UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CARACTERIZAÇÃO DE UM DATASET DE MOBILIDADE DOS ÔNIBUS DA CIDADE DE SÃO PAULO

CAROLINA JUNQUEIRA FERREIRA

ORIENTADOR: PROF. DR. HERMES SENGER

São Carlos – SP

Janeiro/2021

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CARACTERIZAÇÃO DE UM DATASET DE MOBILIDADE DOS ÔNIBUS DA CIDADE DE SÃO PAULO

CAROLINA JUNQUEIRA FERREIRA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Sistemas de Computação
Orientador: Prof. Dr. Hermes Senger

São Carlos – SP

Janeiro/2021

Ferreira, Carolina Junqueira

Caracterização de um dataset de mobilidade de ônibus da cidade de São Paulo: Characterization of a São Paulo city bus mobility dataset / Carolina Junqueira Ferreira -- 2021.
95f.

Dissertação (Mestrado) - Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador (a): Hermes Senger
Banca Examinadora: Hermes Senger, Paulo Matias, Fabio Kon
Bibliografia

1. Redes Veiculares. 2. Traços de mobilidade. 3. Caracterização. I. Ferreira, Carolina Junqueira. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Ronildo Santos Prado - CRB/8 7325

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Carolina Junqueira Ferreira, realizada em 28/01/2021.

Comissão Julgadora:

Prof. Dr. Hermes Senger (UFSCar)

Prof. Dr. Paulo Matias (UFSCar)

Prof. Dr. Fabio Kon (USP)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.
O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

AGRADECIMENTOS

Gostaria de agradecer à família e amigos pelo apoio durante todo percurso do mestrado. Agradeço também ao meu companheiro Luís por todo apoio nesse trajeto.

Agradeço também ao professor e orientador Hermes por todos os conselhos, ajuda e paciência durante esta etapa.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

RESUMO

A simulação é uma técnica amplamente utilizada para validação e testes de protocolo de roteamento em redes veiculares (VANETs). Essas simulações são baseadas em modelos de mobilidade que devem seguir o padrão real de movimentação dos veículos para produzir avaliações realistas e acuradas de protocolos e outras soluções de redes veiculares. Os modelos de mobilidade e as simulações podem ser baseados e ajustados através de traços reais de mobilidade veicular. Esses traços são registros de posicionamento de veículos durante um período de tempo, e incorporam o comportamento e dinâmica real que ocorrem no dia-a-dia de uma cidade (engarrafamentos, maneira de motoristas dirigir, rotinas da população ou dos tipos de veículos). O entendimento de características de mobilidade desses traços auxilia no estudo de viabilidade de redes veiculares, compreensão de fatores que afetam condições de tráfego, configurar e conduzir simulação realistas para que o projeto de protocolos de roteamento reajam a comportamentos reais da movimentação dos veículos. Portanto, uma ferramenta para produzir resultados mais acurados e realistas nessas simulações é a caracterização de *datasets* de mobilidade veicular. Existem trabalhos que já realizam essa caracterização, porém poucos exploram os dados de grandes cidades como São Paulo. Este trabalho caracteriza um *dataset* de traços reais de mobilidade dos ônibus de São Paulo extraindo e demonstrando métricas de mobilidade e conectividade encontradas na literatura. Previamente à caracterização, o conjunto de dados foi pré-processado através de tratamentos comuns para este tipo de dado. A extração das métricas e o pré-processamento foram feitos através da linguagem Python e a ferramenta de *Big Data* Apache Spark. O objetivo deste trabalho é demonstrar o comportamento de um grande centro urbano, e fornecer um conjunto de dados para fonte de validação, ajustes e condução de simulações de redes veiculares. As características extraídas indicam padrões de comportamento na movimentação dos ônibus que se repetem de acordo com os dias da semana (dias úteis, sábados, e domingos), horário, e regiões da cidade.

Palavras-chave: VANETs, traços de mobilidade, ônibus, São Paulo, caracterização, simulação.

ABSTRACT

Simulation is a widely used technique for validation and testing of routing protocols in vehicular networks (VANETs). These simulations are based on mobility models that must follow the real movement patterns of vehicles to produce realistic and accurate assessments of protocols and other vehicular network solutions. Mobility models and simulations can be based and adjusted using real vehicular mobility traces. These traces are records of vehicles positioning over a period of time, and they incorporate the behavior and dynamics that occur in the day-to-day of a city (traffic jams, the way drivers drive, routines of population or types of vehicles). Understanding the mobility characteristics of these traces helps in feasibility study of vehicular networks, understanding of factors that affect traffic conditions, configuring and conducting realistic simulations so that the design of routing protocols reacts to real vehicle movement behaviors. Therefore, a tool to produce more accurate and realistic results in these simulations is the characterization of vehicle mobility datasets. There are works that already makes this type of characterization, but few of them explore data from large cities like São Paulo. This work characterizes a dataset of real mobility traces of buses in São Paulo, extracting and demonstrating mobility and connectivity metrics found in the literature. Prior to characterization, the dataset was pre-processed using common treatments for this type of data. The metrics extraction and pre-processing were done using the Python language and the big data tool Apache Spark. The objective of this work is to demonstrate the behavior of a large urban centers, and to provide a dataset for the validation, adjustment and conduction of simulations of vehicular networks. The extracted characteristics indicate behavior patterns in the movement of buses that are repeated according to the days of the week (weekdays, Saturdays, and Sundays), period of the day, and regions of the city.

Keywords: VANETs, mobility traces, bus, São Paulo, characterization, simulation.

LISTA DE FIGURAS

GLOSSÁRIO

API – *Application programming interface*

AVL – *Automatic Vehicle Location*

AWS – *Amazon Web Services*

CDF – *Cumulative Distribution Functions*

CSV – *Comma Separated Value*

ECDF – *Empirical Distribution Function*

GPS – *Global Positioning System*

GTFS – *General Transit Feed Specification*

MANET – *Mobile Ad Hoc Network*

OSM – *OpenStreetMap*

PDF – *Probability Density Functions*

S3 – *Amazon Simple Storage Service*

SQL – *Structured Query Language*

VANET – *Vehicular Ad Hoc Network*

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 Contexto

As Redes Veiculares Ad Hoc (Vehicular Ad Hoc Networks - VANETs) são redes auto-organizáveis e distribuídas que surgem da comunicação entre veículos. Esse tipo de rede possui algumas características que a difere de outras redes móveis: mobilidade previsível, nós sem restrições de energia, densidade variável de acordo com número de veículos na via, topologia altamente dinâmica e conexões efêmeras. As VANETs tem como principal área de aplicação prover informações de segurança do tráfego: acidentes, colisões, congestionamento, pedidos de socorro.

Diante de desafios provenientes das características de VANETs (entrega de pacotes em tempo mínimo, perda mínima de pacotes, redução de áreas de interferência por conta de construções, e fragmentação da rede), de desafios logísticos, tecnológicos e financeiros para desenvolver, testar e avaliar protocolos de roteamento em ambientes reais, a simulação de redes veiculares é amplamente adotada (AL-SULTAN et al., 2013) (HARTENSTEIN; LABERTEAUX, 2010). Harri, Filali e Bonnet (2009) e Al-Sultan et al. (2013) salientam que é importante que as simulações utilizem modelos de mobilidade realistas e acurados para refletir o real padrão de movimentação dos veículos. Uppoor et al. (2014) acrescenta que a simulação de redes veiculares pode ser facilmente enviesada pelo *dataset* de mobilidade no qual o modelo de mobilidade é baseado, impactando na avaliação, por exemplo, de protocolos de roteamento.

