UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

ENGENHARIA CIVIL

DEPARTAMENTO DE ENGENHARIA DE TRANSPORTES

JOÃO LUCAS ALBUQUERQUE OLIVEIRA

ANÁLISE DA EVOLUÇÃO DA DEMANDA NO TRANSPORTE COLETIVO POR ÔNIBUS EM FORTALEZA UTILIZANDO *BIG DATA*

FORTALEZA

2019

JOÃO LUCAS ALBUQUERQUE OLIVEIRA

ANÁLISE DA EVOLUÇÃO DA DEMANDA NO TRANSPORTE COLETIVO POR ÔNIBUS EM FORTALEZA UTILIZANDO *BIG DATA*

Monografia apresentada ao curso de Engenharia Civil do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Engenheiro Civil.

Orientador: Prof. Ph.D. Francisco Moraes de Oliveira Neto

FORTALEZA

2019

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

O47a Oliveira, João Lucas Albuquerque.
Análise da evolução da demanda no transporte coletivo por ônibus em Fortaleza utilizando big data / João Lucas Albuquerque Oliveira. – 2019.
68 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Tecnologia, Curso de Engenharia Civil, Fortaleza, 2019.
Orientação: Prof. Dr. Francisco Moraes de Oliveira Neto.

1. Demanda. 2. Dados de bilhetagem. 3. Padrões de deslocamento. 4. Clusterização. 5. Transporte coletivo. I. Título.

CDD 620

JOÃO LUCAS ALBUQUERQUE OLIVEIRA

ANÁLISE DA EVOLUÇÃO DA DEMANDA NO TRANSPORTE COLETIVO POR ÔNIBUS EM FORTALEZA UTILIZANDO *BIG DATA*

Monografia apresentada ao curso de Engenharia Civil do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Engenheiro Civil.

Aprovada em: 29/11/2019.

BANCA EXAMINADORA

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

_____
Prof. Ph.D. Carlos Felipe Grangeiro Loureiro
Universidade Federal do Ceará (UFC)

_____
Prof. MSc. Francelino Franco Leite de Matos Sousa
Centro Universitário Christus

AGRADECIMENTOS

À minha mãe, por ter me dado todo o apoio incodicional para que eu pudesse ter chegado até aqui e por tudo que fez por mim.

Aos meus avós maternos, por terem sido importantíssimos no meu desenvolvimento e criação.

À minha família, pai, irmãos avós, tias, primos, e todos que contribuíram de alguma forma para a caminhada até aqui.

À família da tia Zulene, principalmente à Silvana e Esterfeson, por terem me oferecido moradia e acolhimento em dois momentos da graduação.

Ao Prof. Moraes, pela orientação durante metade da graduação e pelo esforço para contribuir em cada parte deste trabalho, além da disponibilidade. Agradeço também pelos conselhos para formação pessoal e humana.

Ao Prof. Felipe, que eu coloco na conta como de grande importância para minha vinda para a área de transporte, ao me mostrar como a engenharia pode incorporar os aspectos sociais das pessoas e me demonstrar que a docência é foda, e Paulo Freire também.

Ao Prof. Franco, por ter dedicado um tempo para participar da banca examinadora e pelas contribuições à monografia.

Aos amigos Renan, Cassiano, Davi, Kauê, Joana, Alessandro, Julie, Roberto, por terem contribuído de forma direta com esse trabalho, nos mais diversos tipos de auxílio.

Aos amigos Gabriel, Fred, Kaio, Beliza, Israel, Nilso, Renata, Gescilam, David, Diego, Jonas, Mat(h)eus, Jesus por todo o apoio e frescura.

Aos amigos do GTTEMA, Sameque, Nara, Gustavo, Moisés, Diego, Vanessa, Caio, Marília, Wendy, por terem feito com que os dias se tornassem mais leves.

À Jéssica, por ter sido uma da pessoa muito importante na minha caminhada na graduação, por todos os auxílios e momentos vividos.

Aos amigos da Transitar, por terem me dado oportunidade e ensinado grandes coisas e tornado os momentos de trabalho mais descontraídos.

RESUMO

O transporte coletivo constitui modo de transporte essencial para uma parte significativa da população. O cenário atual na demanda do transporte coletivo é de queda expressiva nas grandes metrópoles. Como oportunidade, há uma disponibilidade crescente de dados que podem auxiliar no entendimento sobre essa demanda. Esta pesquisa propõe utilizar os dados disponíveis que compõem o *big data (smart card* - bilhetagem, *gps* e cadastro dos usuários) como forma de analisar a evolução da demanda do transporte coletivo por ônibus em Fortaleza. O método está dividido em três macro etapas: análise descritiva temporal, em que se analisaram indicadores temporais comparando os anos de 2014 e 2018, com o intuito de avançar em uma compreensão inicial sobre o comportamento da demanda; análise descritiva espacial, na tentativa de entender se é possível estimar os embarques dos usuários a partir de dados de validações, pois assim como na maioria das cidades brasileiras, em Fortaleza o usuário não necessariamente valida seu cartão ao embarcar e tal peculiaridade deve ser levada em consideração antes de se estudar a espacialidade da demanda por meio das validações; e detecção e análise de grupos, realizada por meio de técnicas de clusterização, a partir das quais buscou-se unir diferentes atributos de uso dos usuários a fim de caracterizar esses grupos e compará-los, verificando quais são os mais frequentes e esporádicos, além de observar sua evolução entre os anos de 2014 e 2018. Como resultados, verificou-se que a demanda de transporte coletivo por ônibus está declinando, sendo o pagamento em inteira e em vale transporte os mais afetados. Na análise espacial, verificou-se que em aproximadamente metade dos casos amostrados, os usuários costumam validar a uma distância maior que 2 quilômetros de distância de sua residência, apresentando alta dispersão a partir dessa distância. Na identificação de agrupamentos foi possível perceber a presença de grupos mais frequentes e grupos mais esporádicos, além de grupos que apresentaram o uso mais intermediário ao longo do ano. Além disso, notou-se uma tendência de aumento dos grupos mais esporádicos e que possuem uso intermediário ao longo do ano. Também há evidências para acreditar na diminuição dos grupos mais cativos.

. **Palavras-chave**: Demanda. Transporte coletivo. Dados de bilhetagem. Clusterização. Padrões de deslocamento.

ABSTRACT

Public transport is an essential mode of transport for a significant part of the population. The current scenario in the demand of public transport is of significant decreasing in big cities. As opportunity, there is a growing availability of data that can help understand this demand. This research proposes to use the available amount of data (smart card, gps and user registration) as a way to analyze the evolution of the demand for transit in Fortaleza. The method is divided into three macro stages: temporal descriptive analysis, in which temporal indicators were analyzed comparing the years 2014 and 2018, in order to understand the demand behavior; spatial descriptive analysis, in an attempt to understand if it is possible to estimate users' boarding from validation data, once like most Brazilian cities, in Fortaleza users does not necessarily validate his card when boarding and such peculiarity should be taken into account before studying the spatiality of demand through validations; and group detection and analysis, performed through clustering techniques, from which we sought to unite different attributes of users' travel patterns in order to characterize each group and compare them, verifying which are the most frequent and sporadic. It was found that the demand for transit is declining in Forteleza, with payment in cash and with smart card being the most affected. In the spatial analysis, it was found that in approximately half of the sampled cases, users usually validate at a distance greater than 2 kilometers from their home, presenting high dispersion from this distance. In the identification of clusters it was possible to notice the presence of more frequent groups and more sporadic groups, as well as groups that presented the most intermediate use throughout the year. In addition, there was a tendency of increasing in the demand of sporadic groups that have intermediate use throughout the year. There is also evidence to believe in the decline of the most captive groups.

Keywords: Demand. Public transport. Ridership. Smart card. Travel Pattern. Clustering.

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

No período pós-guerra, a maior parte da sociedade foi constituída à luz do modelo do Bem-Estar Social, que estabelece a noção de universalização da condição de cidadania e da constituição de responsabilidade social do Estado. Esse pensamento representa a noção capitalista para inclusão social, que busca, teoricamente, ocasionar igualdade social (SPOSATI, 1998). No entanto, o processo de formação das grandes cidades, principalmente as emergentes, geralmente é acompanhado de um crescimento desordenado, que provoca uma série de externalidades negativas. O desenvolvimento econômico brasileiro, principalmente nos grandes centros urbanos, é marcado por um mecanismo de criação de moradias precárias e pela indisponibilidade de serviços essenciais à população. A cidade, portanto, torna-se ambiente de contradições sociais, políticas e econômicas em vários sistemas, gerando exclusão social (BOARETO, 2003).

A exclusão social, além de incluir insuficiência de renda (mais ligada ao conceito de pobreza), é composta ainda pela discriminação social, segregação espacial, pela não-equidade e negação dos direitos sociais. O conceito de exclusão social, dessa forma, tem como premissa a universalização da cidadania, ou seja, a ocorrência dela revela a negação da cidadania (SPOSATI, 1998). Um tipo de exclusão social pode ser evidenciado no processo de urbanização das cidades, que se caracteriza, predominantemente, pela ocupação das regiões periféricas, acrescentando significativamente a necessidade de transporte e oferta de serviços públicos. A população residente dessas áreas tem suas condições de mobilidade e acessibilidade bastante reduzidas e por isso geralmente não tem suas demandas atendidas (IPEA, 2003). Essas inadequações nos padrões de uso do solo e ineficiência na oferta do transporte coletivo provêm muitas vezes da deficiência ou ausência do planejamento urbano, fazendo com que a população mais vulnerável, cativa desse modo de transporte público, tenha seus acessos a oportunidades e serviços deficitários (GARCIA *et al.*, 2018).

Nesse sentido, o transporte coletivo assume papel central na tentativa de melhoria da acessibilidade e mobilidade da população mais vulnerável no espaço urbano, permitindo o acesso às oportunidades de trabalho e serviços básicos, como saúde, educação, lazer, agindo, dessa forma, como importante aparelho de combate à pobreza e agente promotor de igualdade

social. Para atingir tais objetivos, é necessário que o transporte coletivo caminhe no sentido de se tornar mais acessível, mais eficiente e de maior qualidade (IPEA, 2003).

Para que se torne mais eficiente, são necessárias medidas que implementem políticas de incentivo ao transporte coletivo, como é o caso do Plano Nacional de Mobilidade Urbana, Lei Federal no. 12.587 de 2012, que está fundamentado em princípios de desenvolvimento sustentável das cidades, eficiência e eficácia na prestação de serviço de transporte, voltado a diretrizes de priorização de modos de transportes ativos e coletivos. A nível de cidades e metrópoles, é importante que os planejadores conheçam os padrões de deslocamento dos usuários do transporte coletivo, bem como os principais fatores que afetam a sua utilização, com o intuito de melhorar o serviço e, portanto, caminhar no sentido de satisfazer as necessidades e preferências dos usuários (AGARD; TRÉPANIER, 2013; KIEU *et al.*, 2014; MAHRSI *et al.*, 2014).

Atualmente, o cenário nas cidades brasileiras é de queda expressiva na quantidade de usuários que utilizam o transporte coletivo, principalmente em grandes centros urbanos. Segundo a Associação Nacional das Empresas de Transportes Urbanos (2018), em 5 grandes capitais brasileiras, incluindo Fortaleza, entre 1997 e 2017, houve redução de aproximadamente 35,6% de usuários, sendo essa queda mais acentuada nos últimos 5 anos, a partir de 2014, o que culminou numa redução média de 25,9% dos usuários pagantes.

1.2 Problemática

Nas últimas décadas, as pesquisas sobre demanda foram realizadas através de métodos que em sua maioria são manuais, árduos, têm custos elevados e consomem grandes massas de dados. As pesquisas domiciliares são o grande exemplo disso, além de existirem pesquisas mais operacionais, como contagens volumétricas e sobe-desce, entre outras. A última pesquisa origem-destino realizada em Fortaleza foi no ano de 1996 e, com o passar dos anos, a demanda evoluiu em diferentes aspectos e por diferentes causas, que podem estar relacionadas às mudanças ocorridas no uso do solo e na oferta do sistema de transportes, a inserção de novas formas de deslocamento, novas tecnologias e o incentivo ao uso de outros modos, entre outros fatores. No ano de 2019, a prefeitura de Fortaleza noticiou a criação do Plano de Acessibilidade de Fortaleza (PAS-For), em que está inserida uma pesquisa

domiciliar com o intuito de compreender os padrões de deslocamento da cidade e sua área de influência, com custo de pesquisa de R$ 11,3 milhões (NORDESTE, 2018).

No entanto, há uma mudança de paradigma na coleta e análise de dados. Os computadores vêm ganhando cada vez mais poder de processamento e o custo com armazenamento de dados se tornando significativamente menor. Essa facilidade na aquisição e armazenamento tem enorme impacto no sistema de transportes. Há atualmente um movimento crescente e popularizado que conta com as chamadas tecnologias de informação e comunicação (*Information and Communications Technology* – ICT), que podem auxiliar em várias etapas do processo de planejamento. No que diz respeito ao planejamento do transporte público, dados de bilhetagem eletrônica, Especificação Geral de Feeds de Transporte Público (GTFS) e *Global Positioning System* (GPS), que são fontes importantes de *big data* do transporte público (BD – TP), fornecem várias informações indiretas quase instantâneas, altamente desagregadas e quase populacionais dos usuários de transporte público, a custo significativamente menor em relação àqueles métodos mais tradicionais.

No contexto de Fortaleza, esses dados são disponibilizados e, portanto, podem ser utilizados para estudos nos diversos níveis de planejamento. O BD – TP conta com dados relativos às validações dos usuários - momento em que o usuário efetua o pagamento e passa pela catraca, GPS dos carros que compõem as frotas, Especificação Geral de Feeds de Transporte Público (GTFS), a base de endereço de alguns usuários e o banco de bicicletas compartilhadas. Braga (2019), reconhecendo a complexidade de tratamento e integração do BD-TP, realizou um esforço no sentido de consolidar e tratar essas bases de dados, tendo como resultado de um dos esforços a estimação o local da validação, uma vez que a base de validações é diferente da base do GPS, mas são relacionáveis.

Com a base do BD-TP consolidada, com dados disponíveis e considerando que as validações dos usuários podem ser um bom indicador da demanda realizada dada a oferta do sistema, é possível realizar algumas análises com o intuito de compreender melhor a demanda do transporte coletivo por ônibus. Dessa forma, surge a questão mais abrangente**: como a demanda de transporte coletivo por ônibus vem evoluindo na cidade de Fortaleza?**

Adentrando nessa perspectiva mais geral, é possível entender a evolução das validações a partir de diferentes agregações temporais, considerando grupos de usuários que utilizam diferentes tipos de cartões – estudante, vale transporte, bilhete único, gratuidade. É

interessante verificar se/como esses grupos estão emergindo ou decaindo na composição da demanda ao longo tempo, podendo contribuir, por exemplo, com a variação de arrecadação experimentada pelo sistema nos últimos anos, levando às questões de **como varia a dinâmica da demanda realizada para usuários com diferentes cartões considerando diferentes agregações temporais? Qual a composição desses grupos na demanda realizada? É possível perceber a queda ou aumento na demanda realizada de algum grupo?**

Além do aspecto temporal, é interessante também estudar a espacialidade da demanda, sendo possível entender como varia a dinâmica espacial de uso do sistema. O entendimento da dinâmica espacial de uso passa pela estimação do embarque/zona de origem ou destino dos usuários. Vários estudos realizam a estimação dos embarques dos usuários por meio dos dados de validações (Munizaga et al. 2012; TRÉPANIER *et al.* 2007). No entanto, apesar de ser possível estimar a localização da validação da grande maioria dos usuários e considerando que a primeira validação se aproxima da produção de viagens na origem dos usuários na maioria dos casos, não se pode afirmar que a localização da primeira validação é próxima da residência do usuário, ou seja, que é próxima ao embarque. Isso acontece pois na maioria das cidades do Brasil não é obrigatório a passagem na catraca ao embarcar no veículo e há espaço de acomodação na porção traseira, além de que existem casos em que o veículo está com alto nível de lotação e o usuário prefere se acomodar na parte traseira. Então, um passo necessário antes de estudar a espacialidade das validações propriamente dita, é necessário responder a seguinte questão: **é possível considerar que a localização da primeira validação é próxima ao embarque/residência dos usuários?**

Existem muitos estudos após a revolução da coleta de dados a partir de bilhetagem eletrônica que se esforçam na identificação e caracterização de agrupamentos temporais e espaciais de usuários do transporte coletivo (Briand *et al.*, 2015; MA *et al.*, 2013; Morency *et al.,*2006). A busca das respostas das questões anteriores (análise descritiva) pode auxiliar no estabelecimento de hipóteses e em um entendimento preliminar sobre possíveis padrões de uso pelos diferentes tipos de cartão. A identificação de padrões de forma sistemática e menos arbitrária possível pode contribuir para o entendimento dos padrões da demanda do transporte coletivo, que por sua vez podem auxiliar a redesenhar políticas e intervenções, surgindo a questão de **como os usuários podem ser agrupados homogeneamente em**

relação às suas características e seu padrão de uso (temporais e espaciais)? Qual a composição dos cartões utilizados pelos usuários nos agrupamentos? Como esses grupos podem ser interpretados?

1.3 Objetivos

O objetivo geral deste trabalho é analisar a evolução da demanda de transporte coletivo por ônibus utilizando ferramentas de *big data*, a fim de caracterizar a demanda e oferecer alguns subsídios a fim de auxiliar na tomada de decisão por parte dos gestores. Os objetivos específicos, alinhados com as questões de pesquisa, são:

1. Realizar análise descritiva temporal de indicadores obtidos a partir da *big data* do transporte coletivo que tentem representar a demanda realizada e comparar períodos distintos;
2. Verificar se é possível estimar embarque dos usuários a partir da localização das validações;
3. Identificar agrupamentos de usuários do transporte coletivo por ônibus com relação às suas características e padrões de uso para períodos distintos;
4. Interpretar os agrupamentos dos usuários do transporte coletivo por ônibus.

2 PADRÕES DE DESLOCAMENTO A PARTIR DE SMART CARD

Neste capítulo é apresentada a revisão da literatura sobre a utilização de *smart card* para obtenção de padrões de deslocamento. Na seção 2.1 foi feita uma breve discussão sobre as vantagens e desvantagens da tecnologia de *smart card* e suas principais aplicações no contexto do transporte coletivo. Na seção 2.2 são apresentados conceitos básicos sobre *data mining* e técnicas de *machine learning*, além de alguns trabalhos que versam sobre a aplicação dessas ferramentas em *smart card* para detecção de padrões de deslocamento de usuários. Por fim, a seção 2.3 foca na análise crítica e descrição dos principais métodos utilizados pelos trabalhos revistos.

2.1 Smart Card no planejamento urbano

O objetivo desta seção é aprender sobre as diversas aplicações de *smart card* no planejamento urbano com ênfase em transporte público, bem como suas vantagens e desvantagens em relação a outros métodos de pagamento.

A tecnologia de *smart card* foi desenvolvida na segunda metade do século 20 e vem sendo amplamente utilizada no planejamento de transportes (PELLETIER *et al.*, 2011). Ela auxilia na coleta de dados, proporcionando a oportunidade de uma coleta contínua que pode subsidiar o entendimento sobre os comportamentos dos usuários, que por sua vez pode resultar em uma melhora do sistema de transportes. Além disso, o *smart card* constitui um método mais seguro de pagamento que, além de substituir o pagamento por *ticket* ou dinheiro, proporciona a oportunidade de estruturação de modelos diferentes de tarifas e integração com outros modos de transportes (BAGCHI; WHITE, 2005; PELLETIER *et al.*, 2011).

Pelletier *et al.* (2011) construiu uma ampla revisão sobre as várias aplicações do *smart card* no planejamento de transportes, contemplando também vantagens e desvantagens, principalmente no contexto do transporte público. Bagchi e White (2005) descrevem o potencial de dados advindos do *smart card* e ressaltam a possibilidade e capacidade de uso para coletar dados dos usuários do sistema, bem como as vantagens e desvantagens desse método de pagamento.

O método tradicional de pagamento por *ticket* ou dinheiro dificulta algumas análises sobre o comportamento dos usuários, fazendo com haja uma compreensão limitada da demanda por algumas razões. Eles não conseguem identificar os usuários de forma individual

e fornecem dados apenas de forma agregada, apesar de ser sabido em geral que alguns usuários são cativos e representam uma porcentagem significativa do total de usuários. Somado a isso, não é possível identificar viagens com algum tipo de integração entre linhas e horários em que o sistema de transporte é mais requisitado (BAGCHI; WHITE, 2005). Operacionalmente, o *smart card* proporciona redução de fraudes, além de um atraso menor na interação com o equipamento de leitura no veículo, cerca de 1,5 segundos segundo White (2010), em relação aos outros tipos de pagamento como o dinheiro, o que facilita a implementação de modelos de autoatendimento, mais comuns em cidades europeias. Contudo, a implantação desses sistemas geralmente possui baixa aceitação social e necessitam de uma mudança institucional significativa (DEAKIN; KIM, 2001).

Em geral, a redução de custos a longo prazo, flexibilidade na definição de diferentes tarifas, o potencial de gerar informação sobre os usuários do sistema, melhor capacidade de gerir a receita são pontos positivos do uso do *smart card*. Os pontos negativos em geral estão ligados ao alto custo monetário e tecnológico de implementação, somada a possível não aceitação social e institucional (PELLETIER *et al.*, 2011). Long; Thill (2015) também pontuaram algumas limitações desses dados, como a não integração destes com outros modais, já que sua aplicação está mais concentrada no transporte público. Os atores citam que é importante incorporar outros modais em estudos no futuro. Outra limitação é que a informação gerada pelo *smart card* muitas vezes é incompleta, pois não necessariamente engloba a informação acurada de embarque/desembarque dos usuários, além de que não há informações advindas diretamente do cartão sobre os usuários que efetuam pagamento em dinheiro. Eles ainda afirmam, assim como Pelletier *et al.* (2011), que estudos comportamentais a nível de usuários são difíceis por conta do anonimato do cartão e pela ausência de dados socioeconômicos do usuário.

Existem vários estudos utilizando *smart card* presentes na literatura, variando entre níveis estratégico, tático e operacional. O critério de diferença entre esses níveis é normalmente o horizonte temporal de planejamento (PELLETIER *et al.*, 2011). No nível estratégico, Agard *et al.* (2006) buscou compreender o comportamento do usuário ao longo das semanas e dos dias da semana, utilizando os diferentes tipos de cartões existentes (Adulto, estudante e idoso) e a hora de embarque dos usuários. Zhou; Murphy; Long (2014) utilizaram dados de *smart card* juntamente com dados de pesquisa domiciliar para avaliar a quantidade de viagens pendulares por ônibus e carro em Pequim, China. Eles destacam que os dados advindos do *smart card* podem ser utilizados para verificação de padrões recentes

dos usuários, ao contrário em geral das pesquisas domiciliares, o que pode incorporar análises mais dinâmicas. Long; Thill (2015) também analisaram movimentos pendulares em Pequim utilizando *smart card* juntamente com pesquisa domiciliar e sugerem que esses dados podem ser um importante complemento para as pesquisas realizadas convencionalmente.

No nível tático, são mais comuns estudos relacionados ao ajuste da operação, como mudanças adaptativas nos horários programados, que consideram as variações diárias do número de viagens (Utsunomiya *et al.* 2006; Wilson *et al.* 2009), reconstrução de viagens (Bagchi *et al.* 2004), estimação de embarque, entre outros. Os estudos sobre a integração também são relevantes nesse nível de planejamento. Jang (2010) analisou tempo de viagem e integrações utilizando dados da cidade de Seoul, Coréia do Sul. Nessa cidade, os dados de embarque e desembarque são disponíveis, o que facilita esse tipo de análise. Ele verificou a distribuição espacial do tempo de viagem e das integrações, realizando antes uma espécie de consolidação de dados. O autor ressalta ainda a facilidade de coletar dados sobre integração com *smart card* e cita que compreender o comportamento das integrações é importante para identificar e priorizar áreas que necessitam de melhorias. Também há estudos nesse nível com o intuito de estimação de embarque ou desembarque, como o de Trépanier *et al.* (2007), que utiliza um algoritmo para estimar o local mais provável de desembarque mais provável dos usuários, considerando uma série temporal de uso e Munizaga *et al.* (2012), que estima o potencial destino dos usuários e constrói uma matriz origem destino. A maior parte dos estudos com *smart card* estão voltados para a estimação de matrizes origem destino, podendo ser essa estimação em vários níveis de planejamento. No nível operacional, os trabalhos estão mais focados na obtenção de indicadores operacionais da oferta e demanda, com o intuito de mensurar a performance do sistema (exemplo: Trépanier *et al.* (2009)).

