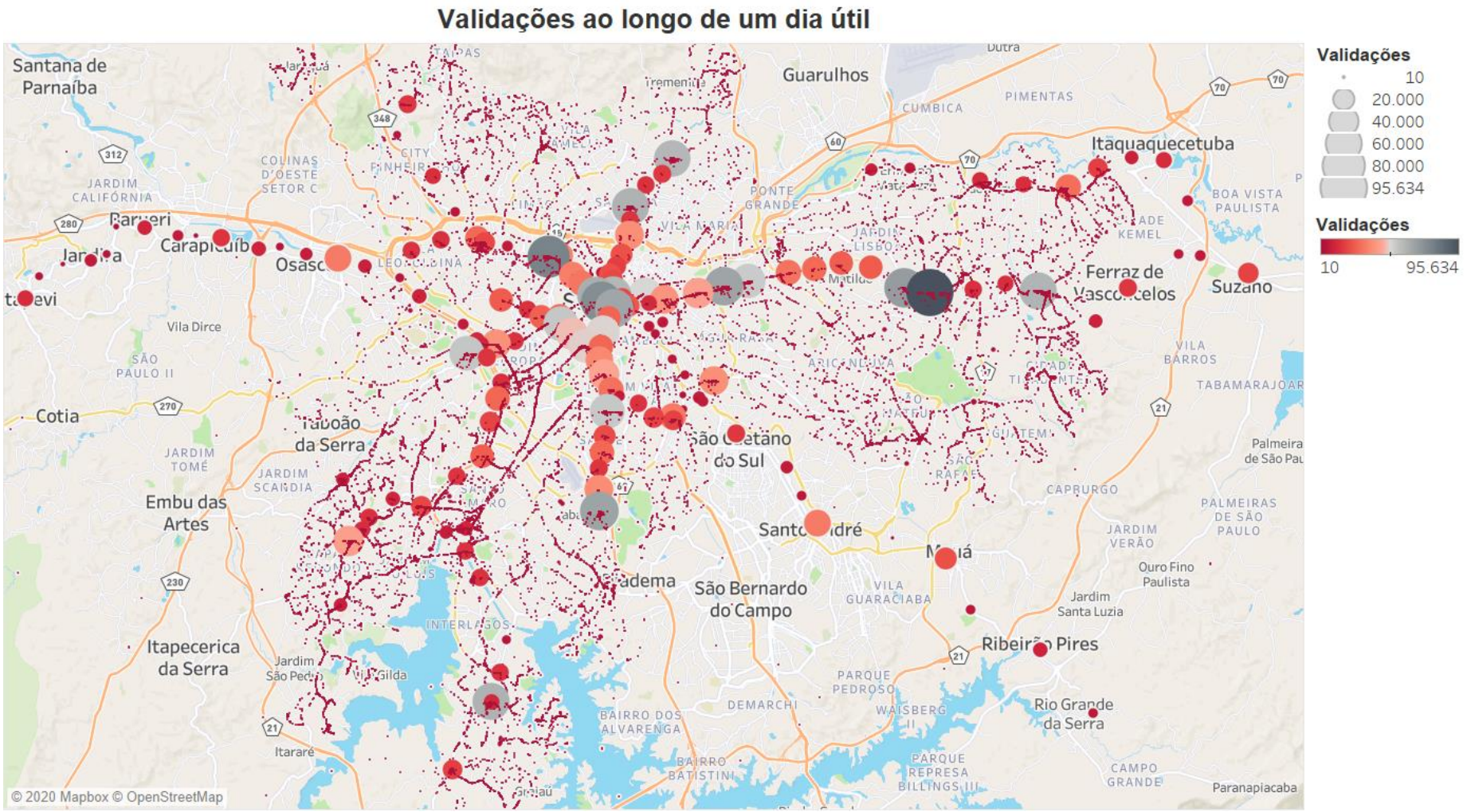
Para efeitos de comparação, criou-se um mapa com a distribuição espacial de todas as validações do mesmo dia útil para que foi realizado o gráfico das primeiras validações. Este mapa se encontra na Figura 13. Como no mapa anterior, o tamanho dos círculos é proporcional ao número de localizações. Nota-se que a distribuição espacial das validações ao longo de todo o dia é parecida com a distribuição das primeiras bilhetagens do dia (eixos das linhas de metrô e trem). Porém, boa parte das validações agora é realizada mais próxima ao centro da cidade de São Paulo.

Figura 12 – Distribuição espacial – Primeiras validações do dia (03-08-2016)



Fonte: Elaboração própria

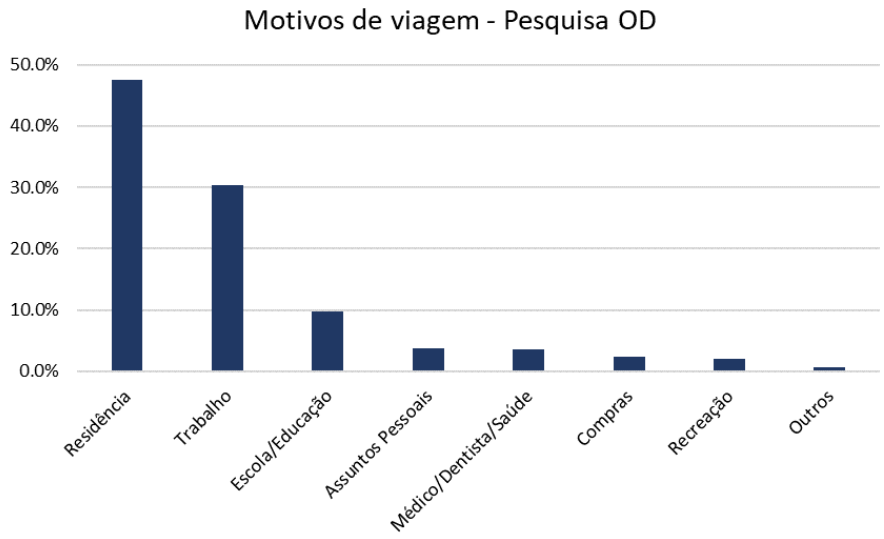
Figura 13 – Distribuição espacial – Todas as validações do dia (03-08-2016)



Fonte: Elaboração própria

Agora com relação aos dados da pesquisa Origem e Destino, o conjunto inclui características sociodemográficas e atributos de viagem de indivíduos. A Figura 14 apresenta o histograma dos motivos das viagens de transporte público no conjunto de dados da pesquisa OD domiciliar de 2017, que incluem Residência, Trabalho, Educação, Compras, Assuntos Pessoais[2], Saúde e Recreação.

Figura 14 – Distribuição de motivos de viagem – Pesquisa OD (2017)



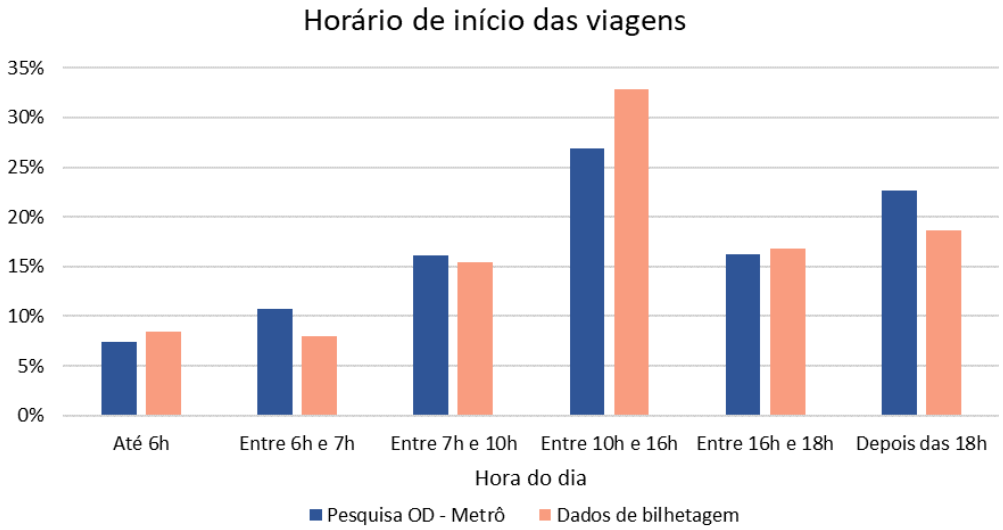
Fonte: Elaboração própria

As figuras a seguir comparam os atributos temporais entre os dois conjuntos de dados – Bilhetagem eletrônica e Pesquisa OD. Para fazer essa comparação, utilizou-se os intervalos de horário de início de viagem e intervalo de tempo entre viagens apresentados nas Tabela 8 e na Tabela 9. Na Figura 15, apresenta-se a comparação das duas bases para o horário de início da viagem. A análise comparativa revelou discrepâncias mínimas, sendo a maior diferença de apenas 6 pontos percentuais no intervalo entre 10h e 16h. Contudo, essa diferença é considerada não significativa, indicando uma relativa concordância entre as estimativas nos dois conjuntos de dados durante esse período. A consistência observada sugere que as fontes de dados apresentam uma correspondência aceitável, o que pode ser interpretado como uma validação mútua das informações.

[2] Relacionado com assuntos particulares ou realizados para terceiros, como, tirar carteira de trabalho, consultar advogado, ir ao banco etc. (Pesquisa Origem Destino, 2017)

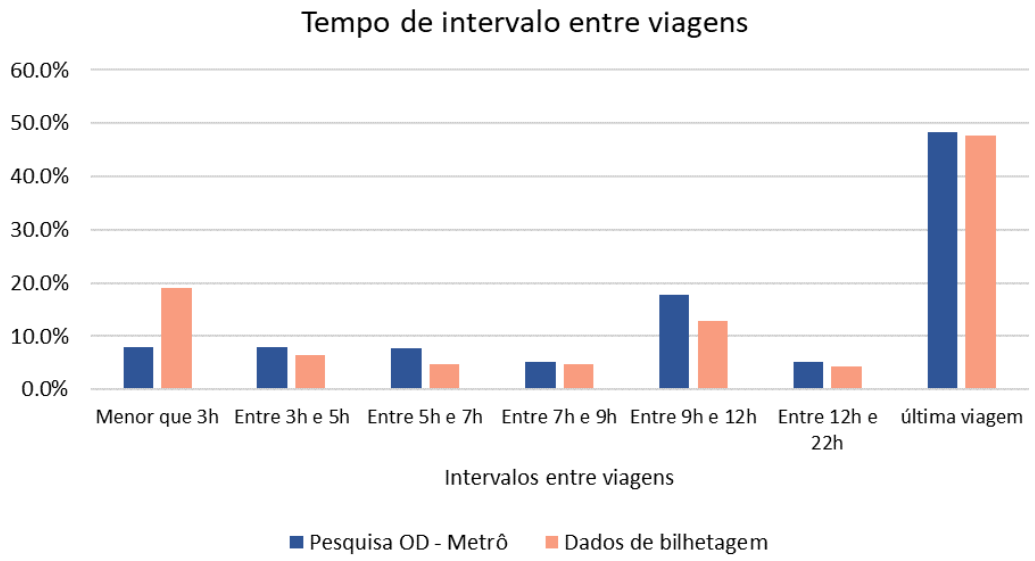
Na Figura 16, apresenta-se a comparação das bases de dados para os intervalos de tempo entre viagens. A maior discrepância entre os conjuntos de dados ocorre na categoria "Menor que 3h", onde a Pesquisa OD no Metrô registra 7.1%, enquanto os Dados de Bilhetagem indicam 19.1%. Essa diferença pode ser atribuída à dinâmica do sistema de transporte coletivo de São Paulo, onde não há cobrança de transferência para intervalos menores de 3 horas. Em situações em que os passageiros realizam deslocamentos muito curtos e não efetuam a transação no início da viagem, eles podem não ser contabilizados nos dados de bilhetagem, resultando em uma subestimação nas estatísticas para essa categoria na Pesquisa OD no Metrô. Esta observação destaca a influência das políticas operacionais do sistema no registro de dados e ressalta a necessidade de considerar tais particularidades ao interpretar as discrepâncias entre fontes de dados, sugerindo a importância de uma análise mais aprofundada para uma compreensão abrangente das diferenças observadas. No geral, observa-se que os atributos temporais das viagens de transporte público no conjunto de dados da pesquisa OD seguem um padrão semelhante aos atributos temporais no conjunto de dados do cartão inteligente.

Figura 15 – Atributo de início da viagem – Dados de bilhetagem e Pesquisa OD



Fonte: Elaboração própria

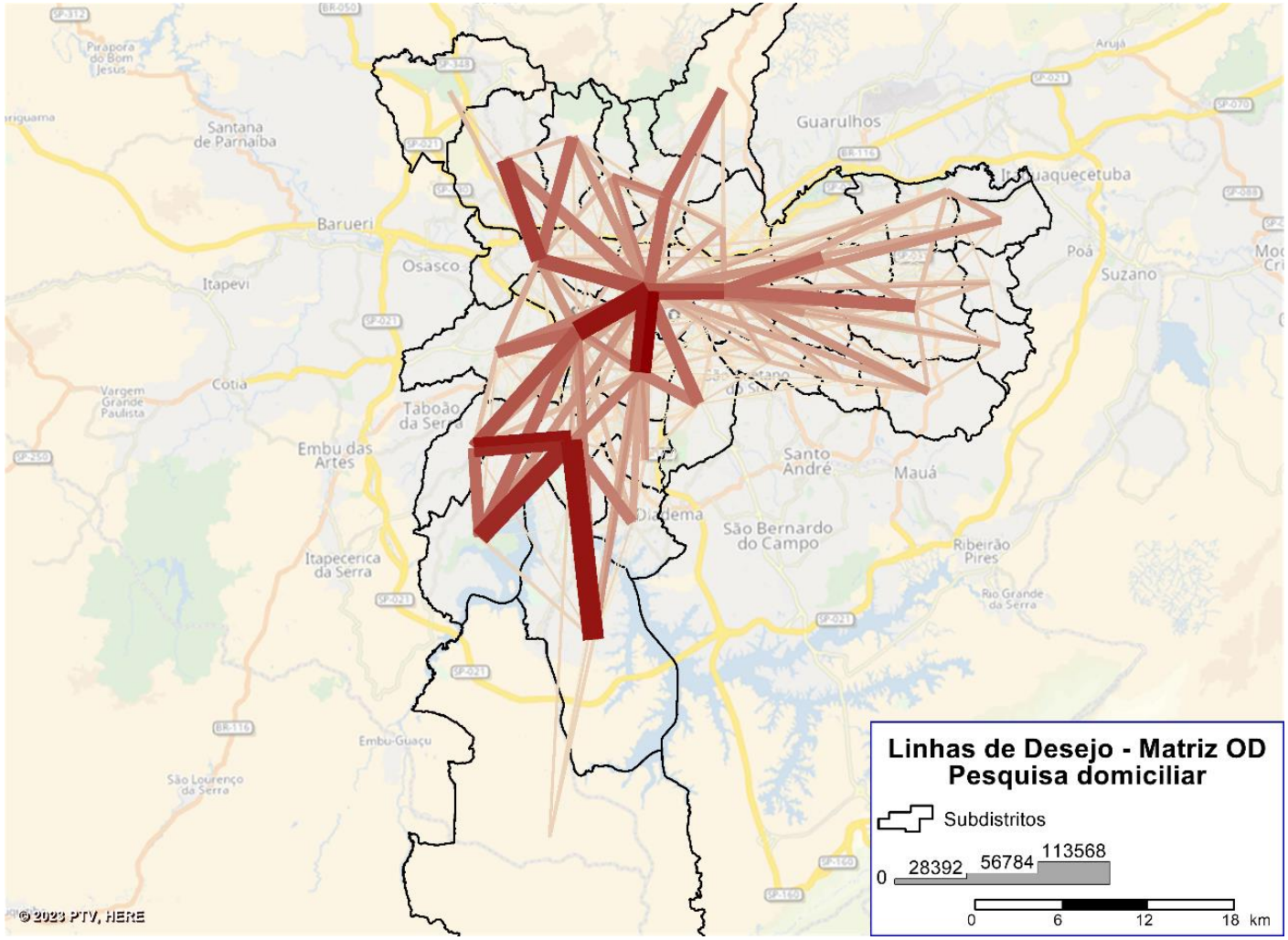
Figura 16 – Atributo de intervalo entre viagens para construção de sequência temporal – Dados de bilhetagem e Pesquisa OD



Fonte: Elaboração própria

A construção de uma matriz Origem-Destino (OD) a partir de dados de pesquisa OD domiciliar envolveu a organização e agregação dos dados, consolidando informações sobre origens e destinos para criar uma contagem de viagens entre cada par correspondente. Neste caso, construiu-se a matriz Origem-Destino dos usuários de transporte coletivo para a cidade de São Paulo e a partir dessa matriz criou-se o mapa da Figura 17, com as principais linhas de desejo desses usuários. Esse mapa destaca visualmente as conexões entre origens e destinos, proporcionando uma representação gráfica das tendências de mobilidade identificadas na análise da pesquisa OD domiciliar. Além disso, essa matriz será utilizada como referência mais adiante, para comparação com a matriz a ser estimada a partir dos dados de bilhetagem eletrônica.

Figura 17 – Linhas de Desejo - Pesquisa OD domiciliar



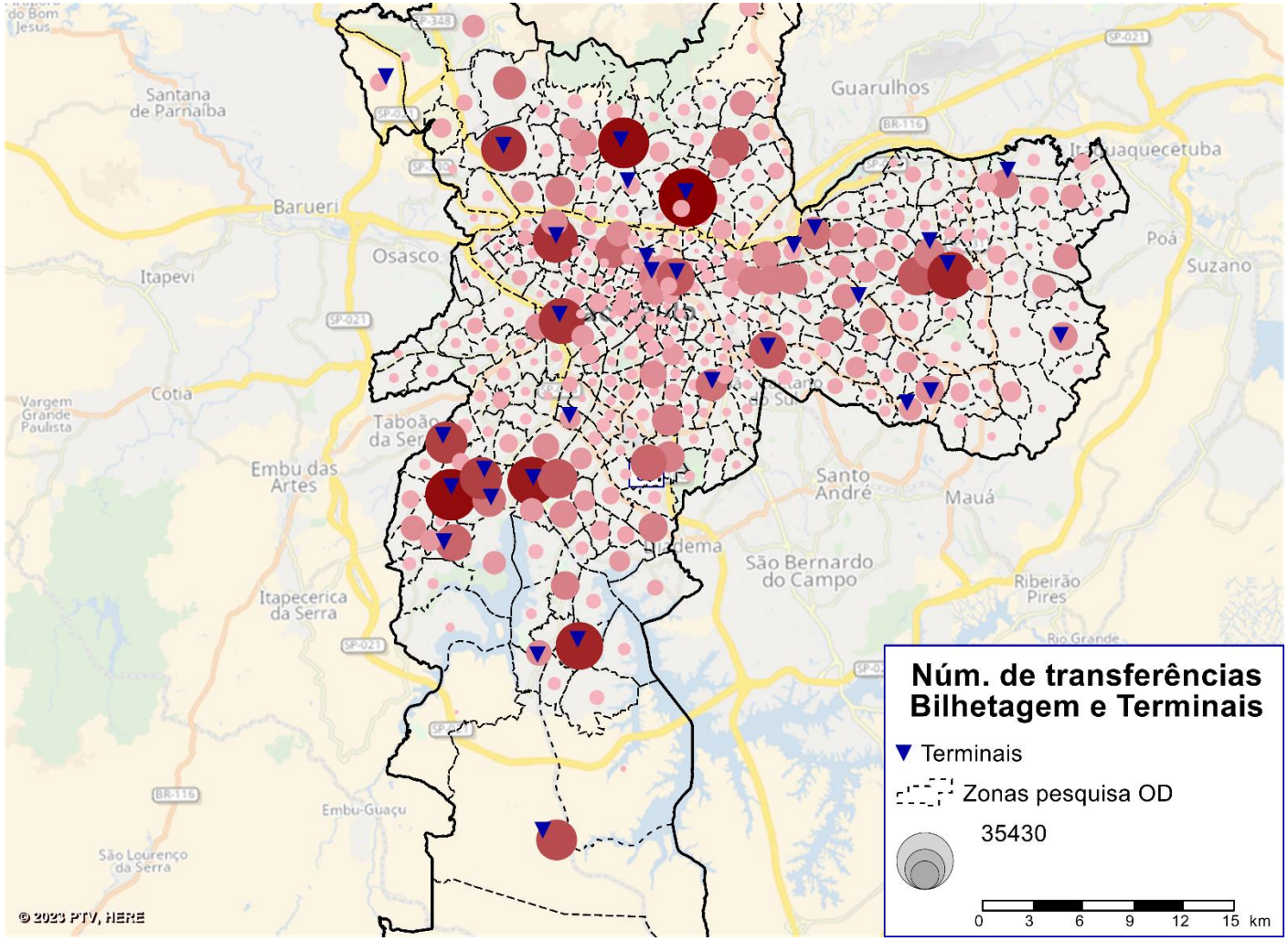
Fonte: Elaboração própria

4.1.1 Detecção de transferências e atividades

Na Figura 18, é apresentado um mapa que ilustra o número total de transferências realizadas pelos usuários de transporte coletivo para cada zona da pesquisa OD domiciliar da cidade. A representação visual utiliza um gráfico de bolha, onde o tamanho das bolas é proporcional à quantidade de transferências ocorridas em cada região. Além disso, na mesma figura, são destacadas as localizações dos terminais de ônibus na cidade. Essa informação é crucial para compreender o contexto das transferências, uma vez que os terminais geralmente são os principais de conexão entre diferentes linhas de transporte coletivo.

Observa-se que a região dos terminais concentra um significativo número de transferências, fornecendo uma validação visual para o método de encadeamento de viagens empregado na inferência dessas transferências.