Os modelos de mobilidade utilizados pelas simulações de VANETs podem ser: baseados em pesquisa, sintéticos, baseados em simuladores de tráfego ou baseados em traços reais de mobilidade. Por exemplo, o trabalho de Uppoor et al. (2014) utiliza um modelo de mobilidade baseado em pesquisa e simulador de tráfego para gerar traços de mobilidade sintéticos para

que possam ser usados em futuras pesquisas com VANETs. Já o estudo de Wen et al. (2018) processa traços reais dos ônibus de São Paulo para melhorar e ajustar o modelo de mobilidade do transporte público utilizado pelo InterSCSimulator, que é um simulador de cidades inteligentes que abrange diversos aspectos de grandes cidades (veículos, ônibus, metrô, sensores, pedestres).

Os traços reais de mobilidade são registros de posicionamento de veículos durante um período de tempo, e incorporam o comportamento e dinâmica real que ocorrem no dia-a-dia de uma cidade (engarrafamentos, maneira de motoristas dirigir, rotinas da população ou dos tipos de veículos). Já os traços sintéticos de mobilidade que também podem ser base para simulações de redes veiculares podem ser gerados por: simuladores de tráfego e/ou mobilidade, imagens, pesquisa de Origem e Destino (O/D), modelos matemáticos, dentre outros. O uso de traços sintéticos em simulações é comum, pois são mais rápidos de serem gerados e obtidos em grande escala comparados aos reais. O uso de traços reais nessas simulações é importante, pois podem: fazer parte do processo de validação dos resultados das simulações e de protocolos; servir de comparação e ajuste de modelos de mobilidade e *datasets* sintéticos; incorporar dinâmicas reais na simulação para que o projeto de protocolos de roteamento reajam a comportamentos reais da movimentação de veículos.

Caracterizar traços de mobilidade veicular em VANETs tem algumas aplicabilidades, como: estudo de viabilidade de redes veiculares, compreensão de fatores que afetam as condições do tráfego, configurar e conduzir simulações realistas, estimar futuras oportunidades de comunicação entre carros em movimento (DOMINGUES; SILVA; LOUREIRO, 2018) (UPPOOR; FIORE, 2012) (DOERING; WOLF, 2015). Na literatura, alguns desses *datasets* são amplamente explorados para projeto de soluções de redes veiculares, como: São Francisco (MARTINS; CUNHA, 2018) (DOMINGUES; SILVA; LOUREIRO, 2018), Shangai (WU; ZHU; LI, 2011), Beijing (WENG et al., 2016), Roma (ALVARENGA et al., 2014), dentre outros. Ainda sim, Santana, Kanashiro e Kon (2018) mencionam a dificuldade em encontrar traços de mobilidade que correspondam ao comportamento de grandes cidades. São Paulo é uma cidade com milhões de pessoas e veículos, porém pouco explorada em trabalhos na área de VANETs, tendo traços de mobilidade do transporte público abertos para consumo através da API Olho Vivo[1] da organização pública SPTrans que gerencia o transporte público de ônibus de São Paulo.

Considerando a importância que traços reais de mobilidade tem na condução de simulações realistas de redes veiculares, a lacuna da cidade de São Paulo ser pouco explorada nessa área, e a importância de caracterizar esses traços para entendimento da dinâmica de mobilidade, este trabalho caracteriza a mobilidade de um *dataset* de ônibus da cidade de São Paulo. Os da-

[1] https://www.sptrans.com.br/desenvolvedores/

dos dados processados resultantes desta pesquisa estão disponíveis no Kaggle ⟨https://www.kaggle.com/caroljunq/sao-paulo-bus-mobility-traces-oct-2015⟩. Já os códigos que foram utilizados para caracterizar e tratar o conjunto estão no Github ⟨https://github.com/caroljunq/sptrans-data-analysis⟩ permitindo a reprodutibilidade do trabalho para outras pesquisas na área de VANETs ou mobilidade.

1.2 Motivação e Objetivos

Harri, Filali e Bonnet (2009) e Al-Sultan et al. (2013) salientam a importância de que modelos de mobilidade aplicados em simulações sigam o real padrão de movimentação dos veículos para produzir avaliações realistas e acuradas de soluções para redes veiculares. Usar traços reais de mobilidade é um dos meios para conduzir simulações realistas e acuradas, e para que protocolos de roteamento reajam a real dinâmica de uma cidade. Entender características desse tipo de conjunto de dados é essencial para entender quais fatores afetam o tráfego, identificar oportunidades de conexão entre veículos, estudo de viabilidade de VANETs. As características extraídas podem ajudar a configurar e ajustar as simulações e modelos de mobilidade. Encontrar traços reais de mobilidade disponíveis que apresentem comportamentos de grandes cidades é uma dificuldade. São Paulo é uma cidade com milhões de habitantes e veículos, e apesar de possuir conjunto de dados de ônibus disponível historicamente e em tempo real, é pouco explorada. Portanto, considerando a importância do estudo de conjuntos de dados de mobilidade veicular para condução de simulações realistas e estudos de protocolos de roteamento em VANETs, e a lacuna de São Paulo, este trabalho tem como objetivo principal caracterizar o *dataset* de mobilidade dos ônibus de São Paulo a partir de métricas extraídas de trabalhos da literatura.

Os objetivos secundários dessa pesquisa são:

- identificar e divulgar métricas para caracterização de mobilidade veicular;
- identificar e mostrar sucintamente o funcionamento de ferramentas de manipulação e processamento de traços reais de mobilidade;
- pré-processar os dados para as métricas necessárias;
- extrair as métricas identificadas do *dataset*;
- divulgar publicamente os métodos utilizados para cálculo das métricas;
- divulgar publicamente o *dataset* após o pré-processamento dos dados.

O principal público desta pesquisa são pesquisadores da área de redes veiculares ou de

mobilidade. Os pesquisadores terão acesso a caracterização de um *dataset* que pode ser aplicado para estudo de protocolos de roteamento e simulações em VANETs, e também pode fornecer informações sobre a dinâmica de mobilidade dos ônibus. A cidade de São Paulo é uma cidade com grande número de ônibus, e diversos fatores que podem afetar a mobilidade do tráfego, portanto o *dataset* e este trabalho tornam-se relevantes para aqueles que precisam entender a dinâmica de uma grande cidade.

Durante o trabalho, o processo de extração das métricas e os pré-processamentos foram documentados. Todos os códigos, e parte do *dataset* processado e sem pré-processamento estão disponíveis publicamente. Esses fatores contribuem para que outros pesquisadores repliquem e reproduzam caracterizações como a desta pesquisa para outros *datasets*. Já o conjunto de dados resultante pode ser aplicado no desenvolvimento de soluções de redes veiculares, como as simulações e testes de protocolos de roteamento.

1.3 Metodologia

No processo de desenvolvimento deste trabalho, foi realizada uma pesquisa bibliográfica para determinar o referencial teórico e os trabalhos relacionados. A metodologia da pesquisa bibliográfica permitiu que fosse estabelecida a fundamentação teórica necessária para conhecer o objeto de pesquisa, além da identificação de lacunas a serem preenchidas.

Na pesquisa bibliográfica foram explorados trabalhos que caracterizassem *datasets* de mobilidade com foco em testes, simulações ou estudo de viabilidade de VANETs, e trabalhos que faziam pré-processamento de dados dessa natureza. Foram considerados autores que utilizassem a caracterização desses *datasets* como foco principal do trabalho ou como apoio para outras etapas. Foram encontrados tanto trabalhos cujo objetivo principal é a descrição e caracterização de *datasets*, quanto trabalhos que utilizam outros conjuntos de traços de mobilidade para gerar novos *datasets*. Esses trabalhos que geram novos conjuntos de dados possuem uma etapa de caracterização dos dados de mobilidade gerados. Já em relação ao pré-processamento, parte dos trabalhos de caracterização ou geração explicam brevemente como fizeram o pré-processamento dos dados, então uma literatura específica na área de *Trajectory mining* (em português, mineração de trajetórias) teve que ser pesquisada para apoiar a fase de pré-processamento dos dados de ônibus de São Paulo.