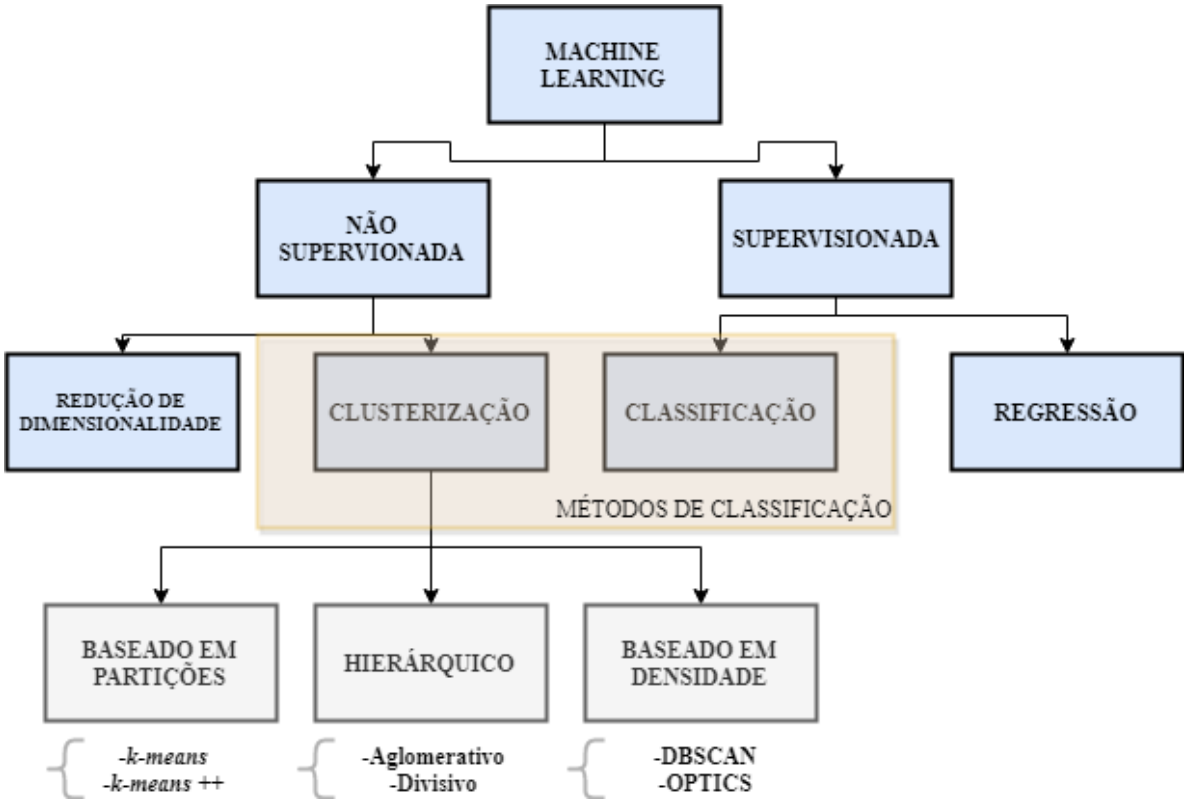
2.2 Análise de padrões utilizando *smart card*

Nesta seção, serão discutidos alguns trabalhos que têm como intuito a análise de padrões de viagens utilizando *smart card*, a fim de compreender quais são os métodos e as características dos dados utilizados, bem como os principais resultados obtidos. Em geral, para análise de padrões de viagens, se utilizam ferramentas de *data mining* que, segundo Witten *et al.* (2005), trata da extração de informações implícitas, anteriormente desconhecidas e potencialmente úteis dos dados. A base técnica para *data mining* está nas ferramentas e algoritmos de *machine learning*.

As ferramentas de *machine learning* podem ser classificadas em duas categorias: supervisionadas e não supervisionadas. A categoria supervisionada diz respeito à detecção de padrões com base em uma amostra rotulada ou de treinamento para previsão de novos padrões, enquanto no modo não supervisionado não há informações prévias sobre o padrão ou rótulo na amostra, ou seja, os padrões são desconhecidos. Dentro da técnica supervisionada, está a ferramenta de regressão e no modelo não supervisionado estão as ferramentas de clusterização e redução de dimensionalidade. Há também a técnica de classificação, que pode ser supervisionada ou não (JAIN *et al.*1999).

Nos trabalhos com ênfase em detecção de padrões utilizando *smart card*, é comum a aplicação de técnicas de clusterização. Os métodos de clusterização se dividem basicamente em três tipos: de partição, hierárquico e baseado em densidade. A maioria dos trabalhos vistos aplicam algum desses métodos. Dentre os algoritmos mais comuns estão o *k-means* e o algoritmo *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), que serão mais bem detalhados na seção seguinte. Na Figura 1 é mostrada uma visão geral sobre os tipos de técnicas dentro de *machine learning,* além dos principais algoritmos de clusterização utilizados.

Figura 1 – Visão geral sobre *machine learning* e algoritmos de clusterização



Fonte: Adaptado de Fahad *et al.* (2014) e Pieroni (2018).

Os estudos de análises de padrões começam a figurar após o início do século XXI, quando o sistema de tarifa por meio de *smart card* começou a se tornar mais difundido em alguns países. Agard *et al.* (2006) e Morency *et al.* (2006) são pioneiros na tentativa de analisar padrões por meio de *smart card*. Agard *et al.* (2006) estudou o comportamento dos usuários no transporte coletivo, aplicando técnicas de clusterização, assim como Morency *et al.* (2006), que também aplicou essa técnica para estudar a variabilidade das viagens.

Agard *et al.* (2006) buscou definir grupos de usuários com hábitos similares durante os diferentes dias da semana e durante diferentes semanas. Os dados utilizados foram de Gatineau (cidade da província de Quebec, Canadá) e, depois de compilados, resultaram em uma amostra de aproximadamente 5 meses com pouco mais de 2,1 milhões de embarques realizados por cerca de 25.500 usuários. Pouco mais de 80% dos deslocamentos eram realizados por *smart card* na cidade. Os autores agregam as validações por horários pré-definidos por eles, que remetem aos períodos de pico da manhã e noite e outros dois períodos fora pico. A partir disso, aplicando dois métodos de clusterização em conjunto (*k-means e hierarchical ascending clustering method*) eles detectam 4 grupos de características similares. Dois dos quatro grupos apresentam comportamentos pendulares, mas com diferença significativa no horário de embarque. Dos grupos restantes, um apresenta características de uso no horário de fora pico mais acentuado e o outro apresenta intensidade de uso dispersa e significativamente baixa em relação aos outros. Há também uma análise composição dos tipos de cartão (*Adult, Student, Elderly*) em cada um dos grupos, além de uma análise de variabilidade desses grupos ao longo das semanas.

De forma similar, Morency *et al.* (2006) utilizou os dados da mesma cidade para identificar padrões de uso em utilizando o dia da semana e hora do dia como parâmetros, fazendo diferença entre o tipo de cartão, no caso idosos e adultos. Os autores se preocuparam em realizar uma análise descritiva antes de aplicar as técnicas de clusterização (*k-means*) de alguns indicadores como número de embarques, com o intuito de entender melhor as variáveis de interesse, para que isso possa subsidiar e melhor a definição dos critérios para definição dos padrões. Os resultados mostram que os idosos possuem dispersão maior em relação ao dia e horário de uso, enquanto os adultos possuem uma maior regularidade no uso ao longo do dia e ao longo da semana, além de usarem mais em horários de pico.

Agard e Trépanier (2013) desenvolveram um trabalho com o intuito de caracterizar os usuários em grupos com similaridade em relação à frequência e horário de uso utilizando clusterização. Eles utilizaram dados do ano de 2008 que contam com cerca de 9,4 milhões de

transações com *smart card* da cidade de Gatineau (cidade da província de Quebec, Canadá) e com o uso do método *k-means* identificaram três grupos bem definidos: os que realizam deslocamentos pendulares e os que realizam preferencialmente deslocamentos no período de pico da manhã ou tarde. Eles também observaram a proporção desses grupos nos diferentes dias da semana e nos diversos tipos de cartão (*studant, regular adults, express adult*, entre outros), além de realizar uma análise relacionada a grupos dominantes, no sentido de que é possível que um mesmo usuário possa mudar de comportamento ao longo dos dias, diferente dos trabalhos anteriores.

Ma *et al.* (2013) utilizaram dados de Pequim (China), que conta com mais de 16 milhões de transações de *smart card* por dia e onde mais de 90% da população utiliza a ferramenta, para gerar dois tipos principais de agrupamentos: individual e de grupos. A ideia do agrupamento individual é verificar a constância de uso de um usuário no tempo e espaço. Se um usuário então possui um agrupamento em determinado horário e área, funciona como se fosse uma geração de viagem típica daquele usuário, que é um cluster. O método utilizado por ele nesse caso (DBCSAN) também permite identificar períodos em que o usuário se comportou de forma atípica. No agrupamento em grupos, ele reúne alguns atributos de todos os usuários, mais relacionados à aspectos de regularidade na frequência e no uso da oferta, e agrupa utilizando método *k-means* ++, para definir usuários com comportamento semelhante. Diferente dos outros trabalhos, nesse é definido a quantidade de grupos que se deseja alcançar e são estabelecidos níveis de regularidade no uso (muito alta, alta, média, baixa e muito baixa). Uma semana típica de dados do mês de julho de 2010 foi a utilizada para as análises.

Kieu *et al.* (2014) estudaram padrões de deslocamento dividindo em espaciais e temporais, utilizando 3 meses de dados da agência de trânsito da Austrália, com um total de 34,8 milhões de transações com *smart card*. Antes disso, eles propuseram um método para reconstruir os itinerários dos usuários de *smart card*, estimando as paradas de embarque e desembarque. Os padrões espaciais e temporais foram extraídos por meio do método DBSCAN, considerando os atributos de localização e tempo de embarque. Após isso, os autores fizeram uma análise conjunta e definiram os grupos de usuários com padrões de deslocamentos distintos, que resultaram em quatro grupos: usuários origem destino (que possuem hábitos espaciais semelhantes porém hábitos temporais mais irregulares), usuários com tempo habitual (que possuem hábitos temporais semelhantes porém hábitos espaciais mais irregulares), usuários regulares no tempo e no espaço, além de usuários irregulares em seus padrões temporais e espaciais de uso. Posteriormente, os autores analisaram a

contribuição de cada grupo na receita, verificando que os usuários regulares no tempo e no espaço são os que mais contribuem na receita, e que a maior parte dos passageiros se comporta de forma irregular, 64%. Os autores também analisaram o tempo de viagem, escolha de rotas, frequência de uso de todos os grupos.

Pieroni (2018) analisou os padrões de deslocamento de pessoas residentes em assentamentos precários em oito comunidades na cidade de São Paulo, Brasil. Ele utilizou primeiro o método DBSCAN com o intuito de inferir a origem do usuário e depois, utilizou outros três métodos (*k-means, TwoStep e Self-Organizing Map*- também conhecido como SOM) para definir agrupamentos de uso similar temporal e espacial. Nessa etapa, ele comparou três métodos de clusterização e incluiu, além de atributos temporais para identificar padrões, critérios espaciais e de uso do solo para definir os agrupamentos. Após o processamento, oito grupos foram obtidos e há grupos que possuem caráter pendular temporal e estão divididos entre áreas de diferentes usos como residencial, comercial, industrial e em áreas que possuem mais empregos informais. Agregar informações de uso do solo aos agrupamentos, dessa forma, faz com que haja mais interpretações sobre o motivo da viagem e podem auxiliar no entendimento sobre quais grupos de pessoas ocupam determinados empregos. No entanto, utilizar somente a validações para estimar a origem e destino dos usuários pode causar vieses, uma vez que o usuário não necessariamente valida onde realiza o embarque.

Mahrsi *et al.* (2014) também incorporou as características socioeconômicas na identificação de padrões de deslocamento, utilizando como estudo de caso a cidade de Rennes e sua área metropolitana, na França. Diferentemente de Pieroni (2018), os autores não utilizaram os dados de deslocamento com o de uso do solo englobados dentro do método de identificação de agrupamentos. Foi usado um modelo de mistura gaussiana com dados textuais da hora do dia e dia da semana de uso para encontrar grupos com diferentes padrões de deslocamento. Para encontrar agrupamentos de uso do solo com informações socioeconômicas similares, foi utilizado o método *Hidden Random Markov Field* (HRMF). Os resultados mostram que as pessoas que possuem horários bem definidos de uso, provavelmente pessoas que viajam por motivo trabalho ou educação (com dois ou três picos de uso ao longo dia) geralmente moram em locais de baixa renda e alta densidade de moradores.

2.3 Técnicas de clusterização

Nesta seção, os métodos de clusterização mais utilizados nos trabalhos serão detalhados, buscando uma melhor compreensão destes e dos seus parâmetros. O objetivo também é realizar uma análise crítica da métodos aplicados nos trabalhos anteriores, levantando vantagens e desvantagens, assim como os atributos ou variáveis utilizadas nas aplicações desses métodos.

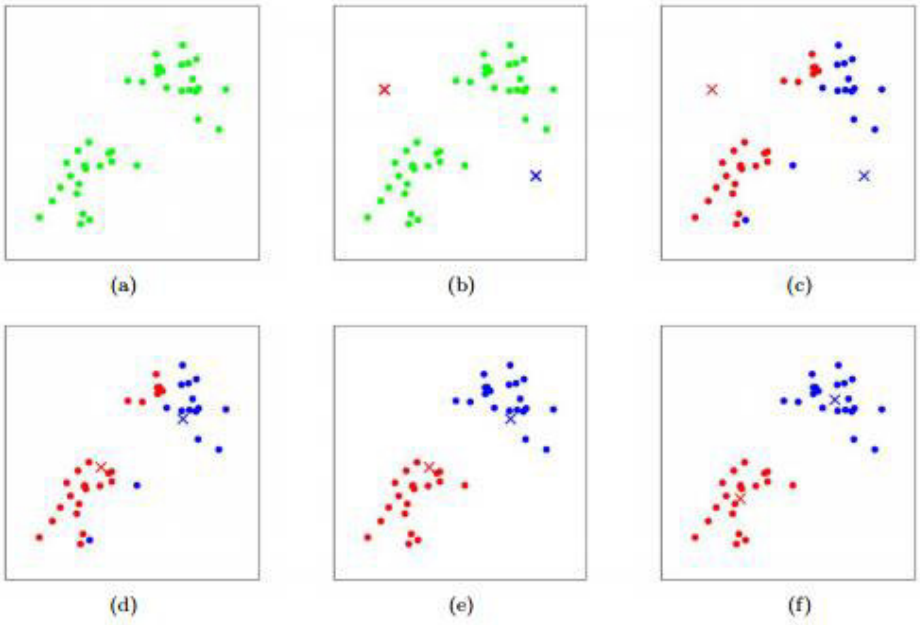
Clustering é basicamente a atividade de classificar pontos,em grupos naturais que são denominados *clusters*, de modo que elementos de um mesmo grupo sejam bastante semelhantes, enquanto pontos de *clusters* distintos sejam o mais diferente possível (ZAKI; MEIRA, 2014). A clusterização é uma importante ferramenta para análise exploratória de padrões em várias áreas, como em visão computacional, biologia, economia, com forte aplicação tomada de decisão (JAIN *et al.*1999). A capacidade de unir vários atributos para compor algum tipo de padrão é uma vantagem significativa das técnicas de clusterização em relação a análises exploratórias mais comuns. Apesar de o sistema de *smart card* atualmente contar com diferentes tipos de cartões utilizados para caracterizar diferentes usuários (gratuidade, vale transporte, estudante) remetendo muitas vezes a um possível motivo da viagem, não se tem informações mais detalhadas sobre o padrão de deslocamento desses usuários.

Como citado anteriormente, os métodos de clusterização podem ser particionais, hierárquicos e por densidade. Nos trabalhos revistos, nota-se uma maior utilização principalmente de dois métodos: *k-means* (particional) e DBSCAN (por densidade). O algoritmo *k-means* é um dos mais conhecidos, simples de ser aplicado e implementado computacionalmente (MORENCY *et al.* 2007).

Na Figura 2 é resumido como o algoritmo funciona. Dado o conjunto de dados distribuídos no espaço, que pode ser multidimensional (a) e assumindo que existem 2 grupos, 2 centroides aleatórios são gerados (b) e então, calcula-se a distância de cada ponto até os dois centroides. O ponto será atribuído ao centroide mais próximo e, portanto, ao grupo mais próximo (c). A partir da primeira definição dos grupos, calcula-se novamente o centroide (agora de um grupo já definido) e o processo se repete iterativamente quanto se queira (d), (e) e (f). A ideia do algoritmo é buscar uma função que minimize a distância quadrática dos pontos aos *clusters* (ZAKI; MEIRA, 2014). Ele requer, como parâmetro de entrada, o número

de agrupamentos desejados. Uma das desvantagens desse algoritmo é justamente a inicialização aleatória do centroide, que pode levar a uma minimização local dos erros quadráticos e não global. Uma opção a isso pode ser uma escolha exaustiva do ponto de partida e a aplicação de um número vasto de iterações (MORENCY *et al.,* 2007; Pieroni, 2018).

Figura 2 – Exemplificação do funcionamento do algoritmo -*k-means*



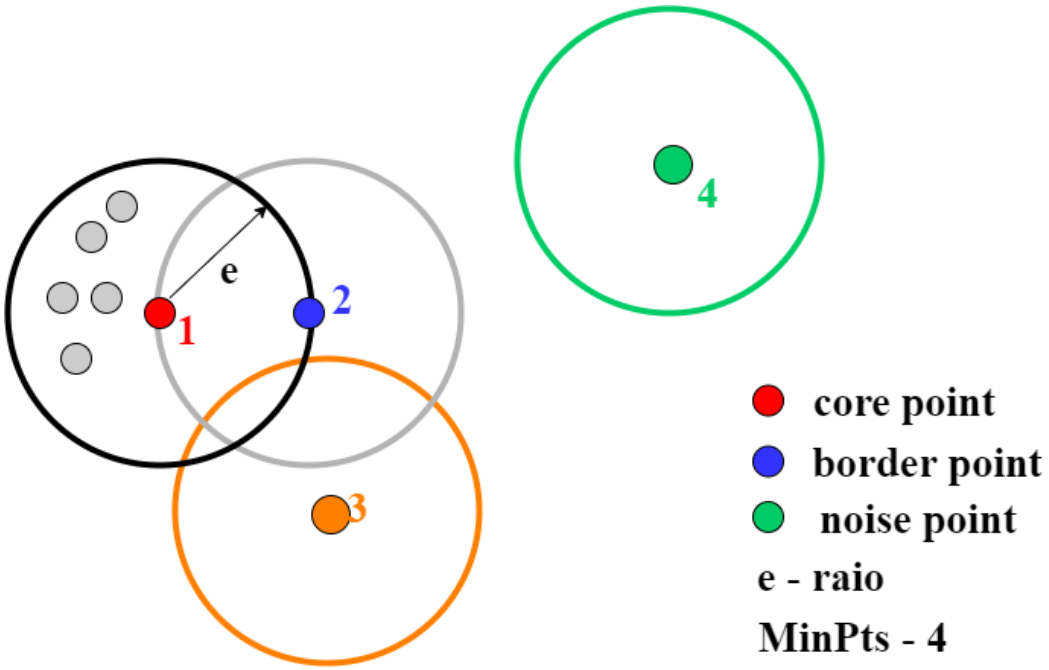
Fonte: Chris Piech, baseado em um trabalho de Andrew Ng [1].

O algoritmo DBSCAN se baseia na identificação de agrupamentos por densidade de pontos e foi proposto inicialmente por Ester *et al.* (1998). Basicamente, pontos que estão em regiões densas formam clusters, enquanto pontos que não estão nessas regiões são classificados como ruídos ou *outliers*. São necessários basicamente dois parâmetros globais para a execução do algoritmo: *MinPts* e ε. O parâmetro ε trata de uma espécie de raio de cobertura de cada ponto e o parâmetro *MinPts* trata da quantidade mínima de pontos consideráveis dentro do raio (ε) para existência de um *core point*.

Na Figura 3 é mostrado um exemplo do processo de funcionamento do algoritmo. Após adoção de um raio (ε) e do *MinPts*, somente um ponto 1 (Vermelho) se torna *core point*, pois quantidade de pontos na sua vizinhança maior do que *MinPts*. O ponto 2 (azul) é chamado de ponto de borda, ou seja, ele não possui na sua vinhança uma quantidade de pontos maior do que *MinPts*, mas é alcançado pelo *core point* 1, então faz parte do mesmo agrupamento. O ponto 3, por sua vez, não é alcançado pelo 1, mas é pelo 2. Esse processo é

chamado de alcance por densidade, o que faz com que também seja alocado para o mesmo agrupamento. O ponto 4, no entanto, não é alcançável por nenhum grupo, sendo caracterizado dessa forma como ruído.

Figura 3 – Exemplificação do funcionamento do algoritmo DBSCAN



Fonte: Elaborado pelo autor.

A vantagem do DBSCAN é que ele é capaz de identificar valores discrepantes e não necessita da definição do número de agrupamentos a *priori*, o que pode ser interessante quando se objetiva descobrir padrões desconhecidos. No entanto, o resultado doa agrupamentos é muito sensível aos parâmetros. O *k-means* possui alta performance, é escalável e simples, com baixo tempo de execução. Como mencionado, a inicialização aleatória dos centroides, além da definição previa do número de agrupamentos são desvantagens desse algoritmo. Para essa última desvantagem, alguns trabalhos utilizam indicadores para estimar o número ótimo de *clusters*, como coeficiente o *average silhouette,* que é uma medida de separabilidade e compacidade dos grupos (ZHAO *et al.*, 2017) ou o método denominado *gap statistic,* proposto por (Tibshirani *et al.*2001), que realiza simulação de Monte Carlo e compara uma dada formação de grupos com outra formação totalmente aleatória. Para tentar minimizar o problema da inicialização aleatória do centroide, Arthur e Vassilvitskii (2007) propuseram o algoritmo *k-means*++, utilizando uma técnica de propagação aleatória para garantir que a melhor solução será obtida.

Para aplicar as técnicas de clusterização, é necessário selecionar variáveis ou atributos que melhor representem o fenômeno a ser estudado, podendo ser esses atributos quantitativos ou qualitativos (JAIN *et al.*1999). Os métodos de clusterização também requerem uma medida de similaridade entre os componentes da amostra, também conhecida como matriz de distância ou matrix de similaridade (JAIN; DUBES, 1988). Para medir a similaridade entre os componentes, pode-se utilizar diferentes medidas de similaridades, como a distância Euclidiana, Minkowski, Mahalanobis, Person, entre outros (XU *et al.*, 2005). Cada tipo de distância tem sua vantagem específica, sendo a mais comum e mais utilizada a distância Euclidiana.

Agard et al. (2006) utilizaram dois métodos em conjunto para detecção de padrões. Eles estabeleceram a existência de 20 grupos e aplicaram o algoritmo k-means. Posteriormente, utilizaram o output do k-means para aplicar o Hierarchical Ascending Clustering (HAC). Os atributos definidos pelos autores têm relação com o dia da semana de utilização e quatro períodos do dia, que remetem dois intervalos de pico e dois de entre pico (5:30h às 8:59h, 9:00h às 15:29h, 15:30h às 17:59h, e 18:00h em diante). Se o tipo de análise é mais operacional, há uma limitação em utilizar períodos longos, uma vez que alguma dinâmica de deslocamento que acontece de forma mais clara dentro desses períodos pode ser omitida e, dessa forma, padrões de deslocamentos mais específicos podem ser omitidos. Nesse trabalho, não há nenhum detalhamento sobre o tipo de distância de similaridade utilizada nesse trabalho.

Morency et al. (2006), usando como atributos a hora do dia de embarque (atribuindo variável binária para o intervalo em que o indivíduo realizou o embarque), além do dia da semana em que o indivíduo utilizou, aplicou a técnica *k-means* para obtenção dos agrupamentos, com uma amostra de 277 dias. Há no trabalho uma análise descritiva prévia de algumas variáveis, no sentido de auxiliar na definição dos melhores atributos. No entanto, não há nenhuma informação sobre a medida de similaridade adotada e sobre os critérios ou hipóteses estabelecidas para a definição dos 29 grupos definidos.

Agard e Trépanier (2013) utilizaram o *k-means* para extrair padrões de viagens ao longo do tempo. No início, eles atribuíram variáveis binárias para o uso ou não do indivíduo nas 24 horas do dia. No entanto, eles realizaram uma discussão sobre como em alguns casos, como o deles, as medidas de similaridade comumente utilizadas (distância Euclidiana, Manhattan, entre outras) podem não ser uma boa opção para oferecer separabilidade aos grupos, pois indivíduos que usam em horários diferentemente espaçados (07:00h/08:00h e

07:00h/09:00h, por exemplo) podem ser significativamente similares. Para a correção desse fator, eles utilizaram uma transformação da distância euclidiana para coordenada polar, incorporando outras variáveis no modelo como a frequência de uso (que não necessariamente estaria incorporada no modo anterior), além de uma medida de distância temporal entre os usos, definindo assim, uma nova matriz de similaridade. Posteriormente, eles definiram o número de agrupamentos por meio de visualização dos pontos plotados em um eixo cartesiano. Uma limitação desse método é que em alguns casos, não é simples definir visualmente a quantidade de agrupamentos de acordo com os critérios estabelecidos.

Ma *et al.* (2013) usam o DBSCAN para identificar os padrões temporais e espaciais de uso de cada usuário. Como atributos, eles utilizaram a diferença entre os tempos de embarque e a distância espacial entre registros. Para que o indivíduo possa possuir um padrão de viagens, ele deve possuir pelo menos três registros não variando mais do que 1 hora e a uma distância menor do que 200 m. Dessa forma, pode-se construir o perfil típico de cada usuário. A vantagem da utilização desse método é que alguns dias atípicos de uso podem ser desconsiderados, sendo identificados pelo DBSCAN como ruído.