Figura 18 – Número de transferências e localização dos terminais

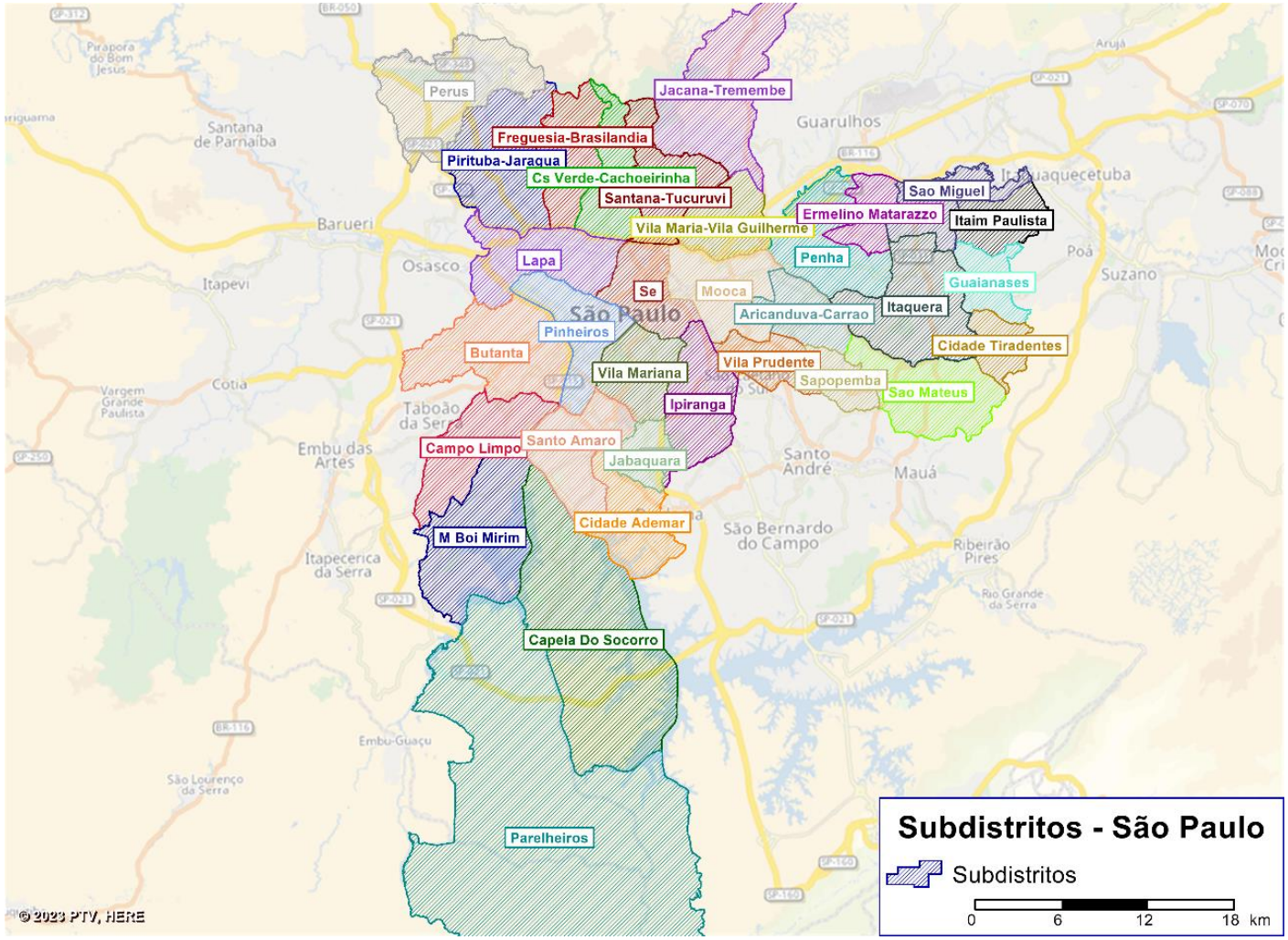


Fonte: Elaboração própria

4.1.2 Matriz Origem Destino

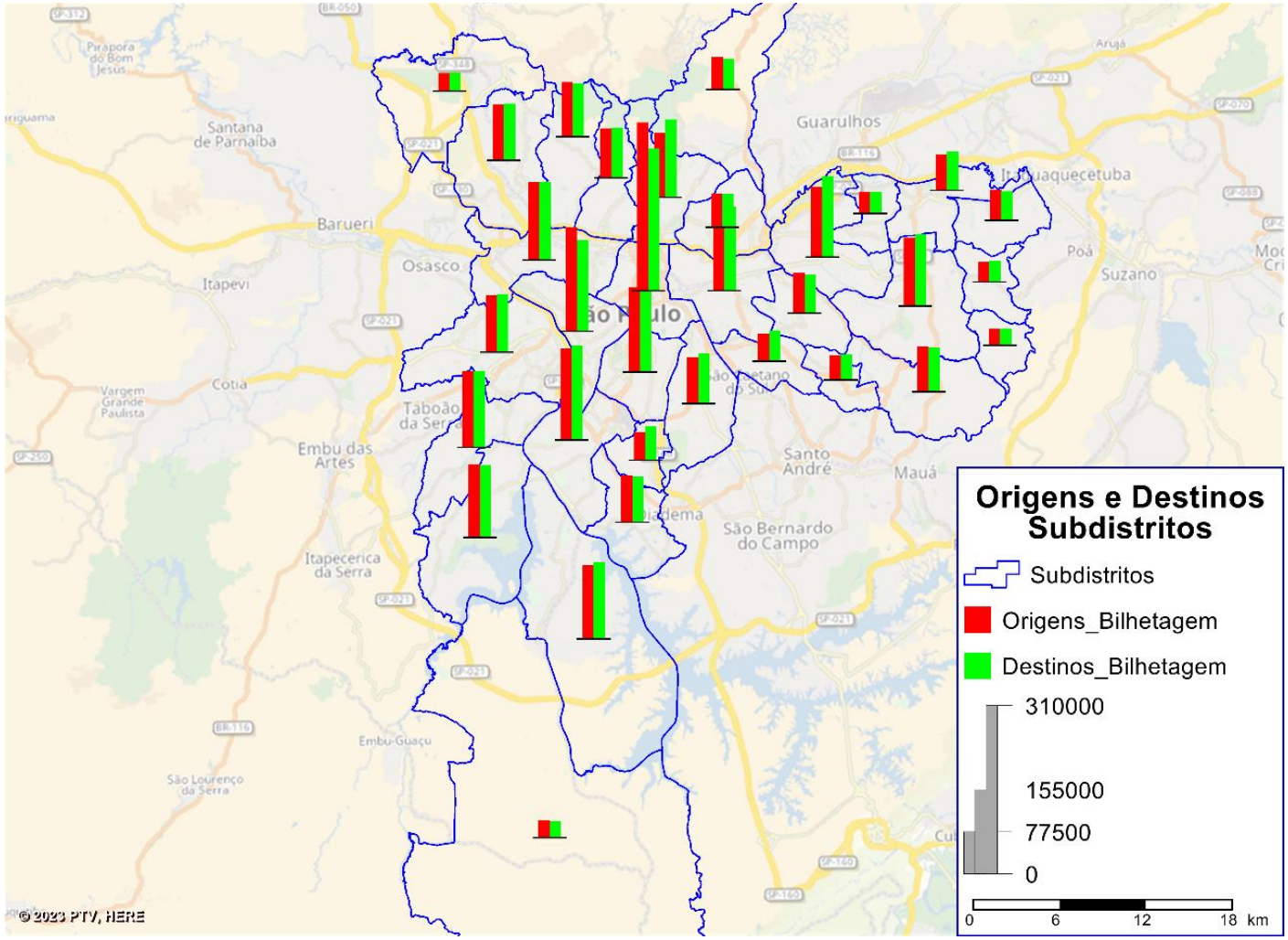
Após a detecção de atividades, inferência de localizações de embarque e desembarque, e a aplicação de restrições específicas para o encadeamento de viagens é possível estabelecer uma visão abrangente da matriz OD para os subdistritos específicos da cidade. Na Figura 19, apresenta-se o mapa com os 32 subdistritos da cidade de São Paulo (Prefeitura, São Paulo). Já na Figura 20, apresentam-se os totais de origens e destinos por subdistrito, essa informação será utilizada na comparação com a matriz Origem-Destino extraída da pesquisa OD domiciliar.

Figura 19 – Mapa dos 32 subdistritos do município de São Paulo



Fonte: Elaboração própria

Figura 20 – Totais de origens e destinos – OD estimada



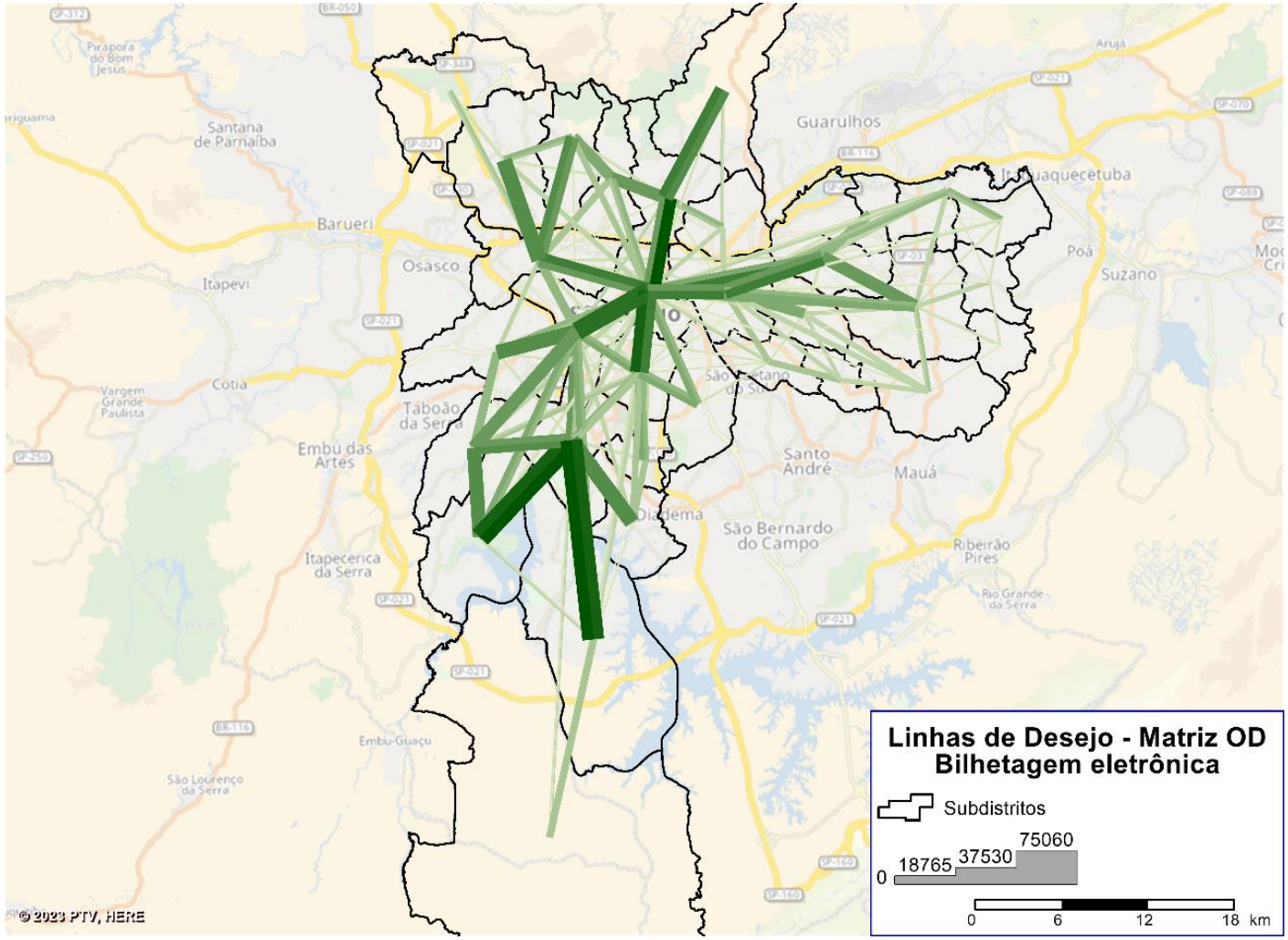
Fonte: Elaboração própria

Na abordagem da matriz OD, as linhas de desejo são representações gráficas das preferências de viagem dos usuários entre pares específicos de origens e destinos dentro do sistema de transporte. Essas linhas indicam as escolhas de viagem mais frequentes ou desejáveis, refletindo a demanda por transporte entre diferentes áreas geográficas. Ao analisar as linhas de desejo na matriz OD, é possível identificar padrões de movimento e avaliar a eficiência e do sistema de transporte em atender às demandas específicas de deslocamento.

Uma limitação significativa nas análises das linhas de desejo surge quando se trata do par intrazonal. A intrazona refere-se a viagens que ocorrem dentro da mesma zona geográfica, como deslocamentos locais dentro de um bairro ou área residencial. Nessas situações, as linhas de desejo não representam essas viagens, tornado esse tipo de análise menos mais sensível a variações devido à proximidade geográfica.

Na Figura 21, apresenta-se as linhas de desejo dos usuários de transporte coletivo entre os subdistritos da cidade de São Paulo, estimadas a partir dos dados de bilhetagem eletrônica. visualização das linhas de desejo contribui para uma compreensão aprofundada dos fluxos de viagens e pode orientar estratégias de otimização do transporte público e planejamento urbano.

Figura 21 – Linhas de Desejo - Estimativa OD bilhetagem eletrônica



Fonte: Elaboração própria

4.1.3 Comparação com a matriz OD – Pesquisa domiciliar

Na presente seção, será conduzida uma análise comparativa das matrizes origem-destino (OD) derivadas das duas fontes de dados citadas anteriormente, matriz OD estimada pela bilhetagem eletrônica e matriz OD extraída pesquisa origem-destino domiciliar do Metrô de São Paulo. Egu e Bonnel (2020) investigam a comparabilidade de matrizes origem-destino (OD), abordando critérios como estrutura, taxa de transferência, distribuição geográfica e participação modal, estimadas a partir de diferentes fontes de dados, nomeadamente coleta automática de tarifas, pesquisas origem-destino e pesquisas de viagens domésticas em Lyon, França.

O intuito desta etapa é examinar as divergências e semelhanças entre essas matrizes, destacando implicações para a estimativa da demanda por viagens. As proporções de distribuição de origens e destinos, além dos pares de viagem serão avaliados, assim como os volumes totais e as taxas de transferência, proporcionando essa comparação.

Estatísticas descritivas referentes a cada conjunto de dados apresentado são fornecidas na Tabela 11. A tabela mostra que, em termos de volume, existem diferenças substanciais. O número total de segmentos de viagem (transações) na base de dados já filtrada de bilhetagem é de cerca de 4,56 milhões em comparação com 10,0 milhões na base de dados da pesquisa domiciliar. Com relação ao total de viagens, a base de dados da pesquisa domiciliar apresenta 6,77 milhões de viagens, enquanto a estimativa de viagens pelos dados de bilhetagem é de 3,2 milhões de metrô (cerca de 52% menor). Essa diferença pode ser vista pelos erros de inferência no processo de encadeamento de viagens e também no cruzamento com camadas espaciais de dados socioeconômicos.

Calculou-se também a taxa de transferência para as duas bases de dados apresentadas. Observa-se que os valores são mais próximos um ao outro em relação ao volume absoluto de viagens. A base da pesquisa domiciliar apresenta uma taxa de transferência de 0,48, ou seja, em média, o usuário realiza esse número de transferências por viagem. Já a estimativa do número de transferência pela base de dados de bilhetagem é de 0,41 transferências por viagem.

Tabela 11 – Estatísticas dos conjuntos de dados

	Pesquisa OD - Metrô	Dados de bilhetagem
Volume de transações	10,039,386	4,560,404
Volume de viagens	6,773,316	3,231,254
Taxa de transferência	**0.4822**	**0.4113**

Fonte: Elaboração própria

A seguir, compara-se a distribuição de origens e destinos das matrizes por subdistrito. Como o fluxo total estimado em cada matriz é diferente, os resultados são apresentados na Figura 22 como uma porcentagem de viagens por zona e por matriz. Este gráfico confirma que, no geral, as distribuições são bastante semelhantes. A maior diferença é encontrada no subdistrito da Sé (maiores volumes), com esse subdistrito representando cerca de 12% de origens e destinos na pesquisa domiciliar e, 10% e 8% na OD de bilhetagem, respectivamente.

Utilizou-se uma métrica global que resuma a diferença entre os conjuntos de dados, a Diferença Total Absoluta. Essa métrica é a soma das diferenças absolutas em todo os subdistritos. Calculado da seguinte maneira:

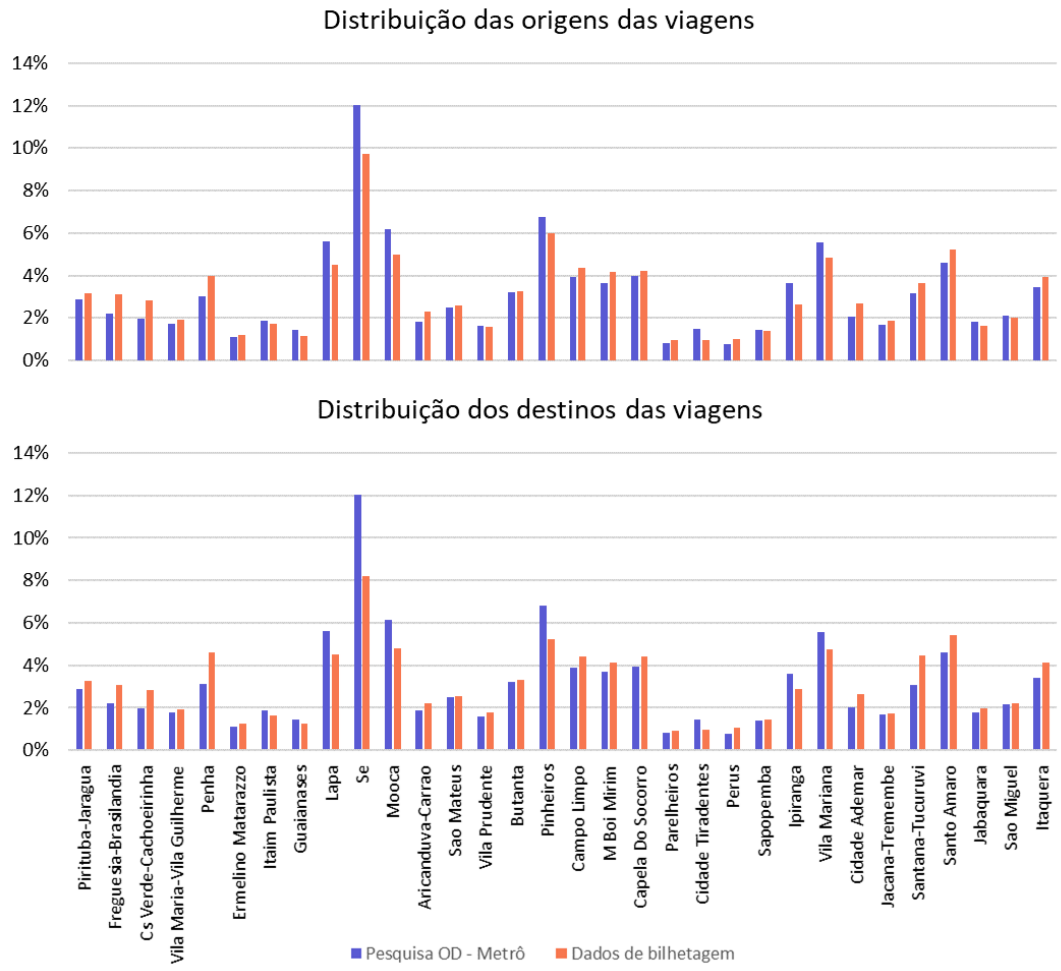
$$Diferença\ total\ absoluta_{Ori} = \sum_{subdistritos} |Pesquisa\ OD_{Ori} - Dados\ bilhetagem_{Ori}| \quad (5)$$

$$Diferença\ total\ absoluta_{Dest} = \sum_{subdistritos} |Pesquisa\ OD_{Dest} - Dados\ bilhetagem_{Dest}| \quad (6)$$

O resultado da diferença total para as origens e destinos foi de 16,6% e 20,7%, respectivamente.

Uma análise comparativa adicional entre as matrizes é realizada ao examinar os pares de Origem-Destino (OD). Esta comparação é conduzida de forma gráfica, explorando as linhas de desejo de cada matriz.

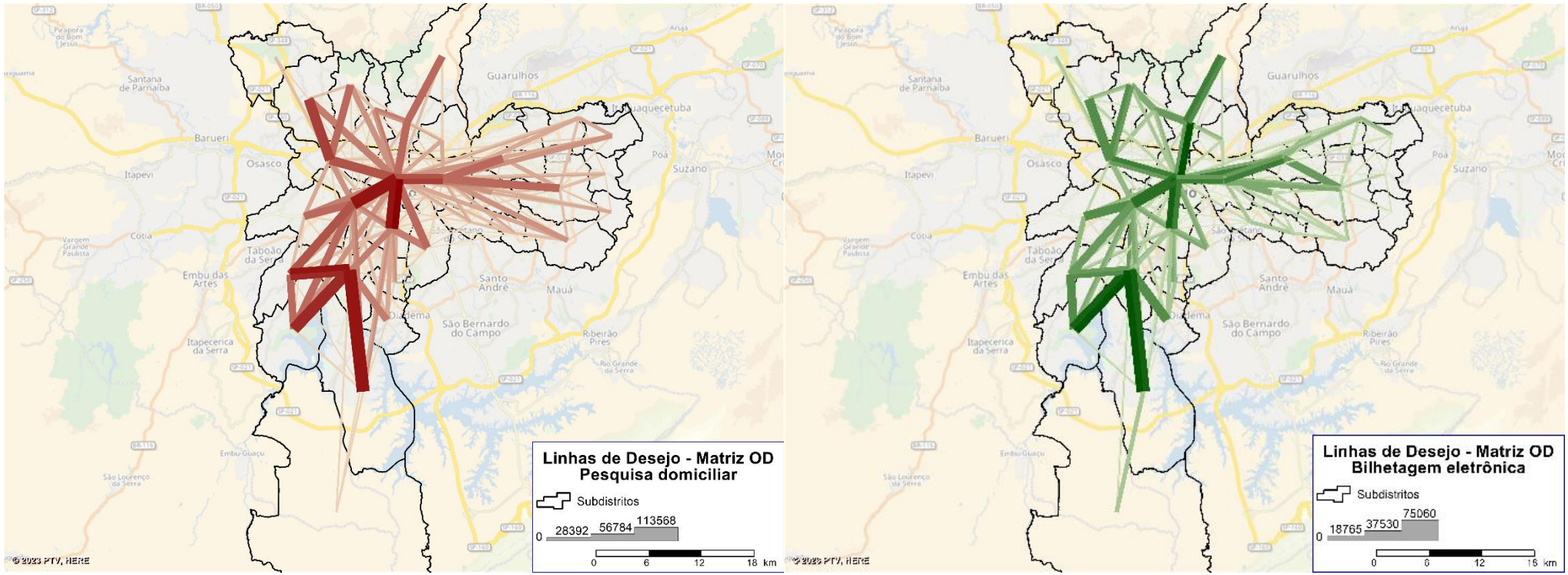
Figura 22 – Comparação da distribuição das origens e destinos por subdistrito



Fonte: Elaboração própria

A Figura 23 ilustra as linhas de desejo das matrizes lado a lado. Notavelmente, os principais pares de viagens entre zonas exibem semelhanças nas matrizes OD. No entanto, os pares com volumes menores concentram-se mais na matriz proveniente da pesquisa domiciliar. Essa disparidade pode ser atribuída à maior variabilidade nos pares Origem-Destino advindos dos dados de bilhetagem eletrônica. Esta distinção também é evidente na quantidade de pares OD distintos com pelo menos uma viagem. Dos 1.024 pares de viagem possíveis (32 x 32 subdistritos), a matriz OD da pesquisa domiciliar registrou pelo menos uma viagem em 825 pares. Em contraste, a matriz OD estimada pela bilhetagem eletrônica apresentou pelo menos uma viagem em 978 pares diferentes.