A partir dos trabalhos relacionados, foram identificadas métricas de mobilidade e de conectividade que foram extraídas do *dataset* de mobilidade desta pesquisa a fim de caracterizá-lo. E também foram identificadas técnicas de pré-processamento para dados de mobilidade.

Este trabalho é uma pesquisa exploratória, pois foram apresentadas métricas, e partir delas foram formuladas hipóteses com base na observação empírica da dinâmica da cidade de São Paulo (horários comerciais conhecidos; algumas áreas atrativas conhecidas como regiões de terminais e centrais). Além disso, a proposta desta pesquisa é aumentar a familiaridade de pesquisadores com o *dataset* de ônibus de São Paulo a fim de clarificar características desses dados para aplicação em pesquisas futuras (LAKATOS EVA MARIA; MARCONI, 2003). A ideia principal deste trabalho é descobrir características do *dataset*, ou seja, demonstrar que valores essas características assumem e como podem ser utilizadas em redes veiculares. Demonstrou-se o método de cada um dos cálculos das métricas para que possam ser replicados a outras janelas tempo do *dataset* de São Paulo ou outros *datasets* de traços de mobilidade.

As etapas de execução desta pesquisa se deram com a identificação das métricas de mobilidade e conectividade nos trabalhos relacionados. Em seguida, foram explorados trabalhos na área *Trajectory mining* que proporcionaram etapas de pré-processamento necessárias para traços de mobilidade. Posteriormente, foram feitas etapas de pré-processamento com o Spark e bibliotecas do Python para tornar o *dataset* apto para extração de métricas. Nos passos seguintes, extraiu-se as métricas de mobilidade e conectividade dos traços de mobilidade de ônibus através do Spark. Os resultados foram demonstrados através de recursos gráficos e da exploração e apresentação dos intervalos estatísticos dos dados. Por fim, foi feita uma breve associação entre os valores das métricas e o cenário de redes veiculares na cidade de São Paulo.

1.4 Organização do trabalho

Este trabalho está organizado da seguinte forma: o Capítulo 2 possui o referencial teórico e os conceitos chave base para esta pesquisa, o Capítulo 3 possui os trabalhos relacionados, o Capítulo de 4 possui as etapas de pré-processamento que foram necessárias para tratar os dados, o Capítulo 5 possui a caracterização do *dataset* a partir de métricas de mobilidade e conectividade e o Capítulo 6 possui a conclusão e trabalhos futuros desta pesquisa.

Capítulo 2

REFERENCIAL TEÓRICO

Este capítulo apresenta fundamentos teóricos para o entendimento da pesquisa, e do processo de caracterização dos traços de mobilidade dos ônibus de São Paulo. Para isto são abordados: o que são VANETs, o roteamento nesse tipo de rede, o que são traços de mobilidade, o processo de caraterização realizado neste trabalho, e que ferramentas de dados podem ser utilizadas para processar o conjunto de dados desta pesquisa.

2.1 VANETs

As Redes Veiculares Ad Hoc (Vehicular Ad Hoc Networks - VANETs) são um tipo de rede móvel ad hoc (Mobile Ad Hoc Network - MANET) auto-organizável e distribuída que surge da comunicação sem fio entre veículos (vehicle-to-vehicle -V2V) e/ou veículos e infraestruturas da via (conhecido como vehicle-to-infraestructure - V2I ou vehicle-to-road- V2R) (MOUSTAFA; ZHANG, 2009) (AL-SULTAN et al., 2013). A Figura 2.1 mostra os tipos de comunicação em redes veiculares onde RSU (*Road Side Unit*) é a infraestrutura da via.

Por serem um tipo de MANET, as VANETs possuem topologia dinâmica e nós que podem agir como terminais de dados ou de computação. Segundo Al-Sultan et al. (2013) e Moustafa e Zhang (2009), as caraterísticas que diferem as redes veiculares de outras redes móveis são:

- **Mobilidade previsível:** diferente de outras MANETs cuja mobilidade dos nós é difícil de predizer, a movimentação de veículos é limitada pela topologia das vias, sinais de trânsito, semáforos e movimento de outros veículos (ex: congestionamento). No caso de ônibus, há maior previsibilidade devido aos horários e rotas estabelecidas;
- **Nós sem restrições de energia:** os veículos fornecem energia constante para os dispositivos que permitem a computação e comunicação com a rede;

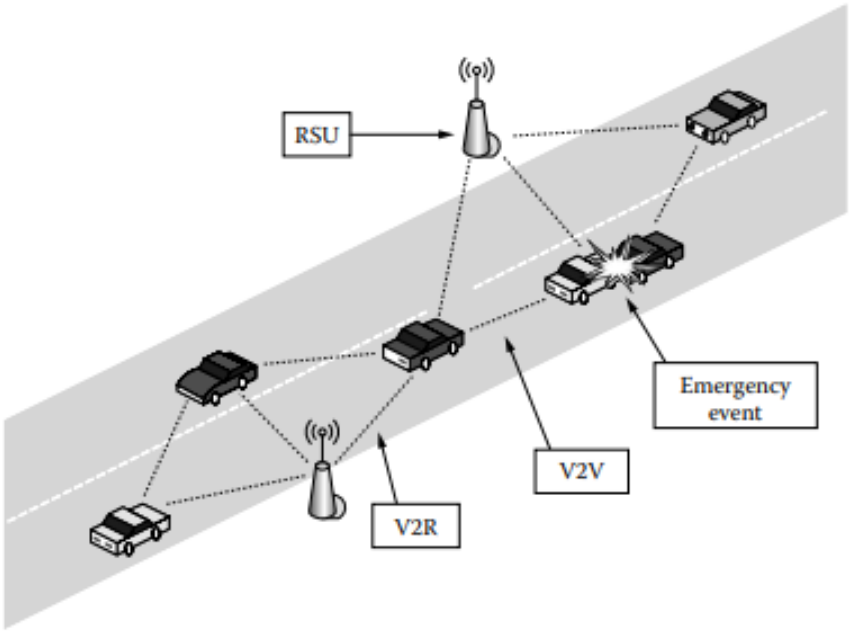


Figura 2.1: Comunicações em VANETs (AYYASH; ALSBOU; ANAN, 2015).

- **Densidade variável:** a densidade de nós varia de acordo com a quantidade de veículos no tráfego. Por exemplo, pode haver a alta quantidade de veículos em zona urbana e baixa quantidade em zona rural;
- **Topologia altamente dinâmica e conexões efêmeras:** a topologia deste tipo de rede muda rapidamente devido a alta velocidade dos veículos. Isso provoca tempo de vida curto (efêmero) para as conexões entre os nós;
- **Alta capacidade computacional:** devido aos nós de uma VANET serem veículos, eles podem ser equipados com sensores e recursos computacionais (processadores, memórias, antenas, GPS, etc.) para o crescimento do seu poder computacional a fim de obter uma comunicação sem fio confiável e adquirir informação precisa sobre a sua posição, direção e velocidade.

As áreas de aplicação das redes veiculares estão centradas em conforto e segurança. Na primeira área estão aplicações para prover aos motoristas e passageiros informações, como: clima, tráfego, restaurantes próximos, postos de gasolina, entre outros. Já na área de segurança estão aplicações que proveem informações sobre: acidentes, colisões, veículos que ultrapassaram sinais, congestionamento, pedidos de socorro, veículos estacionados, entre outros.