Na segunda parte do trabalho, os autores realizam a clusterização com foco no comportamento similar de todos os usuários e utilizam os seguintes critérios: número de dias de uso, número de embarques em horas similares, número de rotas similares e número de utilização em paradas similares. Sobre a escolha dos atributos, os autores ainda discutem que é possível que haja uma correlação significativa entre eles, no entanto, por problemas de registros no sistema de coleta de dados, eles defendem que é importante ter um dado redundante para melhor classificar os grupos. Nessa parte, eles utilizam o algoritmo *k-means*++, que é uma alternativa para tentar mitigar o problema da inicialização aleatória do centroide, um dos problemas do *k-means* tradicional. Como os atributos possuem magnitudes diferentes, há uma normatização das variáveis, para fazer com o que todas possuam o mesmo peso da definição dos grupos. Os autores definem a distância euclidiana como medida de similaridade e centroides de inicialização, representando níveis de regularidade diferentes (MA et al., 2013).

Kieu (2015) utilizaram o método DBSCAN para analisar padrões espaciais e temporais. Nos dois tipos de análise, os autores utilizam um único atributo. Na espacial, a localização da parada de embarque/desembarque foi utilizada, com os parâmetros sendo ε = 1 quilômetro e *MinPts* = 8 e, para análise temporal, os autores utilizaram o tempo de embarque, variando entre 0 e 1440 minutos, representando os minutos ao longo do dia, com

os atributos sendo ε = 5 minutos e MinPts = 6. Nos dois casos, os autores realizaram uma análise de sensibilidade dos parâmetros. O parâmetro ε possui maior relação com a literatura, enquanto o MinPts pode ter sua definição de forma mais arbitrária. Dessa forma, os autores escolheram o valor de MinPts que geravam uma quantidade de usuários regulares maiores, tanto espaciais quanto temporais.

Pieroni (2018) utiliza em duas etapas do seu trabalho métodos de clusterização. Na primeira, o objetivo é estimar o local mais provável de residência dos usuários, por meio da localização da validação, sendo DBSCAN o método utilizado. Nesse caso, os parâmetros utilizados foram ε = 1 quilômetro e MinPts = 2. A autor cita que 1 quilômetro é uma distância razoável de caminhada da residência até o embarque e do embarque até a passagem do cartão propriamente dita. Em relação ao parâmetro MinPts, ao autor realizou uma análise de sensibilidade desse parâmetro, testando valores de 3,4,5 e 11. O maior erro encontrado foi de 3,5%, indicando que não há muita variação dos resultados dada a variação do parâmetro MinPts.

No segundo momento, o autor utilizou o método *k-means* para classificar os usuários quanto aos padrões de uso, considerando vários critérios temporais de deslocamento (mediana da hora da primeira validação, dispersão da hora da primeira validação, frequência de uso semanal, dispersão da frequência de uso semanal), espaciais de deslocamento (mediana da máxima distância entre validações, dispersão da máxima distância entre validações), padrões de atividade (mediana da máxima duração diária das atividades, dispersão da máxima duração de atividades), características socioeconômicas (renda média), além da caracterização quanto ao tipo de do uso do solo do desembarque dos usuários (residencial baixa renda, residencial média/alta renda, comercial/serviços/industrial, residencial/comercial, outros usos) (PIERONI, 2018).

Na Tabela 1 é mostrado um resumo dos trabalhos revistos, descrevendo os atributos utilizados pelos autores para aplicação do método de clusterização, o tipo de variável, o dado utilizado, a distância de similaridade, além do método de clusterização utilizado e seus parâmetros de entrada.

Tabela 1 – Resumo dos estudos que utilizam técnicas de clusterização

Autor	Atributos	Tipo de variável	Dado utilizado	Distância de similaridade	Método	Parâmetros de entrada
Agard et al. (2006)	4 períodos do dia x 5 dias da semana (20 variáveis binárias)	Binária	5 semanas	Não especificada	HAC/ *k-means*	Número de grupos (20 grupos)
Morency (2006)	Hora do dia de utilização e dia da semana	Binária	277 dias	Não especificada	*k-means*	Número de grupos (sem explicação das hipóteses que levaram a escolha da quantidade de grupos)
Agard e Trépanier (2013)	Frequência, hora do dia, variabilidade da hora de uso, distância temporal entre usos	Contíua e binária	1 ano	Distância euclidiana	*k-means*	Experiência do pesquisador na definição dos grupos
Ma et al. (2013)	Número de dias de uso, número de embarques em horas similares, número de rotas similares, número de utilização de paradas similares	Contínua e normatizada	Semana típica	Distância Euclidiana	*k-means* ++	Níveis de similaridade (definição não aleatória dos centroides)
Kieu et al. (2015)	**DBCSAN-** Localização da validação ***K-means***- Minuto de utilização	Contínua	4 meses	Não especificada	DBSCAN	Análise de sensibilidade dos parâmetros e literatura
Pieroni (2018)	**DBSCAN-**Localização da validação ***K-means***-Hora da primeira validação,frequência semanal, ***d***istância especial e temporal entre validações**,** desvio padrão da primeira validação e da distância temporal e especial das validações, frequência, **,*r***enda media**,*u***so do solo	Continua e normatizada	11 semanas	Distância euclidiana	DBSCAN/ *K-means*	**DBSCAN-** Análise de sensibilidade dos parâmetros ***K-means-*** *average silhouette*

Fonte: Elaborado pelo autor.

2.4 Considerações finais

De forma geral, com a leitura dos trabalhos, percebe-se que a aplicação de técnicas de clusterização podem auxiliar no entendimento sobre o padrão de deslocamento, seja ele temporal, espacial ou espaço-temporal, além de permitir considerações mais significativas do que as obtidas por meio dos dados brutos.

Sobre os métodos utilizados, a maioria dos trabalhos utilizou os mais conhecidos, *k-means* (particional) e DBSCAN (baseado em densidade). Em geral, os atributos como hora de utilização são amplamente utilizados pelos autores, além da frequência de utilização e da localização das validações/embarques. O intervalo temporal entre usos também é usado em alguns trabalhos e é importante pois pode remeter à duração da atividade dos usuários. O intervalo espacial entre usos também é utilizado e pode ser importante, uma vez que faz luz a distância entre atividades e pode auxiliar na detecção de grupos que residem em áreas diferentes, com características socioeconômicas diferentes e diferentes níveis de acessibilidade e oferta do transporte coletivo. No entanto, como há a possibilidade de as validações diárias não serem realizadas na origem e no destino dos usuários, a distância espacial entre validações pode perder esse caráter explicativo.

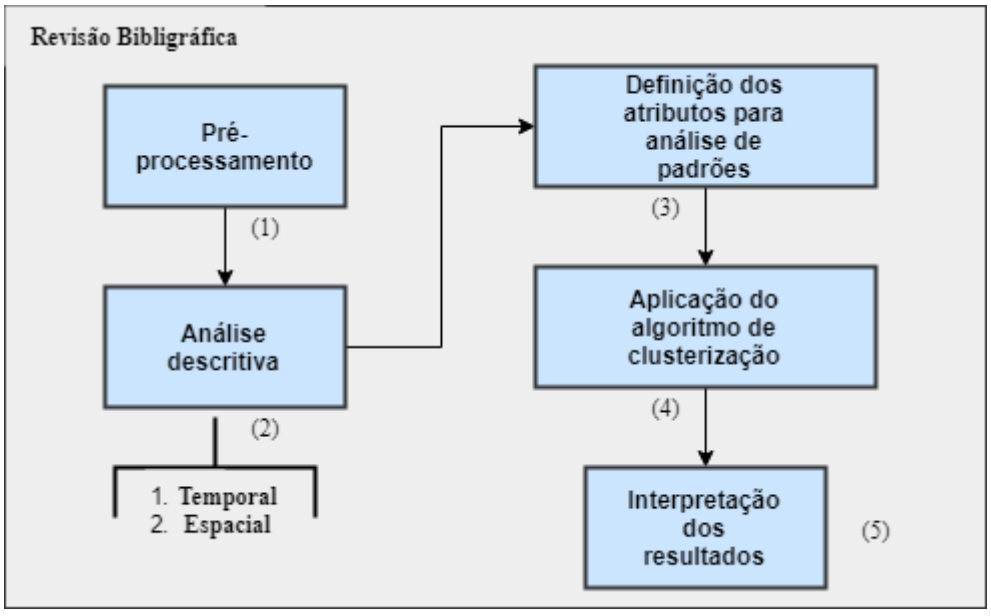
Dos trabalhos que utilizam o método *k-means*, somente Pieroni (2018) utiliza o coeficiente *average silhouette*, como tentativa de estimar a quantidade ótima de agrupamentos. Os outros trabalhos que utilizam esse método definem a quantidade de agrupamentos baseado somente na experiência do pesquisador, o que pode limitar a análise, no sentido de que podem haver grupos que não sejam do conhecimento do pesquisador. Quando utilizam o DBSCAN, os trabalhos inserem os parâmetros do algoritmo baseado na literatura e em uma análise de sensibilidade, o que pode tornar a detecção dos grupos mais confiável, além desse algoritmo não necessitar da definição dos grupos *a priori*.

Como limitação, apesar de analisarem a mudança dos padrões ao longo do tempo, a maioria dos trabalhos não analisam os padrões com foco na variação da demanda realizada, ou seja, nos tipos e padrões de usuários que estão possivelmente ingressando ou deixando de utilizar o sistema de transporte coletivo. Isso pode ser significativamente importante para fornecer subsídio aos tomadores de decisão, haja vista o cenário de queda da demanda.

3 MÉTODO

Esta seção apresentará o método utilizado no trabalho. O método, subsidiado pela revisão bibliográfica, está representado na Figura 4. A primeira etapa se concentra basicamente em consolidar os dados que serão usados. Como dito anteriormente, o BD – TP conta com dados relativos às validações dos usuários - momento em que o usuário efetua o pagamento e passa pela catraca, GPS dos carros que compõem as frotas e GTFS. No entanto, as validações em sua forma original não contêm o dado de localização. Braga (2019) consolidou a base de validações, estimando a localização por meio do GPS. Essa etapa então tem como principal objetivo estimar as localizações das validações de todos o período de análise. É importante ressaltar que não se obtêm a localização de todas as validações, mas no geral a proporção é bastante significativa, girando em torno de 90% para os anos analisados. Os dados utilizados para análise são relativos ao período entre os anos de 2014 e 2018 e foram disponibilizados pela Prefeitura Municipal de Fortaleza (PMS) e todas as consolidações de dados, bem como resultados, foram produzidos utilizando o *software* QGIS 2.18 e a linguagem R de programação.

Figura 4 – Sequência do método



Fonte: Elaborado pelo autor.

3.1 Tipologia dos dados

Como sobredito, os dados utilizados para análise serão as validações georreferenciadas. Os dados de validações possuem informações relativas ao código de identificação do usuário (id), onde esse código possui valor 0 se o pagamento for realizado

na forma de dinheiro, também chamado de inteira. No mais, o código id remete ao código do cartão utilizado pelo usuário. Dessa forma, é possível verificar o uso do transporte público dos usuários continuamente, para aqueles que possuem id diferente de 0.

Além do código de identificação, existem informações como o prefixo do carro e o id do veículo, sentido da viagem, tipo de cartão utilizado, se a validação foi resultado de uma integração ou não, além da localização da validação. Uma amostra dos dados utilizados é representada na Tabela 2.

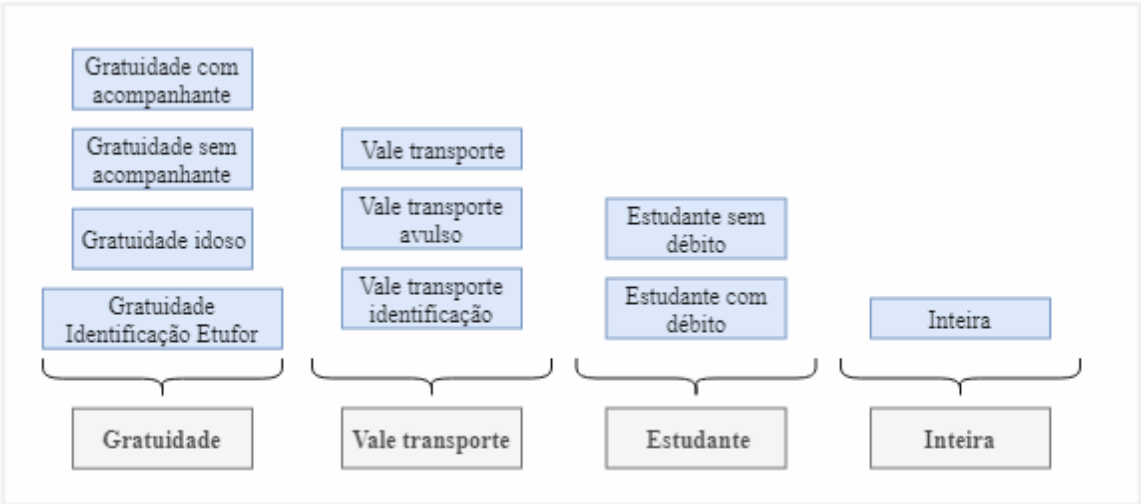
Tabela 2 – Amostra dos dados de validações utilizados

id	linha	prefixo_carro	vehicleid	tipo_cartao	integração	dia	hora	longitude	latitude
6644879	632	35263	33551	Vale Transporte	N	01/10/2018	06:09:37	-38.479	-38.574
5142964	632	35263	33551	Vale Transporte	N	01/10/2018	06:10:06	-3.848	-3.857
5432765	632	35263	33551	Gratuidade	N	01/10/2018	06:10:13	-3.848	-3.857
0	632	35263	33551	Inteira	N	01/10/2018	06:11:24	-38.484	-3.856
0	632	35263	33551	Inteira	N	01/10/2018	06:11:28	-38.484	-3.856

Fonte: Elaborado pelo autor

Em relação aos tipos de cartão, originalmente existem 10 categorias. Contudo, por entender que essas várias categorias podem ser reduzidas e representam os usuários que experimentam modelos de tarifas semelhantes, foram definidas 4 categorias, são elas: Estudante, Vale transporte, Inteira e Gratuidade. A Figura 5 ilustra como as categorias de cartões foram reduzidas. É interessante ressaltar que dentro da categoria vale transporte, estão majoritariamente os indivíduos que possuem emprego formal. Contudo, a categoria vale transporte avulso é relativa a usuários que podem ou não possuir empregos formais.

Figura 5 – Agregação dos diferentes tipos de cartões adotada



Fonte: Elaborado pelo autor

3.2 Análise descritiva – temporal

O objetivo desta etapa é propiciar um melhor entendimento sobre indicadores temporais da base de *smart card* do BD-TP. Os indicadores analisados, bem como seu objetivo são:

- **Quantidade de usuários por dia**

Esse indicador mensura a quantidade de passageiros distintos que usaram o transporte coletivo por dia e por tipo de cartão (Vale transporte, Estudante, Gratuidade e Inteira). O objetivo é verificar a variação desse indicador em agregações temporais distintas, como o ano, mês e dia da semana, verificando se há diferenças significativas de uso entre os anos. É interessante ressaltar que quando o usuário realiza pagamento por meio de dinheiro, chamado de pagamento em inteira, não é possível obter qualquer identificação desse usuário, não sendo possível, portanto, calcular a quantidade de usuários por dia que realizaram pagamento em dinheiro. Dessa forma, para esse tipo de cartão, utilizou-se número bruto de validações.

- **Frequência de uso**

A frequência de uso é calculada baseado na quantidade de vezes que o indivíduo utilizou o sistema em determinada agregação temporal. O objetivo é obter a distribuição de frequência dos usuários em um dia, na semana, dependendo do tipo de cartão, o número de dias de uso na semana, bem como os dias que os usuários deixam de utilizar. Nesse caso, não é possível obter medidas de frequência do pagamento em dinheiro, sendo essa análise então realizada para os tipos de cartões restantes. Em todas essas análises, realizou-se comparação médias entre os anos de 2014 e 2018.

- **Intervalo temporal entre validações**

Esse indicador mensura o intervalo temporal entre a primeira e a última validação de um mesmo indivíduo, subtraídas as validações que representam integração temporal. Também foram retiradas da amostra para essa análise as validações de um mesmo usuário em um intervalo menor que 2 horas, pois podem ser possíveis integrações não registradas e viagens muito curtas, sendo a ideia representar melhor com esse indicador os percursos para as atividades de trabalho educação. O objetivo é avaliar de forma descritiva a distribuição desse indicador ao longo dos anos e relacioná-lo com a possível duração das atividades dos usuários, pois

o intuito é aproximar cada vez mais o evento da validação de predominantemente uma produção ou atração de viagem, dependendo do horário do dia. Nesse caso, foram utilizados dados dos anos de 2014 e 2018 para estimar esse indicador. O objetivo é comparar as distribuições visualmente e verificar alguma mudança.

Essa análise inicial é importante e pode auxiliar na definição dos melhores critérios para identificar padrões na etapa posterior, além de oferecer informações mais detalhadas sobre demanda (possível variação de grupos específicos, quedas de demanda mais acentuada em alguns grupos e em alguns locais) para as agências de planejamento e operação de transporte público.

3.3 Análise descritiva – Espacial

Nesta etapa, tem-se como objetivo principal verificar se é possível considerar se a localização da primeira validação pode representar uma produção de viagem de um usuário, ou seja, se a primeira validação é significativamente próxima do embarque/endereço de moradia do usuário. Para isso, com o mês de outubro para cada ano de análise (2014 e 2018), foram amostrados cerca de 1.200 usuários que utilizaram o sistema de transporte público mais 10 vezes no mês, com o intuito de garantir a variabilidade do uso ao longo do tempo de cada usuário. O objetivo de analisar os anos de 2014 e 2018 vem também da necessidade de avaliar o impacto das mudanças de localização da catraca de trás para frente dos veículos, somado à diminuição do espaço disponível para acomodação antes da validação. A hipótese é que os usuários no ano de 2018 devem validar mais próximo à origem do que anteriormente.

Posteriormente, utilizando hexágonos de diagonal de aproximadamente 1100 m (BRAGA, 2019) como nível de agregação espacial, atribuíram-se as primeiras validações dos usuários aos hexágonos em todo o período de análise. O intuito é utilizar um nível de agregação intermediário entre o setor censitário e bairro, a fim de permitir a visualizações do uso dos usuários a nível de corredor e em um nível aceitável de caminhada. Para estimar o endereço dos usuários utilizando os dados de cadastro, utilizou-se a ferramenta do *google* de georreferenciamento. É importante ressaltar que não é possível estimar o endereço de moraria de todos os usuários, sendo possível somente obter uma amostra, devido geralmente à falha no cadastro dos endereços. Dessa forma, o tamanho da amostra escolhido foi de 1.200 para que ao final fosse possível obter resultados de aproximadamente 1.000 usuários.

Com o endereço estimado e o padrão de validações nos hexágonos ao longo do mês,

calculou-se a distância euclidiana do endereço estimado para o hexágono que possuía mais validações realizadas pelos usuários. O intuito é verificar a distribuição dessas distâncias dos usuários e comparar os anos de 2014 e 2018.

3.4 Obtenção dos padrões de deslocamentos

A partir da revisão bibliográfica, bem como da análise descritiva, foram definidos os atributos para a obtenção dos grupos. Os atributos, bem como sua definição e justificativa são:

1. **Frequência média diária de uso**: esse atributo representa o quão os usuários usam o transporte coletivo (AGARD E TREPANIER, 2013). Nesse sentido, é possível separar aqueles usuários que apresentam características de uso diferentes com relação à frequência. O cálculo foi realizado por meio da contagem do número de validações diárias realizadas pelos usuários.
2. **Desvio padrão da frequência**: esse atributo diz respeito a variabilidade da frequência dos usuários, ou seja, o objetivo é tentar captar usuários que possuem comportamentos diferentes ao longo do ano. A variável é obtida calculando-se o desvio padrão das observações de frequência diárias. Esse atributo também é utilizado por Pieroni (2018).
3. **Número de dias típicos de uso durante o ano**: com essa variável, é possível perceber como a utilização dos usuários está distribuída ao longo do ano, assim como feito por Ma *et al.* (2013). Como pôde ser observado na análise descritiva, alguns usuários usam poucas vezes na semana e esse efeito pode ser escalado para o ano inteiro. Essa variável foi calculada a partir da contagem do número de dias em que o usuário utilizou o sistema, independente das frequências e excluindo os finais de semana.
4. **Número de dias em que o usuário realizou uma validação**: Essa variável tem como objetivo captar o efeito de um uso mais esporádico do usuário, ou seja, o usuário realizou somente parte da viagem utilizando transporte público, se sua viagem é pendular. Nesse sentido, usuários que têm esse comportamento frequente, podem ser mais propensos a deixar de usar o sistema.
5. **Distância temporal média entre embarques (horas)**: Essa variável é calculada como sendo o intervalo temporal entre a primeira e a última validação. A ideia é que nessa variável seja possível perceber uma estimativa do tempo da atividade dos usuários, assim como feito por Pieroni (2018), bem como perceber algum efeito da operação no tempo de viagem deste, uma vez que o usuário pode não validar na

origem e no destino.

6. **Proporção de validações realizadas pela manhã, tarde e noite:** Nesses indicadores são calculadas as proporções de vezes em que os usuários realizaram a primeira validação pela manhã, tarde e noite. O objetivo é distinguir os usuários que teoricamente geram viagens em horários distintos. Os períodos de manhã, tarde e noite foram definidos a partir da observação do perfil horário de validações.

A partir dos atributos definidos, realizou-se a normalização das variáveis, fazendo variar de zero à um, para que todas a exerçam pesos iguais sobre a definição dos grupos (Pieroni, 2018; Ma *et al.*,2013). A medida de similaridade/distância utilizada foi a euclidiana, que é utilizada na grande maioria dos trabalhos e o método de clusterização utilizado foi o *k-means*, sendo utilizado o critério do *gap statistic* para estimar o número ótimo dos grupos. Acredita-se que esse método de estimação seja mais capaz de definir grupos menores com características distintas. Para aplicar o *gap statistic*, foram usadas amostras de 2000 usuários para cada ano, devido ao esforço computacional requerido nesse método.

O *k-means* foi utilizado pois, além de possuir boa rapidez para a detecção de grupos grandes, ele não necessita de parâmetros além do número de agrupamentos para sua operacionalização. Quando se usam muitos atributos, como é o caso, as variáveis de entrada de outros métodos como o DBSCAN, por exemplo, perdem o significado físico e podem levar à um mal ajuste do modelo de agrupamentos, além de necessitar de uma análise de sensibilidade, para que os resultados se tornem mais confiáveis.

3.5 Interpretação dos padrões

Os agrupamentos foram gerados para os anos de 2014 e 2018 para todos os usuários que utilizaram cartão naquele ano. Como esforço inicial, verificou-se a proporção de cada cartão (Vale Transporte, Estudante e Gratuidade) nos agrupamentos obtidos, com o intuito de perceber se existem agrupamentos em que alguns tipos de cartões são dominantes.

Posteriormente, foi observado um resumo de cada atributo para cada agrupamento e para cada ano, utilizando a mediana como medida de tendência central, uma vez que a distribuição da maioria dos atributos não é simétrica. Foram mostrados a também quantidade de usuários de cada grupo e porcentagem de cada grupo em relação ao total de usuários que utilizaram cartão no respectivo ano. O objetivo nessa etapa é comparar os grupos entre si e identificar quais grupos são mais relevantes, frequentes ou esporádicos, levantando hipóteses

sobre o comportamento desses grupos, por meio dos valores dos atributos obtidos.

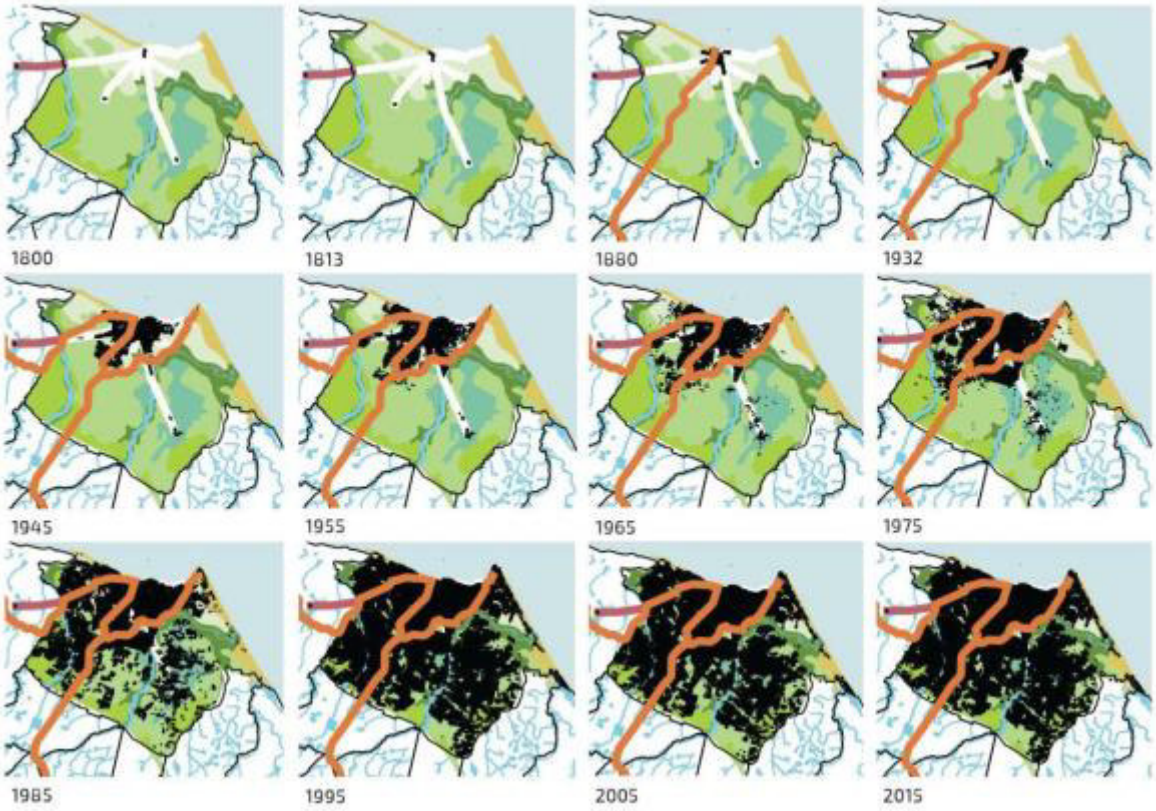
Por fim, fez-se uma análise de comparação entre os grupos obtidos em 2014 e 2018, verificando quais grupos são mais semelhantes. Essa semelhança é calculada utilizando a mediana de todos os atributos. O intuito é verificar quais grupos são semelhantes nesses anos e verificar uma estimativa da variação do tamanho do agrupamento entre os anos. Dessa forma, é possível levantar hipóteses sobre a variação dos grupos. No entanto, não é possível garantir se os grupos mais semelhantes são formados pelos mesmos usuários, sendo essa alternativa somente uma forma de levantar hipóteses sobre a as mudanças de comportamento ao longo do tempo.