Figura 23 – Comparação das linhas de desejo das matrizes OD (Pesquisa domiciliar à esquerda e bilhetagem eletrônica à direita)



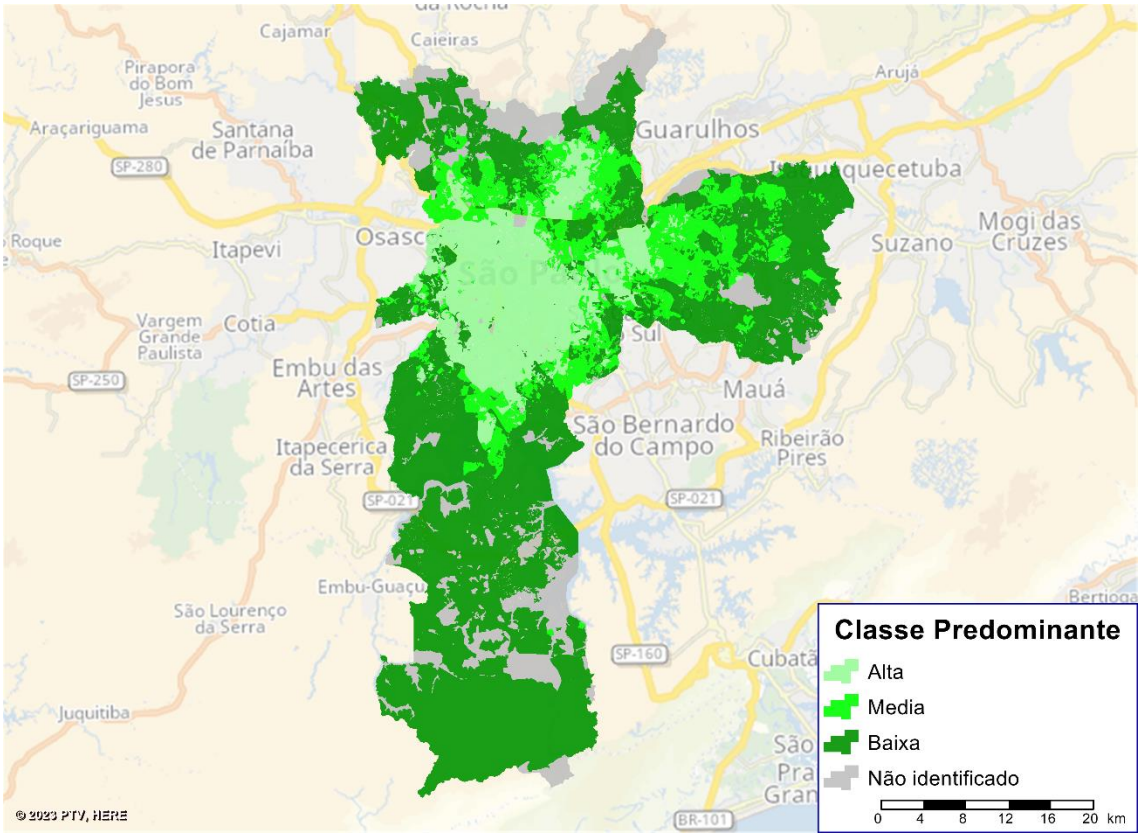
Fonte: Elaboração própria

4.1.4 União de dados socioeconômicos

O cruzamento dos dados de bilhetagem eletrônica com dados socioeconômicos adiciona uma camada de complexidade à análise, permitindo uma compreensão melhor dos padrões de viagem em relação a fatores sociais e econômicos. Neste estudo, esta integração envolve a fusão de informações provenientes da base de dados de bilhetagem eletrônica com dados de classe social, raça/cor e uso predominante do solo.

A inclusão da classe social oferece o entendimento sobre as preferências de viagem e comportamentos de deslocamento de diferentes estratos sociais. Utilizou-se neste estudo uma camada de classe social predominante para os setores censitários da cidade de São Paulo (IBGE, 2010), apresentada na Figura 24. As classes utilizadas foram: Classe Alta, classe média e classe baixa.

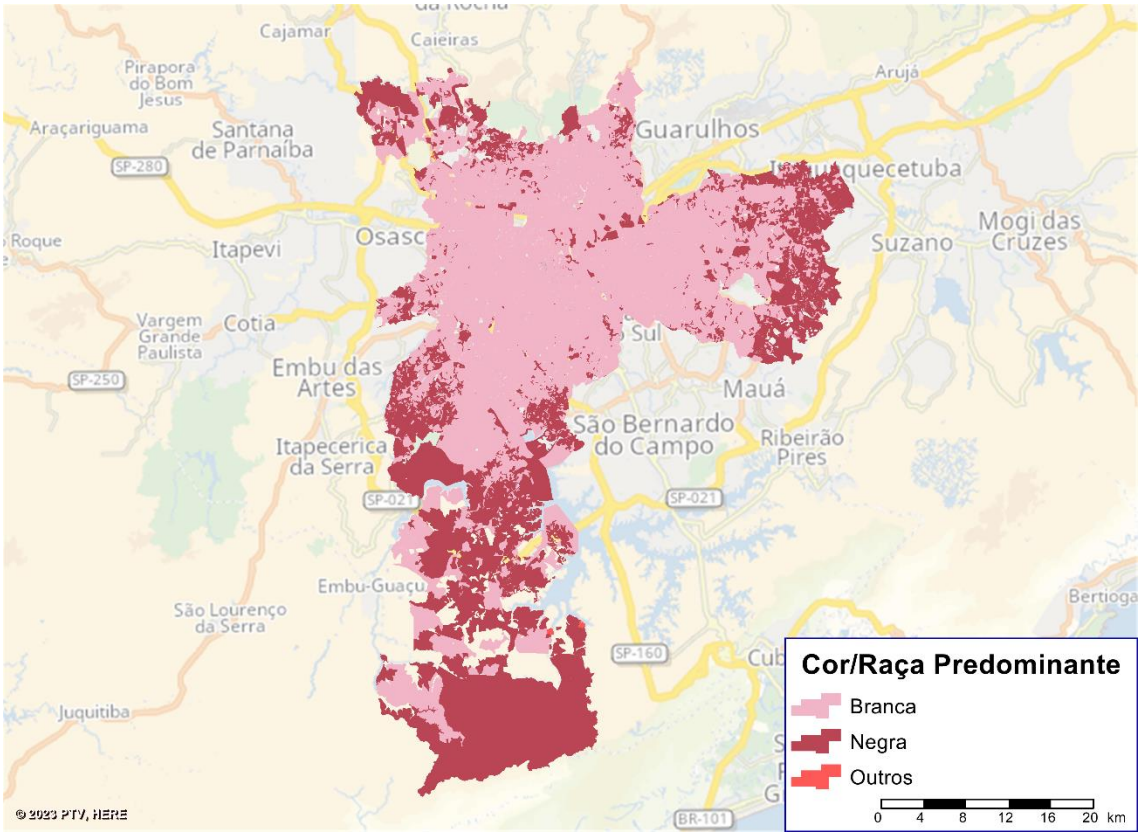
Figura 24 – Classe social predominante por setor censitário



Fonte: Elaboração própria

Da mesma forma, a inclusão de dados sobre raça/cor pode lançar luz sobre questões de equidade e acessibilidade no sistema de transporte público (Bittencourt; Giannotti; Marques, 2021). A análise das viagens em relação a variáveis raciais pode destacar padrões específicos de mobilidade e identificar áreas onde intervenções podem ser necessárias para garantir um acesso equitativo aos serviços de transporte público. Utilizou-se uma camada de raça/cor predominante para cada setor censitário (Figura 25), de maneira análoga à classe social. As classificações utilizadas foram as seguintes: Branca, Negra e Outros.

Figura 25 – Cor/Raça predominante por setor censitário

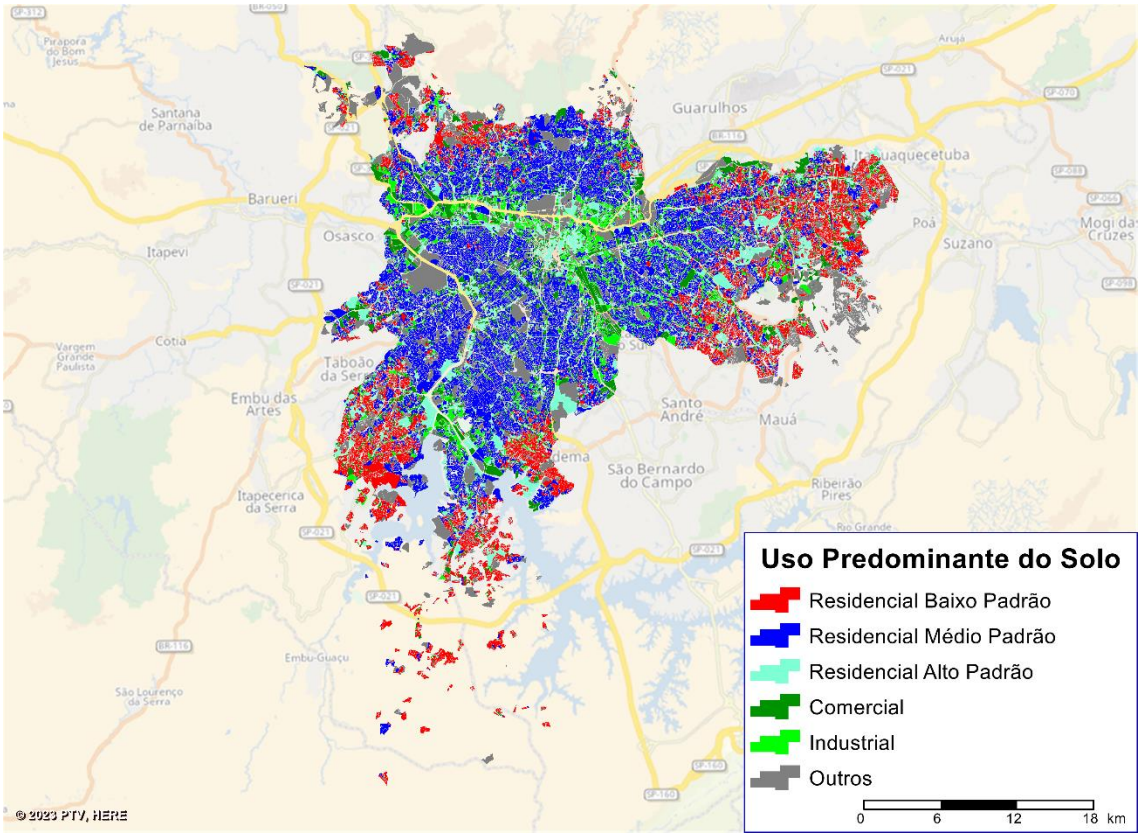


Fonte: Elaboração própria

Além disso, o uso predominante do solo fornece um contexto adicional para as viagens, considerando o ambiente urbano em que os usuários embarcam e desembarcam. A diferenciação entre áreas residenciais, comerciais e industriais pode impactar os padrões de deslocamento. As informações sobre o uso do solo foram obtidas pela prefeitura de São Paulo. A partir dos dezesseis tipos originais de uso do solo do município, realizou-se um agrupamento para

reduzi-los a seis tipos, apresentado no mapa da Figura 26. Os grupos foram classificados como: Residencial de baixo padrão; residencial de médio padrão, residencial de alto padrão, comercial, industrial e outros.

Figura 26 – Usos Predominantes do Solo



Fonte: Elaboração própria

4.1.5 Inferência do motivo de viagem

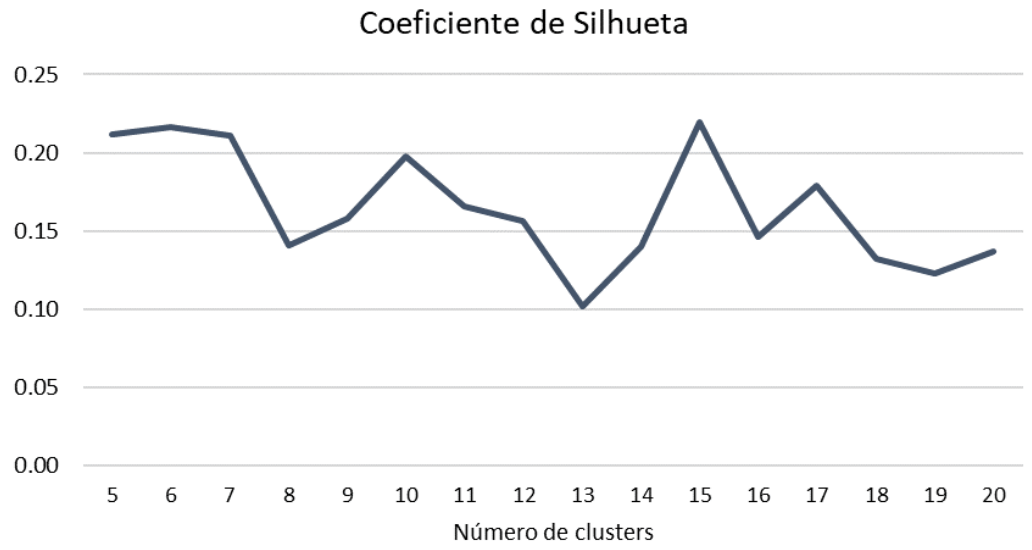
O motivo de viagem é um dos principais atributos ausentes nos conjuntos de dados de bilhetagem eletrônica. Neste estudo, realiza-se a inferência do motivo das viagens dos usuários a partir dos dados da pesquisa OD domiciliar do Metrô de São Paulo, com o intuito de auxiliar na avaliação dos padrões de viagem.

Conforme explicado na seção 3.2.2, metodologia para inferência do motivo de viagem, utiliza-se atributos temporais das viagens dos usuários. Primeiro, para cada usuário de transporte coletivo da pesquisa OD domiciliar, concatena-se a sequência temporal de eventos (horários de embarque e intervalos entre viagens).

O passo seguinte é o agrupamento (clusterização) dessas sequências temporais e seus respectivos motivos de viagem, a fim de se obter uma correlação entre esses dois atributos e a posterior classificação das sequências temporais da base de bilhetagem eletrônica. Neste caso, o método de agrupamento k-means foi escolhido.

Para efetuar o agrupamento, inicialmente é necessário escolher o número ideal de clusters e, para isso, utilizou-se o coeficiente de Silhueta (Rousseeuw, 1987). Com base na literatura (Pieroni et al., 2021; Faroqi; Mesbah, 2021), o número ideal de clusters nos estudos de planejamento de transportes estão geralmente na faixa de 5 a 20 clusters. A Figura 27 apresenta os valores de coeficiente de silhueta para essa faixa adotada. O coeficiente de silhueta é uma métrica utilizada para avaliar a coesão e a separação dos clusters em uma análise de agrupamento. Ele atribui a cada ponto de dados um valor que varia de -1 a 1, indicando o quão similar o ponto é ao seu próprio cluster em comparação com outros clusters próximos. Um valor mais alto sugere que o ponto está bem ajustado ao seu cluster. Neste estudo, o agrupamento com 15 clusters apresentou o maior valor para o coeficiente.

Figura 27 – Coeficiente de Silhueta – Inferência de motivo de viagem



Fonte: Elaboração própria

A Tabela 12 apresenta as sequências temporais de viagens para os grupos descobertos, juntamente com suas etiquetas de motivos de viagem

resultantes para os dados da pesquisa OD domiciliar. Por exemplo, o cluster 1 contém 6588 indivíduos com a sequência de viagem BeEg (esses indivíduos tiveram duas viagens durante o dia: a primeira iniciou no intervalo de tempo "B" (entre 6h e 7h) com uma lacuna de "e" horas antes de sua segunda viagem do dia, que iniciou no intervalo de tempo "E" (entre 16h e 18h), sendo a última viagem do dia - lacuna "g"), que são rotulados como passageiros que saem de sua residência, vão ao trabalho e voltam para casa. A sequência de motivos de viagem 'Residência-Trabalho-Residência' é o rótulo que mais se repete nos clusters.

Tabela 12 – Médias dos clusters de inferência de motivo de viagem

Cluster	Sequência Temporal	Transferências	Motivos de Viagem	Tamanho Cluster
1	BeEg	0	Residencia-Trabalho-Residencia	6588
2	Dg	0	Residencia-Trabalho	1537
3	BeFg	3	Residencia-Trabalho-Residencia	495
4	CeFg	2	Residencia-Trabalho-Residencia	4491
5	BcDg	0	Residencia-Escola/Educacao-Residencia	1827
6	BfFg	0	Residencia-Trabalho-Residencia	337
7	AfFg	1	Residencia-Escola/Educacao-Residencia	385
8	FbFg	0	Residencia-Escola/Educacao-Residencia	425
9	CeFg	0	Residencia-Trabalho-Residencia	1070
10	CeFbFg	2	Residencia-Trabalho-Escola/Educacao-Residencia	758
11	CbDg	0	Residencia-Medico/Dentista/Saude-Residencia	1104
12	DdFg	0	Residencia-Trabalho-Residencia	306
13	DaDg	0	Residencia-Compras-Residencia	958
14	CeEg	0	Residencia-Trabalho-Residencia	905
15	DaDg	2	Residencia-Assuntos Pessoais-Residencia	378

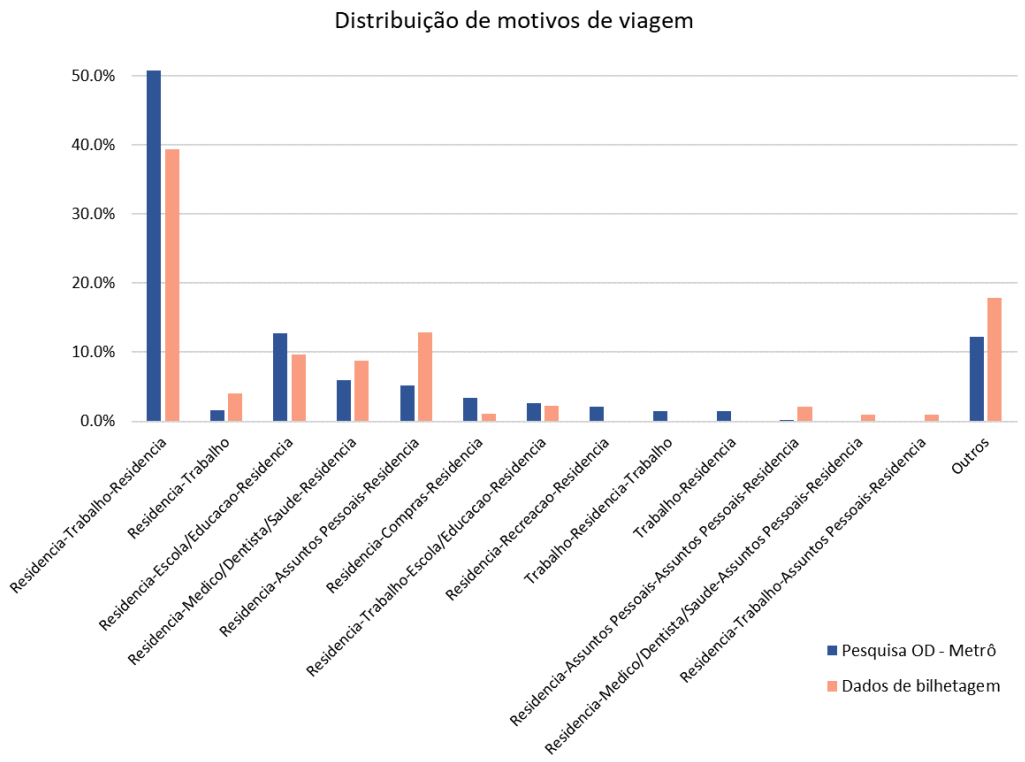
Fonte: Elaboração própria

Após a realização do agrupamento dos dados da pesquisa OD domiciliar, esses clusters são utilizados para classificar as sequências temporais dos usuários na base de bilhetagem eletrônica. Essa inferência é determinada pelo índice de similaridade de Jaccard, que compara as sequências de viagens dos passageiros com as sequências de viagens de cada cluster. Dessa forma, o motivo da viagem inferido para uma transação corresponde ao rótulo do cluster ao qual ela foi associada.

Para as sequências temporais que não foram inicialmente associadas aos clusters, devido à maior variabilidade de padrões existentes nos dados de bilhetagem eletrônica, utiliza-se também o índice de Jaccard para inferir os motivos de viagem para cada sequência a partir da maior quantidade de elementos em comum.

A Figura 28 apresenta a distribuição de motivos de viagem para a base da pesquisa Origem-Destino domiciliar e para a base de dados de bilhetagem eletrônica, após o processamento de inferência descrito anteriormente.

Figura 28 – Comparação da distribuição de motivos de viagem



Fonte: Elaboração própria

4.2 Estrutura da base de dados

A estrutura do banco de dados de bilhetagem eletrônica, processado e enriquecido com as análises realizadas anteriormente para a avaliação dos padrões de viagem, é apresentada abaixo.