2.1.1 Desafios de VANETs

Em razão das características de uma rede veicular, existem desafios a serem considerados para implantação desse tipo de rede. Al-Sultan et al. (2013), Ayyash, Alsbou e Anan (2015) e Moustafa e Zhang (2009) salientam que a topologia altamente dinâmica, os nós com alta velocidade de deslocamento e a densidade variável de nós resultam em questões como fragmentação na conectividade da rede e *broadcast storm*. A conectividade numa VANET sofre frequente fragmentação devido à curta duração dos contatos entre os veículos e densidade variável de nós. Devido a essa fragmentação na rede e na tentativa de alcançar o maior número possível de veículos, a taxa de transmissão e retransmissão em redes veiculares é alta, provocando *broadcast storm* que, em áreas densas, reduz o número de mensagens entregues e aumenta o tempo para entrega de mensagens. Existem outros desafios como a necessidade de entrega de pacotes em tempo mínimo, perda mínima de pacotes, redução de áreas de interferência por conta de construções.

Estes desafios junto a desafios tecnológicos, logísticos, e financeiros fazem com que as simulações sejam amplamente utilizadas para testes e desenvolvimento de soluções de roteamento em VANETs.

2.2 Roteamento em VANETs

O roteamento é uma das principais questões de pesquisa em redes veiculares. Como vimos na Seção 2.1, as VANETs possuem propriedades únicas, como: número grande de veículos, topologia da rede e densidade de nós altamente dinâmicos, influências por congestionamento, mobilidade previsível. Essas propriedades fazem com que pesquisadores desenhem protocolos de roteamento que podem se adaptar a diferentes situações.

Segundo Moustafa e Zhang (2009), Al-Sultan et al. (2013), e Singh e Agrawal (2014), os protocolos para este tipo de rede podem ser projetados para suportar: as variações da densidade de veículos, situações onde há maior ou menor número de veículos; fragmentação da rede; redução de áreas de interferência causadas por construções; predição de eventos; mudança no deslocamento de veículos por influência de fatores de mobilidade (horários do dia, estilos de direção, tipo de veículo); dentre outros. Para estas adaptações, esses protocolos podem usar técnicas, como: regular a taxa de transmissão de pacotes, descoberta de vizinhança, trajetória pré-computadas, disseminação de mensagens com níveis de prioridade de transmissão, entrega de pacotes em tempo mínimo principalmente em situações de emergência, perda mínima de pacotes, uso de informações adicionais.

2.2.1 Paradigmas de roteamento

De acordo com Singh e Agrawal (2014) e Moustafa e Zhang (2009), os paradigmas de roteamento em VANETs são:

- **Paradigma baseado em topologia (ad hoc):** os algoritmos de roteamento dependem da topologia da rede. Por exemplo, os protocolos proativos como o OLSR (*Optimized Link State Routing*) constroem tabelas de roteamento mesmo que não haja mensagem para encaminhar. Já os protocolos reativos como o AODV (*On-Demand Distance Vector*) determinam a rota para um destino somente na requisição;
- **Paradigma baseado em informação geográfica:** algoritmos baseiam-se na posição geográfica dada por algum dispositivo como, por exemplo, o GPS. Exemplos de protocolos dessa abordagem são o GAMER (*Geocast Adaptive Mesh Environment for Routing*) e o GPSR (*Greedy Perimeter Stateless Routing*);
- **Paradigma baseado em hierarquia:** a rede é composta por *clusters* que são definidos como grupos de nós que compartilham características. As mensagens são propagadas de *cluster* para *cluster*. O HSR (*High availability Seamless Redundancy*) é um exemplo de protocolo que segue este paradigma;
- **Paradigma baseado em *broadcast*:** um nó recebe a mensagem e retransmite para todos os seus vizinhos, garantindo que o maior número possível de nós recebam a mensagem. Um exemplo deste tipo de protocolo é o V-TRADE (*Vector-based Tracking Detection*);
- **Paradigma baseado em movimento:** a mensagem é propagada até o destino de acordo com a movimentação dos nós. Portanto, um nó carrega uma mensagem até encontrar o destino.

2.2.2 Simulação e modelos de mobilidade

Para avaliar os diferentes cenários de mobilidade e adaptações dos protocolos, e diante dos desafios logísticos, tecnológicos e financeiros de testar e avaliar protocolos de roteamento em ambiente reais, a simulação é amplamente adotada para testes, avaliação de performance e validação de protocolos em redes veiculares (AL-SULTAN et al., 2013) (HARTENSTEIN; LABERTEAUX, 2010). Harri, Filali e Bonnet (2009) e Al-Sultan et al. (2013) realçam a importância dessas simulações utilizarem modelos de mobilidade realistas e acurados a fim de refletirem o real padrão de movimentação dos veículos.

Os modelos de mobilidade são representações dos movimentos reais de veículos que mudam de velocidade e direção ao longo do tempo. Esses modelos são utilizados por simulações para reproduzir padrões de mobilidade do mundo real (HARTENSTEIN; LABERTEAUX, 2010). O desafio na simulação de VANETs está no uso de modelos de mobilidade realistas e acurados. Segundo Harri, Filali e Bonnet (2009), um modelo de mobilidade deveria apresentar os seguintes blocos de construção para ser realista: mapa topológico preciso e realista, obstáculos para movimentação do carro e comunicação *wireless*, pontos atrativos e repulsivos, características dos veículos, viagens (origem e destino), caminhos utilizados nas viagens, aceleração e desaceleração suaves dos veículos, padrões humanos ao dirigir, intersecções das estradas, padrões de tempo e influência externa (acidentes, obras, etc.). Quanto mais blocos um modelo apresentar, mais realista e acurado ele será.

Segundo Harri, Filali e Bonnet (2009), as classes de modelos de mobilidade são:

- **Modelo baseado em traços de mobilidade:** extrai padrões de movimento de traços de mobilidade, podendo criar novos modelos ou calibrar existentes através de dados estatísticos de velocidade, tempos de pausa e número de contato entre veículos (HARTENSTEIN; LABERTEAUX, 2010). De acordo com Harri, Filali e Bonnet (2009), as limitações deste modelo envolvem disponibilidade de traços de mobilidade e especificidade do ambiente onde o estudo é realizado. Outra dificuldade acrescentada por Hartenstein e Laberteaux (2010) é inferir padrões que não são observados diretamente no conjunto de dados analisado, por exemplo tentar modelar movimento de veículos pessoais a partir de informações de GPS de ônibus. Uma aplicação possível para o modelo baseado em traços de mobilidade é prever matematicamente onde um veículo estará em determinado momento. Um dos benefícios desse tipo de modelo é incorporar comportamentos reais de mobilidade (congestionamento, padrões de direção humanos, influência da vias e horários do dia, semáforos) (HARTENSTEIN; LABERTEAUX, 2010);
- **Modelo baseado em *survey* (pesquisa):** coletam-se comportamentos de indivíduos através de pesquisas (Origem e Destino O/D, questionários) para desenvolver um modelo genérico de mobilidade capaz de reproduzir pseudo-aleatoriamente ou deterministicamente o comportamento observado no real cenário urbano. A limitação desses modelos é que as estatísticas reproduzidas são grosseiras e não refletem a realidade com grande grau de detalhes (HARTENSTEIN; LABERTEAUX, 2010). Esse tipo de modelo pode ser utilizado para calibrar modelos sintéticos a fim de obter modelos mais realistas;
- **Modelo sintético:** tenta reproduzir matematicamente movimentos veiculares (modelos estocásticos, modelos de tráfego, modelos comportamentais, dentre outros). Harri, Filali

e Bonnet (2009) propõem que um processo de validação para este tipo de modelo pode ser feito coletando e medindo traços reais de mobilidade em grande escala e comparando-os com os padrões desenvolvidos pelo modelo sintético. A maior limitação desse tipo de modelo é não modelar o comportamento humano detalhadamente, pois os humanos respondem a estímulos e perturbações locais que podem ter um efeito global na modelagem de tráfego. O modelo sintético pode ser calibrado com traços de movimento reais ou pesquisas;

- **Modelo baseado em simuladores de tráfego:** refinando modelos sintéticos e validando-os por um processo de comparação com traços de mobilidade reais e pesquisas, algumas companhias e pesquisadores desenvolveram simuladores de tráfego. Esses simuladores seguem modelos de mobilidade e são capazes de simular, por exemplo, comportamento de pedestres, consumo de energia, poluição, entre outros. Entretanto, simuladores de tráfego não são facilmente ou imediatamente conectáveis a simuladores de rede, necessitando, em alguns casos, converter dados ou comprar licenças de outros programas. Esses simuladores podem ser alimentados também por traços reais ou sintéticos de mobilidade.