4 PADRÕES DE DESLOCAMENTO EM FORTALEZA-CE

4.1 A cidade de Fortaleza-CE

O objetivo desta seção é contextualizar a região de estudo, exibindo como algumas variáveis relacionadas ao uso do solo, como renda, população, empregos, escolas estão distribuídas ao longo do território, assim como descrever brevemente como está organizado e opera o sistema de transporte coletivo. A cidade de Fortaleza possui área de aproximadamente 300 quilômetros quadrados e abriga, em 2019, segundo o IBGE, 2.669.342 habitantes. Sendo uma cidade litorânea, Fortaleza começou a se desenvolver nas proximidades do litoral, na porção que hoje se localiza a região mais desenvolvida economicamente, se expandindo mais tarde para as áreas mais periféricas, como é mostrado na Figura 6.

Figura 6 – Expansão urbana de Fortaleza em seus eixos viários

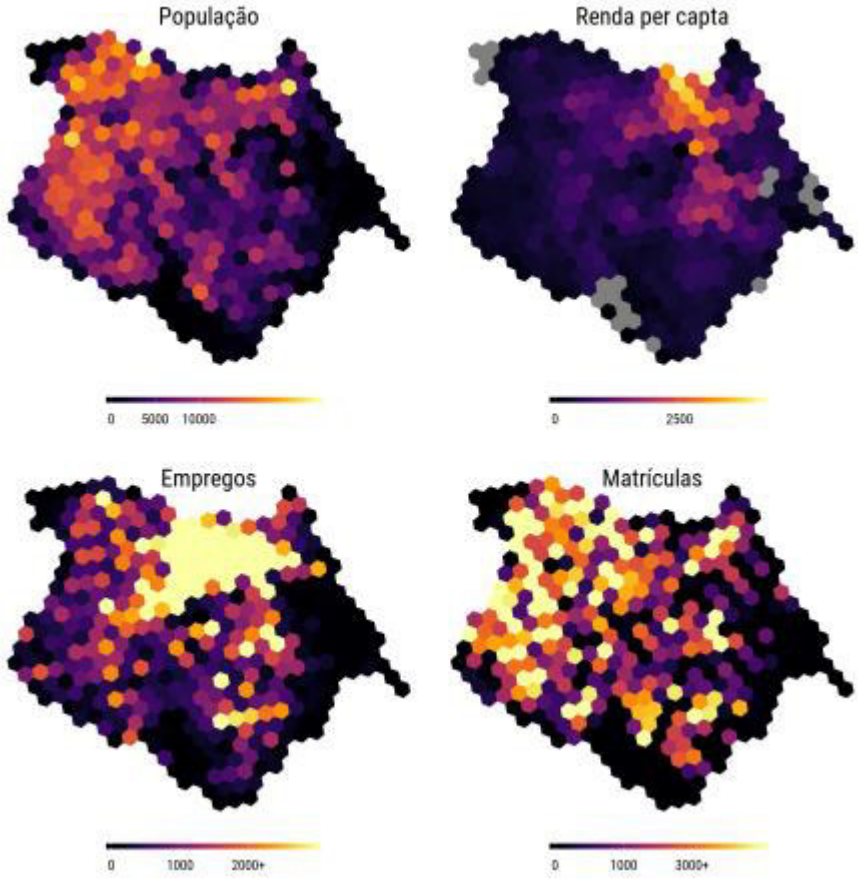


Fonte: Fortaleza 2040

A distribuição espacial de algumas variáveis socioeconômicas e de uso do solo é mostrada na Figura 7. A população está mais concentrada na região noroeste e sudoeste da cidade. A renda, no entanto, tem uma distribuição significativamente diferente e está mais reunida na área central da cidade. Também é basicamente nessa área que se concentram a

maior parte dos empregos de Fortaleza, sendo essa área com uso do solo bastante comercial. Em relação as matrículas de escolas, percebe-se que essas estão concentradas na região oeste da cidade.

Figura 7 – Distribuição de variáveis socioeconômicas e de uso do solo na cidade de Fortaleza



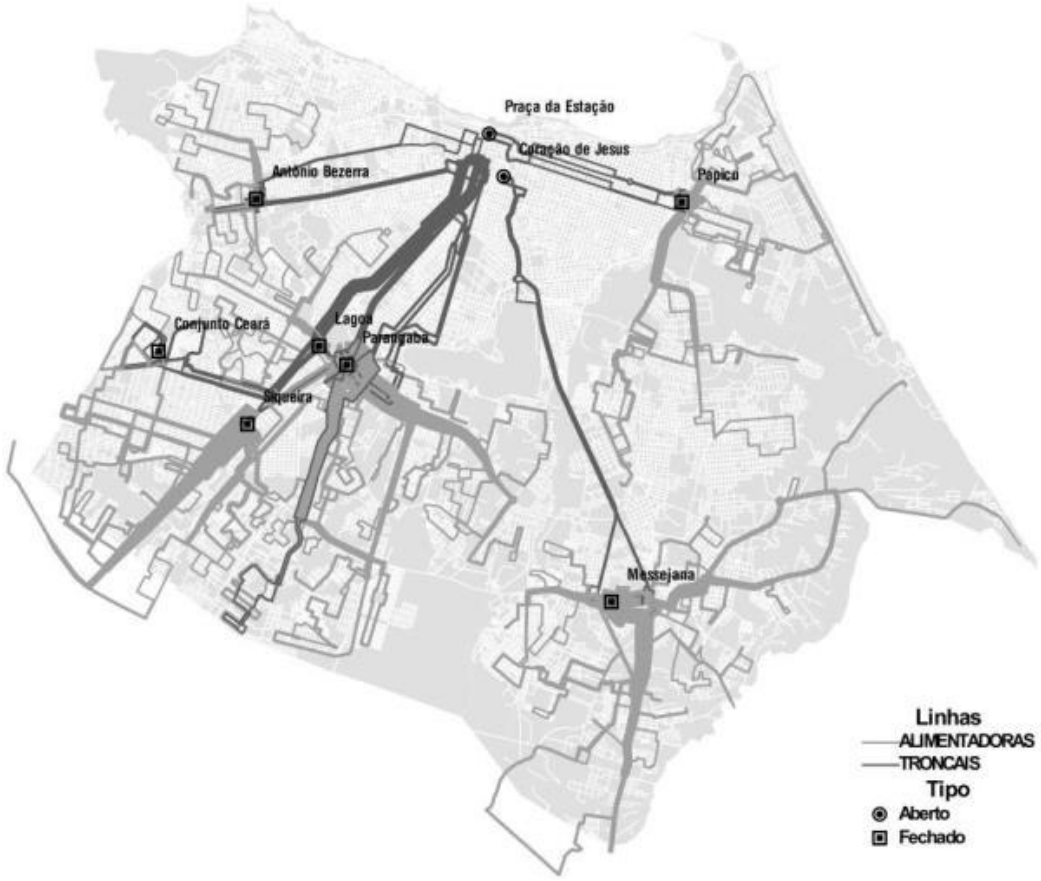
Fonte: Braga (2019)

É basicamente nesse contexto que, em 1992, surge o Sistema Integrado de Transporte de Fortaleza (SIT-FOR). Nesse sentido, foi estabelecido que o sistema de transporte público coletivo de Fortaleza seria do tipo tronco-alimentador, com dois tipos de linhas: as que fazem a integração bairro-terminal, denominadas alimentadoras, e as que integram o terminal ao centro da cidade ou ainda a outro terminal, denominadas troncais. Também existem as linhas circulares e inter-bairros, que oferecem integração nos terminais. Atualmente, o sistema conta com 7 terminais fechados e 2 abertos, localizados em bairros periféricos e no Centro da cidade. Nos terminais fechados, a transferência é gratuita para qualquer linha que sirva o terminal, uma vez que nos terminais existem bilheterias para o acesso, mediante pagamento da tarifa, para usuários predominantemente provenientes de áreas adjacentes. A disposição espacial dos terminais, bem como a configuração das linhas troncais e alimentadoras, é apresentada na Figura 8. Em 2010, o SIT-FOR contava com uma

frota de aproximadamente 1.800 veículos, distribuídos entre 300 linhas, tendo esses veículos idade média de cerca de 4 anos (Anuário 2010). Em 2018, segundo a prefeitura de Fortaleza, 27% da frota possuía ar-condicionado e 56% oferecem acesso gratuito à internet (Nordeste, 2018).

Em relação à integração temporal, em 2014, foi implantado o sistema de bilhete único (denominada aqui vale transporte avulso) em Fortaleza, sendo o atual modelo de integração. Com esse sistema, o usuário pode embarcar quantas vezes se queira em um período de 2 horas. É importante ressaltar que todas as pessoas podem requerer o bilhete único.

Figura 8 – Disposição espacial dos terminais e das linhas troncais/alimentadoras em Fortaleza



Fonte: Fortaleza (2010)

4.2 Análise descritiva (temporal)

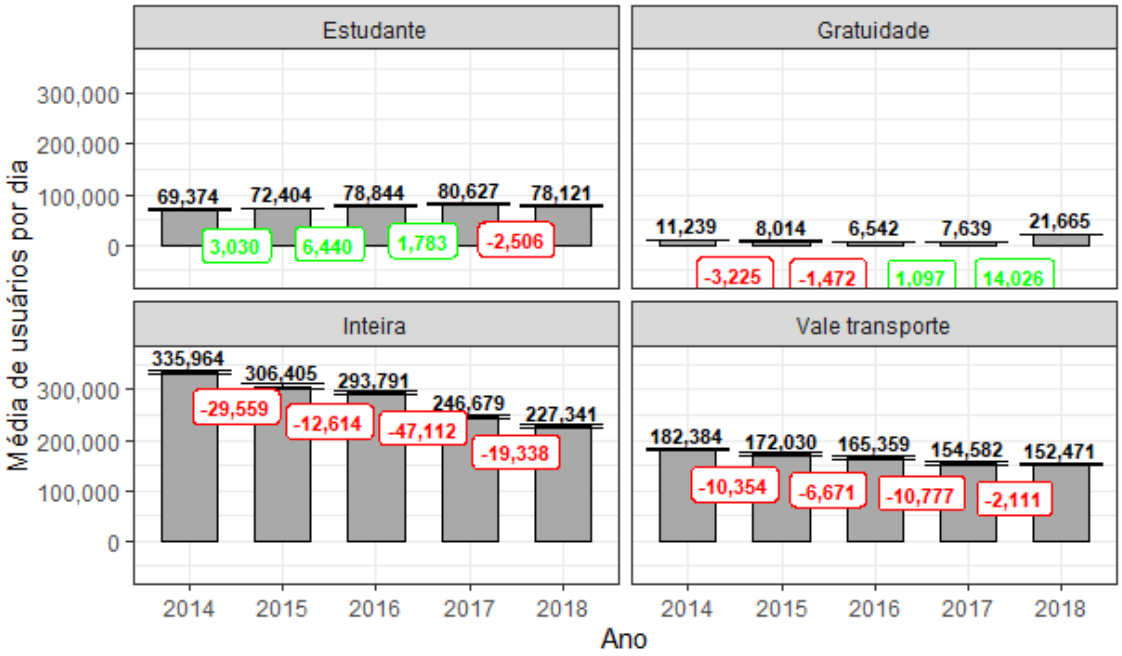
Nesta seção é realizada a primeira parte da análise descritiva, dando ênfase ao aspecto temporal das validações, analisando os indicadores propostos.

4.2.1 Usuários por dia

Com o intuito de subsidiar informações sobre a variação da demanda nos últimos

anos, calculou-se o indicador de usuários por dia para cada tipo de cartão e para os anos de análise (2014 a 2018), calculando a diferença das médias entre os anos, que resulta na estimativa média de usuários que deixaram ou ingressaram no sistema (Figura 9).

Figura 9 – Quantidade média de usuários por ano



Fonte: Elaborado pelo autor

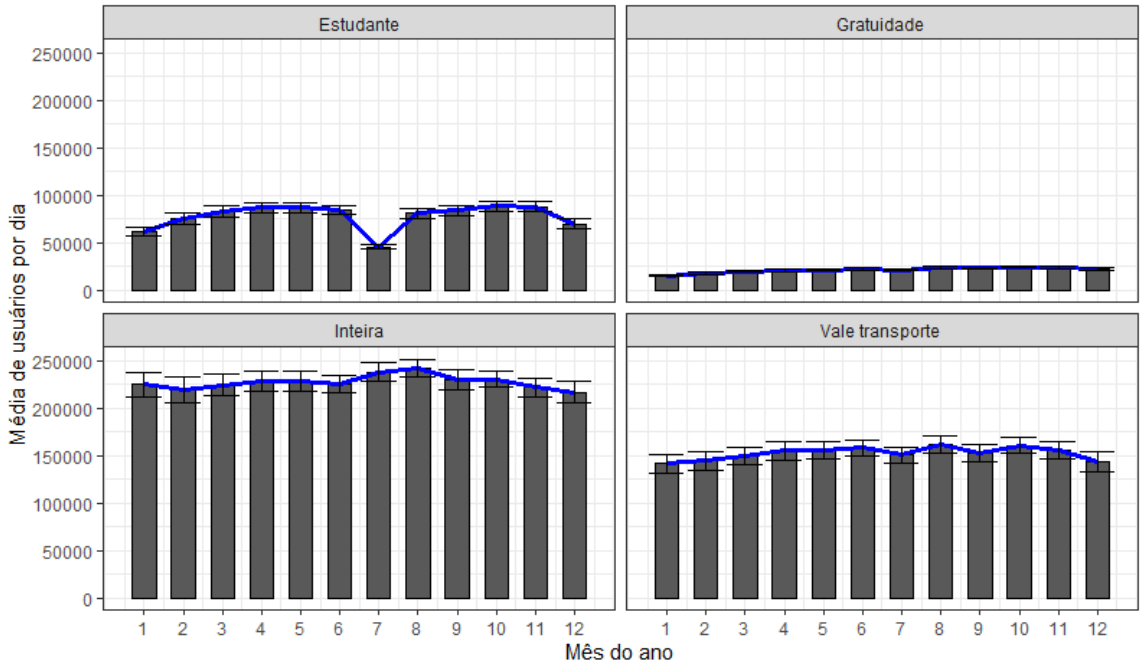
É possível perceber que o número de estudantes não variou significativamente dentre os anos, ganhando mais usuários do que perdendo no período de análise, comportamento este relativamente semelhante aos usuários com gratuidade, que perderam usuários entre os anos de 2014 e 2016, mas recuperaram no decorrer dos anos restantes, com destaque para o crescimento de cerca de 14.000 usuários entre os anos de 2017 e 2018. A hipótese principal é que esse crescimento pode ter sido causado pelo aumento de registros de idosos, aliado ao envelhecimento cada vez mais acentuado da população.

A categoria inteira é que a perde mais significativamente validações no período, com uma perda média estimada entre os anos de 2014 e 2018 de 110.000 validações diárias, sendo essas perdas mais acentuadas entre os anos de 2014/2015 e 2016/2017. Essa tendência maior de queda nesses anos também se manifesta nos usuários de vale transporte, que, do total de usuários perdidos, 71% foram entre esses anos. Entre todo o período analisado, estima-se em média que houve a perda de 30.000 usuários com vale transporte por dia.

Com o intuito de verificar se há diferença na quantidade de usuários que utiliza o transporte de forma mais desagregada, calculou-se o indicador para os diferentes meses do ano de 2018, junto com o intervalo de confiança, como é mostrado na Figura 10. Nota-se que

a presença de sazonalidade é mais acentuada para os estudantes, onde a quantidade de usuários é consideravelmente diferente nos meses de janeiro, julho e dezembro, período em que ocorrem as férias escolares. Nas outras categorias, a sazonalidade parece afetar pouco a quantidade de usuários que utiliza o sistema, não sendo possível afirmar que esse indicador pode ser considerado diferente em alguns meses do ano, o que pode denotar que essas viagens precisam ser realizadas continuamente, indicando possivelmente viagens com motivo trabalho.

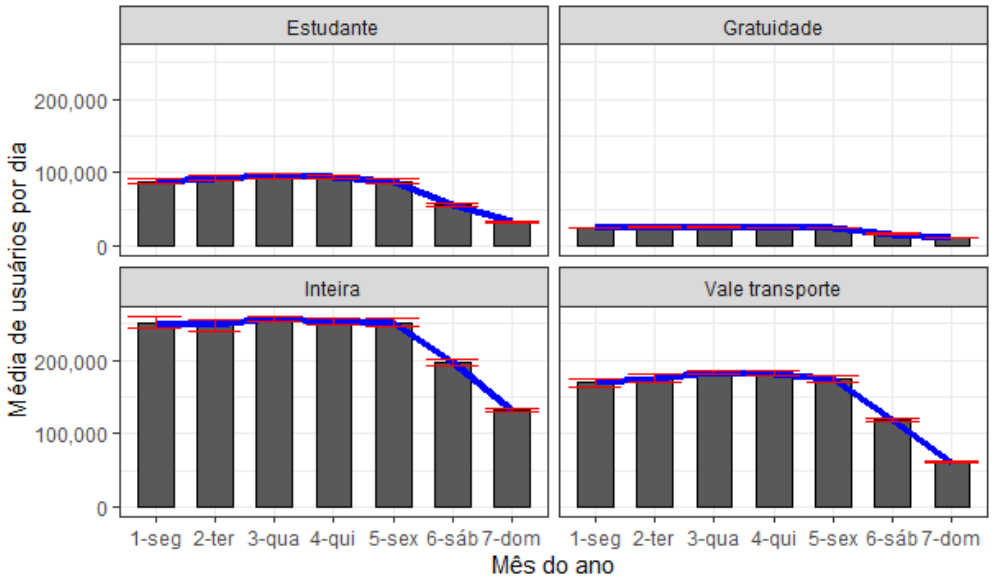
Figura 10 – Quantidade média de usuários por mês



Elaborado pelo autor

A quantidade de usuários por dia também foi calculada para os diferentes dias da semana e para os diferentes tipos de cartão, no ano de 2018, como representado na Figura 11. Em geral, o número de usuários para todos os tipos de cartão, exceto gratuidade, não se altera ao longo dos dias úteis da semana. Nos finais de semana o indicador é significativamente menor do que nos dias úteis, sendo menos acentuada essa tendência para os estudantes e usuários com gratuidade. Dessa forma, é possível assumir que a quantidade de usuários é significativamente diferente nos finais de semana.

Figura 11 – Quantidade média de usuários por dia da semana



Fonte: Elaborado pelo autor

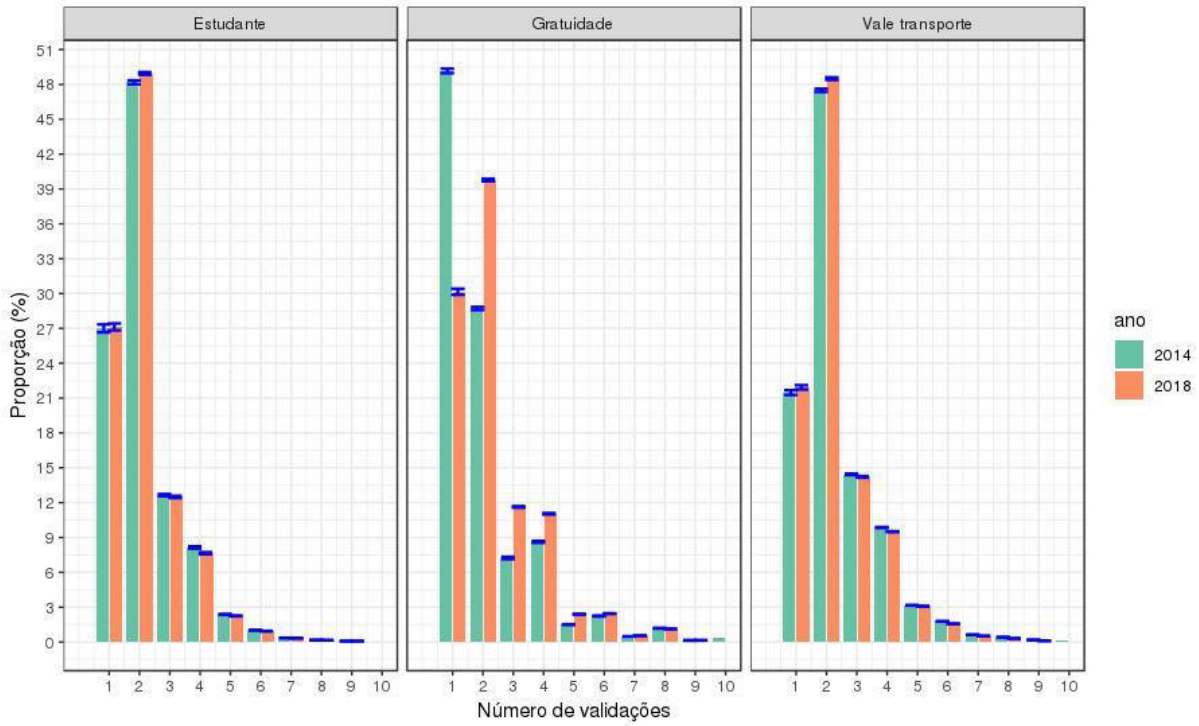
4.2.2 Frequência de uso

Para entender como se comporta a intensidade de uso dos usuários do sistema, calculou-se o a distribuição de frequência média diária de todos os dias típicos para os diferentes tipos de cartões nos anos de 2014 e 2018, como representado na Figura 12. Para o ano de 2014, nota-se que aproximadamente metade dos estudantes e usuários que utilizam vale transporte validam duas vezes por dia, indicando possivelmente um caráter pendular mais forte, com esse comportamento se repetindo no ano de 2018.

Grande parte dos indivíduos com gratuidade em 2014 realizam em média uma validação por dia, o que pode indicar a realização de viagens por outros motivos, como viagens para consultas, compras, entre outros. Porém, esse padrão se altera um pouco em 2018, onde essa categoria passa na maior proporção a realizar duas validações por dia, o que mostra que esses usuários podem ter se tornado mais frequentes, levando a impactos na tarifa, por exemplo. Ademais, cerca de 20% dos usuários dos diferentes tipos de cartão realizam 3 ou 4 validações. Esse público pode ser o que realiza viagens encadeadas ao longo dia.

Para os usuários que utilizam Vale Transporte, nota-se que a distribuição de frequência não se altera de forma significativa entre os anos e que aproximadamente metade dos usuários realiza cerca de 2 validações por dia, assim como os estudantes, indicando, portanto, um forte caráter pendular nas viagens.