Essa estrutura foi concebida a partir de estudos anteriores, incluindo trabalho de Pieroni et al. (2021), Morency et al. (2007), Ortega-Tong (2013) e Zhao et al. (2017). Embora existam variações na definição de atributos, todos esses estudos organizaram seus conjuntos de dados com base nos perfis dos passageiros. Esses perfis são caracterizados pela distribuição de todas as suas validações ao longo do período de análise e por variáveis específicas que descrevem as rotinas de viagem de cada usuário. A estrutura adotada nessa pesquisa serve como conjunto de variáveis para o agrupamento, visando

identificar segmentações de passageiros em grupos específicos. Cada linha na base de dados representa um usuário da base de bilhetagem eletrônica, destacando suas características de viagem durante o dia analisado.

O conjunto de atributos deste estudo é composto pelas seguintes 6 variáveis categóricas:

- Sequência temporal: A sequência temporal representa o padrão de deslocamento temporal de um usuário específico ao longo do dia com base nos horários de início de suas viagens e nos intervalos de tempo entre essas viagens. Cada caractere na sequência corresponde a uma classificação específica de horário de início da viagem ou intervalo de tempo entre viagens, conforme definido nas tabelas fornecidas.
- Sequência de Motivos de viagem: A sequência de motivos de viagem descreve os principais propósitos ou destinos associados às atividades de deslocamento de um usuário específico ao longo do dia, inferido a partir da sequência temporal de cada indivíduo. Cada segmento na sequência representa uma etapa ou atividade específica durante o percurso do usuário.
- Quantidade de transferências: O atributo de quantidade de transferências representa a soma, para todas as viagens, do número de vezes que o usuário realiza uma mudança de modo de transporte (ônibus para metrô, por exemplo) ou de linha de ônibus, dentro da mesma viagem.
- Sequência de zonas de classes sociais: A sequência de zonas de classes sociais indica as características socioeconômicas das áreas por onde o usuário transitou durante suas viagens. Cada segmento na sequência representa uma zona específica associada a uma predominância de um determinado nível de renda. Ficaram assim associados dados de classe social por setor censitário, onde a codificação criada indica qual a predominância da classe naquele setor, com A para alta, M para média e B para baixa.
- Sequência de zonas de raça/cor: A sequência de predominância de raça/cor em zonas percorridas por um usuário, reflete as

características demográficas das áreas por onde o usuário viajou. Cada segmento na sequência representa uma zona específica associada a uma categoria de raça ou cor. A caracterização demográfica do setor censitário onde ocorreu embarque ou desembarque foi feita indicando se o setor possui uma proporção de autodeclaração como negro acima da média de todos os setores da cidade, com S (sim) para acima ou igual a média e N (não) para abaixo da média.

- Sequência de zonas de uso do solo: A sequência de predominância do uso do solo em zonas percorridas por um usuário representa as características das áreas urbanas atravessadas durante suas viagens. Cada segmento na sequência reflete a predominância do uso do solo em uma determinada zona, sendo RB para residencial baixa renda, RM para residencial média renda, RA para residencial alta renda, C para comércio, I para uso industrial e de armazéns e O para outros.

A Tabela 13 ilustra um exemplo da estrutura da base de dados, com os atributos de alguns usuários. Por exemplo, o usuário 1000067 (ID do cartão) realiza 3 viagens ao longo do dia. Este usuário entra no sistema de transporte coletivo entre 7h e 10h (C), tem o primeiro intervalo entre viagens entre 7h e 9h (d), ou seja, atividade nesse destino dura cerca de 8 horas. Já a segunda viagem tem início entre 16h e 18h, com um tempo da segunda atividade menor de 3h. Já a terceira e última viagem se inicia após as 18h. Este usuário tem sua primeira origem em um local predominantemente de baixa renda, tem seu destino da primeira viagem em uma área de alta renda, destino da sua segunda viagem em região de baixa renda e regressa para sua região de origem. Com relação à característica demográfica, esse usuário tem origem em uma área de predominância de raça/cor negra acima da média, tem seu primeiro destino numa região de predominância de raça/cor branca, seu segundo destino numa região de predominância de raça/cor negra acima da média volta para sua região inicial. Em relação aos motivos de viagem deste usuário, a sequência realizada foi Residência, Trabalho, Assuntos Pessoais e Residência.

O usuário 1000166, por outro lado, realiza duas viagens ao longo do dia. Este usuário entra no sistema antes das 6h e inicia a sua segunda e última viagem do dia após as 18h. Este usuário tem sua primeira origem em um local predominantemente de baixa renda e tem seu destino da viagem em uma área de alta renda. Já sua segunda viagem é no sentido contrário, origem em região de alta renda e destino em região de baixa renda. Com relação à característica demográfica, esse usuário tem origem em uma área de predominância de raça/cor negra acima da média, destino numa região de predominância de raça/cor branca e seu destino final numa região de predominância de raça/cor negra acima da média novamente. Em relação aos motivos de viagem deste usuário, a sequência realizada foi Residência, Trabalho e Residência, com os usos do solo das regiões transitadas na seguinte sequência: residencial de baixa renda, residencial de alta renda e residencial de baixa renda.

Tabela 13 – Exemplo de usuários da estrutura do banco de dados para avaliação dos padrões de viagem

ID_Cartão	Seq_temporal	Transferências	Seq_UsoSolo	Seq_PredRaça	Seq_PredClasse	Seq_Motivos
1000067	CdEaFg	1	O-O-RM-RA-RA-RA	S-N-N-S-S-N	B-A-A-B-B-A	Residencia-Trabalho-Assuntos Pessoais-Residencia
1000070	BeEaFg	2	RM-RA-O-O-I-O	N-N-N-N-N-N	A-M-M-M-M-A	Residencia-Trabalho-Assuntos Pessoais-Residencia
1000089	CfFg	0	RM-RM-RM-O	N-N-N-N	A-A-A-A	Residencia-Trabalho-Residencia
1000091	AfFg	2	RB-RA-RA-RM	S-S-S-S	M-M-M-M	Residencia-Trabalho-Residencia
1000112	AeEg	1	RM-C-C-RA	N-N-N-N	M-A-A-B	Residencia-Trabalho-Residencia
1000147	BcDcFaFg	0	RM-RA-C-O-O-O-O-RA	N-N-N-N-N-N-N-N	A-A-M-M-M-M-M-A	Residencia-Escola/Educacao-Residencia-Recreacao-Residencia
1000155	DaDaDg	3	RM-O-O-C-RA-RM	N-N-N-N-N-N	M-M-M-A-A-M	Residencia-Assuntos Pessoais-Assuntos Pessoais-Residencia
1000158	DdFg	0	RM-RM-C-RB	N-S-S-S	B-B-B-B	Residencia-Trabalho-Residencia
1000164	CaCg	4	C-RM-RM-RM	N-N-N-N	A-A-A-A	Residencia-Medico/Dentista/Saude-Residencia
1000166	AfFg	0	RB-RA-RA-RB	S-N-N-S	M-A-A-M	Residencia-Trabalho-Residencia
1000168	CfFg	1	O-O-O-O	S-S-S-S	B-B-B-B	Residencia-Trabalho-Residencia
1000176	DaDg	0	RM-O-RM-RM	N-N-N-N	A-A-A-A	Residencia-Assuntos Pessoais-Residencia
1000180	Fg	0	RM-RA	N-S	A-B	Trabalho-Residencia
1000184	DdFg	4	RM-RM-RA-RB	S-N-N-S	B-A-A-B	Residencia-Trabalho-Residencia
1000204	Bg	1	I-RA	N-S	A-B	Residencia-Trabalho
1000235	Dg	0	RA-RA	N-N	A-A	Escola/Educacao-Residencia

Fonte: Elaboração própria

5. Análise dos padrões de viagem e desigualdades

Este capítulo concentra-se nas principais análises dos padrões de viagem emergentes a partir dos dados do cartão inteligente, explorando possíveis disparidades e desigualdades. Será descrito o método de clusterização no item 5.1 e será realizada uma investigação dos agrupamentos identificados, examinando como diferentes grupos de usuários apresentam padrões de viagem distintos. Além disso, serão abordadas questões relacionadas a desigualdades observadas nos padrões de mobilidade, considerando variáveis como classe social, raça/cor e uso do solo.

5.1 Método de clusterização

A técnica de agrupamento, também conhecida como clusterização, envolve a divisão de uma base de dados em grupos naturais denominados clusters. Dentro de um cluster, os elementos são altamente semelhantes, enquanto elementos entre clusters são distintos (Zaki; Meira, 2013). Essas técnicas de agrupamento são valiosas na análise exploratória de padrões, categorização, tomada de decisões e aprendizado de máquina, abrangendo áreas como mineração de dados e classificação de padrões (Jain; Murty; Flynn, 1999).

Como a base de dados do cartão inteligente não possui informações prévias sobre categorias de passageiros com base em seus padrões de viagem, é necessário realizar um processo de agrupamento, uma forma de classificação não supervisionada, para identificar grupos distintos de usuários com padrões de viagem semelhantes nos dados de bilhetagem eletrônica.

Como citado anteriormente, os atributos da base de dados de bilhetagem eletrônica processados para avaliação dos padrões de viagem são categóricos, ou seja, são atributos que podem assumir um número limitado de valores, sendo esses, categorias ou grupos pré-definidos. Por este motivo, se utilizará o método k-modes de clusterização.

O método de clusterização k-modes é uma extensão do k-Means (Huang, 1997; Chaturvedi; Green; Caroll, 2001), desenvolvida para lidar com dados

categóricos, sendo particularmente útil quando a maioria das variáveis são nominais. Utilizando distâncias específicas, como a de Hamming, o k-modes agrupa instâncias de dados com características semelhantes, utilizando a moda do conjunto de dados, minimizando a dissimilaridade entre esses grupos. O algoritmo começa escolhendo k centroides iniciais (representantes do cluster), que podem ser selecionados de maneira aleatória ou usando alguma heurística. Em seguida, atribui cada instância de dados ao cluster cujo centroide é o mais próximo em termos de dissimilaridade. Posteriormente, recalcula os centroides dos clusters e itera esse processo até alcançar a convergência. (Huang, 1997).

Khan e Ahmad (2013) definem o modelo matemático de Hamming. Sejam X e Y dois objetos de dados categóricos descritos por m atributos categóricos. A medida de dissimilaridade $d\ (X, Y)$ entre X e Y pode ser definida pelo total de incompatibilidades das categorias de atributos correspondentes dos dois objetos. Quanto menor o número de incompatibilidades, mais similares são os dois objetos, expressado desta maneira:

$$d\ (X, Y) = \sum_{j=1}^{m} d(x_j, y_j) \qquad (7)$$

$$onde, \quad d(x_j, y_j) = \begin{cases} 0\ se\ x_j = y_j \\ 1\ se\ x_j \neq y_j \end{cases}$$

5.1.1 Número de clusters

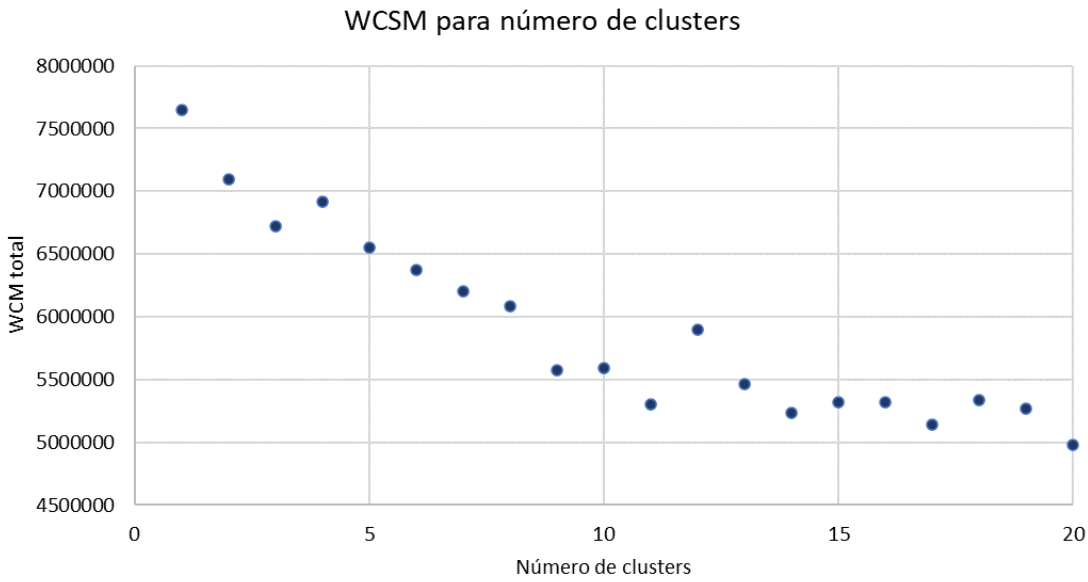
Determinar o número ideal de clusters é uma etapa necessária no processo de aplicação do algoritmo k-modes. Em contraste com o método do cotovelo tradicional que utiliza a medida Within-Cluster Sum of Squares (WCSS) em dados numéricos, para dados categóricos, como no k-modes, emprega-se uma abordagem que considera a medida Within-Cluster Simple Matching (WCSM).

O WCSM reflete a soma das diferenças entre as modas dos clusters (distância de Hamming explicada anteriormente), proporcionando uma métrica adequada para a compactação dos modos dentro de cada agrupamento. Para determinar o número ótimo de clusters, executa-se o algoritmo k-modes para diferentes valores de k, calcula-se o WCSM correspondente a cada configuração e observamos o ponto em que o ganho marginal na redução do WCSM diminui.

Neste estudo, o algoritmo k-modes foi executado para 20 números de clusters diferentes (1 a 20).

A Figura 29 apresenta o WCSM para cada número de clusters. Observa-se que, para um número de cluster acima de 12, os valores de WCSM são parecidos. Portanto, escolheu-se o número de 15 clusters para a análise de padrões de viagem por apresentar um dos menores valores de WCSM e também por ser o número de clusters ótimo utilizado na inferência de motivo de viagem.

Figura 29 – WCSM para escolha do número de clusters



Fonte: Elaboração própria

5.2 Análise dos agrupamentos de padrões de viagem

Nesta seção, são apresentados distintos aspectos dos resultados do agrupamento realizado. O objetivo central é compreender os padrões de viagem distintos a partir da análise de similaridades e variabilidades dos atributos e, mais especificamente, explorar nuances relacionadas à desigualdade social. Ao concentrar em clusters específicos, busca-se investigar como diferentes grupos de usuários contribuem para as dinâmicas de mobilidade.

Ao considerar a variabilidade, busca-se capturar a riqueza e a heterogeneidade inerentes aos comportamentos de viagem dos usuários de

transporte público, reconhecendo que as diferenças individuais podem ser tão instrutivas quanto as semelhanças.

Ao aplicar o k-modes ao conjunto de dados, obtém-se um conjunto de grupos, cada um representando um perfil ou padrão presente nos dados. Cada grupo é caracterizado pelas modas para os 6 atributos categóricos selecionados, indicando as categorias mais frequentes em cada grupo. A Tabela 14 apresenta as modas para cada um dos grupos que foi criado pelo algoritmo.

Tabela 14 – Resultado dos clusters e suas modas para cada atributo da base

Cluster	Seq_temporal	Transferências	Seq_UsoSolo	Seq_PredRaça	Seq_PredClasse	Seq_Motivos	Prop_Cluster
1	Dg	0	RA-O	N-N	A-A	Escola/Educacao-Residencia	7.2%
2	AeEg	2	RB-RM-RM-RB	S-N-N-S	B-A-A-B	Residencia-Trabalho-Residencia	4.7%
3	Cg	1	RM-RM	N-N	M-M	Outros	2.7%
4	Dg	1	O-O	S-S	B-B	Escola/Educacao-Residencia	3.8%
5	CeFg	0	RM-RM-RM-RM	N-N-N-N	A-A-A-A	Residencia-Trabalho-Residencia	6.5%
6	Dg	1	RM-RM	N-N	M-M	Outros	2.0%
7	CeFg	1	RB-O-O-RB	S-N-N-S	B-A-A-B	Residencia-Trabalho-Residencia	14.6%
8	Fg	1	RA-RA	N-N	A-A	Trabalho-Residencia	2.2%
9	Fg	0	RA-O	N-N	A-A	Trabalho-Residencia	13.9%
10	DcFg	0	RB-RA-RA-RB	S-N-N-S	B-M-M-B	Residencia-Escola/Educacao-Residencia	4.3%
11	CdEg	2	RA-RM-RM-RA	N-N-N-N	B-M-M-B	Residencia-Trabalho-Residencia	3.8%
12	CeFg	0	RM-RA-RA-RM	N-N-N-N	A-A-A-A	Residencia-Trabalho-Residencia	17.6%
13	BdEg	2	C-O-O-C	S-N-N-N	B-A-A-A	Outros	2.5%
14	CaDg	0	C-RA-RA-C	S-N-N-S	B-A-A-B	Residencia-Medico/Dentista/Saude-Residencia	4.0%
15	DaDg	0	RB-RB-RB-RB	S-S-S-S	B-B-B-B	Residencia-Assuntos Pessoais-Residencia	10.3%

Fonte: Elaboração própria

Da Figura 30 à Figura 44, apresenta-se a distribuição das características de cada grupo derivado da clusterização utilizando k-modes. Os gráficos para cada um dos grupos representam a proporção dos valores dos atributos que predominantes dentro de cada grupo, sendo que o maior valor corresponde à moda apresentada na tabela anterior.

Os passageiros classificados nos Clusters 2, 5, 7, 11 e 12 estão associados a padrões de viagem de passageiros que se deslocam de sua residência para o trabalho e de volta para sua residência, realizando duas viagens durante o dia. O cluster 2 apresenta a hora mais precoce da primeira transação do usuário – antes das 6h. Já os outros 4 clusters apresentam o mesmo intervalo de tempo para a primeira transação – entre 7h e 10h. Com relação ao intervalo entre viagens, o cluster 11 apresenta o menor intervalo - entre 7h e 9h. Os outros 4 clusters apresentam intervalo entre viagens de 9h a 12h. Esse atributo está vinculado ao tempo de atividade dos usuários e sugere um expediente de trabalho em tempo integral. Já em relação à segunda viagem, os clusters 2 e 11 realizam essa transação entre 16h e 18h e, os outros clusters, realizam a transação após as 18h. As diferenças mais significativas entre esses clusters começam a aparecer ao analisar a sequência de classes sociais e predominância de raça das áreas por onde os usuários passam durante o dia. Enquanto os clusters 2, 7 e 11 têm origem em locais de classe social baixa pela manhã, destinos em locais de classe média/alta, fazem pelo menos uma transferência durante o dia e dois desses clusters apresentam predominância de raça negra na origem, os clusters 5 e 12 têm origem e destino em regiões de classe alta, apresentam predominância de raça/cor branca e não realizam nenhuma transferência ao longo do dia. Com relação à representatividade do cluster, o cluster 2 apresenta maior assertividade dos usuários em relação à moda do cluster, principalmente os atributos de motivo de viagem (~83% dos usuários tem a sequência ‘Residência-Trabalho-Residência’), atributo de raça/cor (cerca de 74% dos usuários moram em região de predominância negra acima da média e trabalham em região de predominância branca), e classe predominante (56% moram em região de classe predominantemente baixa e trabalham em região de classe alta). Com relação à sequência temporal, 26% dos usuários são iguais à moda, tendo a primeira transação antes das 6h. Já em relação ao número de transferências, aproximadamente 45% dos usuários desse

grupo realizam duas transferências durante o dia. Observando o uso do solo relativo às áreas percorridas pelos usuários nesse cluster 2, tem-se que os usuários têm origem em uma região classificada como residência de baixa renda e trabalham numa área residencial de média/alta renda, onde a maioria dos empregos disponíveis envolve funções como porteiro, faxineiro, empregada doméstica e ocupações semelhantes. Um cluster semelhante a este é visto em Pieroni et al. (2021), validando as metodologias entre si.

Outro cluster com mais assertividade em relação à moda, a ser detalhado nesse conjunto é o cluster 5. Como o cluster 2, grande parte dos usuários (~65%) atribuídos a esse cluster tem o trabalho como atividade principal, se deslocam apenas em regiões de predominância de raça/cor branca (cerca de 89%) e predominância de classe social alta (70%). Com relação à sequência temporal, cerca de 18% dos usuários são iguais à moda, tendo a primeira transação antes entre 7h e 10h. Já em relação ao número de transferências, aproximadamente 37% dos usuários desse grupo não realizam nenhuma transferência durante o dia.

As diferenças entre o Cluster 2 e o Cluster 5 podem ser atribuídas a diversos fatores que refletem nuances nos padrões de deslocamento e características sociodemográficas dos usuários. A disparidade na composição socioeconômica das áreas de residência e trabalho é um ponto crucial, com o Cluster 2 apresentando uma dinâmica em que os usuários residem em regiões de classe predominantemente baixa e desempenham atividades, que podem estar relacionadas a empregos manuais de baixa remuneração em áreas de média/alta renda. Por outro lado, o Cluster 5 revela uma concentração em regiões de alta classe social. Além disso, as discrepâncias nos padrões de transferência e horários de transação sugerem variações nos estilos de vida e necessidades de mobilidade entre os dois clusters, com implicações diretas nas escolhas de rota e na frequência de transferências ao longo do dia. Essas divergências ressaltam a importância de considerar não apenas os padrões de movimentação, mas também o contexto socioeconômico e de desigualdades sociais no planejamento de transportes do município.

. Diferentemente do primeiro conjunto de clusters analisados, o segundo formado pelos Clusters 10, 14 e 15, se associam a padrões de viagem de usuários que se deslocam para fins de educação e assuntos pessoais durante o

dia (médico, saúde, entre outros). Em comum, tem-se que os intervalos entre viagens é menor do que o encontrado nos clusters analisados anteriormente, sugerindo que o tempo despendido nessas atividades é menor e parcial em relação ao total do dia. Outro ponto em comum é a não realização de transferências por parte desses usuários.

Analisando-os separadamente, o cluster 10 apresenta o intervalo de tempo da primeira transação entre 10h e 16h, tendo aproximadamente 72% dos usuários com escola/educação como motivo de viagem. Observa-se para esse cluster o tempo de intervalo entre viagens de 5h a 7h, o que leva a crer que estes usuários podem ser estudantes do período vespertino (cerca de 40% dos usuários desse cluster possuem essa sequência temporal). Com relação às informações socioeconômicas, esses usuários apresentam origem em região de classe de predominância baixa e destino em região de classe média (38%). Já a predominância de cor/raça é classificada como negra na região de origem e branca no destino (48%).