2.3 Traços de mobilidade

Os traços de mobilidade (*mobility traces*) são dados de geolocalização capturados ou gerados a partir do deslocamento de objetos móveis, como táxis, ônibus, metrô. Esses traços podem ser reais ou sintéticos.

2.3.1 Traços reais de mobilidade

Os traços reais de mobilidade são dados de geolocalização capturados de um veículo ou objeto que se desloca ao longo do tempo. Esse tipo de dado pode ser coletado por GPS ou algum sistema ou sensor AVL ((Automatic Vehicle Location - Localização automática de veículos) que detecta o posicionamento do veículo e reporta-o a um sistema (UPPOOR et al., 2014). Esses traços, normalmente, têm cobertura, em km^2, e duração (dias) maior do que os traços sintéticos. Alguns conjuntos de dados desse tipo amplamente utilizados em VANETs podem ser encontrados na Tabela 2.1.

Alguns exemplos de aplicações para os traços reais de mobilidade são os trabalhos de Zheng et al. (2017) e Xu, Li e Chen (2018). Zheng et al. (2017) utilizam os traços de mais de 10000 táxi da cidade de Chongqing na China para determinar distribuição do tráfego e áreas atrativas da cidade através de algoritmos de clusterização. Já Xu, Li e Chen (2018) utilizam os traços de

táxi de Shanghai, Beijing e Nanjing na China para analisar a topologia oportunística de imensas redes de táxi focando no número, localização e evolução de componentes conectados para revelar as relações entre a dinâmica da topologia e parâmetros chave relacionados à mobilidade.

Apesar de não encontrar muitos trabalhos ainda explorando esses conjuntos de dados, são dados publicamente disponíveis e que talvez possam ser explorados futuramente no âmbito de VANETs: viagens de Uber[1] e de táxi[2] de Nova Iorque.

Os dados caracterizados nesta pesquisa são traços reais de mobilidade coletados dos ônibus da cidade de São Paulo. Estes dados podem ser coletados em tempo quase real pela API Olho Vivo ou dados históricos disponibilizados pela organização SPTrans que coordena o transporte público de ônibus em São Paulo.

Em relação a outros *datasets* populares na área de VANETs, o conjunto de dados de São Paulo possui mais veículos que alguns deles. Além disso, como São Paulo é uma metrópole com milhões de pessoas, a dinâmica da cidade em variadas horas do dia pode incorporar comportamentos diferentes dos *datasets* apresentados na Tabela 2.1. No quesito tempo de cobertura, os traços de mobilidade de São Paulo permitem escolher a janela de captura (mês, dia, hora, etc.).

2.3.2 Traços sintéticos de mobilidade

Os traços sintéticos de mobilidade são gerados: por simuladores de tráfego ou mobilidade, análises de imagens e filmagens, pesquisas de Origem-Destino, ou por detectores nas estradas (UPPOOR et al., 2014). Os dados sintéticos, geralmente, apresentam menor cobertura em km^2 e duração (horas) comparados aos reais. Um exemplo de *dataset* sintético amplamente utilizado em redes veiculares é o da cidade de Colônia, como demonstra a Tabela 2.1 e é abordado pelo trabalho de Uppoor et al. (2014).

Um exemplo de aplicação para os traços sintéticos foi o trabalho de Uppoor et al. (2014) que descreve o processo de criação do traço de Colônia na Alemanha com base numa pesquisa de Origem-Destino e no simulador SUMO (Simulation of Urban Mobility) produzindo aproximadamente 600 mil viagens durante 24 horas numa área de 400 km^2.

2.3.3 Importância dos traços reais de mobilidade

Os traços reais incorporam comportamento e dinâmica real que ocorrem no dia-a-dia de uma cidade (engarrafamentos, maneira de motoristas dirigir, rotinas da população ou dos tipos

[1] https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city
[2] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Cidade	Fonte de dados	Veículos	Granularidade	Duração	Tipo Tráfego	Trabalhos
São Paulo	AVL	mais de 14000	45s		ônibus	(WEN et al., 2018) (YAI, 2016)
São Francisco	GPS	500	média 39s	30 dias	táxi	(MARTINS; CUNHA, 2018) (DOMINGUES; SILVA; LOUREIRO, 2018)
Shanghai	GPS	2500	30s	24h	táxi	(WU; ZHU; LI, 2011)
Beijing	GPS	20000	30s	24h	ônibus	(WENG et al., 2016)
Roma	GPS	320	7s	30 dias	táxi	(ALVARENGA et al., 2014)
Seattle	AVL	765	26-120 s	16 dias	ônibus	(DOERING; WOLF, 2015)
Chicago	GPS	1647	20-40 s	18 dias	ônibus	(DOERING; WOLF, 2015)
New York	GPS	4.5 bilhões de *pickups* e 19 milhões de viagens			Uber	
New York	GPS	1,1 bilhão de viagens de táxi			táxi	
Cologne				24h	sintético	(POLAT; SOYTURK, 2016) (UPPOOR et al., 2014) (UPPOOR; FIORE, 2012)

Tabela 2.1: *Datasets* de mobilidade

de veículos, padrões de horários). Já os traços sintéticos, podem apresentar essas características, mas através da calibragem e do refinamento do processo que os gerou (simulador, modelo matemático, pesquisa).

O uso de traços reais nas simulações de redes veiculares é importante, pois podem: fazer parte do processo de validação dos resultados das simulações e de protocolos; servir de comparação e ajuste de modelos de mobilidade e *datasets* sintéticos; incorporar dinâmicas reais na simulação para que o projeto de protocolos de roteamento reajam a comportamentos reais da movimentação de veículos.

2.4 Caracterização de datasets de mobilidade veicular

A mobilidade de um veículo é caracterizada pela topologia das vias, alta velocidade, mudanças repentinas de direção e aceleração, e densidade variáveis nas vias dependendo da hora do dia e da área de localização. Entender essas e outras características de conjuntos de dados de mobilidade veicular é essencial para o planejamento e desenvolvimento de protocolos para as redes veiculares cuja algumas das principais preocupações são a mobilidade e densidade de nós (UPPOOR; FIORE, 2012) (POLAT; SOYTURK, 2016) (DOERING; WOLF, 2015). Uppoor e Fiore (2012) adverte que a maior parte dos trabalhos em VANETs, como o desenvolvimento e testes de protocolos, tem considerações simplistas acerca da mobilidade dos veículos: mobilidade unidimensional, contato atômico entre veículos, densidade de nós constante ao longo do tempo, fluxos de tráfego aleatórios, tempo exponencial de chegada e de permanência dos veículos nas regiões.

Em relação à caracterização de conjuntos de dados de traços reais de mobilidade, Domingues, Silva e Loureiro (2018) ressaltam que é uma grande oportunidade para realizar estudos de viabilidade e identificar comportamentos de mobilidade que podem aperfeiçoar redes veicu-

lares, evitando também considerações simplistas sobre a movimentação dos veículos. Doering e Wolf (2015) acrescentam que o estudo da mobilidade a partir de traços reais de mobilidade podem aprimorar a acurácia e confiança das simulações. Além disso, esses autores apontam que os traços de mobilidade podem apresentar características e situações variáveis de tempo de contato entre veículos permitindo serem aplicadas no desenvolvimento e avaliação de protocolos em VANETs.