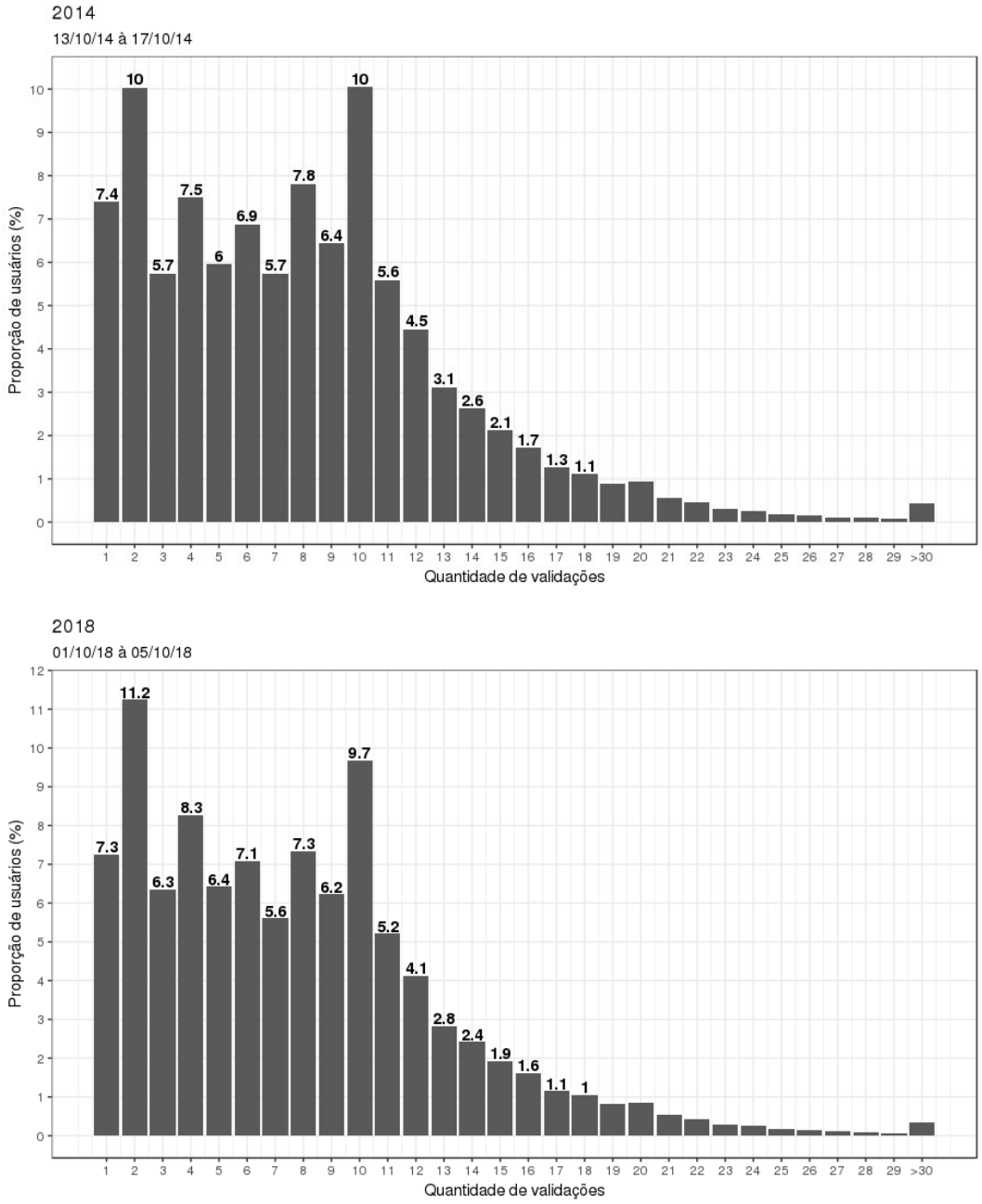
Figura 12 – Frequência diária média dos usuários



Fonte: Elaborado pelo autor

Também se calculou a distribuição de frequência semanal dos usuários, com o intuito de observar de forma mais clara o uso esporádico. Na Figura 13 está a representada distribuição de frequência semanal para os anos de 2014 e 2018, relativos a uma semana típica do mês de outubro. A principal diferença entre a distribuição de frequência diária é que existe uma porcentagem significativa de usuários, cerca de 18%, que valida uma ou duas vezes na semana. Um fato interessante é que as proporções de usuários que realizam até 6 validações aumentaram em 2018. Dessa forma, tem-se evidências para acreditar que os usuários estejam se tornando mais esporádicos.

Figura 13 – Frequência semanal média dos usuários

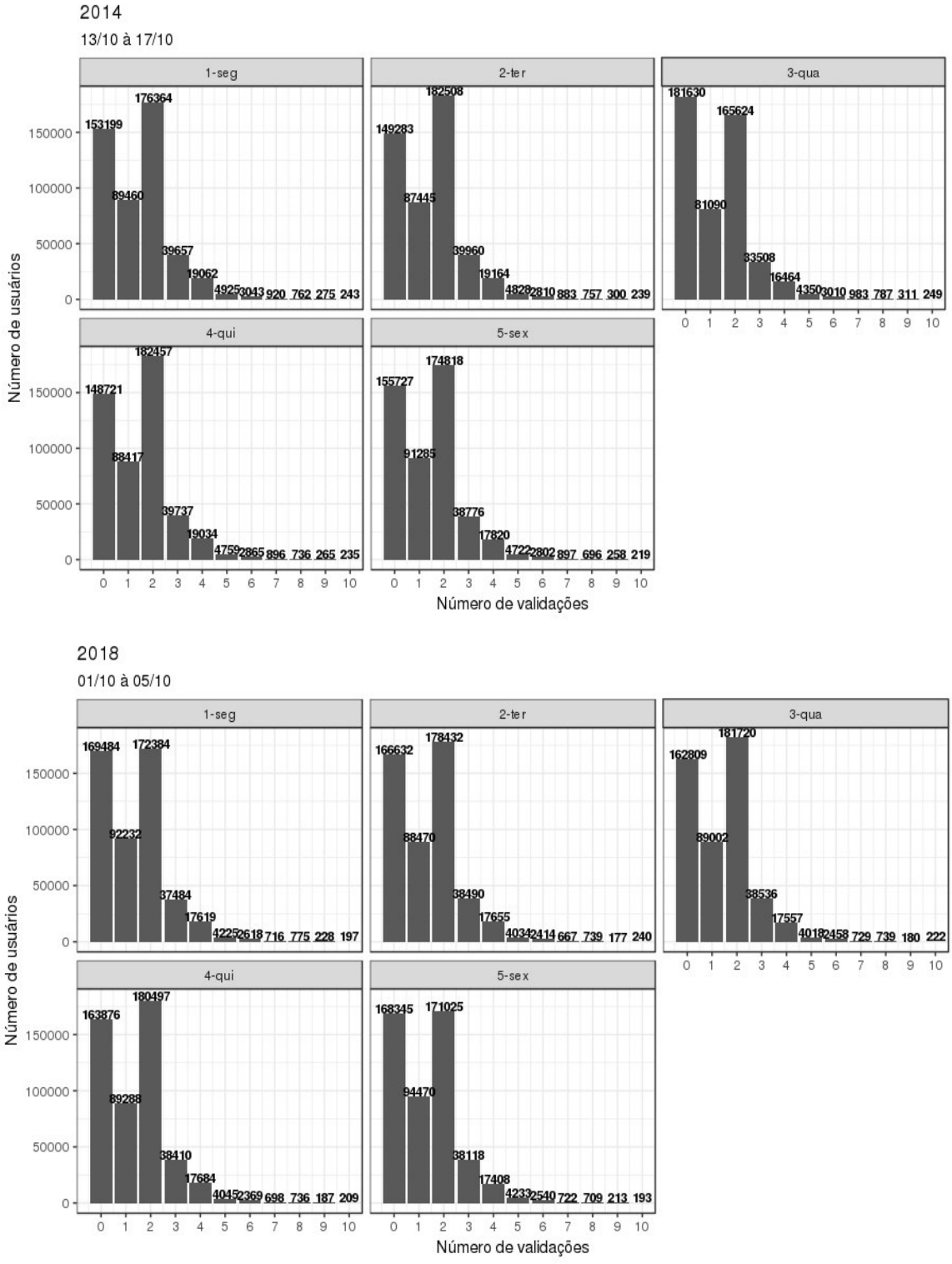


Fonte: Elaborado pelo autor.

Com o intuito de verificar como se dá a intensidade do uso de cada usuário ao longo da semana, utilizando a mesma semana típica de 2014 e 2018, todos os usuários foram rastreados e calculou-se o número de validações do usuário no dia, como representado na Figura 14. O número de validações zero significa que, dentre todos os usuários que utilizaram nessa semana, uma quantidade não utilizou nesse dia específico. A distribuição nesse caso fica muito semelhante à distribuição diária de frequência, no entanto é possível perceber se existem dias em que é mais comum não se usar o transporte coletivo. Para o ano de 2014, os usuários que deixam de usar em algum dia da semana giram em torno de 150 mil pessoas,

exceto para quarta-feira, onde esse número chega a cerca de 180 mil pessoas, o que pode ter sido ocasionado somente nessa semana em específico. Para 2018, esse número cresce e passa para cerca de 170 mil pessoas por dia que deixam de utilizar em algum dia da semana, não havendo aparentemente um dia em específico em que os usuários deixam de utilizar mais.

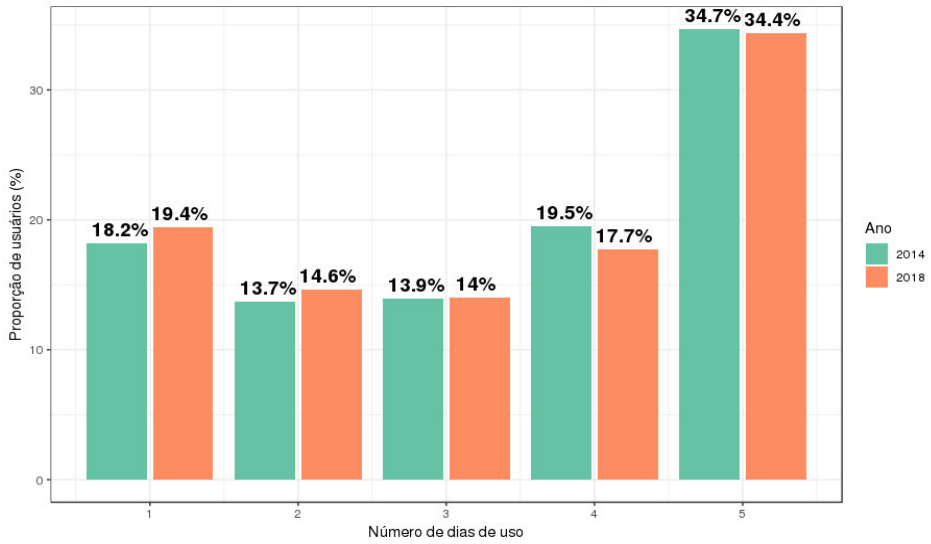
Figura 14 – Utilização ao longo da semana para os anos de 2014 e 2018



Fonte: Elaborado pelo autor.

Por fim, é importante verificar também o número de dias de uso dos usuários dentro da semana típica, para servir de informação complementar junto à informação de validação. Nesse sentido, na Figura 15 são representados a proporção de usuários, bem como os dias de uso desses nas duas semanas típicas analisadas anteriormente de 2014 e 2018. É possível perceber uma tendência para o ano de 2018, dos usuários utilizarem cada vez menos o transporte coletivo dentro da semana.

Figura 15 – Média do número de dias de uso para os anos de 2014 e 2018



Fonte: Elaborado pelo autor.

4.2.3 Distância temporal entre validações

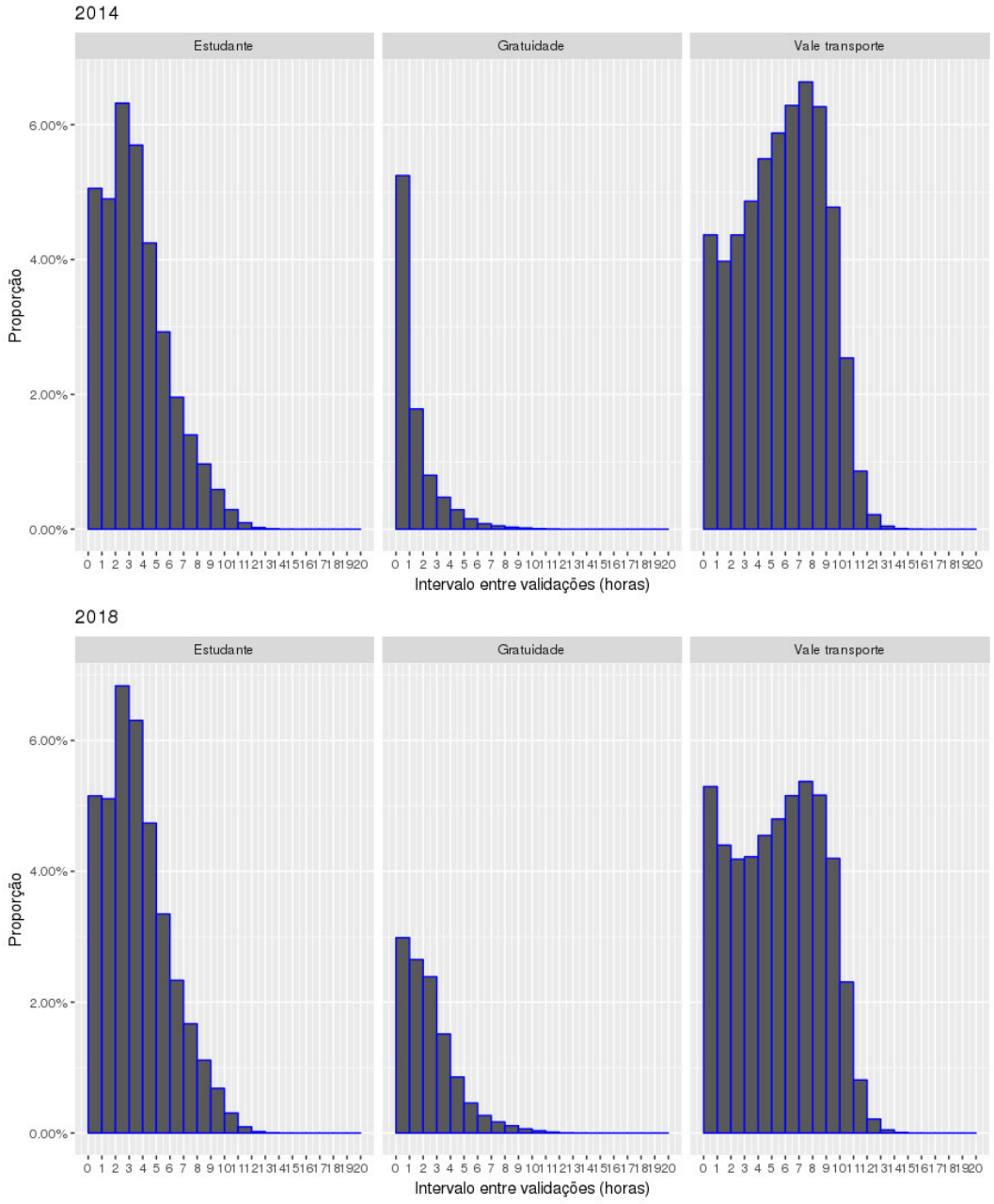
Por meio do cálculo do intervalo temporal entre validações, é possível analisar uma estimativa da duração da atividade dos usuários. Nesse sentido, na Figura 16 é mostrado um histograma com o tempo entre validações para cada tipo de cartão e para 2014 e 2018. Verifica-se que esse indicador em 2014 para os estudantes está entre 3 e 5 horas na maior porção da amostra, o que pode ter relação com a duração de um turno escolar comum. A distribuição para 2014 é bem semelhante quando comparada à 2018.

Para os indivíduos com gratuidade, a maior proporção entre 0 e 1 indica que estes em 2014 realizam somente uma validação (quando isso acontece, considera-se distância temporal entre validações igual a zero). Já em 2018, há um deslocamento da distribuição para a direita e um aumento da dispersão, indicando que esses usuários estão realizando alguma atividade com duração maior.

Os usuários de vale transporte concentram a duração das atividades entre 7 e 9 horas,

indicando que possivelmente eles possuem turnos maiores de trabalho. Esse padrão para esses usuários se altera em 2018, onde essa proporção de atividades entre 7 e 9 horas diminui, indicando talvez que esses usuários deixaram de realizar viagens com essa duração. Também para os usuários de vale transporte, há uma proporção significativa que realiza somente uma validação por dia. Acredita-se que isso pode ser explicado pela inclusão das pessoas que têm bilhete único na classe de vale transporte, pois essa não necessariamente tem alguma atividade fixa de trabalho.

Figura 16 – Intervalo temporal entre validações para os anos de 2014 e 2018

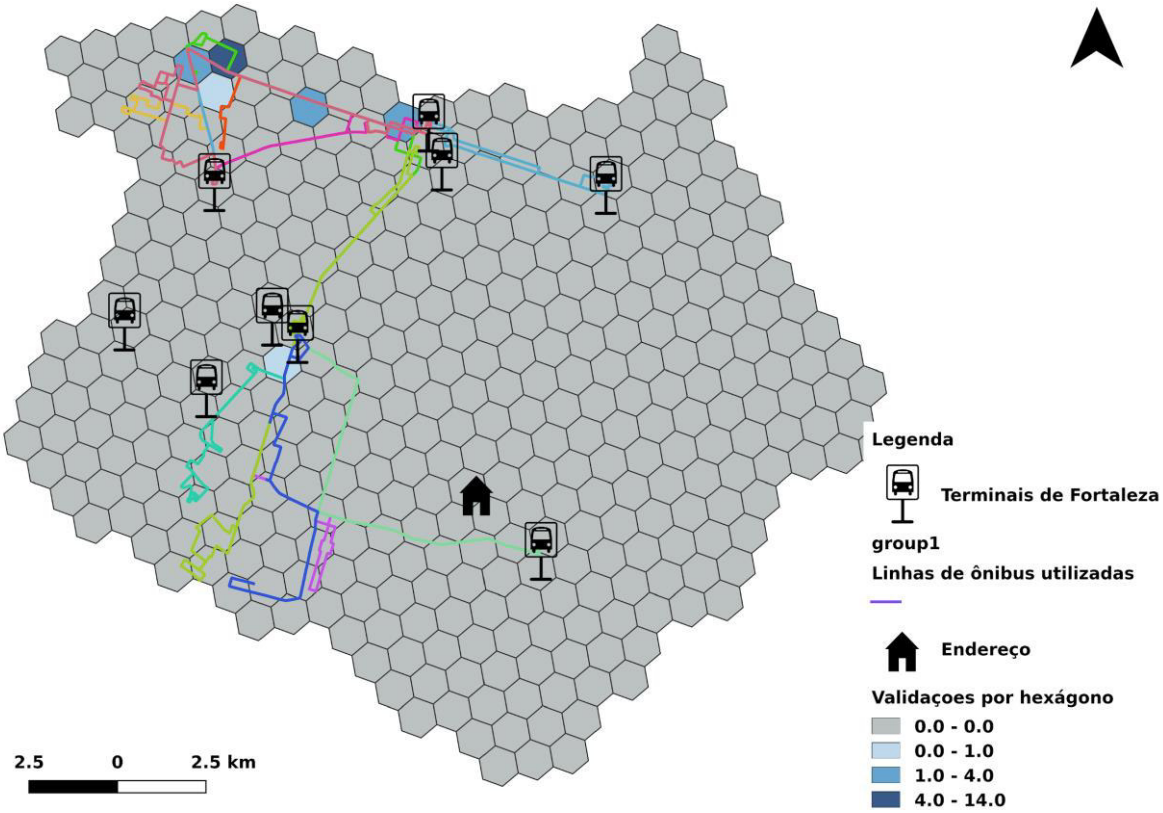


Fonte: Elaborada pelo autor.

4.3 Análise descritiva espacial

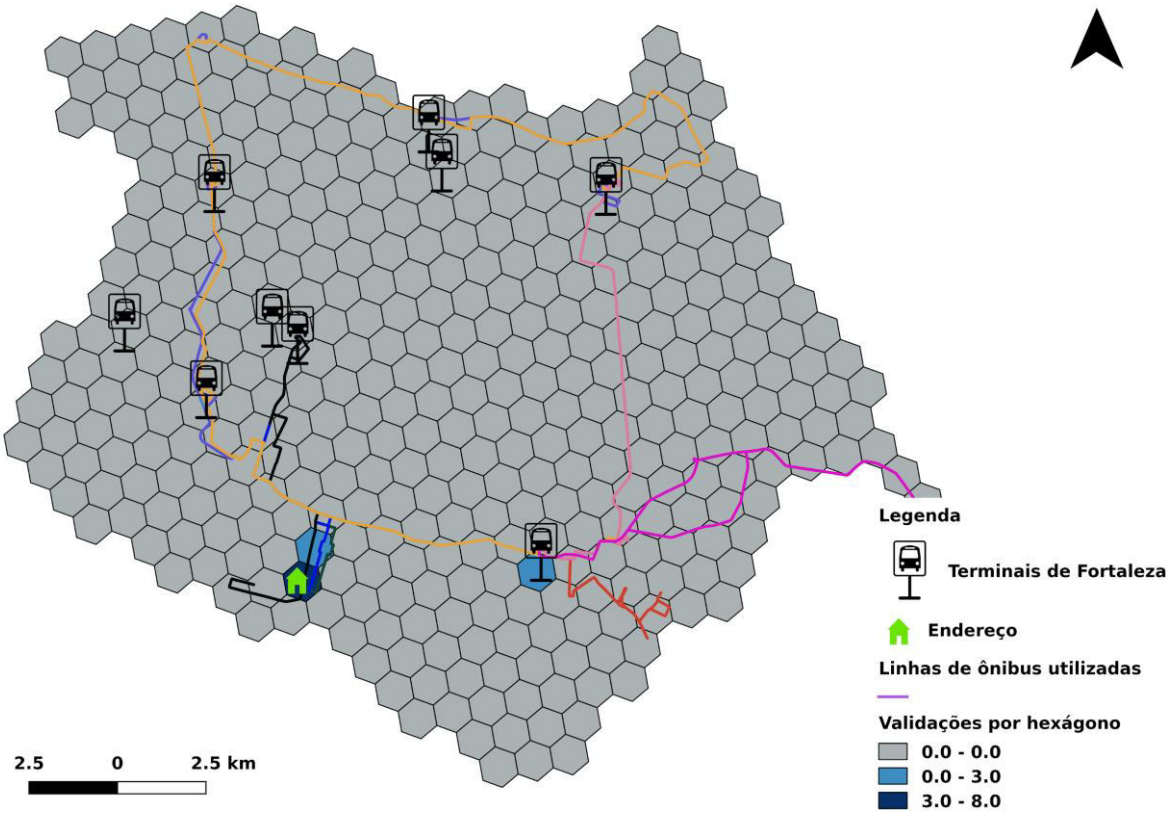
Com o intuito de demonstrar o método empregado para avaliar se é possível considerar que as primeiras validações são próximas ao embarque dos usuários, nas Figura 17 Figura 18 são mostrados dois usuários que apresentam comportamentos distintos de validação. O primeiro é o usuário que comumente não valida próximo a residência e provavelmente mais próximo ao seu destino. Esse usuário geralmente passa por terminais e valida mais frequentemente próximo ao terminal da Parangaba e na região central e noroeste da cidade. Pode-se perceber também que ele usa linhas geralmente trocais, ou seja, que vão dos terminais para as áreas centrais. O segundo usuário (Figura 18), por sua vez, costuma validar a primeira vez no dia próximo a sua residência. As linhas utilizadas por ele são mais circulares e ele também costuma validar próximo ao terminal de Messejana.

Figura 17 – Exemplo de usuário que valida comumente longe da residência



Fonte: Elaborada pelo autor.

Figura 18 – Exemplo de usuário que valida comumente próximo da residência



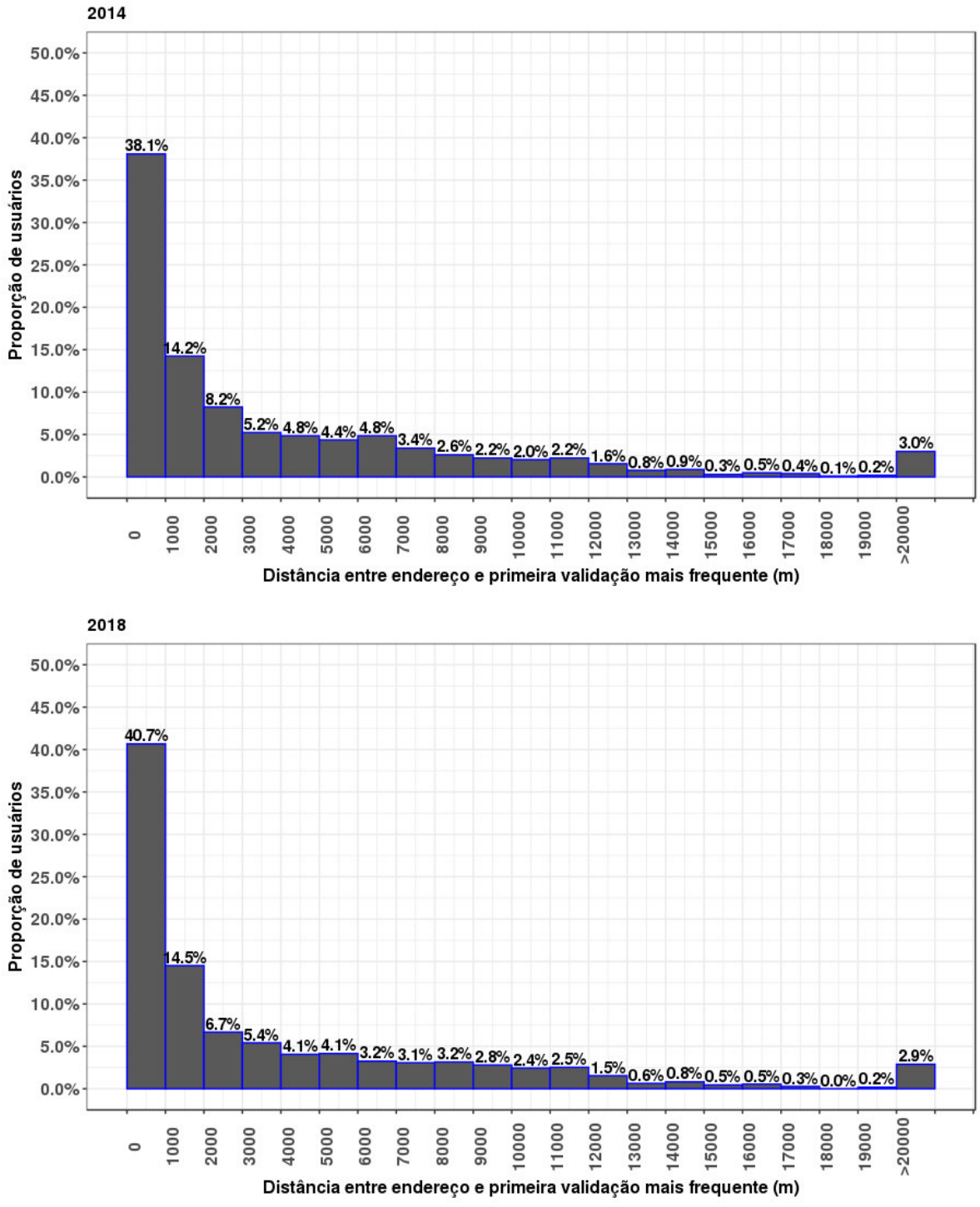
Fonte: Elaborada pelo autor.