O cluster 14 apresenta a saúde como motivo principal de viagem, abrangendo aproximadamente 80% dos usuários. Notavelmente, a maioria dos usuários desse cluster realiza viagens com intervalos de até 3 horas, indicando uma dinâmica de deslocamento mais concentrada em um curto espaço de tempo. A análise da origem e destino revela uma disparidade socioeconômica, com a região de classe social baixa predominando na origem e a região de classe social média/alta sendo predominante no destino. A mudança na composição racial/cor, com uma maioria negra na origem e branca no destino (cerca de 42%), sugere que as viagens relacionadas à saúde podem envolver a busca por serviços médicos em áreas com melhores infraestruturas de saúde. Esses padrões indicam que o acesso a cuidados de saúde pode ser mais limitado em regiões de baixa renda, ressaltando desafios de equidade no acesso a serviços médicos para comunidades em situações socioeconômicas desfavorecidas.

O Cluster 15 destaca-se por apresentar o motivo de viagem principal como 'Assuntos pessoais', representando aproximadamente 62% dos usuários. Uma característica notável é a realização dessas viagens geralmente fora do horário de pico, entre 10h e 16h, indicando que os usuários desse cluster podem estar envolvidos em atividades não vinculadas a compromissos de trabalho ou estudo.

Além disso, a presença de um intervalo de até 3 horas entre viagens para cerca de 48% dos usuários sugere uma dinâmica mais flexível e espaçada em comparação com clusters com padrões de deslocamento mais regulares. Em contraste com o Cluster 14, analisado anteriormente, os usuários do Cluster 15 concentram suas atividades em locais que mantêm a predominância de sua origem, residindo em regiões de classe social baixa (61% dos usuários) e de raça/cor negra acima da média (57% dos usuários). Essa preferência por permanecer em áreas similares à sua origem pode indicar a presença de atividades sociais e familiares nessas regiões.

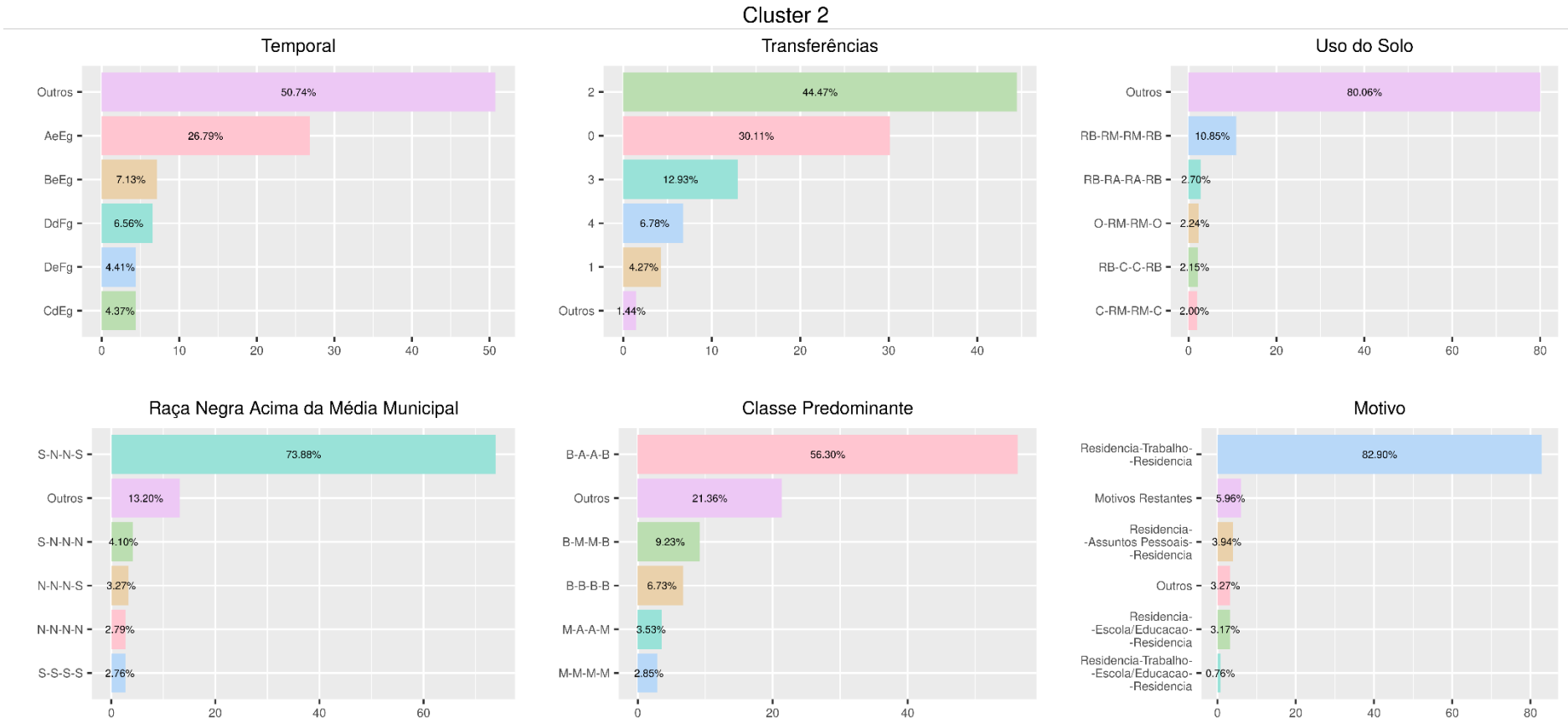
Outro conjunto de clusters analisado (clusters 1, 4, 8 e 9) apresentou usuários com apenas uma viagem de transporte coletivo identificada ao longo do dia. A análise dessas viagens aponta para motivos de viagem predominantemente relacionados à educação e trabalho, conforme indicado pelos horários de deslocamento, classificados com os dados da pesquisa OD domiciliar. Notavelmente, esses clusters também refletem a relação previamente observada em outros agrupamentos, onde regiões de classe alta e predominância de raça/cor branca estão associadas, contrastando com regiões de classe baixa e predominância de raça/cor negra acima da média.

Por último, os clusters 3, 6 e 13 examinados compartilham o motivo de viagem designado como 'Outros' como sua atividade principal. Todos esses clusters apresentam a categoria ‘Outros’ como predominante em todos os atributos analisados. Para além da classificação ‘Outros’, os clusters 3 e 6 se aproximam em relação à predominância de classe social média e predominância de raça/cor branca pelas regiões por onde se deslocam, o cluster 13 apresenta uma viagem entre regiões diferentes com relação a dados socioeconômicos.