Algumas das aplicações da caracterização de *datasets* de mobilidade veicular em VANETs são:

- estudo de viabilidade de implantação de VANETs (DOMINGUES; SILVA; LOUREIRO, 2018);
- compreensão de que fatores afetam as condições do tráfego (ex: horário e engarrafamento) (DOMINGUES; SILVA; LOUREIRO, 2018);
- configurar e conduzir simulações realistas (DOERING; WOLF, 2015);
- determinar o potencial das transferências de dados entre veículos (UPPOOR; FIORE, 2012);
- estimar futuras oportunidades de comunicação entre carros em movimento (UPPOOR; FIORE, 2012);
- planejamento da implantação de infraestrutura de rede na via de tráfego (UPPOOR; FIORE, 2012);
- alocação dinâmica de recursos para usuários móveis em redes de acesso (UPPOOR; FIORE, 2012);
- estudo sobre protocolos e outras aplicações.

2.4.1 Pré-processamento de traços reais de mobilidade

Os sensores de localização, como o GPS, inevitavelmente possuem erros e, ocasionalmente, incluem ruídos devido a fatores como sinais de posicionamento ruins em ambientes urbanos (ZHENG; ZHOU, 2011). Portanto, traços reais de mobilidade captados por esses sensores não são totalmente acurados, sendo necessário pré-processar esses dados antes de realizar qualquer estudo, como a caracterização da mobilidade de objetos (ZHENG, 2015).

Zheng (2015) aponta que um dos ruídos mais comuns que os sensores de localização podem ter é quando as coordenadas registradas de um veículo desviam-se em muitos metros da

trajetória que está sendo desempenhada, causando descontinuidade e acarretando dificuldade em derivar informações úteis e acuradas, como a velocidade, como é apresentado na Figura 2.2. Os pontos p_5, p_{10}, p_{11} e p_{12} são pontos de ruído da trajetória que interferem na extração de informações, como as velocidades v_1, v_2, v_3, v_4 e v_5. Esse tipo de ruído pode ser ocasionado, por exemplo, devido ao sinal fraco do GPS dentro de túneis. Há estudos que descrevem os passos para o pré-processamento desse e de outros tipos de ruídos.

No trabalho de Zheng et al. (2017), o primeiro passo do pré-processamento dos traços de mobilidade de táxis é eliminar os registros cujas coordenadas não estejam dentro dos limites de latitude e longitude da área analisada. Considerando a taxa de amostragem para registrar a posição de um veículo entre 15 segundos a 20 segundos do trabalho, calculou-se que um veículo teria de 3000 a 5000 registros por dia, se num determinado dia um táxi tivesse acima desse valor, esse conjunto de dados seria eliminado. O último passo do trabalho foi filtrar os pontos das trajetórias baseando-se num limite de velocidade. Se restasse menos que um terço dos pontos de uma trajetória após a filtragem, então a trajetória seria eliminada.

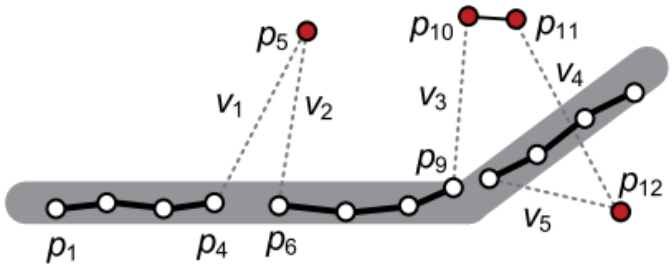


Figura 2.2: Pontos de ruído numa trajetória segundo Zheng (2015).

Em algumas situações os erros e ruídos apresentados pelos sensores de localização são aceitáveis, como em determinar em que cidade uma pessoa está. Entretanto, outras vezes, quando o trabalho leva em conta a trajetória do veículo, os ruídos necessitam ser filtrados a fim de diminuir os erros nas medições (ZHENG; ZHOU, 2011). Para filtrar ruídos, Zheng e Zhou (2011) e Zheng (2015) apontam alguns tipos de filtros:

- **Filtro por média ou mediana:** estimativa das coordenadas baseadas na média ou mediana dos pontos adjacentes ou anteriores a determinado ponto - janelas deslizantes. Além disso, pode ser interpretado como uma *sliding window* o qual um ponto só é incluído na trajetória se o erro acumulado é menor do que o especificado pelo limite especificado pelo usuário. Este filtro é indicado para ruídos em pontos individualmente, como o p_5 na Figura 2.2. Caso haja muitos ruídos consecutivos, como p_{10}, p_{11} e p_{12} na Figura 2.2, pode perpetuar erros ao longo da trajetória. Outro momento em que este filtro não é indi-

cado é quando a taxa de amostragem é muito baixa, ou seja, a distância entre dois pontos consecutivos na trajetória é muito alto;

- **Filtro de Kalman:** modela o ruído e a dinâmica da trajetória, estimando valores com base nessa modelagem e que obedecem as leis da física. Este filtro pode estimar valores de nível superior com base no movimento, como a velocidade e aceleração. Em suma, o filtro faz distinção do que está sendo medido e do que está sendo estimado, formulando uma relação linear entre eles e dependendo de valores anteriores para estimar o atual.
- **Filtro de partículas:** é similar ao filtro de Kalman que utiliza um modelo de medição e um modelo dinâmico. O filtro de Kalman ganha eficiência assumindo modelos lineares adicionados de ruídos gaussianos. Já o filtro de partícula relaxa essas suposições para um algoritmo mais geral. Este filtro recebe o nome de partículas, pois cada uma delas representa um estado estimado. Existe um conjunto de partículas sempre que uma medida se torna disponível. Quando as partículas são combinadas representam a distribuição de probabilidade dos possíveis estados. Uma desvantagem deste filtro é o tempo de computação que é afetado pelo número de partículas. Um vantagem é ser um filtro que consegue descrever mais estados, podendo adicionar mais variáveis do que no filtro de Kalman.
- **Detecção baseada em heurística:** os filtros dos tópicos anteriores substituem os valores de ruído da trajetória com valores estimados, este método remove os pontos de ruído diretamente da trajetória utilizando algoritmos de detecção de ruído. Por exemplo, um algoritmo de detecção baseado em velocidade. Primeiro calcula-se a velocidade de deslocamento entre os pontos de uma trajetória baseado na distância e no *timestamp*, então são eliminados todos os segmentos cuja velocidade está acima de um limite especificado. Na Figura 2.2, poderiam ser eliminados os segmentos $p_4 \rightarrow p_5$ e $p_5 \rightarrow p_6$, dependendo do limite imposto. Estes algoritmos podem ser aplicados em baixo nível, diretamente em coordenadas e *timestamp*, ou em alto nível nas trajetórias, como velocidade e aceleração.

A taxa de amostragem é o intervalo de tempo entre pontos consecutivos numa trajetória, ou seja, o tempo médio de registro de um ponto em uma trajetória. Se essa taxa é alta - o tempo de registro é pequeno (ex: a cada 1, 5 ou 15 segundos) - pode provocar grande volume de traços de mobilidade e/ou redundância de pontos, ou seja, registrar consecutivamente as mesmas coordenadas em tempo diferentes (ZHENG; ZHOU, 2011). Se a taxa de amostragem é baixa (ex: 2 ou 5 minutos), pode criar lacunas de muitos metros na trajetória, interferindo na extração de informações acuradas, como a velocidade. Além disso, essas lacunas também podem ser produzidas ao filtrar ruídos. A redundância de pontos na trajetória pode ser eliminada com filtros, já as lacunas são preenchidas por algoritmos, como o de clusterização de Silva et

al. (2015), e no trabalho de Xu, Li e Chen (2018) que utiliza a interpolação de pontos baseado no tempo consecutivo de duas coordenadas.