A partir do método utilizado, foi possível obter a distribuição da distância euclidiana entre o endereço dos usuários e o polígono em que a primeira validação foi mais frequente para os anos de 2014 e 2018 (Figura 19). Nota-se que aproximadamente metade dos usuários, nos anos de 2014 e 2018, validam a uma distância de até 2000 metros da sua residência. A outra metade por sua vez valida em sua maior porção acima de 2000 metros, possuindo alta dispersão.

Ao comparar os anos de 2014 e 2018, nota-se que há semelhança na distribuição da distância, com base na amostra coletada que, como dito anteriormente, é cerca de 1000 usuários. Dessa forma, há evidências para acreditar que a mudança de localização da catraca em algumas linhas não alterou significativamente o comportamento de validação da população ao embarcar e que não é possível, na maioria dos casos, considerar que a primeira validação pode representar uma produção de viagem.

Além disso, há evidências para acreditar que não é possível considerar que a primeira validação, em cerca de metade da amostra, não representa a zona de embarque ou de endereço dos usuários.

Figura 19 – Distribuição da distância entre endereço e local mais frequente de validações para os anos de 2014 e 2018



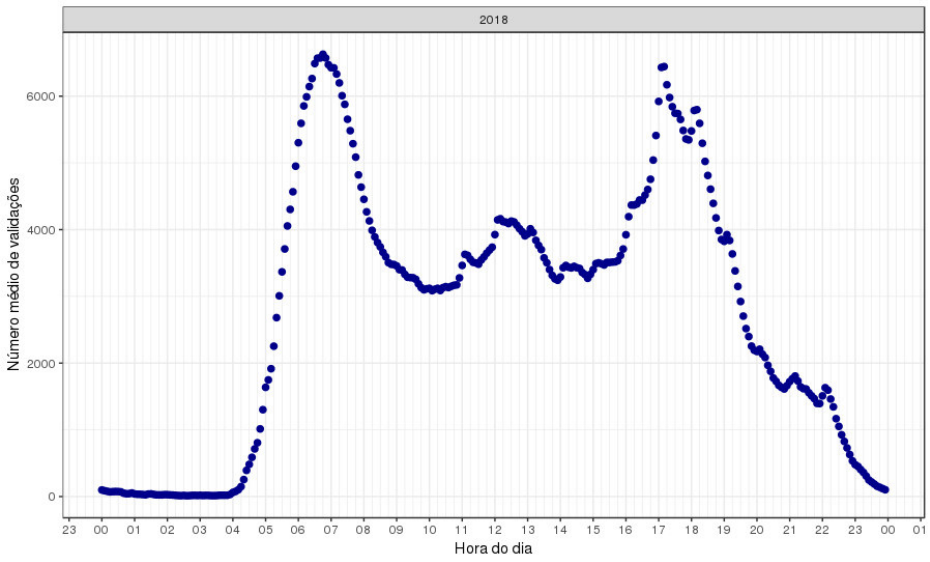
Fonte: Elaborada pelo autor.

4.4 Obtenção dos padrões de deslocamento

4.4.1 Definição dos horários de análise

A partir dos dados coletados, agregou-se as validações em intervalos de 5 minutos, a fim de captar melhor a variabilidade do número de validações ao longo do dia, e construiu-se então o perfil médio horário para o todos os anos de análise. No entanto, verificou-se que o perfil não muda significativamente ao longo dos anos. Na Figura 20 é mostrado perfil horário das validações para o ano de 2018.

Figura 20 – Perfil horário médio das validações para o ano de 2018



Fonte: Elaborada pelo autor

Nota-se que há um período de pico bem definido começando entre 05:00h e 06:00h da manhã, com término praticamente às 09:00h, que possivelmente está representando o começo das atividades na cidade. Posteriormente, há um período fora-pico manhã em que as validações ficam praticamente constantes até às 11:00h, e depois começam a crescer, representando deslocamentos que ocorrem no período de almoço, indo basicamente até às 14:00h. No período entre 14:00h e 16:00h, as validações ficam praticamente estagnadas, voltando a crescer a partir das 16:00h, que dá início ao pico da tarde/noite, indo até aproximadamente às 19:00h, possivelmente representando o deslocamento de volta das atividades. É possível perceber que ainda há algum tipo de atividade significativa até as 23:00h. Dessa forma, os horários definidos para representar os períodos da manhã, tarde e noite no atributo da hora de uso, estão representados na Tabela 3.

Tabela 3 – Períodos de análises definidos

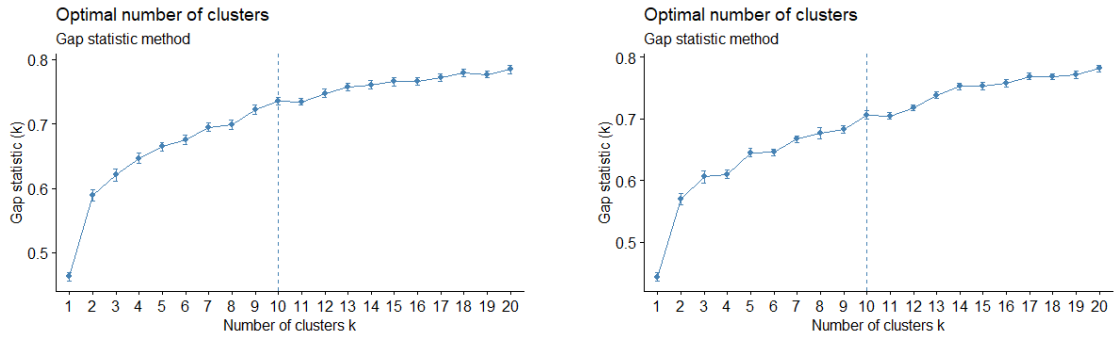
Período	Horário
1-Manhã	05:00h às 11:00h
2- Tarde	11:00h às 16:00h
3- Noite	16:00h às 00:00h

Fonte: Elaborado pelo autor

4.4.2 Extração e discussão sobre os agrupamentos

A partir do critério *gap statistic,* foi possível definir a quantidade ótima de grupos para os anos de 2014 e 2018. Na Figura 21 são mostrados o número ótimo de grupos estimados nos dois anos de análise. Nota-se que o primeiro máximo local é obtido para os dois anos com a quantidade de 10 grupos, que foram então gerados para os anos de análise.

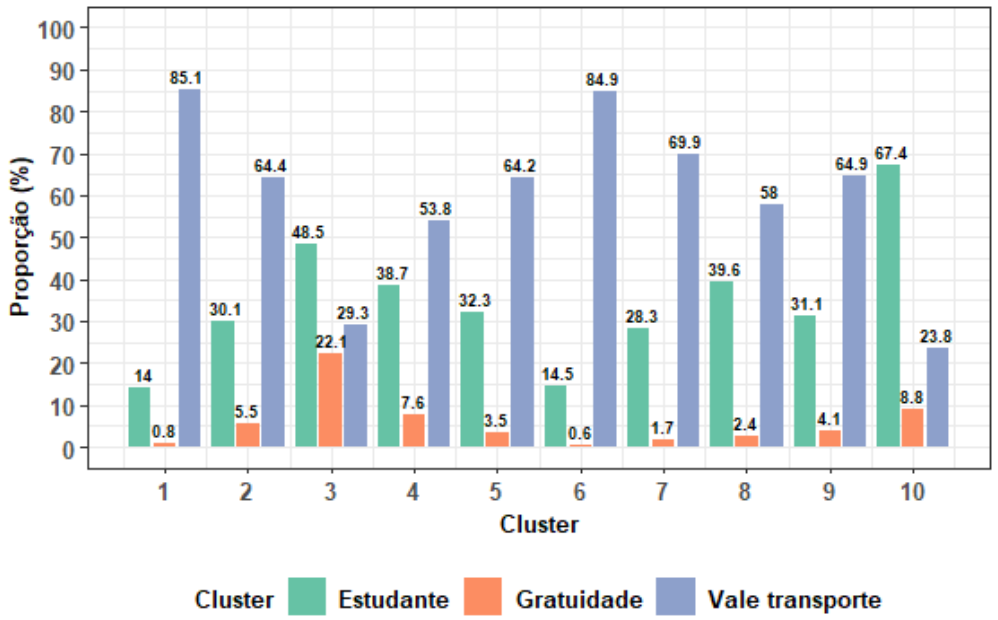
Figura 21 – Número ótimo de *clusters* obtidos para os anos de 2014 e 2018



Fonte: Elaborado pelo Autor.

A participação em forma de porcentagem dos diferentes tipos de cartões em cada agrupamento para o ano de 2014 é mostrada na Figura 22. É possível perceber que na maioria dos grupos (08), há a predominância de cartões do tipo Vale Transporte, com presença significativa em alguns grupos, como o 1 e 6. Os grupos 3 e 10, por sua vez, são compostos predominantemente por estudantes. Nota-se que as gratuidades têm pouca participação na maioria dos grupos, com participação mais acentuada no grupo 3.

Figura 22 – Proporção dos diferentes tipos de cartões nos grupos obtidos para o ano de 2014



Fonte: Elaborado pelo Autor.

A partir de uma análise mais geral, é possível observar as características médias de cada grupo, baseado na mediana como medida de tendência central (Tabela 4). Observa-se que existem basicamente três grupos que são mais assíduos (1, 5 e 7), que compõem aproximadamente 27% dos usuários que utilizam o sistema com cartão, totalizando cerca de 268.300 usuários. A diferença principal entre esses grupos está no intervalo temporal entre embarques, onde o grupo 1 possui intervalo médio com cerca de 10,5 horas, indicando que esse pode ser um grupo que realiza atividade durante todo o período do dia, realizando embarques predominantemente no período da manhã. O grupo 5 já possui o intervalo entre embarques um pouco menor, cerca de 8 horas e realiza sua primeira validação mais frequentemente a tarde, ou seja, pode indicar indivíduos que trabalham no período da tarde a uma carga horária menor, de 6 horas, por exemplo. O grupo 7, por sua vez, realiza embarques mais distribuídos entre os três horários estabelecidos, além de possuir tempo entre embarques um pouco maior, ou seja, esse grupo pode ser composto por pessoas que têm mais flexibilidade de horários em seus empregos.

Além disso, é interessante notar que nos grupos 5 e 7, ditos frequentes, há uma grande quantidade de dias em que os usuários validaram somente no começo/fim da viagem, mais proporcionalmente do que no grupo 1, ou seja, não realizaram uma viagem pendular utilizando transporte coletivo. Isso pode se dar possivelmente pelo horário da volta da viagem ser predominantemente à noite e pode estar ligado com fatores relacionados à segurança pública ou pelo fato de a viagem de ida ou volta poder sido realizada utilizando carona, por

exemplo. Esse comportamento acontece de forma mais significativa nos grupos que realizam mais embarques no período da noite, como o grupo 8, onde em um terço das viagens há somente uma validação, em média.

Os grupos 1 e 6 se assemelham bastante na composição dos cartões, possuindo em sua composição uma grande maioria de usuários com vale transporte. No entanto, a diferença principal entre esses grupos está no número médio de dias típicos de uso, onde o grupo 6 apresenta um uso em média quase 4 vezes menor. Uma hipótese que pode ser levantada é que esse grupo pode ser composto por pessoas que ganharam ou perderam seus empregos ao longo do ano, além de autônomos, já que o bilhete único está incluído na categoria vale transporte. O grupo 4 também apresenta essas características, mas com uso mais distribuído ao longo do dia, o que pode indicar a presença de atividades autônomas. Os grupos 4 e 6 são responsáveis por 25,4% dos usuários que utilizaram cartão nesse ano.

Os grupos 3 e 10, que são grupos onde há maior proporção de estudantes e gratuidades, além do grupo 2, apresentam baixo uso do sistema e, portanto, podem ser considerados usuários esporádicos, representando cerca de 33,7% dos usuários que utilizam cartões, contabilizando um total de 221.250 usuários. A principal mudança entre esses grupos está na distribuição horária de uso ao longo do dia e a hipótese inicial é que esses grupos podem representar estudantes que fazem cursos temporários e pessoas com gratuidade que realizam viagem por outros motivos como saúde, lazer, entre outros.

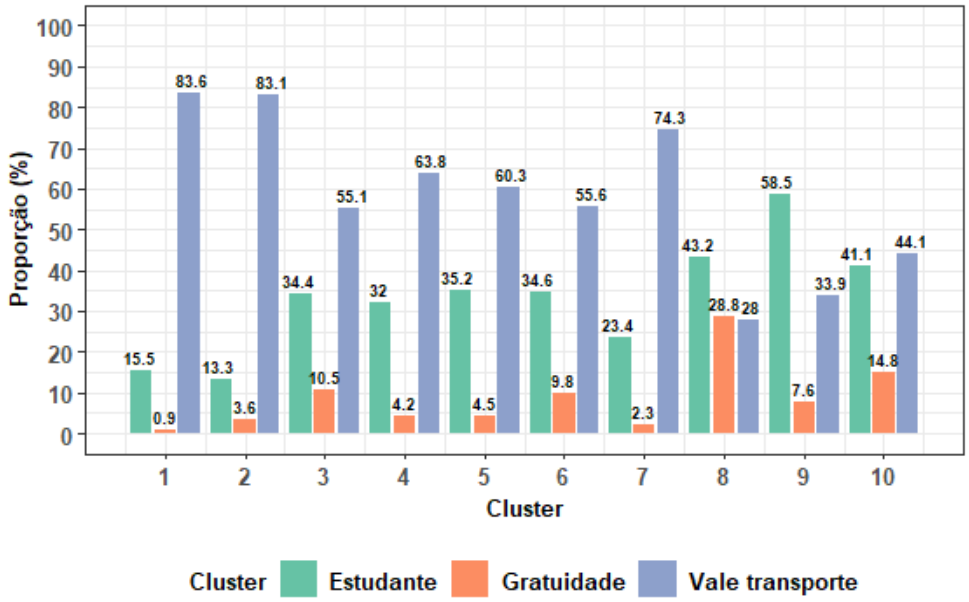
Tabela 4 – Resumo dos atributos obtidos para os grupos do ano de 2014

Cluster	Mediana do intervalo temporal entre embarques (horas)	Nº de dias que ocorreu somente uma validação	Proporção de validações pela manhã	Proporção de validações a tarde	Proporção de validações a noite	Mediana da Frequência (Validações/dia)	Mediana do número de dias típicos de uso	Número de usuários	Porcentagem cluster (%)
1	10,4	31	85,8	6,1	5,9	4,2	193	113.989	11,3
2	8,0	9	55,2	21,4	19,8	2,3	27	119.107	11,8
3	4,4	7	11,1	58,8	25,0	1,3	13	119.475	11,8
4	6,2	19	19,3	40,0	39,3	2,4	45	102.348	10,1
5	7,9	42	14,6	62,1	15,6	3,9	173	59.534	5,9
6	10,3	9	83,1	6,9	7,7	3,6	51	155.117	15,3
7	8,6	47	50,6	23,7	22,8	3,6	143	94.740	9,4
8	6,4	33	7,4	17,4	66,7	3,3	90	65.888	6,5
9	7,3	11	11,1	69,6	13,3	3,2	50	80.220	7,9
10	4,3	8	5,9	25,0	63,6	1,4	14	101.775	10,1
							∑	1.012.193	

Fonte: Elaborado pelo Autor.

Para o ano de 2018, participação em forma de porcentagem dos diferentes tipos de cartões em cada agrupamento é mostrada na Figura 23. Como ocorre no ano de 2014, a maioria dos grupos (07 grupos) têm porcentagem maior em sua composição de cartões do tipo vale transporte, com presença mais acentuada em alguns grupos, como o 1 e 2. Por sua vez, os grupos 8,9 e 10 apresentam composição mais distribuída, com maior porcentagem de estudantes e gratuidades.

Figura 23 – Proporção dos diferentes tipos de cartões nos grupos obtidos para o ano de 2018



Fonte: Elaborado pelo Autor.

As características médias dos atributos de cada agrupamentos do ano de 2018 são apresentadas na Tabela 5. Observa-se que, assim como no ano de 2014, onde havia 3 grupos mais frequentes, para 2018 há 4 grupos que são bastante assíduos no transporte coletivo (2, 4, 6 e 7) representando cerca de 27 % dos utilizadores de cartão, aproximadamente a mesma proporção para o ano de 2014. Dois desses grupos, 2 e 7, apresentam características bastante semelhantes de uso e se assemelham muito com o grupo 1 obtido em 2014. No entanto, os dois parecem ter distribuição de dias típicos de uso ao longo ano mais concentradas, o que talvez tenha dividido o grupo 1 de 2014 nesses dois grupos. A distribuição frequência de todos os atributos para os agrupamentos nos diferentes anos estão mostradas no Apêndice A.

Os grupos 4 e 6, que se assemelham devido a frequência e ao intervalo temporal entre embarques, relativamente menor em relação aos grupos 1 e 7, possuem diferença significativa nos horários de uso ao longo do dia, com o grupo 4 utilizando o sistema de forma mais distribuída ao longo do dia.

Assim como acontece com os grupos em 2014, existem dois grupos que são majoritariamente compostos por usuários com Vale Transporte (grupos 1 e 2) e apresentam várias características semelhantes de uso, o que se assemelha com o que foi verificado entre os grupos 1 e 6 de 2014. Nesse sentido, pode-se afirmar que há um grupo de indivíduos com padrão de uso bem definidos, no entanto esse uso do sistema pode ser realizado de forma estratificada ou concentrada ao longo do ano.

Os grupos 8, 9 e 10, onde há maior proporção de estudantes e gratuidades, nota-se um comportamento semelhante aos grupos com as mesmas características de 2014. Ademais, a diferença entre esses grupos de 2018 é basicamente a distribuição horária de uso, além de uma leve diferença do intervalo temporal entre embarques.

Os grupos 1, 3 e 5 têm características de uso intermediárias entre os mais assíduos e os menos frequentes, com dias típicos de uso variando entre 40 e 100 dias. A diferença entre esses grupos está basicamente no horário da primeira validação, onde o grupo 1 valida de forma mais concentrada mais no horário da manhã junto com o grupo 5, e o grupo 3 no horário da tarde.

Tabela 5 – Resumo dos atributos obtidos para os grupos do ano de 2018

Cluster	Mediana do tempo entre embarques (horas)	Nº de dias que ocorreu somente uma validação	Proporção de validações pela manhã	Proporção de validações a tarde	Proporção de validações a noite	Mediana da Frequência (Validações/dia)	Mediana do número de dias típicos de uso	Número de usuários	Porcentagem cluster (%)
1	10,4	8	85,1	5,6	6,9	3,7	44	87.934	9,6
2	10,5	33	89,2	4,6	4,4	4,4	219	55.281	6,1
3	7,0	18	14,3	61,5	17,8	3,0	66	105.411	11,5
4	8,7	66	49,3	20,5	27,4	3,9	175	58.026	6,4
5	8,7	26	63,7	17,2	15,6	3,0	94	81870	9,0
6	7,6	47	13,0	56,6	20,5	3,9	171	77151	8,4
7	10,2	24	87,5	5,5	5,3	4,0	157	52712	5,8
8	4,3	8	10,5	57,8	27,9	1,4	18	142421	15,6
9	5,1	17	7,4	22,9	62,5	2,0	36	161453	17,7
10	6,5	9	50,0	26,7	21,7	1,9	24	92122	10,1
							∑	914.381	

Fonte: Elaborado pelo Autor.

Por fim, realizou-se uma análise no intuito de comparar os grupos dos anos de 2014 e 2018. Os grupos mais semelhantes de acordo com as medidas de tendências centrais dos atributos, assim como a variação percentual dos grupos e a variação de usuários experimentada são apresentadas na Tabela 6.

É possível notar de forma expedita que há o aumento acentuada de alguns grupos com uso esporádico, como o par de grupos 3/8, que são aqueles grupos que não possuem distribuição de cartões discrepantes e que usam o transporte coletivo a tarde, ou seja, há evidências para acreditar que o uso esporádico aumentou durante o período da tarde. Além disso, esses grupos possuem tempo entre validações curtos, o que indica que a atividade realizada tem período pequeno, se assemelhando com viagens motivo educação e outros motivos.

Por outro lado, os pares de grupos 5/6 e 7/4 mostram uma queda no número de usuários de grupos que eram considerados mais frequentes, onde o par 5/6 tem uso mais concentrado no período na tarde e tempo entre validações que giram em torno de 8 horas, enquanto que o par 7/4 tem maior tempo entre validações e possui uso mais disperso durante o dia. Esse efeito pode ter sido causado pelo desemprego ou pela escolha de outros modos de transporte, como a motocicleta, por exemplo.

Os pares de grupos que apresentam tempo de uso do sistema intermediário (6/1 e 9/3), apresentam respectivamente diminuição e aumento na quantidade de usuários. A diferença entre esses pares está basicamente nos horários do dia e no tempo entre validações, onde o par 6/1 trata de grupos que usam predominantemente no horário da manhã, com mais de 80% das primeiras validações nesse horário e tempo entre validações que giram em torno de 10 horas. Já o par 9/3 utiliza o sistema preferencialmente pela tarde, com tempo entre validações girando em torno de 7 horas.

Tabela 6 – Semelhança entre os grupos obtidos em 2014 e 2018

2014	Proporção do grupo	2018	Proporção do grupo	Diferença entre grupos (%)	Diferença de usuários
1	11,85%	2	9,62%	-2,23%	-31173
10	6,55%	8	10,07%	3,52%	26234
2	5,92%	5	6,05%	0,12%	-4253
3	6,55%	8	11,53%	4,98%	39523
4	10,12%	10	6,35%	-3,78%	-43749
5	15,43%	6	8,95%	-6,47%	-73247
6	11,34%	1	8,44%	-2,90%	-36838
7	10,18%	4	5,76%	-4,41%	-49636
8	10,18%	4	15,58%	5,40%	40073
9	11,88%	3	17,66%	5,77%	41978

Fonte: Elaborado pelo Autor.

5 CONCLUSÕES E RECOMENDAÇÔES

Conforme estabelecido como primeiro objetivo específico deste trabalho, foi possível realizar uma análise descritiva com ênfase em aspectos temporais dos dados de validações. Nesse sentido, concluiu-se que a demanda de transporte coletivo por ônibus vem sofrendo queda na cidade de Fortaleza, principalmente nas categorias inteira e vale transporte, corroborando em uma queda de aproximadamente 110.000 mil validações em inteira e a perda de 30.000 usuários com vale transporte. As categorias estudante e gratuidade não apresentaram oscilação acentuada no período de análise. No entanto, a categoria gratuidade ganhou uma quantidade significativa de usuários entre os anos de 2017 e 2018, cerca de 14.000. É interessante ressaltar que a perda de usuários com categoria inteira é preocupante pois essa categoria impacta fortemente a tarifa do sistema. A queda da categoria vale transporte pode estar relacionada com a taxa de desemprego atual, além de outros fatores como outros modos de transporte.

Além disso, notou-se que a quantidade de usuários não varia de forma significativa entre os meses, exceto para a categoria de estudantes, onde é possível perceber de forma mais acentuada a sazonalidade devido ao período de férias escolares. Entre os dias da semana, não há diferença do número de usuários nos dias típicos (segunda-feira à sexta-feira). No entanto, a quantidade de usuários cai nos finais de semana, devido ao recesso das atividades de trabalho e educação. No que diz respeito à frequência dos usuários, nota-se que boa parte deles utiliza o transporte coletivo utilizando duas validações e, portanto, apresentam características pendulares. Ao observar a frequência semanal dos usuários, percebe-se que há uma significativa proporção de usuários que realiza somente uma validação por dia, o que indica um forte caráter esporádico de alguns usuários. Quando se analisou o tempo entre validações, notou-se que este é uma boa estimativa do tempo da atividade dos usuários, com esse tempo em sua maior parte variando entre 4 e 6 horas para os estudantes (indicando possivelmente um tempo relativo ao turno escolar) e 9 e 11 horas para usuários com vale transporte (indicando possivelmente um período de trabalho e 6/8 horas).

No que diz respeito a consideração sobre o dado de validação em si ser suficientemente bom para estimar embarques, percebeu-se que, em aproximadamente metade dos casos amostrados e com o método utilizado, os usuários costumam validar seu embarque a mais de 2 quilômetros de distância de sua residência, o que indica que usar somente os dados de validações para estimar embarques pode fornecer vieses em análises de padrão de deslocamento espaciais e estimações de embarques utilizando somente o dado de validação.

Também se verificou que não há aparentemente um efeito acentuado no embarque relacionado a mudança de posição das catracas dos ônibus para a porção dianteira do veículo. As principais limitações desse método na verificação desse objetivo foram o tamanho da amostra utilizada (cerca de 1.000 usuários), o georreferenciamento utilizado pelo *google*, que pode estimar endereços incorretos apesar de possuir boa acurácia e a limitação dos dados de cadastro de usuários, que possui alguns erros de nomenclatura de bairros, ruas e cidades.