Figura 30 – Grupo formado pelo algoritmo k-modes – Grupo 1

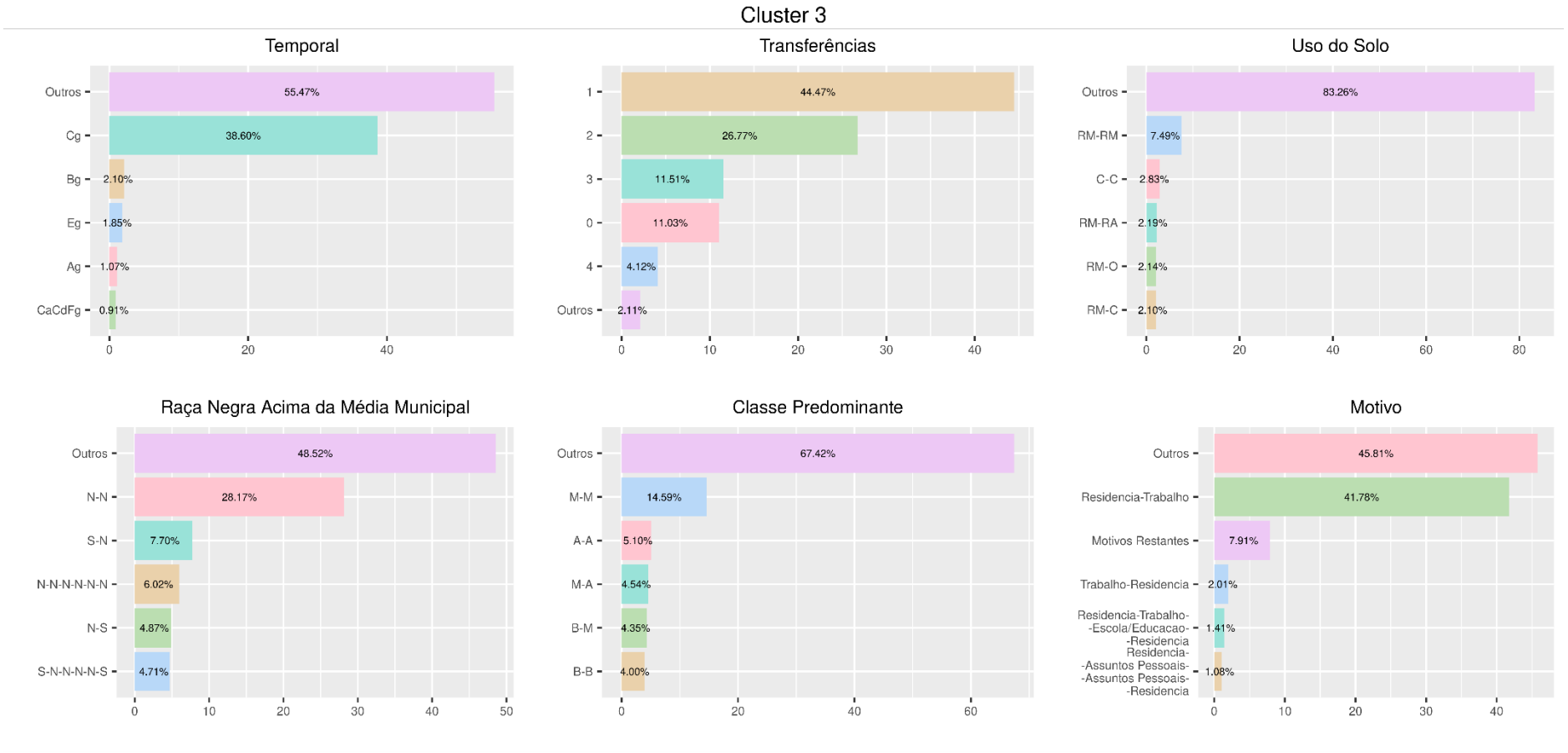
Fonte: Elaboração própria

Figura 31 – Grupo formado pelo algoritmo k-modes – Grupo 2



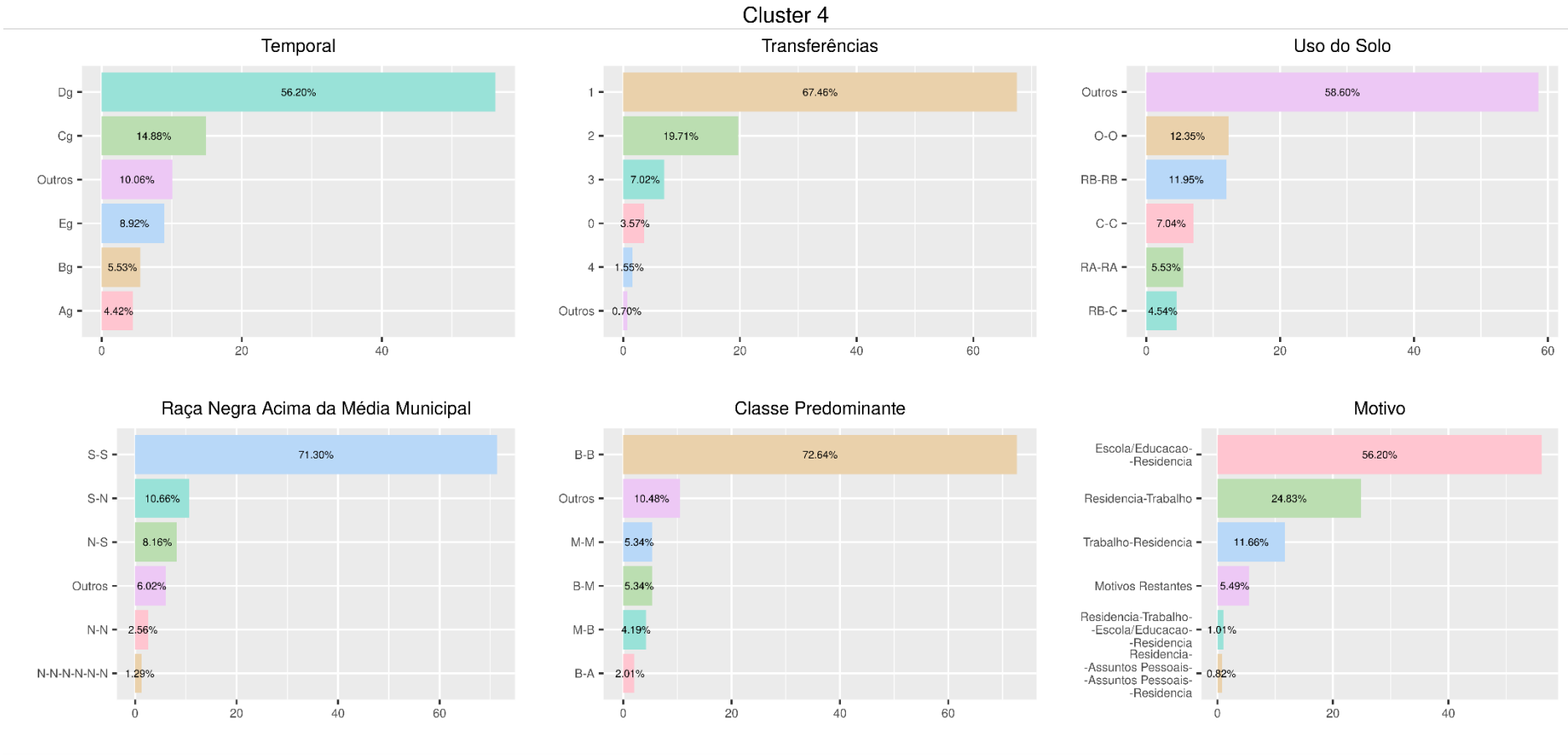
Fonte: Elaboração própria

Figura 32 – Grupo formado pelo algoritmo k-modes – Grupo 3



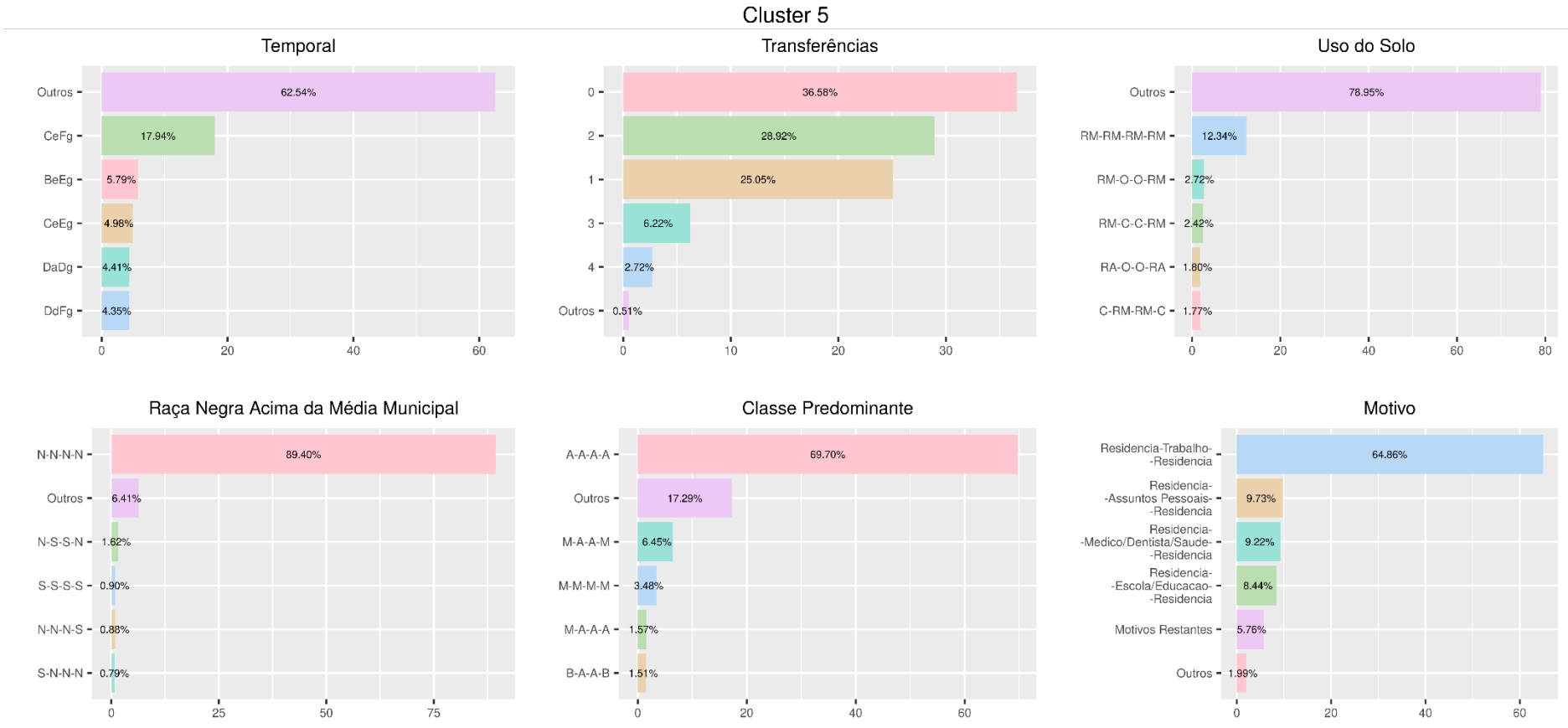
Fonte: Elaboração própria

Figura 33 – Grupo formado pelo algoritmo k-modes – Grupo 4



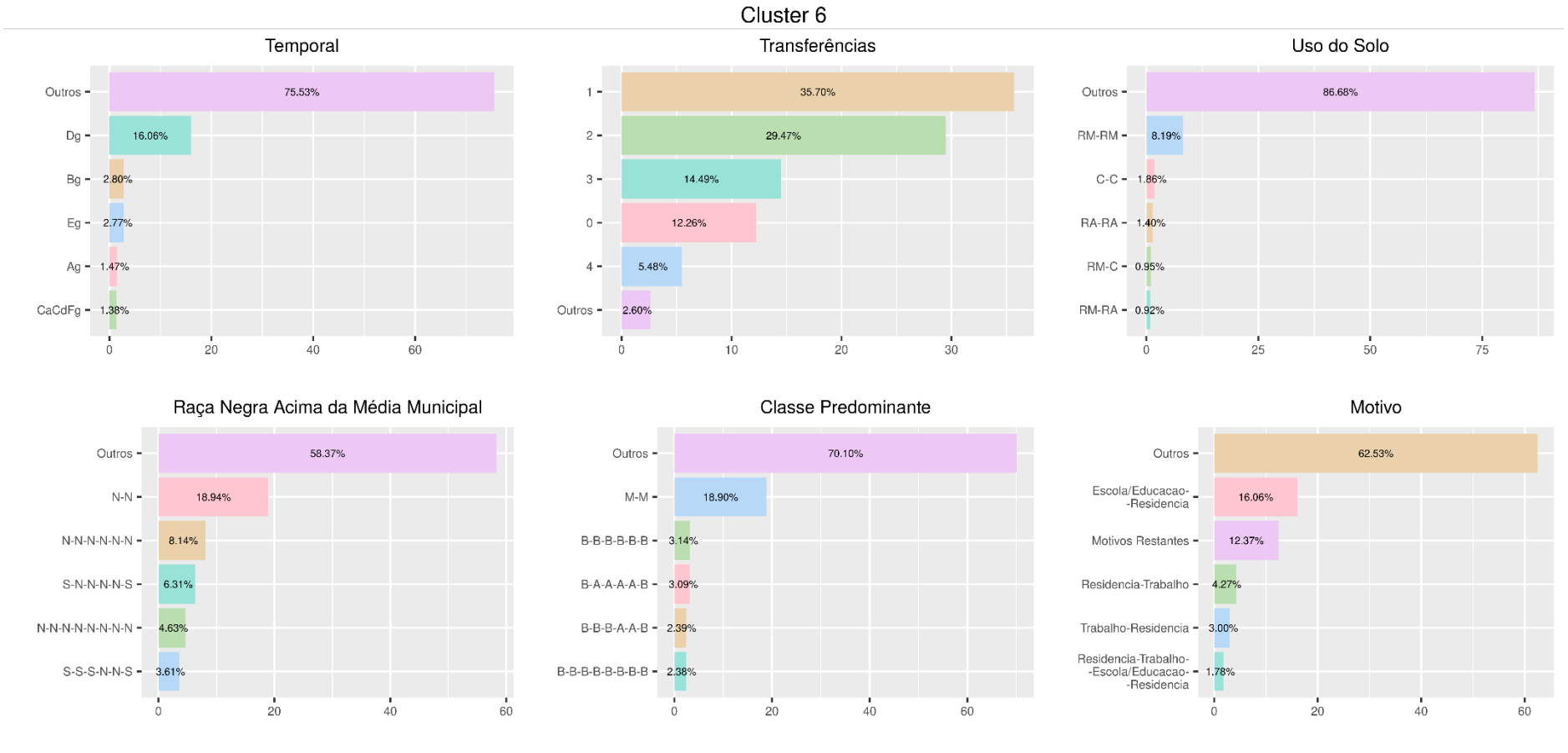
Fonte: Elaboração própria

Figura 34 – Grupo formado pelo algoritmo k-modes – Grupo 5



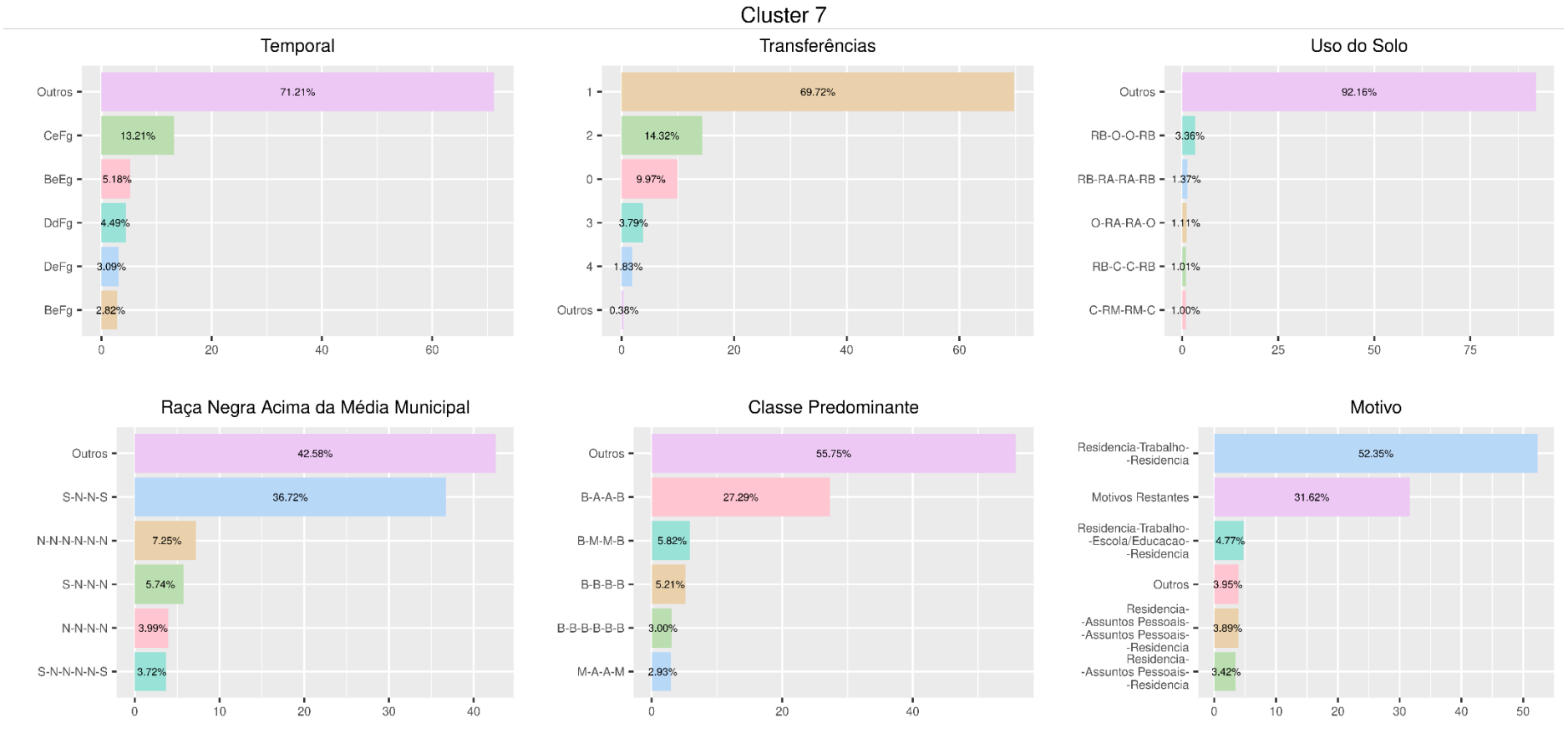
Fonte: Elaboração própria

Figura 35 – Grupo formado pelo algoritmo k-modes – Grupo 6



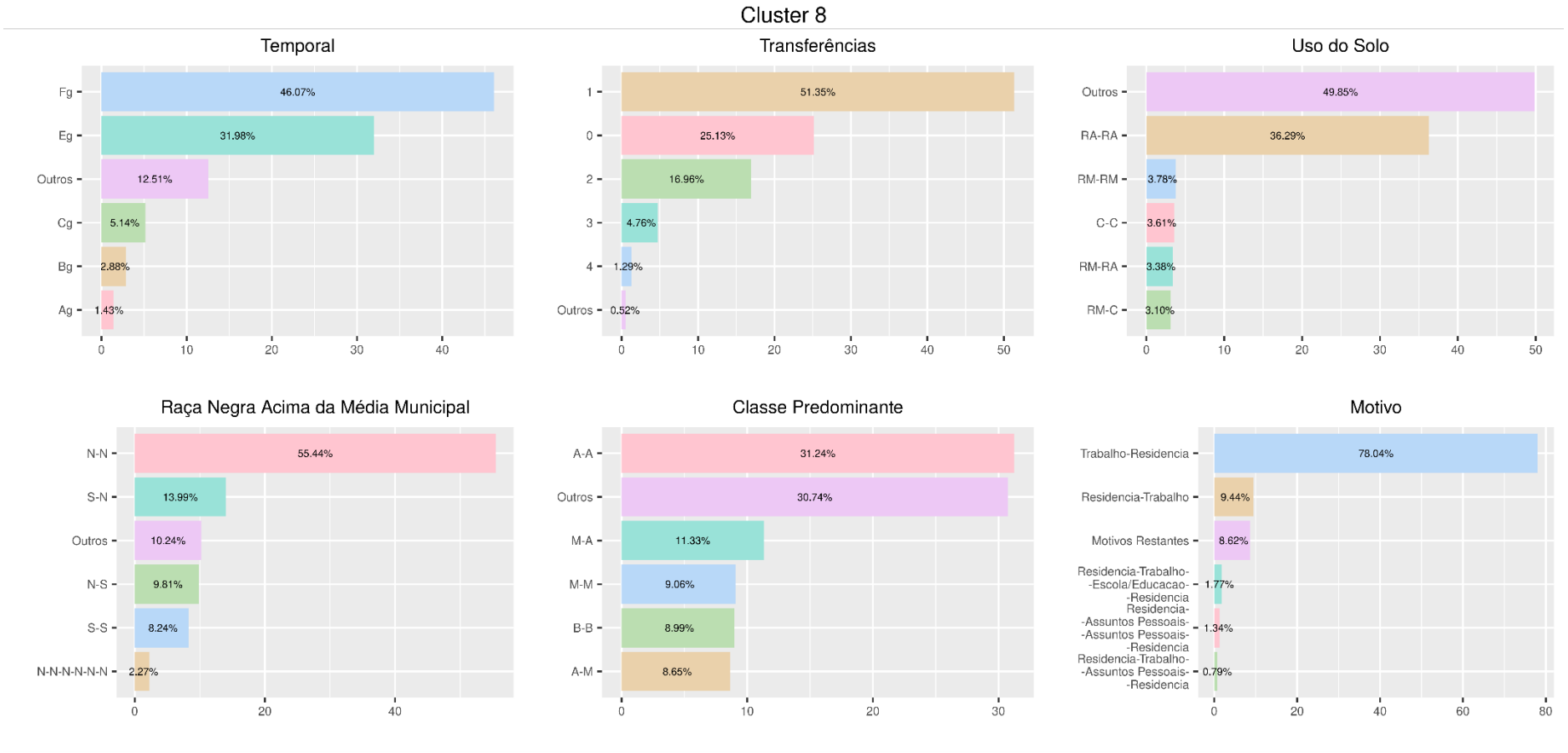
Fonte: Elaboração própria

Figura 36 – Grupo formado pelo algoritmo k-modes – Grupo 7



Fonte: Elaboração própria

Figura 37 – Grupo formado pelo algoritmo k-modes – Grupo 8

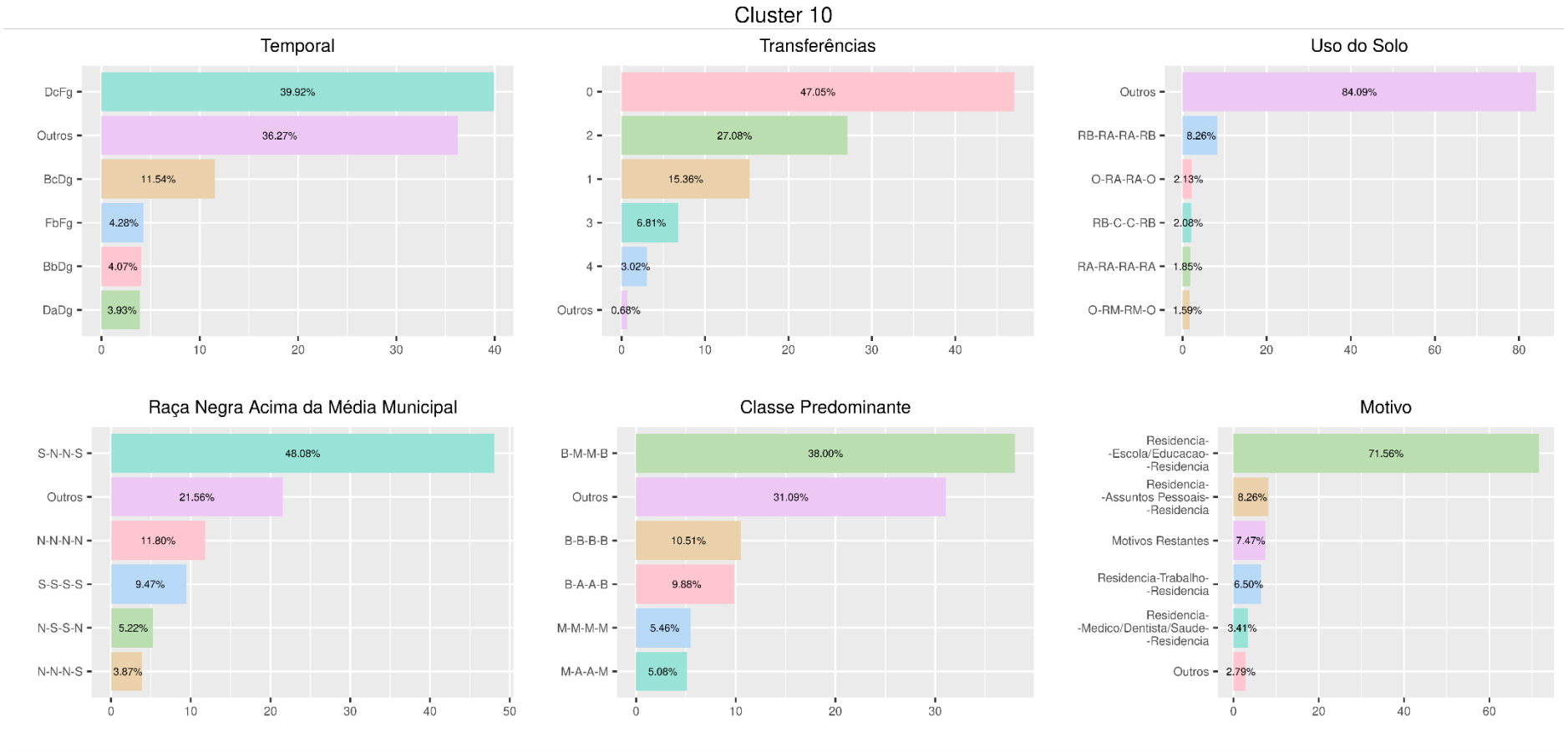


Fonte: Elaboração própria

Figura 38 – Grupo formado pelo algoritmo k-modes – Grupo 9

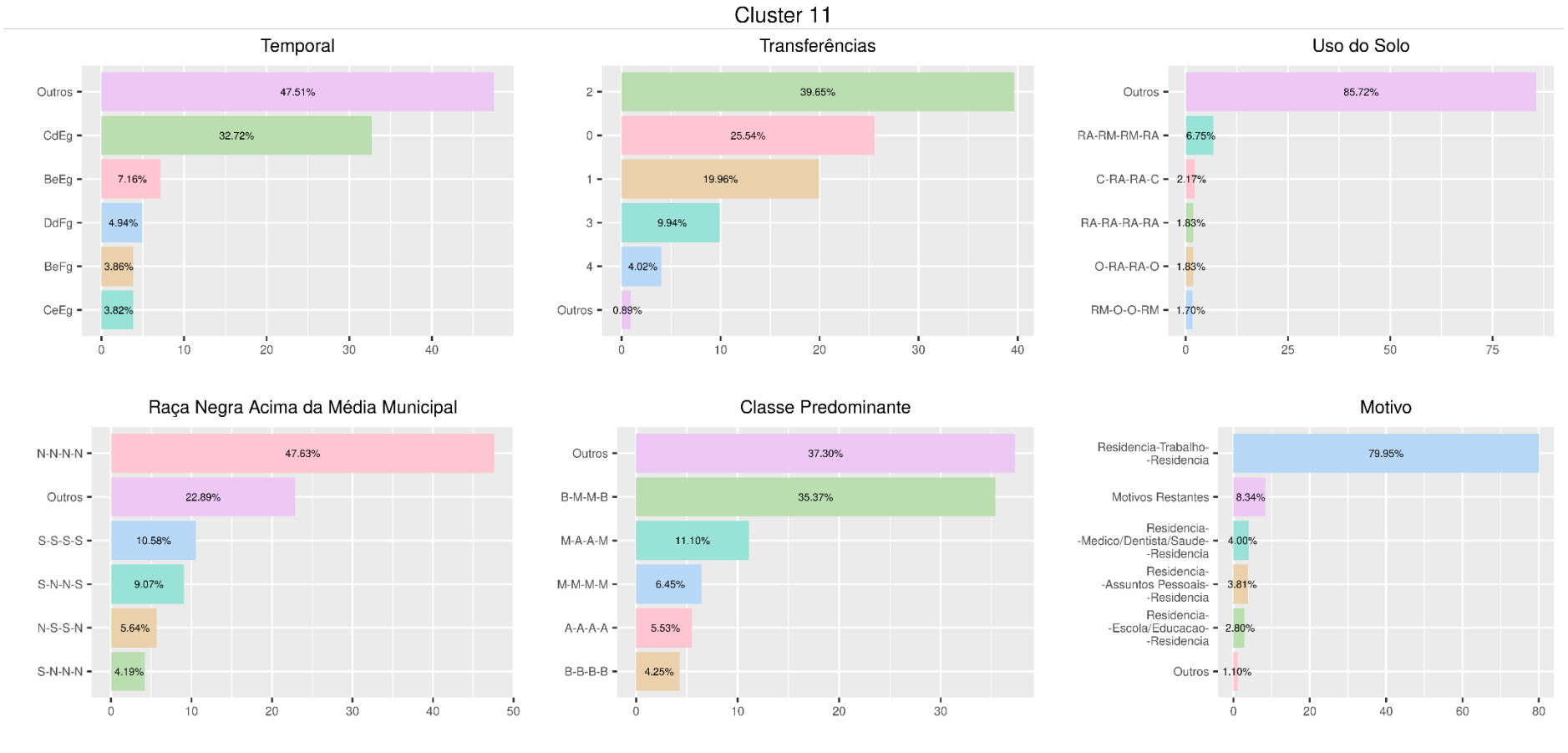
Fonte: Elaboração própria

Figura 39 – Grupo formado pelo algoritmo k-modes – Grupo 10



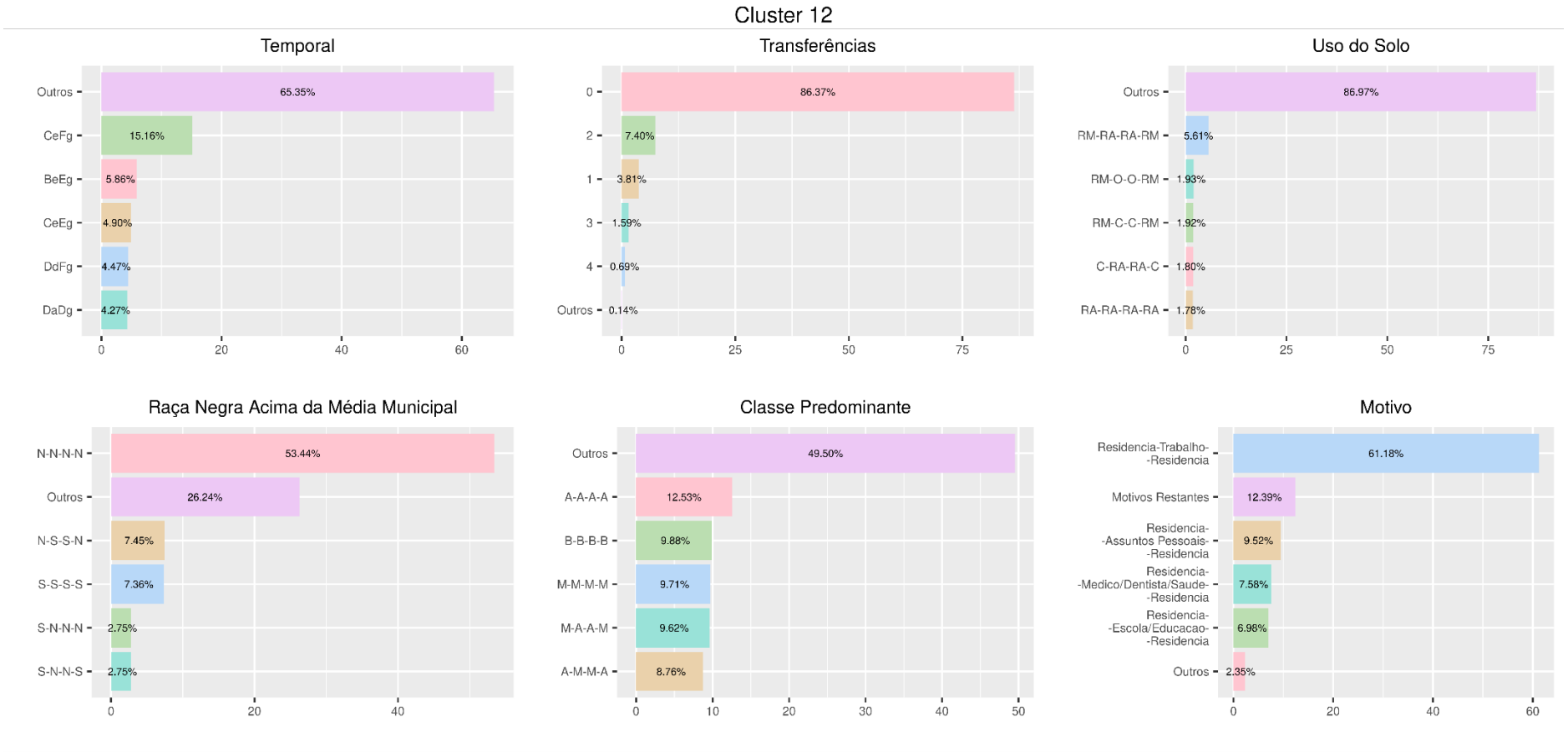
Fonte: Elaboração própria

Figura 40 – Grupo formado pelo algoritmo k-modes – Grupo 11



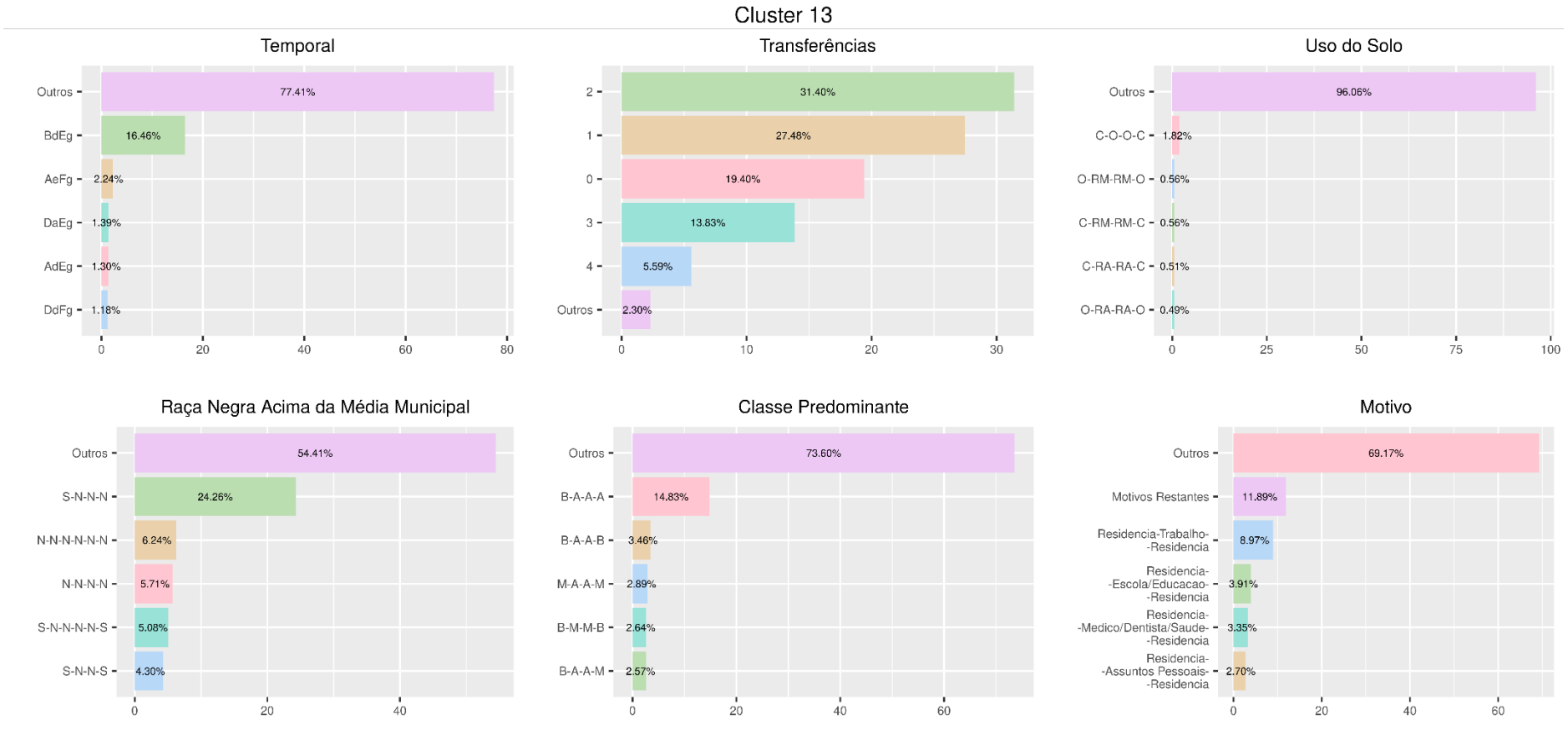
Fonte: Elaboração própria

Figura 41 – Grupo formado pelo algoritmo k-modes – Grupo 12



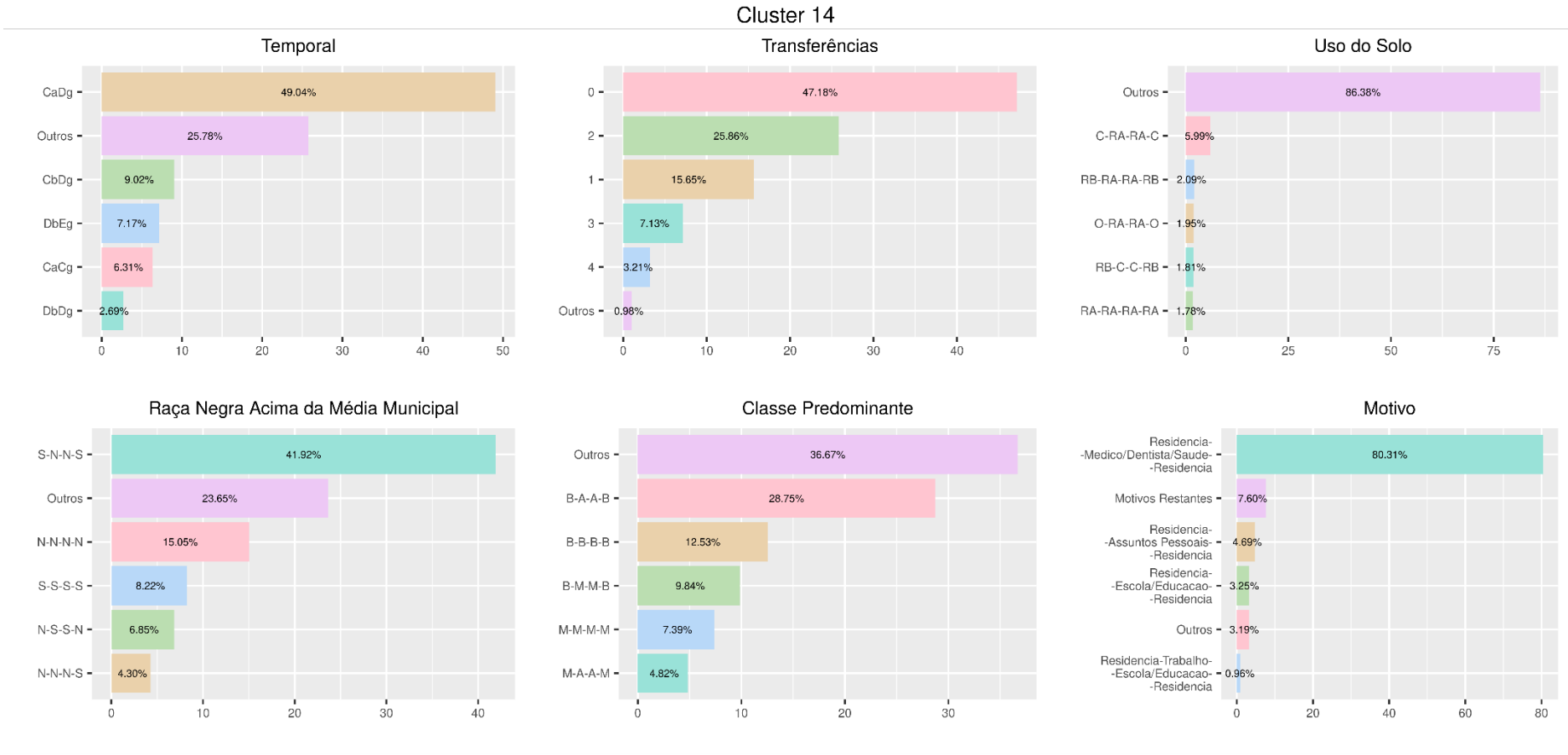
Fonte: Elaboração própria

Figura 42 – Grupo formado pelo algoritmo k-modes – Grupo 13



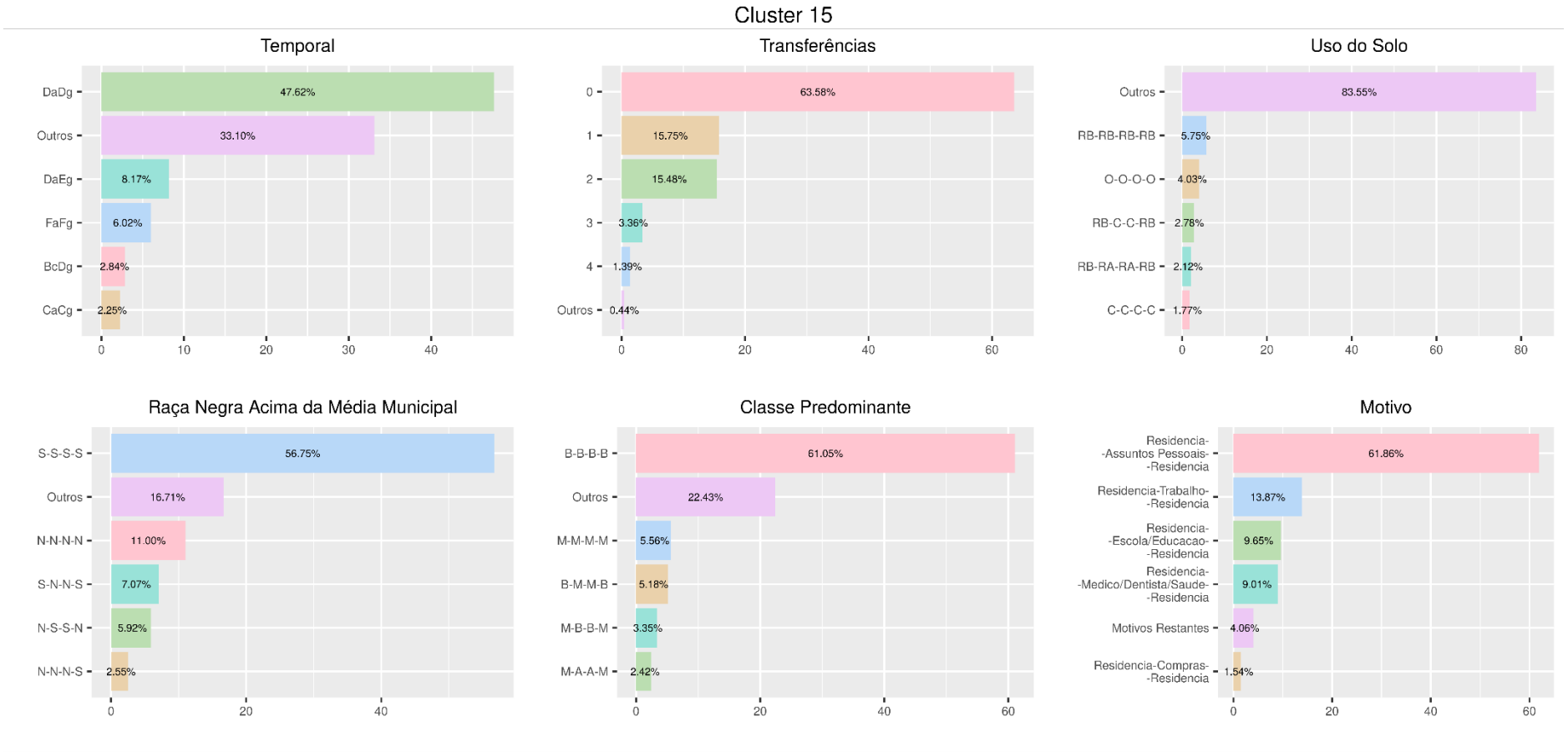
Fonte: Elaboração própria

Figura 43 – Grupo formado pelo algoritmo k-modes – Grupo 14



Fonte: Elaboração própria

Figura 44 – Grupo formado pelo algoritmo k-modes – Grupo 15



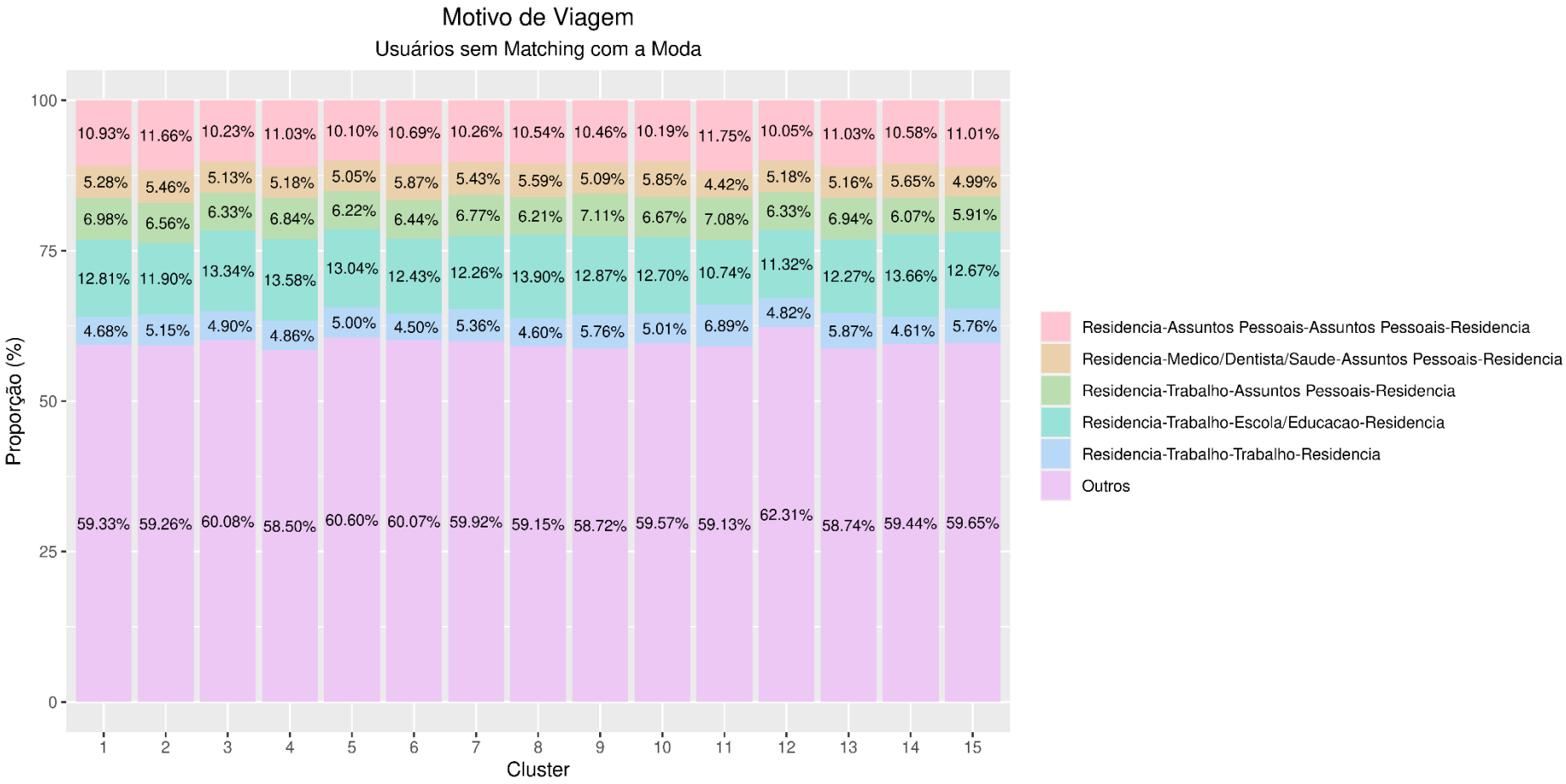
Fonte: Elaboração própria