Outra fase que pode haver no pré-processamento de traços de mobilidade é fase de *map matching* cujo objetivo é mapear os traços de mobilidade para a sequência correta de ruas ou estradas. Esse processo pode ser necessário devido aos desvios produzidos nos pontos de uma trajetória decorrentes dos ruídos e erros dos sensores de localização. O *map matching* é indicado para situações em que deseja-se: saber em que rua um veículo esteve para planejamento do fluxo do tráfego, prever onde um veículo está indo, detectar os caminhos mais frequentemente utilizados por motoristas, entre outros.

Zheng (2015) menciona que os principais algoritmos de *map matching* pode ser baseados em considerações geométricas, topológicas, probabilísticas, dentre outros métodos. Para realizar o *map matching* existem trabalhos que utilizam *Hidden Markov Model* (HMM, em português Modelos Ocultos de Markov) que encontra a rota mais provável considerando medição de ruído e a topologia das vias de tráfego, podendo ser aplicado em cenários esparsos ou taxa de amostragem baixa, ou seja, há grande distância entre os pontos das trajetórias (NEWSON; KRUMM, 2009) (XU; LI; CHEN, 2018). Outros algoritmos mapeiam a trajetória para a via mais próxima utilizando mapas e topologias disponíveis (ex: Open Street Maps), como é o trabalho de Domingues, Silva e Loureiro (2018). Já Weng et al. (2016) utiliza informações de planejamento do transporte público de ônibus para realizar o *map matching* dos traços do ônibus com a parada mais próxima para calcular a velocidade média com base em distâncias conhecidas.

É importante salientar que nem todos os processos descritos nesta subseção sobre pré-processamento de traços de mobilidade são necessários. A escolha dos processos depende do tipo de estudo a ser realizado a partir deles. Por exemplo, se o objetivo for estudar os pontos de origem e destino e identificar as ruas utilizadas para trafegar, enquanto os dados possuem alta esparsidade ou baixa taxa de amostragem, talvez seja necessário filtrar ruídos e realizar o *map matching*. Em casos onde deseja-se apenas identificar pontos de interesse, sem interesse no trajeto, então talvez somente o filtro de ruídos seja requisitado. O pré-processamento de traços de mobilidade tem como objetivo descartar pontos da amostra sem sacrificar a qualidade dos dados de trajetória, gerando trajetórias acuradas e próximas da real (ZHENG; ZHOU, 2011).

Neste trabalho foram utilizados filtros de média e heurística, o processo de *map matching*, e filtro do número registros. Todos esses processos são abordados e aprofundados no Capítulo 4.

2.4.2 Processo de caracterização de datasets de mobilidade veicular

Para realizar a caracterização de *datasets* de mobilidade veicular através de traços reais de mobilidade, o primeiro passo é a coleta de dados. É preciso obter os traços de mobilidade da região desejada. Após a coleta, esses dados são explorados para identificar: o formato em que informações estão disponíveis (latitude, velocidade, aceleração), qual o número de registros e tamanho do arquivos. Em seguida, realiza-se os pré-processamentos dos traços necessários dependendo do objetivo do estudo, como dito na Subseção 2.4.1.

Depois do pré-processamento, é feita a extração de métricas que são estabelecidas de acordo com os objetivos do estudo. Os resultados das métricas podem ser observados através de recursos gráficos, e descrição de diferentes períodos ou quantis do dado. Os gráficos geralmente utilizados para demonstrar as métricas são gráficos temporais e mapas, como pode ser visto em Uppoor et al. (2014) e Doering e Wolf (2015).

2.4.3 Métricas para caracterização de datasets de mobilidade veicular

As métricas utilizadas para caracterizar um conjunto de dados de mobilidade veicular real ou sintético podem ter foco na mobilidade dos veículos, e/ou na conectividade entre eles, como é o caso de trabalhos que desenvolvem soluções para VANETs. A Tabela 2.2 demonstra quais métricas de conectividade são utilizadas para caraterização de traços de mobilidade e em quais trabalhos são utilizadas. Já a Tabela 2.3 mostra as métricas de mobilidade.

Nome	Descrição	Trabalhos
Grau de conectividade médio de veículos	Com quantos veículos um veículo tem conexão em determinado momento	(ALVARENGA et al., 2014) (UPPOOR et al., 2014) (SANTANA; KANASHIRO; KON, 2018)
Encontros repetidos ao longo do dia	Quantas vezes dois veículos se comunicaram ao longo do dia	(MARTINS; CUNHA, 2018)
Encontros repetidos ao longo do dia na mesma localização	Quantas vezes dois veículos se comunicaram ao longo do dia na mesma posição geográfica	(MARTINS; CUNHA, 2018)
Taxa de repetição de encontros	De todos os encontros que um veículo teve no dia, quantos foram repetidos	(MARTINS; CUNHA, 2018)
Duração do contato entre os veículos	Quanto tempo em média dura a conexão entre dois veículos	(UPPOOR; FIORE, 2012)
Tempo de residência em áreas do mapa	Quanto tempo os veículos permaneceram nas áreas demarcadas no mapa	(POLAT; SOYTURK, 2016)(UPPOOR; FIORE, 2012)
Veículo mais conectado ao longo do tempo	Ao longo das horas qual o número de conexões do veículos mais conectado a outros veículos	(POLAT; SOYTURK, 2016)
Número de grupos de veículos se comunicando	Um grupo é dois ou mais veículos conectados em determinado momento. Essa métrica diz quantos grupos existem	(UPPOOR et al., 2014)
Tamanho dos grupos se comunicando	Quantos veículos estão dentro de cada grupo da métrica do item anterior	(UPPOOR et al., 2014)

Tabela 2.2: Métricas de conectividade

Nome	Descrição	Trabalhos
Número de veículos ativos	Número de veículos que estão trafegando. Os trabalhos demonstram essa métrica ao longo do dia, por região, densidade de veículos por zona no mapa. Utilizada tanto em trabalhos de simulação e geração de *datasets* quanto em trabalhos de caracterização de traços de mobilidade	(ALVARENGA et al., 2014) (POLAT; SOYTURK, 2016) (YAI, 2016) (UPPOOR; FIORE, 2012) (DOERING; WOLF, 2015) (SANTANA; KANASHIRO; KON, 2018) (UPPOOR et al., 2014)
Velocidade média dos veículos	Velocidade média dos veículos ao longo das horas, por região, ou num panorama geral. Métrica utilizada tanto para traços sintéticos quanto traços reais, e também em simulações	(SILVA, 2010) (CAMPOS; MORAES; SILVA, 2010) (ALVARENGA et al., 2014) (POLAT; SOYTURK, 2016) (YAI, 2016) (UPPOOR; FIORE, 2012) (DOERING; WOLF, 2015) (SANTANA; KANASHIRO; KON, 2018) (UPPOOR et al., 2014)
Duração média de viagens	Quanto tempo em média dura a viagem de táxis, ônibus ou veículos, ou seja, por quanto tempo eles trafegam. Métricas encontradas em trabalhos de geração de *datasets* sintéticos e simulação	(WEN et al., 2018) (SANTANA; KANASHIRO; KON, 2018) (UPPOOR et al., 2014)
Distância média entre veículos	Quantos metros em média um veículo está do outro	(POLAT; SOYTURK, 2016)
Distância e tempo entre os pontos consecutivos de um traço de mobilidade	Qual o tempo e a distância média entre as posições consecutivas num traço de mobilidade	(DOERING; WOLF, 2015) (YAI, 2016)
Quantas linhas um ônibus esteve associado em um dia	Quantas linhas um ônibus atendeu em um dia	(DOERING; WOLF, 2015)
Número de viagens por dia	Número de viagens que ocorreram num dia ou numa hora. Está relacionado à simulação de VANETs	(SANTANA; KANASHIRO; KON, 2018)
Distância total da viagem	Quantos *km* em média um veículo ficou trafegando. Está relaciona a trabalhos de geração de *traces* sintéticos	(SANTANA; KANASHIRO; KON, 2018) (UPPOOR et al., 2014)

Tabela 2.3: Métricas de mobilidade

2.5 Ferramentas de processamento de traços de mobilidade

Esta seção descreve ferramentas que podem armazenar e/ou processar conjuntos de dados de traços reais de mobilidade. Elas armazenam grande volume de dados e possuem métodos para consultar e agregar dados com base em coordenadas geográficas e dados temporais. As ferramentas testadas foram: MongoDB, PostgreSQL com extensão do PostGIS, Google BigQuery, e Apache Spark.