Em relação à detecção de agrupamentos, o método apresentou resultados satisfatórios, o que permitiu identificar grupos diferentes com diversas particularidades. No entanto, a utilização de muitos atributos, dificulta por vezes a interpretação dos agrupamentos obtidos e a inserção de atributos como o período do dia de utilização do transporte coletivo aumenta significativamente a quantidade de grupos estabelecidos. O atributo relacionado à variabilidade da frequência também não ofereceu um caráter explicativo forte aos grupos, diferentemente de alguns atributos como tempo entre embarques e número de dias típicos de utilização do transporte coletivo, por exemplo. Vale ressaltar que não foi incorporado nessa análise nenhum atributo de caráter socioeconômico dos usuários, como idade, gênero, renda e estes podem oferecer maior poder explicativo à dinâmica de deslocamentos dos diferentes grupos.

Vale ressaltar que nesse método todos os atributos considerados tiveram o mesmo peso na definição dos grupos e não foi realizada nenhuma análise de correlação entre os atributos. Essa análise pode ser importante no sentido de que pode haver atributos altamente correlacionados e que, portanto, contribuirão de forma semelhante na definição dos grupos. A medida de similaridade utilizada (Euclidiana) é a mais simples e existem outras medidas que podem incorporar o efeito dinâmico e temporal dos atributos, como a distância *Mahalanobis*. Outro fato a ser discutido é a utilização do método *k-means*, que pode ocasionar viés em relação à definição aleatória do centroide, como visto anteriormente, o que pode impactar nas medidas de compacidade e separabilidade dos grupos.

Nos agrupamentos obtidos, para os anos de análise foi possível perceber a presença de grupos mais frequentes e grupos mais esporádicos, além de grupos que apresentaram o uso mais intermediário ao longo do ano. Para o ano de 2014, notou-se que, dois 10 agrupamentos, grande parte possui na sua composição majoritariamente o cartão vale transporte. Em relação ao comportamento cativo, três possuíam comportamento bastante frequente e estes representam cerca de 27% dos usuários que utilizaram cartão naquele ano. A principal diferença nesses três grupos frequentes está basicamente no horário de utilização ao longo do dia e no tempo entre validações de cada um dos grupos, sendo o maior grupo

dito mais frequente (114.000 usuários) aquele que possui em média 10,4 horas entre validações, além de realizar estes aproximadamente 86% das vezes no período da manhã.

Além disso, houve também a identificação de grupos menos frequentes, com número de dias típicos de uso no ano variando entre 50 e 100 dias, além da variação dos usos entre os horários do dia. Nesse caso, foi levantada a hipótese de que esses grupos podem ser compostos por indivíduos que ganharam ou perderam seus empregos ao longo do ano ou até por indivíduos que realizam atividades como autônomos. Este último acontecimento é possível já que dentro da categoria vale transporte, estão englobados os usuários que possuem bilhete único. Os grupos que possuem comportamento mais esporádicos são aqueles compostos mais por estudantes e gratuidades, que apresentam intervalo entre validações entre 4 e 5 horas, além de usarem em média de 15 a 30 dias típicos no ano. Como limitação, não foi observada a distribuição temporal e espacial dos grupos, além da composição dos grupos na demanda diária, o que pode oferecer informações mais acuradas sobre a demanda diária.

No ano de 2018, foram detectados grupos com basicamente as mesmas características, no entanto com tamanhos diferentes. Foi possível notar, com a comparação expedita dos grupos entre os anos de 2014 e 2018, uma tendência de aumento dos grupos mais esporádicos e dque possuem uso intermediário ao longo do ano. Também há evidências para acreditar na diminuição dos grupos mais cativos. A limitação do método utilizado nesta etapa está na comparação entre os grupos ter sido realizada baseada somente na medida de tendência central dos atributos. Ademais, a tendência de diminuição dos grupos mais cativos e do aumento dos grupos mais esporádicos pode ter diferentes causas, como o nível de desemprego ou escolha de outros modos e carece de uma investigação mais profunda.

Diante do que foi exposto e do atual estado de desenvolvimento do estudo, foram estabelecidos como sugestão para trabalhos futuros os seguintes temas:

a) Estimação do embarque dos usuários considerando outros bancos de dados além do banco de validações;
b) Definição de um método de agrupamento para detecção de padrões de deslocamento que possa melhorar a capacidade de compacidade e separabilidade dos grupos;
c) Incorporação de características socioeconômicas da detecção dos grupos;
d) Análise da dinâmica espacial e temporal dos agrupamentos obtidos;
e) Verificar quais são os fatores que impactam na mudança de característica da demanda observada.

REFERÊNCIAS BIBLIOGRÁFICAS

AGARD, Bruno; MORENCY, Catherine; TRÉPANIER, Martin. **Mining Public Transport User Behaviour From Smart Card Data**. *IFAC Proceedings Volumes*, v. 39, n. 3, p. 399–404, 2006. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1474667015359310>.

AGARD, Bruno; TRÉPANIER, Martin. **Assessing Public Transport Travel Behaviour From Smart Card Data With Advanced Data Mining**. *13th WCTR*, p. 1–13, 2013. Disponível em: <http://www.mgi.polymtl.ca/agard/fr/publications/doc/2013/2013_WCTR-Vahid.pdf>.

ARTHUR, David; VASSILVITSKII, Sergei. **K-means++: The advantages of careful seeding**. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, v. 07-09-Janu, p. 1027–1035, 2007.

ASSOCIAÇÃO NACIONAL DAS EMPRESAS DE TRANSPORTES URBANOS. **Anuário NTU 2017-2018**. p. 76, 2018.

BAGCHI, M.; WHITE, P. R. **The potential of public transport smart card data**. *Transport Policy*, v. 12, n. 5, p. 464–474, 2005.

BAGCHI, Mousumi; WHITE, Peter R. **What role for smart-card data from bus systems?** *Proceedings of the Institution of Civil Engineers: Municipal Engineer*, v. 157, n. 1, p. 39–46, 2004.

BOARETO, Renato. **Mobilidade Urbana Sustentável**. *Revista dos Transportes Públicos - ANTP*, 2003.

BRAGA, Carlos Kauê Vieira. **Big data de transporte público na análise da variabilidade de indicadores da acessibilidade às oportunidades de trabalho e educação.** 2019. 108 f. Dissertação (Mestrado) - Curso de Engenharia Civil, Universidade Federal do Ceará, Fortaleza, 2019.

BRIAND, Anne Sarah *et al.* **A mixture model clustering approach for temporal passenger pattern characterization in public transport**. *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*, v. 1, n. 1, p. 37–50, 2015.

DEAKIN, ElizabethKim; KIM, Songju. ***Transportation Technologies: Implications for Planning***. . UC Berkeley: [s.n.], 2001. Disponível em: <https://escholarship.org/uc/item/9gt0f9d2>.

ESTER, Martin *et al.* **A Desity-Based Algorithm for Discovering Clusters**. *Data Mining and Knowledge Discovery*, v. 2, n. 2, p. 169–194, 1998. Disponível em: <http://link.springer.com/10.1023/A:1009745219419>.

FAHAD, Adil *et al.* **A survey of clustering algorithms for big data: Taxonomy and empirical analysis**. *IEEE Transactions on Emerging Topics in Computing*, v. 2, n. 3, p. 267–279, 2014.

GARCIA, Camila Soares Henrique Fontenele *et al.* **Strategic Assessment of Lisbon's Accessibility and Mobility Problems from an Equity Perspective**. *Networks and Spatial Economics*, v. 18, n. 2, p. 415–439, 2018.

IPEA. Transporte Urbano e Inclusão Social : **Elementos para Políticas Públicas.Acção para a Expanção do Metro-Ferroviário nas Regiões Metropolitanas**. *Instituto de pesquisa econômica aplicada*, p. 1–26, 2003.
JAIN, A, K; MURTY, M, P; FLYNN, P, J. Data clustering: a review. *ACM Computing Surveys*, v. 31, 1999.

JAIN, Anil K; DUBES, Richard C. **Algorithms for clustering data**. [S.l: s.n.], 1988. Disponível em: <https://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf>.

JANG, Wonjae. **Travel time and transfer analysis using transit smart card data**. *Transportation Research Record*, n. 2144, p. 142–149, 2010.

KIEU, Le Minh *et al.* **Passenger Segmentation Using Smart Card Data**. n. December 2013, p. 1–12, 2014.

LONG, Ying; THILL, Jean Claude. **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.

MA, Xiaolei *et al.* **Mining smart card data for transit riders' travel patterns**. *Transportation Research Part C: Emerging Technologies*, v. 36, n. 2013, p. 1–12, 2013. Disponível em: <http://dx.doi.org/10.1016/j.trc.2013.07.010>.

MAHRSI, Mohamed K El *et al.* **Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data.** *The 3rd International Workshop on Urban Computing (UrbComp 2014)*, 2014.

MORENCY, Catherine; TREPANIER, Martin; AGARD, Bruno. **Analysing the variability of transit users behaviour with smart card data**. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, p. 44–49, 2006.

MUNIZAGA, Marcela A.; PALMA, Carolina. **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, n. October 2012, p. 9–18, 2012.

NORDESTE, D. do. **Após 22 anos, Fortaleza terá novo Plano de Acessibilidade**. 2018. Disponível em: <http://diariodonordeste.verdesmares.com.br/editorias/metro/apos-22-anosfortaleza-tera-novo-plano-de-acessibilidade-1.1962880>. Acesso em: 2019-08-30.

PELLETIER, Marie Pier; TRÉPANIER, Martin; MORENCY, Catherine. **Smart card data use in public transit: A literature review**. *Transportation Research Part C: Emerging Technologies*, v. 19, n. 4, p. 557–568, 2011. Disponível em: <http://dx.doi.org/10.1016/j.trc.2010.12.003>.

PIERONI, Caio de Borthole Valente. ***Analysis of travel patterns from precarious settlements transit users in São Paulo through smart card data mining*.** 2018. 121 f. Universidade de São Paulo, 2018.

SPOSATI, Aldaíza. **Exclusão social abaixo da linha do Equador**. *Seminário Exclusão Social*, p. 1–9, 1998. Disponível em: <http://www.seuvizinhoestrangeiro.ufba.br/twiki/pub/GEC/RefID/exclusao.pdf>.

TIBSHIRANI, Robert; WALTHER, Guenther; HASTIE, Trevor. ***Estimating the number of data clusters via the gap statistic***. *Journal of the Royal Statistical Society: Series B*. [S.l: s.n.]. , 2001

TRÉPANIER, Martin; MORENCY, Catherine; AGARD, Bruno. **Calculation of Transit Performance Measures Using Smartcard Data**. *Journal of Public Transportation*, v. 12, n. 1, p. 79–96, 2009.

TRÉPANIER, Martin; TRANCHANT, Nicolas; CHAPLEAU, Robert. **Individual trip destination estimation in a transit smart card automated fare collection system**. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, v. 11, n. 1, p. 1–14, 2007.

UTSUNOMIYA, Mariko; ATTANUCCI, John; WILSON, Nigel. **Potential uses of transit smart card registration and transaction data to improve transit planning**. *Transportation Research Record*, n. 1971, p. 119–126, 2006.

WHITE, P., 2010. **The role of smartcard data in public transport**. In: Presented at the 12th World Conference on Transport Research, Lisbon, Paper No. 1461.

WILSON, Nigel H.M.; ZHAO, Jinhua; RAHBEE, Adam. **The potential impact of automated data collection systems on urban public transport planning**. *Operations Research/ Computer Science Interfaces Series*, v. 46, n. 1, p. 75–99, 2009.

WITTEN, Ian H; FRANK, Eibe; HALL, Mark a. ***Data Mining: Practical Machine Learning Tools and Techniques***. 2. ed. [S.l: s.n.], 2005.

XU, Rui; WUNSCH, Donald. **Review of Clustering Algorithms**. v. 16, n. 3, p. 7–28, 2005.

ZAKI, Mohammed J.; MEIRA, JR, Wagner. ***Data Mining and Analysis***. Cambridge University Press: [s.n.], 2014.

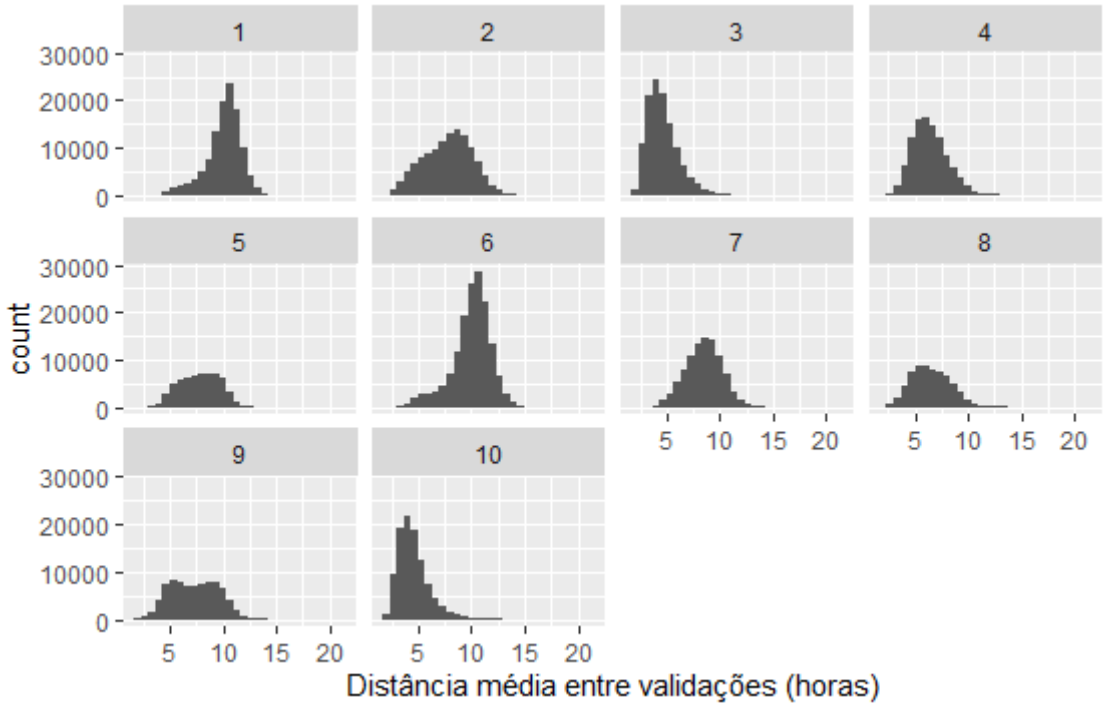
ZHAO, Juanjuan *et al.* **Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data**. *IEEE Transactions on Intelligent Transportation Systems*, v. 18, n. 11, p. 3135–3146, 2017.

ZHOU, Jiangping; MURPHY, Enda; LONG, Ying. **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. Disponível em: <http://dx.doi.org/10.1016/j.jtrangeo.2014.09.006>.

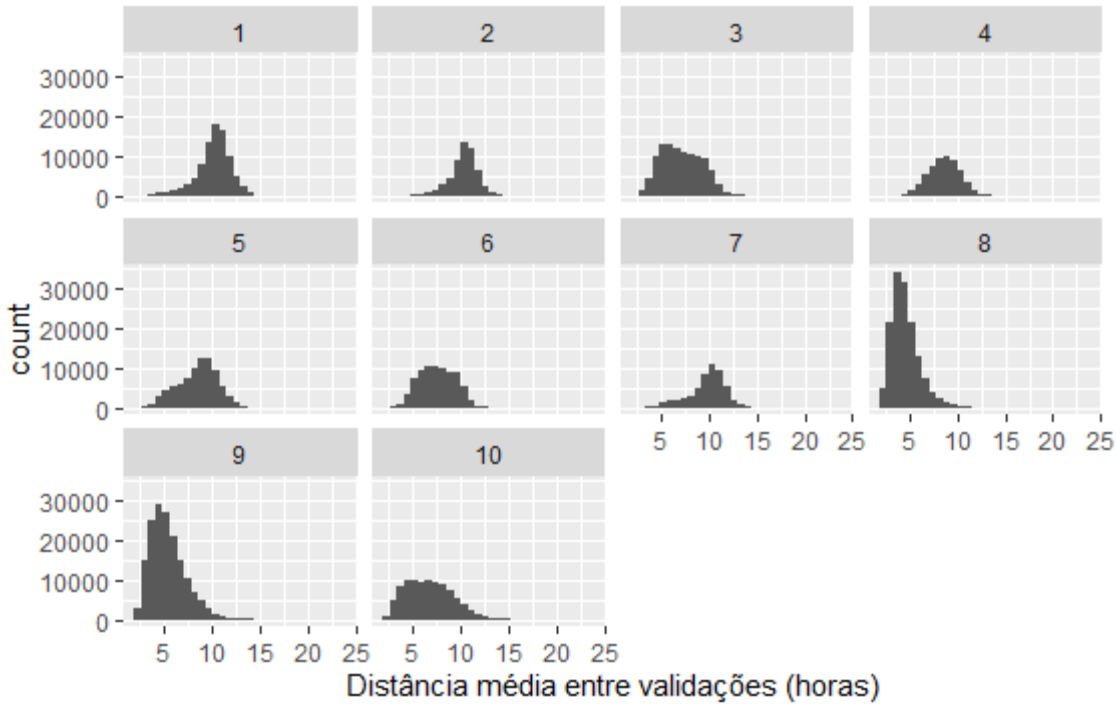
APÊNDICE A – GRÁFICOS DE DISTRIBUIÇÃO DOS ATRIBUTOS PARA OS AGRUPAMENTOS OBTIDOS