A análise dos clusters revela que em alguns deles a moda dos atributos não atinge a marca de 50%. Isso significa que mais da metade dos usuários dentro desses clusters não segue precisamente o mesmo padrão de viagem em relação a todos os atributos analisados. Essa diversidade de padrões sugere uma variabilidade significativa dos dados de bilhetagem eletrônica no comportamento de deslocamento dos usuários. Uma outra análise é conduzida, focalizando especialmente nos usuários que não compartilham nenhum atributo com a moda de nenhum dos clusters. Essa abordagem visa identificar e compreender padrões menos convencionais de mobilidade, contribuindo para uma compreensão mais completa das dinâmicas de deslocamento presentes na cidade.

Da Figura 45 à Figura 48, apresentam as proporções de cada atributo em cada cluster e as comparações entre os agrupamentos para usuários que não possuem nenhum atributo igual às modas do cluster.

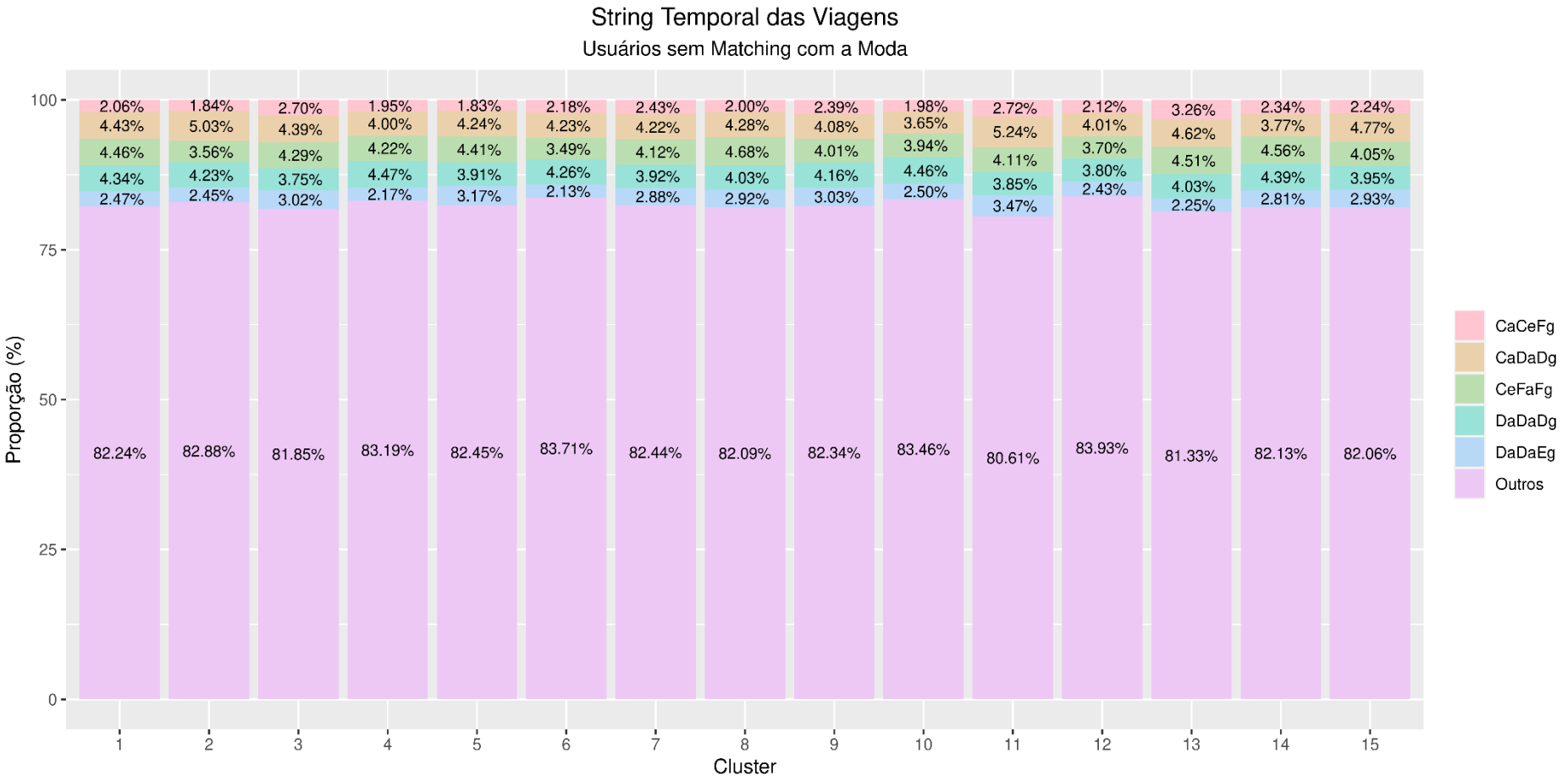
Ao analisar os gráficos, observa-se que, além das classificações como 'Outros', uma parcela significativa dos usuários nessa análise realiza mais de duas viagens por dia. Em relação aos motivos de viagem, as sequências mais predominantes incluem 'Residência-Assuntos Pessoais-Assuntos Pessoais-Residência' e 'Residência-Trabalho-Escola/Educação-Residência', sendo realizadas por uma parte menor da população, o que torna sua identificação mais desafiadora. Essa tendência persiste ao compararmos o atributo de predominância da classe social, destacando que os usuários que realizam três ou mais viagens e têm origem em regiões de classe social baixa são os mais comuns nessa análise. No que diz respeito à predominância de raça/cor, os gráficos indicam que usuários com origem em regiões de predominância de negros acima da média são os mais frequentes nessa análise de usuários que se afastam dos padrões dos clusters.

Figura 45 – Comparação de clusters – Usuários com nenhum Motivo de Viagem igual à moda



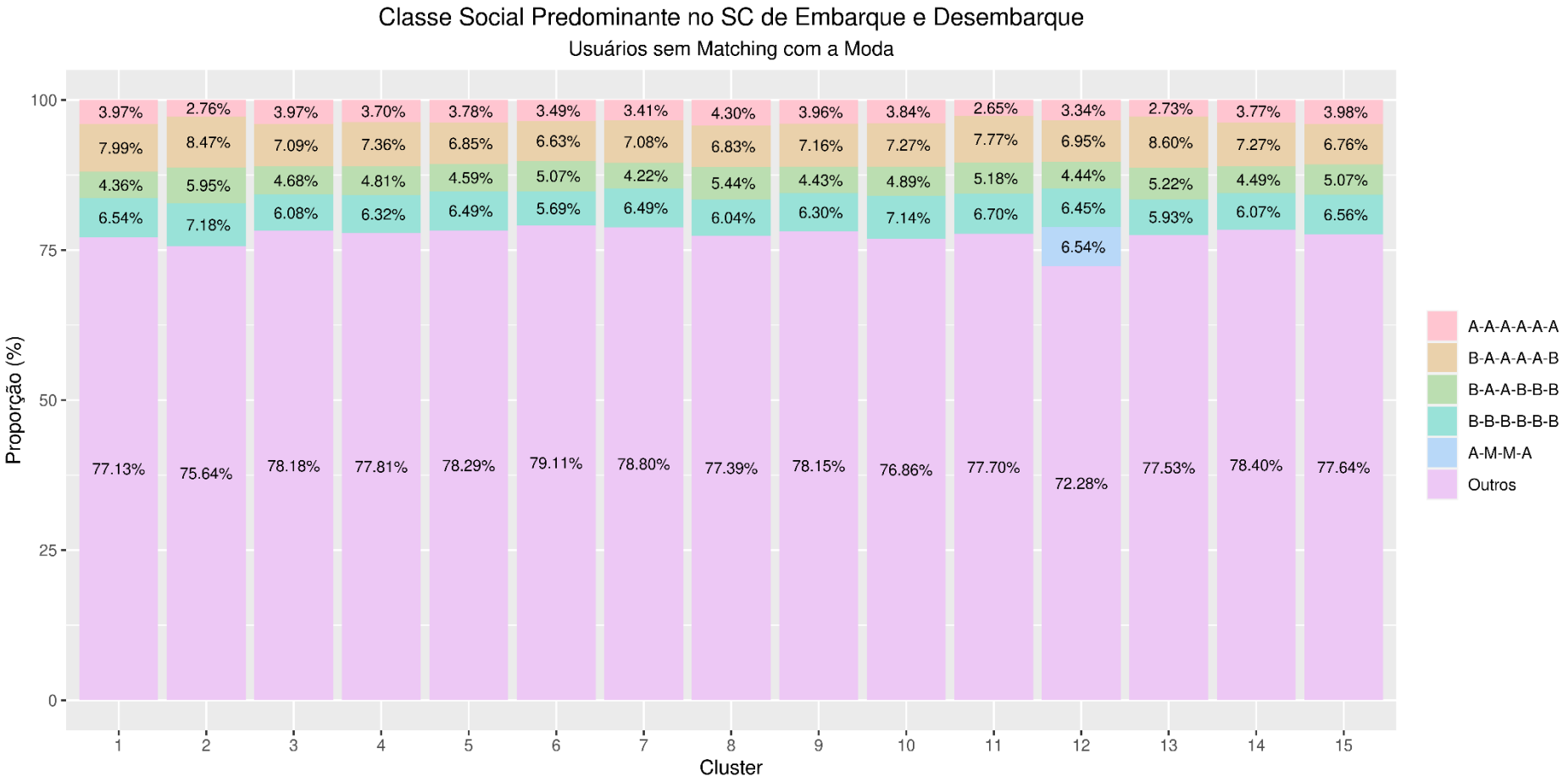
Fonte: Elaboração própria

Figura 46 – Comparação de clusters – Usuários com nenhuma sequência temporal igual à moda



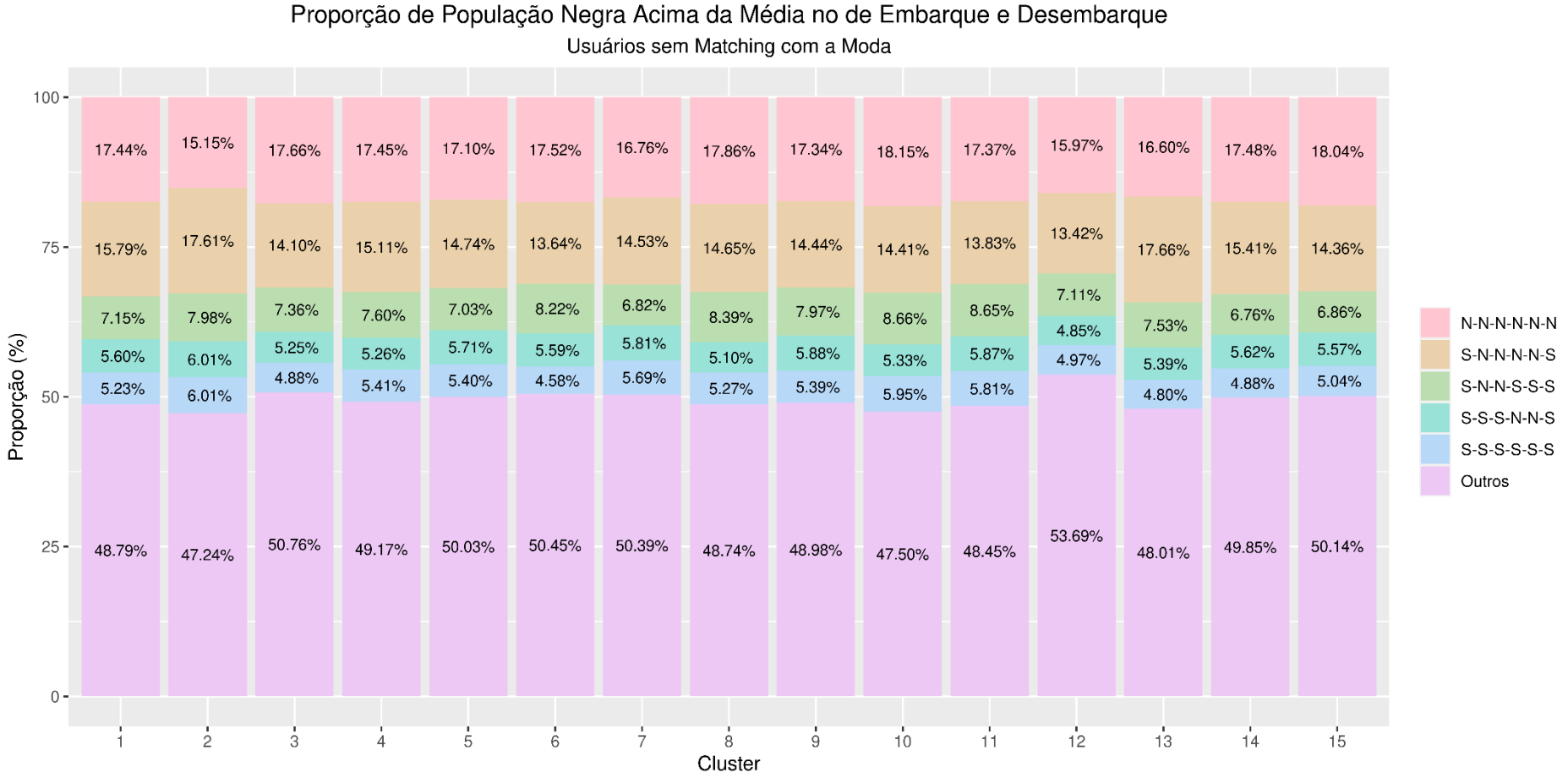
Fonte: Elaboração própria

Figura 47 – Comparação de clusters – Usuários com nenhuma sequência de classe social igual à moda