- TEMPO ENTRE VALIDAÇÔES

2014

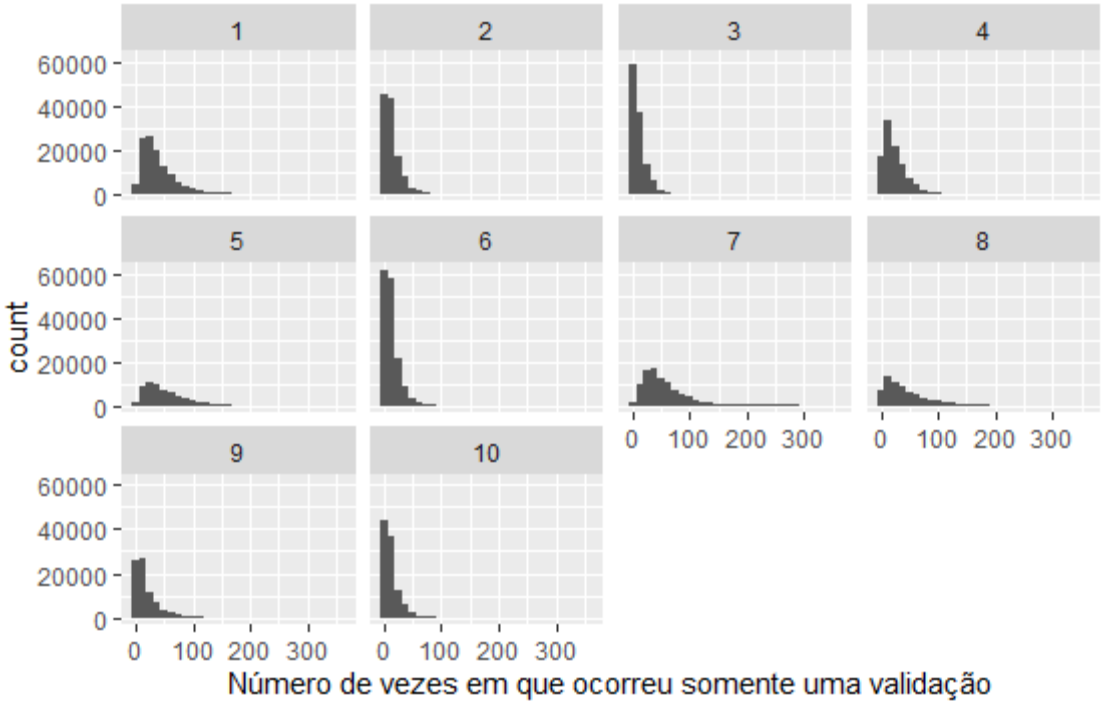


2018

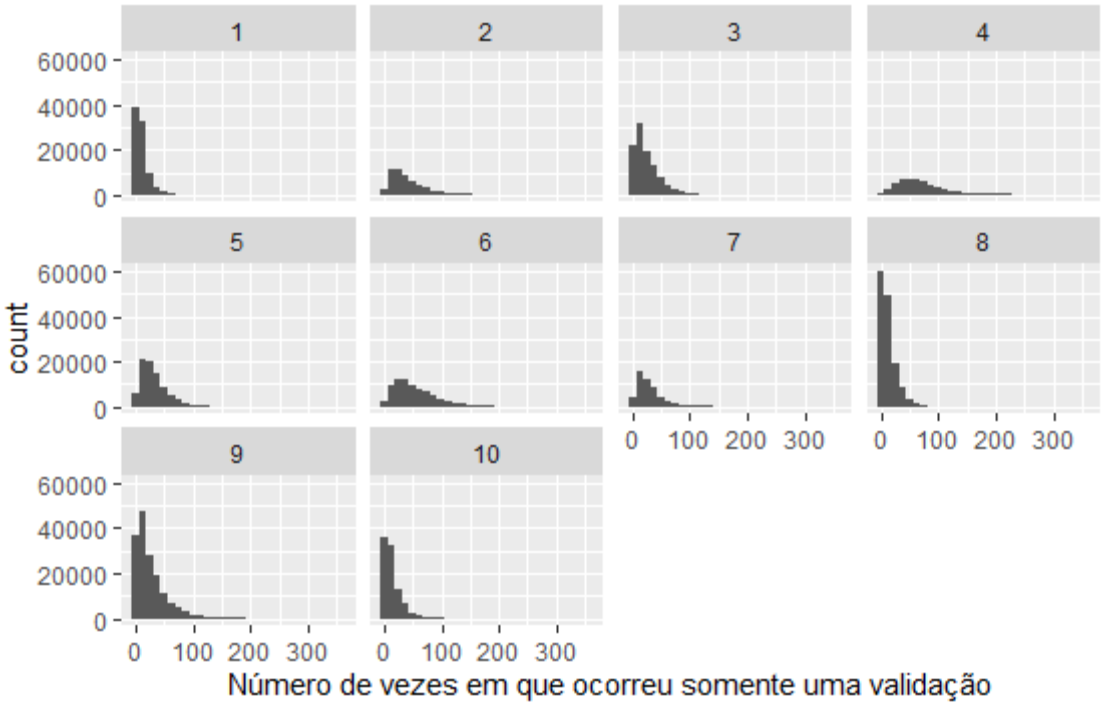


- NÚMERO DE VEZES DE OCORRÊNCIA DE UMA VALIDAÇÃO

2014

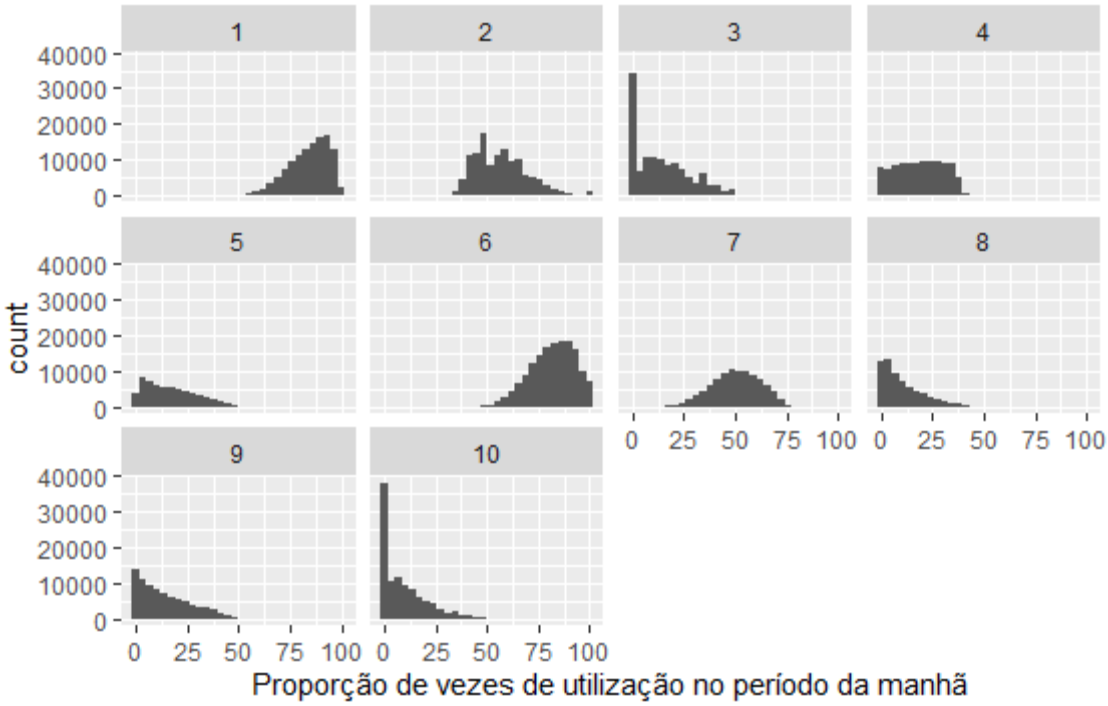


2018

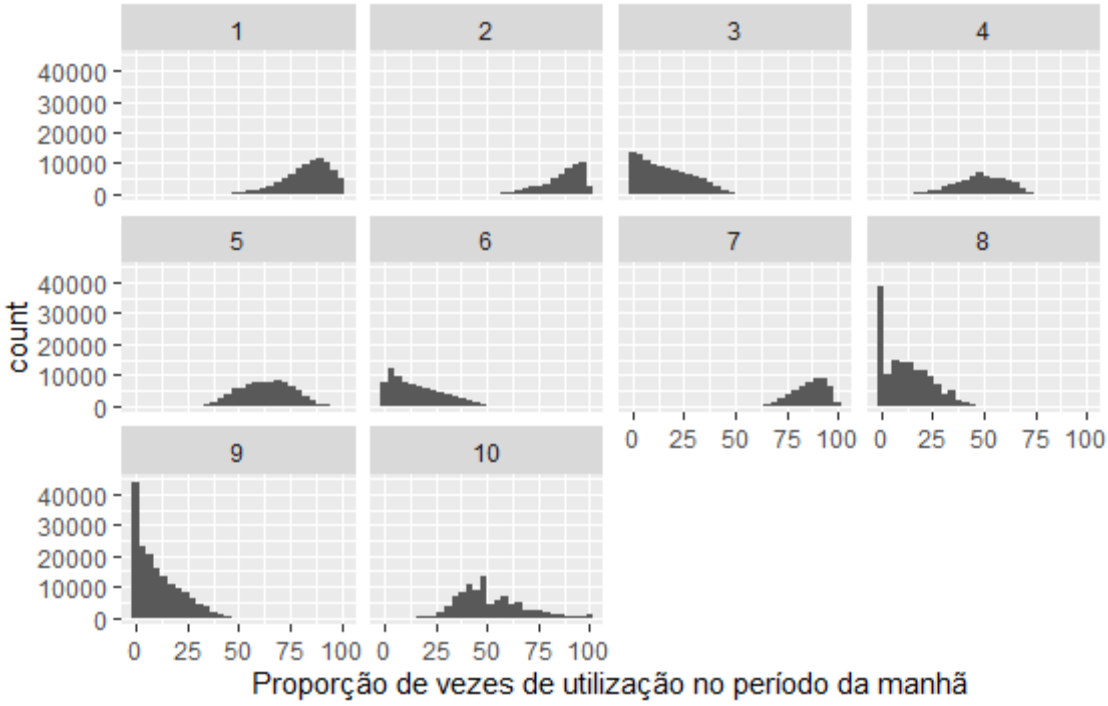


- PROPORÇÃO DE VEZES DE UTILIZAÇÃO NO PERÍODO DA MANHÃ

2014

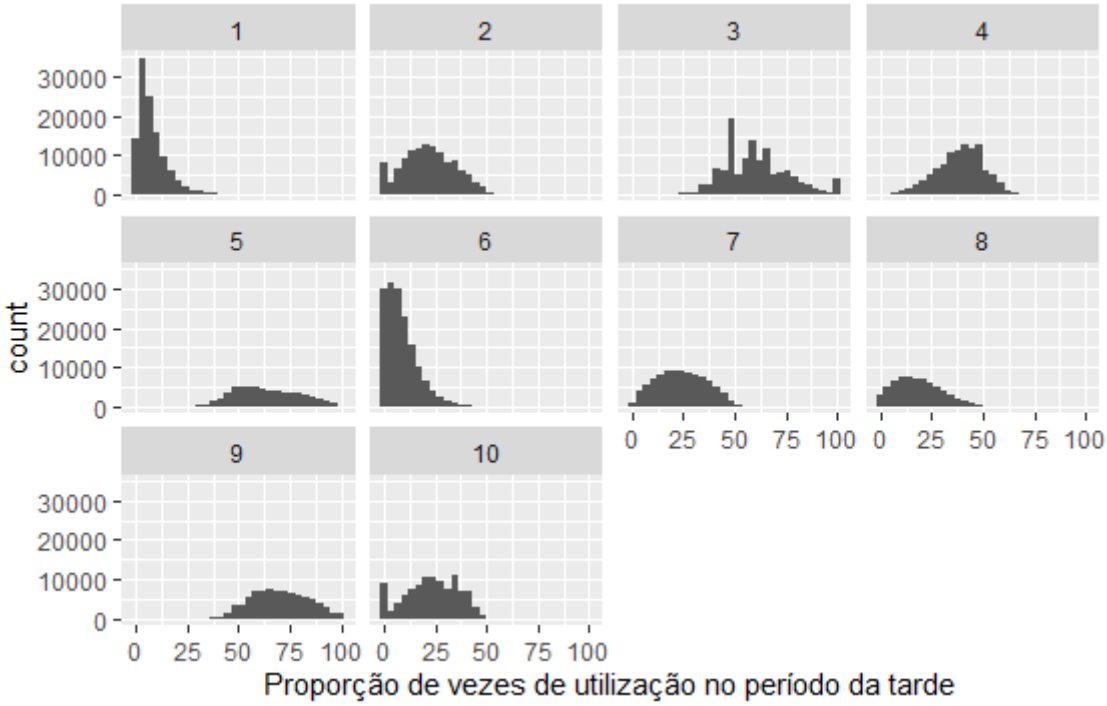


2018

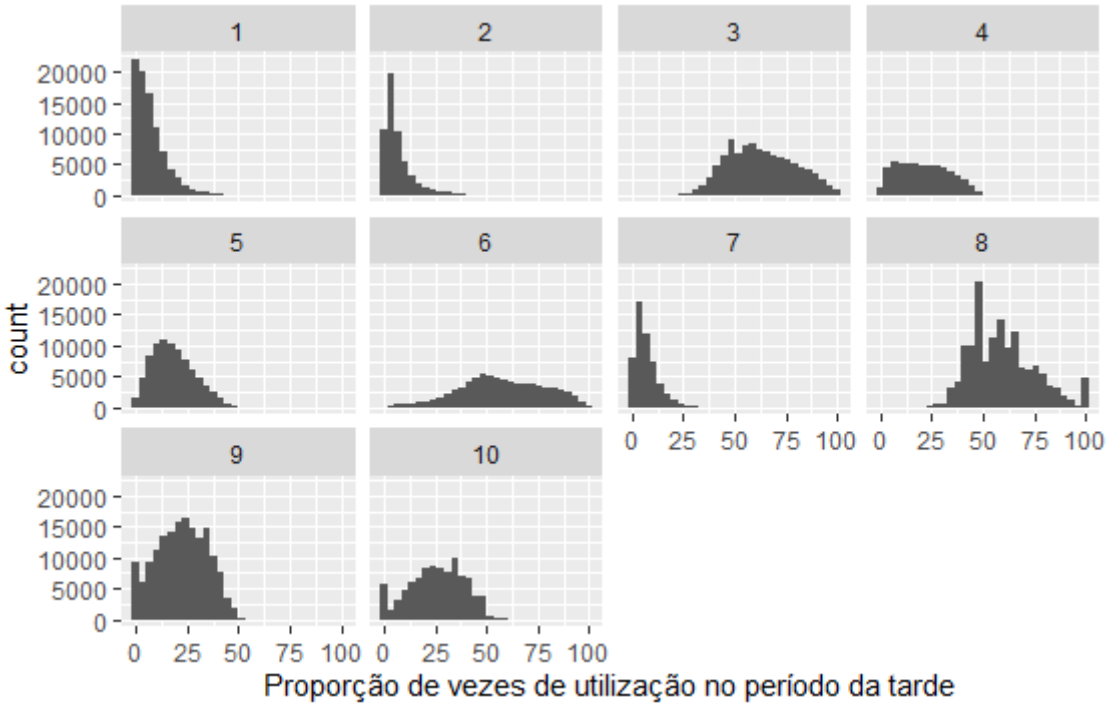


- PROPORÇÃO DE VEZES DE UTILIZAÇÃO NO PERÍODO DA TARDE

2014

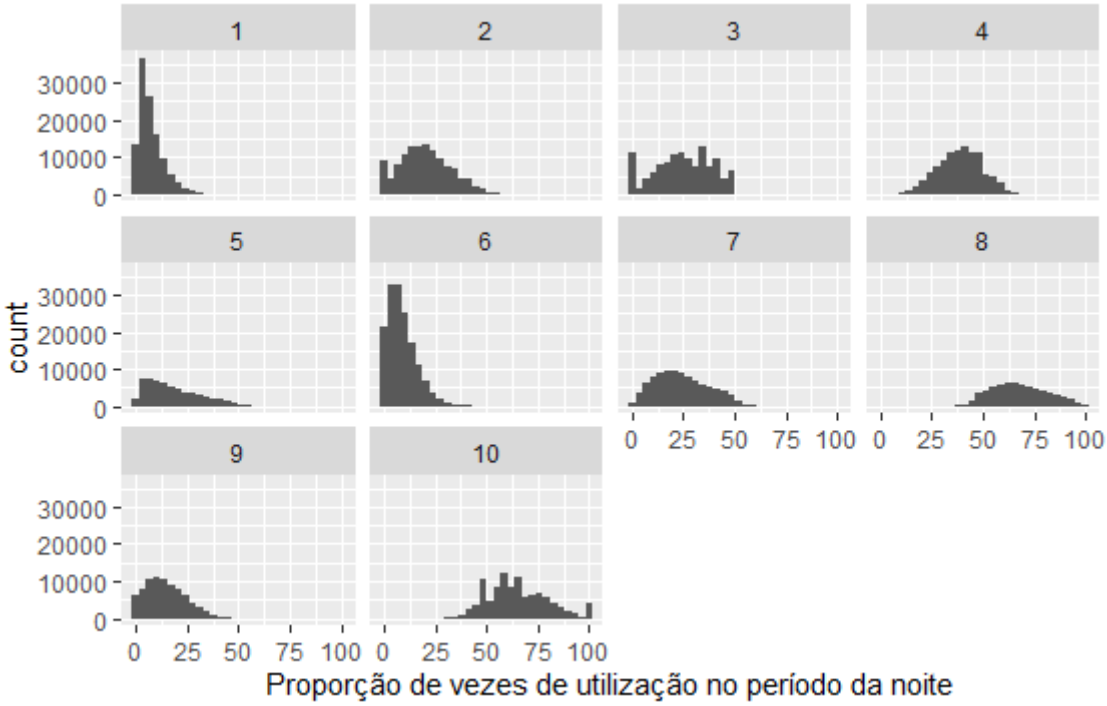


2018

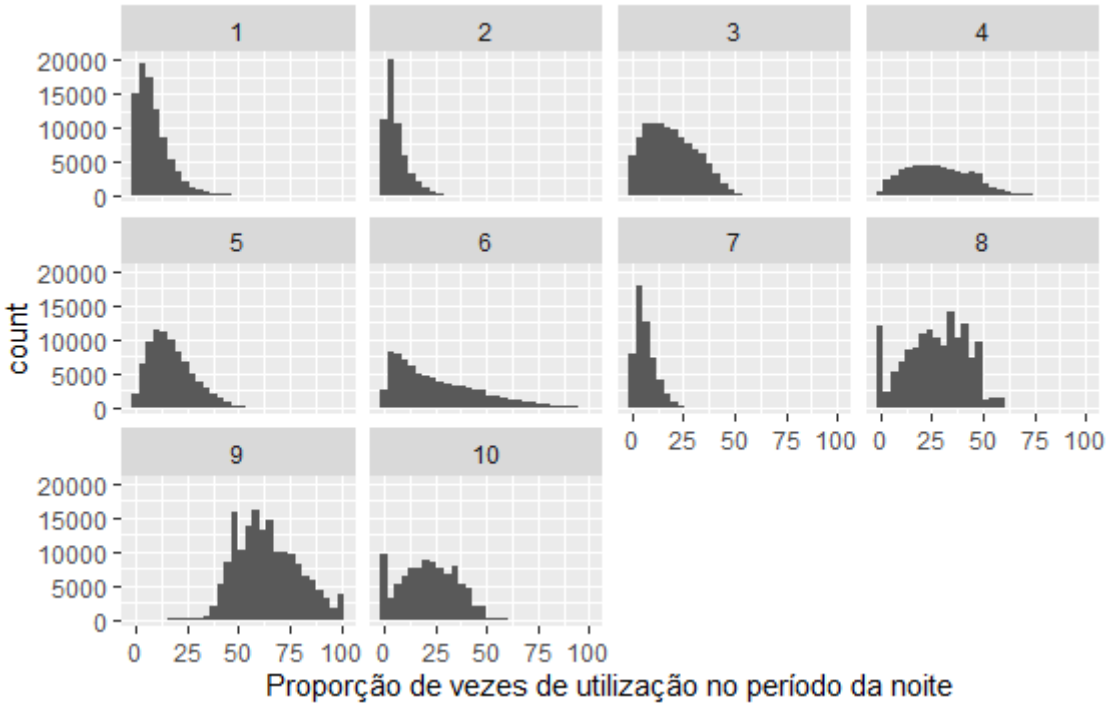


- PROPORÇÃO DE VEZES DE UTILIZAÇÃO NO PERÍODO DA NOITE

2014

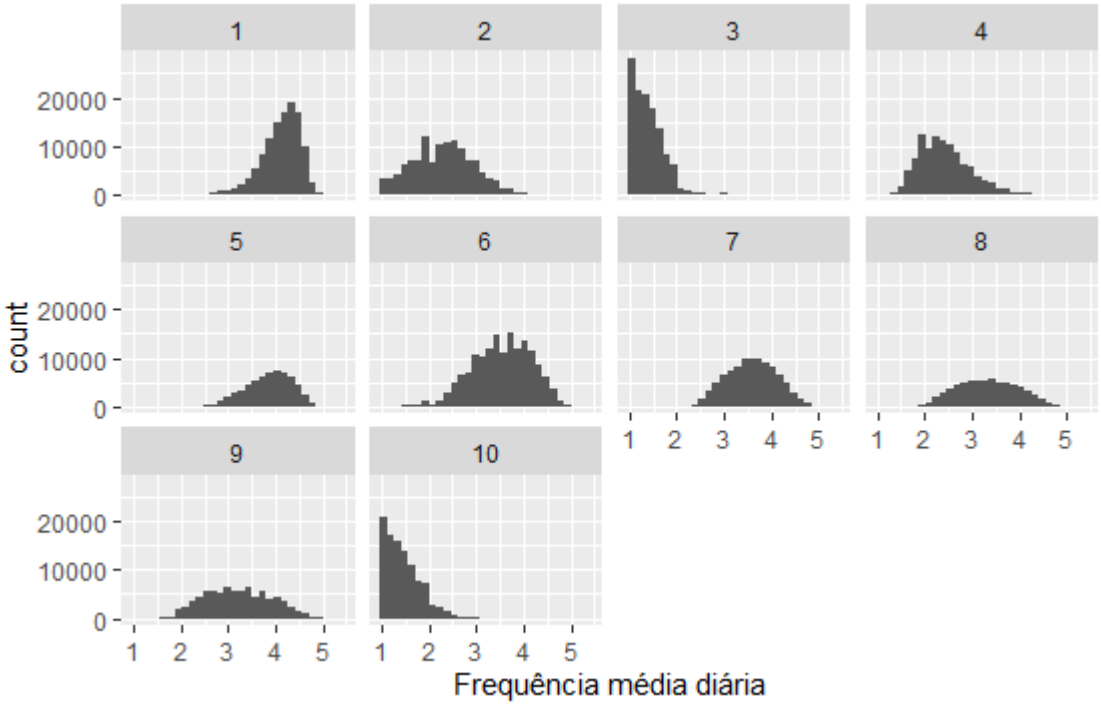


2018

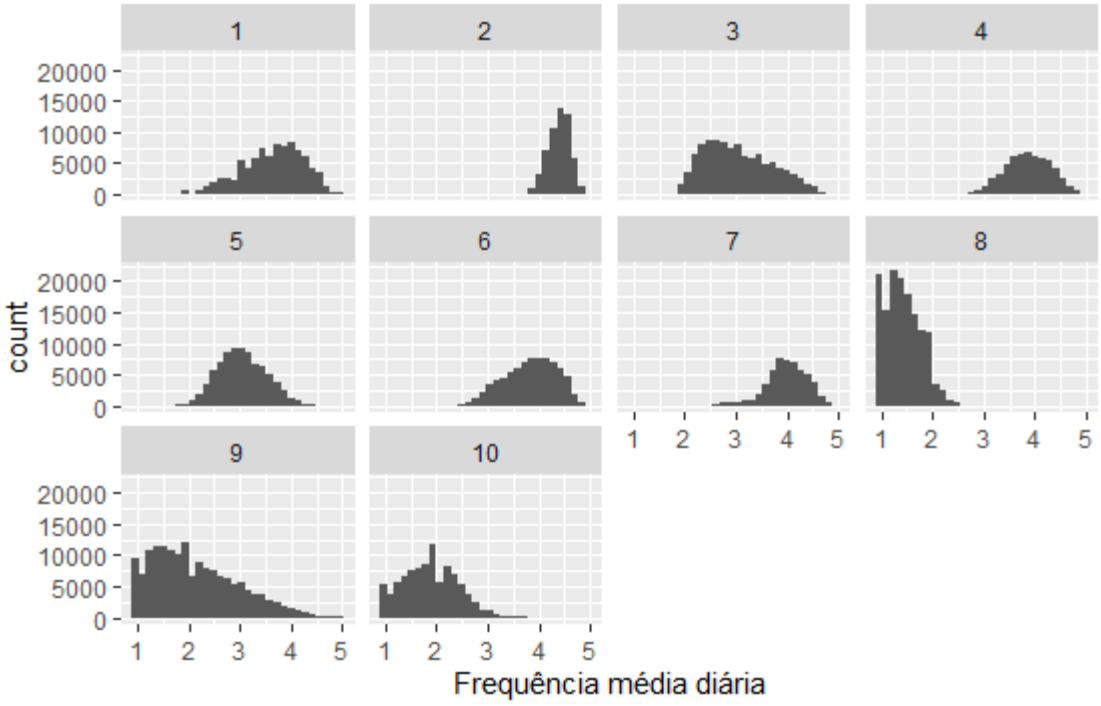


- FREQUÊNCIA DIÁRIA MÉDIA

2014

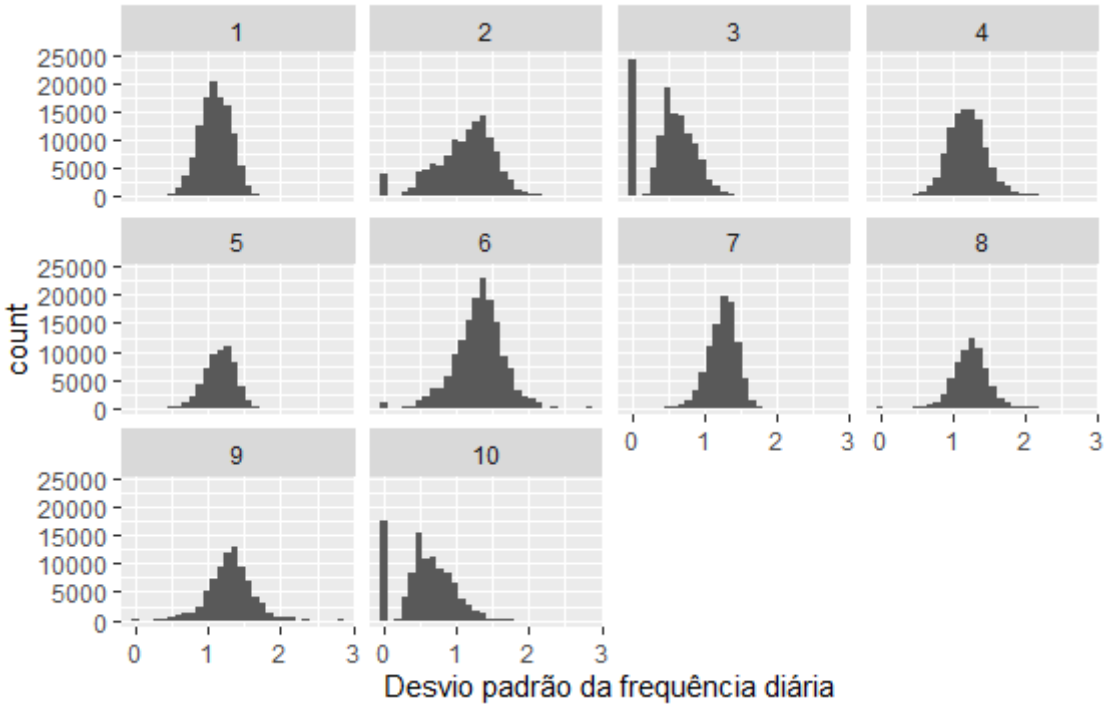


2018

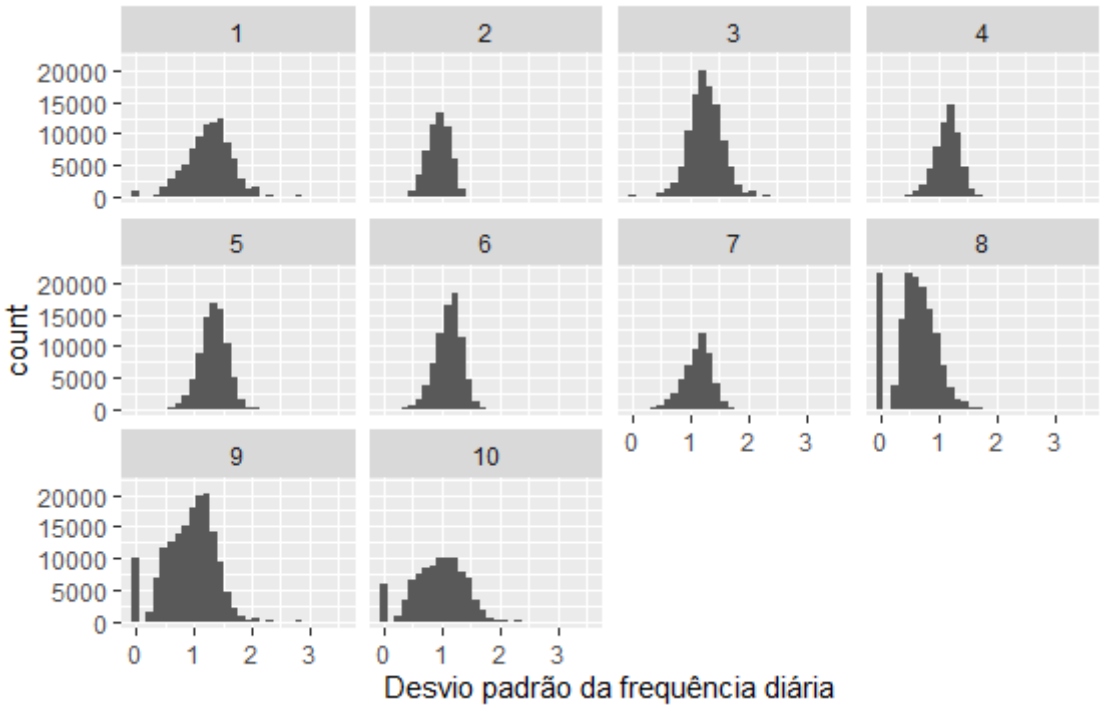


- DESVIO PADRÃO DA FREQUÊNCIA

2014

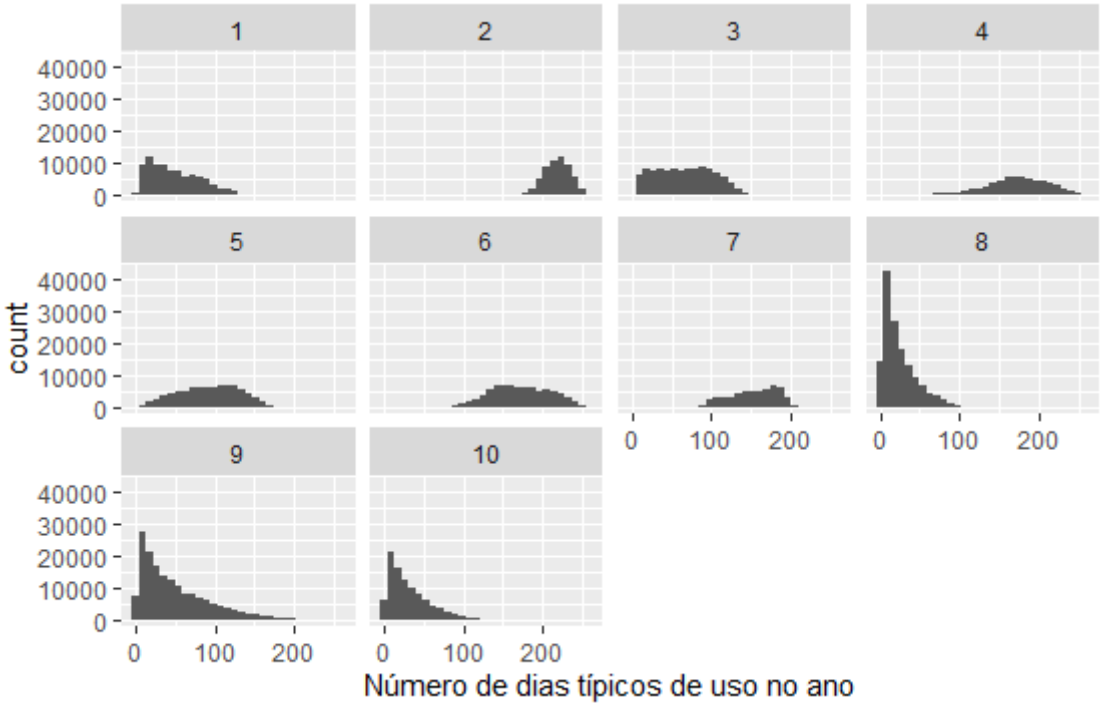


2018



- NÚMERO DE DIAS TÍPICOS DE USO NO ANO

2014



2018