Fonte: Elaboração própria

Figura 48 – Comparação de clusters – Usuários com nenhuma sequência de região de raça/cor igual à moda



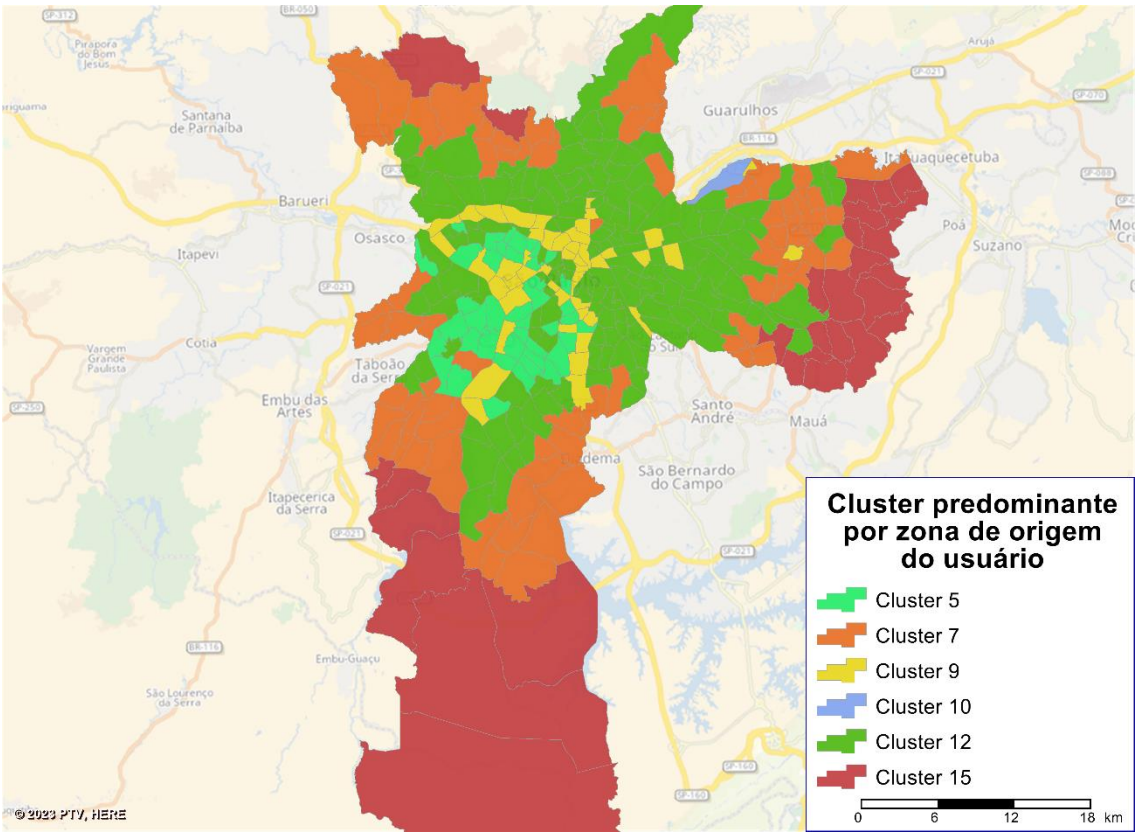
Fonte: Elaboração própria

5.3 Distribuição espacial dos clusters

Nesta seção, realiza-se uma análise espacial ao mapear as origens dos usuários de transporte público em São Paulo para os clusters previamente identificados. Essa abordagem geográfica busca proporcionar uma representação visual da distribuição dos padrões de mobilidade encontrados na cidade. Associar cada ponto de origem a um cluster específico permitirá uma compreensão mais clara da dispersão geográfica dos diferentes perfis de deslocamento, destacando áreas de convergência e divergência.

A Figura 49 apresenta o cluster predominante, ou seja, aquele que apresentou maior número de usuários na zona de origem das viagens. Pela análise realizadas, 6 clusters diferentes foram predominantes em alguma região.

Figura 49 – Predominância do cluster por zona de origem dos usuários



Fonte: Elaboração própria

Examinando cada cluster predominante, as origens do Cluster 5 estão concentradas em uma região central da cidade, caracterizada por uma classe predominantemente média/alta. Este padrão geográfico se alinha com as

características sociodemográficas previamente identificadas para esse cluster, além de serem usuários que não realizam nenhuma transferência durante o dia, por terem origem e destino mais próximos.

No aspecto espacial, as origens do Cluster 7 estão concentradas em regiões mais periféricas da cidade, caracterizadas por uma classe social baixa e ambientes residenciais com padrão baixo. Essa distribuição espacial reflete a realidade socioeconômica dessas áreas, onde os usuários dependem do transporte público para acessar locais de trabalho em regiões de classe média/alta e também reflete os atributos predominantes desse cluster, com origem em áreas de classe social baixa pela manhã e se deslocam para destinos em outros locais, realizando pelo menos uma transferência durante o dia.

O Cluster 9 destaca-se por estar espacialmente próximo a áreas de comércio e alta/média renda. Associados principalmente a motivos de viagem relacionados a trabalho, esses usuários revelam padrões de deslocamento que refletem a interação entre características socioeconômicas e geográficas. A presença em regiões de alta/média renda e a predominância de raça/cor branca seguem as tendências observadas em outros agrupamentos.

O Cluster 10 é proeminente em uma zona específica da cidade de São Paulo, notadamente na região do campus USP-Leste. Este agrupamento, identificado pelo motivo de viagem relacionado à educação/escola, reflete uma distribuição espacial focalizada.

Em contraste, o Cluster 12 abrange predominantemente a região de classe social média/alta, correlacionando-se com padrões de uso do solo que indicam atividade principal relacionada ao trabalho. Os usuários desse cluster parecem ser passageiros regulares do sistema de transporte coletivo.

Por fim, o Cluster 15 é predominante nas regiões mais periféricas da cidade, alinhando-se à análise anterior, onde os usuários concentram atividades em áreas que mantêm as características predominantes de sua origem, como classe social baixa e raça/cor negra acima da média.

6. Conclusões e trabalhos futuros

Este estudo empreendeu a construção de uma matriz Origem-Destino (OD) de viagens, por meio da metodologia de encadeamento de viagens e a análise dos padrões de mobilidade urbana dos usuários de transporte público em São Paulo utilizando o método de clusterização k-modes, valendo-se de uma abordagem que integra dados socioeconômicos e motivacionais às informações de bilhetagem eletrônica. Para tal foi necessário, a partir do embasamento na literatura, construir as matrizes Origem-Destino (OD) a partir de dados de bilhetagem eletrônica para a subsequente análise de desigualdades na mobilidade urbana. Para a construção da matriz foi necessária a inferência de localização de todas as transações da base de dados, a associação de pontos de embarque, inferência do local e ponto de desembarque das transações, a detecção de transferência ou atividade para, finalmente, desenvolver a clusterização e avaliação dos padrões de viagem.

Os clusters analisados revelam distintos padrões de viagem e associações sociodemográficas entre os usuários. O primeiro conjunto foi ligado a deslocamentos de casa para o trabalho, com diferenças notáveis nas características socioeconômicas, horários e padrões de transferência. O Cluster 7 (15% dos usuários), por exemplo, destaca-se pela origem em regiões de classe baixa, enquanto o Cluster 12 (18%) concentra-se em áreas de média e alta classe social. O segundo conjunto foi associado a atividades de educação, saúde e assuntos pessoais, com menor intervalo entre viagens. Destaca-se a disparidade na acessibilidade a serviços de saúde, evidenciada pelo Cluster 14 (4%), que parte de regiões de classe baixa para buscar assistência médica em áreas de classe média/alta. O terceiro conjunto reflete viagens únicas relacionadas principalmente a educação e trabalho, com predomínio em regiões de classe alta e raça/cor branca. Essas análises oferecem uma visão abrangente das complexidades dos padrões de mobilidade urbana e suas implicações sociais.

A observação dos clusters revelou a existência de padrões, mas também a variabilidade notável nos padrões de viagem, com alguns clusters

apresentando modas inferiores a 50%, indicando que mais da metade dos usuários desses grupos não segue precisamente o mesmo padrão.

Essa heterogeneidade é ainda mais evidente ao examinarmos usuários que não compartilham nenhum atributo com a moda dos clusters, ressaltando comportamentos menos convencionais e a necessidade de uma análise mais detalhada para compreendê-los. Foi possível observar que a variabilidade se reflete nas características socioeconômicas, com uma predominância de usuários que realizam três ou mais viagens originárias de regiões de classe social baixa e áreas de predominância de raça/cor negra acima da média fora do padrão dos clusters.

A aplicação da metodologia de encadeamento de viagens e a inferência dos motivos de viagem preenchem lacunas nos dados de bilhetagem eletrônica, permitindo uma visão mais abrangente dos deslocamentos. A identificação de grupos homogêneos de comportamento de viagem destaca-se como um passo crucial para uma análise mais precisa e eficaz da mobilidade urbana. Essa segmentação não apenas auxilia os planejadores de transporte público na adaptação de serviços às demandas específicas dos usuários, mas também proporciona insights valiosos para a compreensão das desigualdades sociais relacionadas ao acesso ao transporte público.

Para futuras investigações, é importante aprofundar a análise da variabilidade nos padrões de viagem, visando aprimorar a acessibilidade para grupos de usuários que não se enquadram nos clusters convencionais. Uma abordagem mais detalhada permitiria uma compreensão mais abrangente das necessidades dos usuários, para o desenvolvimento de soluções mais eficazes no contexto do transporte público. Além disso, aprimorar a inferência de residência, utilizando dados de vários dias de bilhetagem eletrônica, pode enriquecer a precisão das informações demográficas e espaciais. E por fim, aplicação da metodologia para dias de fim de semana e a subsequente comparação de resultados ofereceriam uma perspectiva mais completa das dinâmicas de deslocamento em diferentes momentos da semana, contribuindo para estratégias mais adaptadas e abrangentes no planejamento urbano e no transporte público.

.

Referências Bibliográficas

AGARD, B.; TRÉPANIER, M.; PARTOVI NIA, V. Assessing public transport travel behaviour from smart card data with advanced data mining technique. 2013.

ALSGER, A. et al. Validating and improving public transport origin–destination estimation algorithm using smart card fare data. Transportation research. Part C, Emerging technologies, v. 68, p. 490–506, 2016.

ALSGER, A. et al. Public transport trip purpose inference using smart card fare data. Transportation research. Part C, Emerging technologies, v. 87, p. 123–137, 2018.

ALSGER, A. A. et al. Use of smart card fare data to estimate public transport origin–destination matrix. Transportation research record, v. 2535, n. 1, p. 88–96, 2015.

AMAYA, M.; MUNIZAGA, M. Estimación de zona de residencia en base a sistemas tecnológicos de transporte público, 2013.

ANDA, C.; ERATH, A.; FOURIE, P. J. Transport modelling in the age of big data. International journal of urban sciences, v. 21, n. sup1, p. 19–42, 2017.

ARBEX, R.; ALVES, B.; GIANNOTTI, M. Comparing Accessibility in Urban Slums Using Smart Card and Bus GPS Data. In: Transportation Research Board TRB 95th Annual Meeting. Washington DC, 2016.

ARBEX, R.; DA CUNHA, C. B.; SPEICYS, R. Before-and-after evaluation of a bus network improvement using performance indicators from historical smart card data. Public transport, v. 13, n. 3, p. 483–501, 2021.

ARBEX, R. O.; DA CUNHA, C. B. 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, v. 25, n. 3, p. 166, 2017.

BARRY, J. J. et al. Origin and destination estimation in New York city with automated fare system data. Transportation research record, v. 1817, n. 1, p. 183–187, 2002.

BASSO, F. et al. Accessibility to opportunities based on public transport gps-monitored data: The case of Santiago, Chile. Travel behaviour & society, v. 21, p. 140–153, 2020.

BITTENCOURT, T. A.; GIANNOTTI, M.; MARQUES, E. Cumulative (and self-reinforcing) spatial inequalities: Interactions between accessibility and segregation in four Brazilian metropolises. Environment and planning. B, urban analytics and city science, v. 48, n. 7, p. 1989–2005, 2021.

CHATURVEDI, A.; GREEN, P. E.; CAROLL, J. D. k-modes clustering. Journal of classification, v. 18, n. 1, p. 35–55, 2001.

CURRIE, G. Quantifying spatial gaps in public transport supply based on social needs. Journal of transport geography, v. 18, n. 1, p. 31–41, 2010.

DE ORTUZAR, J.; WILLUMSEN, L. G. Modelling Transport. 4. ed. [s.l.] Wiley, 2011.

DEVILLAINE, F.; MUNIZAGA, M.; TRÉPANIER, M. Detection of activities of public transport users by analyzing smart card data. Transportation research record, v. 2276, n. 1, p. 48–55, 2012.

EGU, O.; BONNEL, P. How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination surveys and household travel survey? An empirical investigation in Lyon. Transportation research. Part A, Policy and practice, v. 138, p. 267–282, 2020.

EL FAOUZI, N.-E.; LEUNG, H.; KURIAN, A. Data fusion in intelligent transportation systems: progress and challenges - a survey. Inform. Fusion, v. 12, n. 1, p. 4–10, 2011.

FAROQI, H.; MESBAH, M. Inferring trip purpose by clustering sequences of smart card records. Transportation research. Part C, Emerging technologies, v. 127, n. 103131, p. 103131, 2021.

FARZIN, J. M. Constructing an automated bus origin–destination matrix using farecard and global positioning system data in São Paulo, Brazil. Transportation research record, v. 2072, n. 1, p. 30–37, 2008.

GORDON, J. B. et al. Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transportation research record, v. 2343, n. 1, p. 17–24, 2013.

HIMANEN, V.; JÄRVI-NYKÄNEN, T.; RAITIO, J. Daily travelling viewed by self-organizing maps. In: Neural Networks in Transport Applications. [s.l.] Routledge, 2019. p. 85–110.

HORA, J. et al. Estimation of Origin-Destination matrices under Automatic Fare Collection: the case study of Porto transportation system. Transportation research procedia, v. 27, p. 664–671, 2017.

HUANG, D. et al. A Method for Bus OD Matrix Estimation Using Multisource Data. Journal of Advanced Transportation, 2020.

HUANG, J. A fast clustering algorithm to cluster very large categorical data sets in data mining. Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

Instituto Brasileiro De Geografia e Estatística - IBGE. Censo demográfico 2010.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. ACM Computing Surveys, v, v. 31, p. 264–323, 1999.

KHAN, S. S.; AHMAD, A. Cluster center initialization algorithm for k-modes clustering. Expert systems with applications, v. 40, n. 18, p. 7444–7456, 2013.

KUJALA, R. et al. Travel times and transfers in public transport: Comprehensive accessibility analysis based on Pareto-optimal journeys. Computers, Environment and Urban Systems, 2018.

KUSAKABE, T.; ASAKURA, Y. Behavioral data mining of transit smart card data: A data fusion approach. Transportation Research Part C, 2014.

LEE, S. G.; HICKMAN, M. Travel pattern analysis using smart card data of regular users. In: Preprint CD-ROM for the 90th Annual Meeting of Transportation Research Board. Washington, DC, 2011.

LI, D. et al. Estimating a transit passenger trip origin-destination matrix using automatic fare collection system. In: Database Systems for Adanced Applications. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 502–513.

LONG, Y.; THILL, J.-C. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. Computers, environment and urban systems, v. 53, p. 19–35, 2015.

LUCAS, K. et al. Transport poverty and its adverse social consequences. Proceedings of the Institution of Civil Engineers, Transport, v. 169, n. 6, p. 353–365, 2016.

MA, X. et al. Mining smart card data for transit riders' travel patterns. Transportation research. Part C, Emerging technologies, v. 36, p. 1–12, 2013.

MA, X. et al. Understanding commuting patterns using transit smart card data. Journal of transport geography, v. 58, p. 135–145, 2017.

MARTENS, K. Justice in transport as justice in accessibility: applying Walzer's 'Spheres of Justice' to the transport sector. Transportation, v. 39, n. 6, p. 1035–1053, 2012.

MORENCY, C.; TRÉPANIER, M.; AGARD, B. Measuring transit use variability with smart card data. Transport Policy, v. 14, n. 3, p. 193–203, 2007.

MUNIZAGA, M. A. et al. Validating travel behaviour estimated from smartcard data. Transport. Res. Part C: Emerg. Technol, 2014.

MUNIZAGA, M. A.; PALMA, C. Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. Transportation research. Part C, Emerging technologies, v. 24, p. 9–18, 2012.

NASSIR, N.; HICKMAN, M.; MA, Z.-L. Activity detection and transfer identification for public transit fare card data. Transportation, v. 42, n. 4, p. 683–705, 2015.

NIEUWENHUIS, J. et al. Does segregation reduce socio-spatial mobility? Evidence from four European countries with different inequality and segregation contexts. Urban studies (Edinburgh, Scotland), v. 57, n. 1, p. 176–197, 2020.

NUNES, A. A.; GALVAO DIAS, T.; FALCAO E CUNHA, J. Passenger journey destination estimation from automated fare collection system data using spatial validation. IEEE transactions on intelligent transportation systems: a publication of the IEEE Intelligent Transportation Systems Council, v. 17, n. 1, p. 133–142, 2016.

OD Pesquisa Origem Destino 2017. (2019). Informações de Pesquisa Domiciliar. São Paulo: Metrô.

ORTEGA-TONG, M. A. Classification of London's Public Transport Users Using Smart Card Data, 2013.

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, 2011.

PEREIRA, R.; BRAGA, C.; SERRA, B. Desigualdades socioespaciais de acesso a oportunidades nas cidades brasileiras. Texto para Discussão Ipea, v. 2535, p. 1–58, 2019.

PIERONI, C. et al. Big data for big issues: Revealing travel patterns of low-income population based on smart card data mining in a global south unequal city. Journal of transport geography, v. 96, n. 103203, p. 103203, 2021.

Prefeitura de São Paulo, Secretaria Municipal de Desenvolvimento Urbano: Distritos do Município de São Paulo, 2015.

ROUSSEEUW, P. J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics, v. 20, p. 53–65, 1987.

SCHWANEN, T. et al. Rethinking the links between social exclusion and transport disadvantage through the lens of social capital. Transportation research. Part A, Policy and practice, v. 74, p. 123–135, 2015.

SEABORN, C.; ATTANUCCI, J.; WILSON, N. H. M. Analyzing multimodal public transport journeys in London with smart card fare payment data. Transportation research record, v. 2121, n. 1, p. 55–62, 2009.

SHEN, L.; STOPHER, P. R. A process for trip purpose imputation from Global Positioning System data. Transportation research. Part C, Emerging technologies, v. 36, p. 261–267, 2013.

SINNOTT, R. W. Virtues of the Haversine. Sky and Telescope, v. 68, n. 2, 1984.

TRÉPANIER, M.; TRANCHANT, N.; CHAPLEAU, R. Individual trip destination estimation in a transit smart card automated fare collection system. Journal of intelligent transportation systems, v. 11, n. 1, p. 1–14, 2007.

YAN, F.; YANG, C.; UKKUSURI, S. V. Alighting stop determination using two-step algorithms in bus transit systems. Transportmetrica A Transport Science, v. 15, n. 2, p. 1522–1542, 2019.

YAN, J.; THILL, J.-C. Visual data mining in spatial interaction analysis with self-organizing maps. Environment and planning. B, Planning & design, v. 36, n. 3, p. 466–486, 2009.

YU, C.; HE, Z. Travel Pattern Recognition using Smart Card Data in Public Transit. International Journal of Emerging Engineering Research and Technology, n. 7, p. 6–13, 2016.

ZAKI, M. J.; MEIRA, M. J. Data Mining and Analysis: Fundamental Concepts and Algorithms. [s.l.] Cambridge University Press, 2013.

ZHAO, J. et al. Spatio-temporal analysis of passenger travel patterns in massive smart card data. IEEE transactions on intelligent transportation systems: a publication of the IEEE Intelligent Transportation Systems Council, v. 18, n. 11, p. 3135–3146, 2017.

ZHAO, P.; CAO, Y. Commuting inequity and its determinants in Shanghai: New findings from big-data analytics. Transport policy, v. 92, p. 20–37, 2020.

ZHONG, C. et al. Measuring variability of mobility patterns from multiday smart-card data. Journal of computational science, v. 9, p. 125–130, 2015.

ZHOU, J.; MURPHY, E.; LONG, Y. Commuting efficiency in the Beijing metropolitan area: an exploration combining smartcard and travel survey data. Journal of transport geography, v. 41, p. 175–183, 2014.