Nesta pesquisa, a ferramenta Spark foi selecionada para tratar os dados e extrair as métricas dos traços de mobilidade. Nas subseções a seguir, além do Apache Spark, algumas funcionalidades das ferramentas testadas foram documentadas para que possam ser aplicadas em trabalhos futuros.

2.5.1 MongoDB

O MongoDB[3] ou Mongo é um banco de dados NoSQL orientado a documentos. Os registros (documentos) são salvos em um formato similar ao JSON e não possuem esquema fixo

[3]https://www.mongodb.com/

(CHODOROW, 2013).

Em relação à representação de coordenadas geográficas, o MongoDB suporta dados no formato GeoJSON que é um padrão de representação de formas espaciais, como ponto, polígono, e *linestring* (conjunto pontos). O documento salvo no banco deve possuir ao menos um campo que esteja no formato GeoJSON, como mostra a Figura 2.3. Neste caso, os campos *geometry_1* está no formato GeoJSON do tipo *Point*, o campo *geometry_2* é do tipo *Linestring* e o campo *geometry_3* é do tipo *Polygon* . Outros tipos de objetos GeoJSON suportados pelo Mongo podem ser encontrados na documentação oficial[4].

```
{
  "name": "ExemploDocumento",
  "geometry_1": {
    "type": "Point",
    "coordinates": [125.6, 10.1]
  },
  "geometry_2": {
    "type": "LineString",
    "coordinates": [ [ 40, 5 ], [ 41, 6 ] ]
  },
  "geometry_3":{
    "type": "Polygon",
    "coordinates": [[[ 0 , 0 ] , [ 3 , 6 ] , [ 6 , 1 ]]]
  }
}
```

Figura 2.3: Exemplo de documento com campos GeoJSON (elaborado pela autora).

Para consultar documentos que possuam campos no formato GeoJSON, o Mongo possui os seguintes operadores (CHODOROW, 2013):

- $*geoIntersects*: seleciona documentos que intersectam o ponto passado como parâmetro;
- $*geoWithin*: seleciona documentos dentro de uma área dada;
- $*near*: seleciona todos os documentos cujo os pontos estão dentro do intervalo de metros passado. A Figura 2.4 mostra uma consulta de documentos pertencentes à coleção *places*. A consulta irá retornar todos os documentos cujas coordenadas (*Point*) estejam no mínimo a 1000 metros e, no máximo a 5000 metros da coordenada geográfica -73,9667 e 40,78;
- $*maxDistance*: limite máximo, em metros, que uma coordenada pode estar para ser retornada pelo $*near*;
- $*minDistance*: limite minímo, em metros, que uma coordenada pode estar para ser retornada pelo $*near*.

[4]https://docs.mongodb.com//manual//reference//geojson//index.html

```
db.places.find(
   {
     location:
       { $near :
          {
            $geometry: {
              type: "Point",
              coordinates: [ -73.9667, 40.78 ]
            },
            $minDistance: 1000,
            $maxDistance: 5000
          }
       }
   }
)
```

Figura 2.4: Consulta de documento utilizando o operador $*near* (elaborado pela autora).

O MongoDB é um banco de dados escalável. Devido ao modelo de dados organizado em documentos há facilidade em dividir os documentos em múltiplas máquinas (*shards*). O *sharding* é um mecanismo que distribui dados e balanceia consultas em múltiplas máquinas num *cluster*.

2.5.2 PostgreSQL

O PostgreSQL[5] (ou Postgres) é um sistema gerenciador de banco de dados objeto relacional cuja estrutura é baseada em tabelas (relações) compostas de linhas e colunas (The PostgreSQL Global Development Group, 2019). Este banco utiliza a linguagem de pesquisa SQL para realizar consulta de dados.

Em relação a representação e consulta de coordenadas geográficas, o Postgres possui uma extensão chamada PostGIS[6] que adiciona o suporte para objetos geográficos (geometrias ou *geometry*), como coordenadas, permitindo ao SQL executar consultas deste tipo de dado. Os objetos geográficos adicionados pelo PostGIS são (POSTGIS, 2019b): *Point* que representa uma localização única na Terra; *Linetring* é uma sequência de pontos que representa um caminho entre esses pontos (ex: ruas e estradas); *Polygon* é a representação de uma área (polígono); *Collections* agrupam múltiplas geometrias (ex: múltiplos pontos, múltiplos *linestrings*, múltiplos *polygons*). Neste banco de dados, as geometrias são representadas no formato WKT(*Well-known text*) que é linguagem de marcação para geometrias. Neste tipo de representação, uma coordenada geográfica seria representada, por exemplo, como $POINT(12.3\ \ 25.0)$.

O PostGIS possui funções específicas para consultar as geometrias e que auxiliam no pro-

[5]https://www.postgresql.org/
[6]https://postgis.net/

cessamento de trajetórias, como (POSTGIS, 2019a):

- **ST_Intersection:** retorna um ponto ou polígono que representa a intersecção entre pontos ou polígonos;
- **ST_DWithin:** retorna verdadeiro se a distância entre dois pontos é igual ou menor a uma certa distância, caso contrário, retorna falso;
- **ST_Distance:** calcula a distância entre dois pontos geográficos.
- **ST_BUFFER:** Retorna um polígono que representa todos os pontos dado um ponto e um raio de distância.

A Figura 2.5 apresenta traços de mobilidade numa tabela do Postgres cujos campos são: *bus_id* representa o identificar de um ônibus, *register_date* é a data de registro, *long* é a longitude, e *lat* é a latitude. Nessa figura, há dois ônibus com os identificadores 111 e 222, e para cada um deles são exibidos 3 coordenadas de suas trajetórias. A Figura 2.6 consulta os registros da tabela *traces* (Figura 2.5) que estejam localizados até 95 metros da coordenada cuja longitude é -23,652 e latitude é -46,776 utilizando a função *ST_DWithin*. O resultado desta consulta está na linha 10-13 da Figura 2.6. O valor *true* na linha 7 significa que utiliza-se o sistema de coordenadas geográficas esferoidais na função *ST_DWithin*.

```
--> Tabela Completa
 bus_id |      register_date      |    long    |    lat
--------+-------------------------+------------+------------
    111 | 2016-01-01 04:00:34     | -23.652967 | -46.776522
    222 | 2016-01-01 04:00:49.347 | -23.651398 | -46.612833
    111 | 2016-01-01 04:00:36.697 |  -23.65232 | -46.776375
    222 | 2016-01-01 04:00:36.67  |   -23.6503 | -46.613433
    111 | 2016-01-01 04:00:36.113 | -23.652047 | -46.776865
    222 | 2016-01-01 04:00:46.887 | -23.651065 | -46.612208
```

Figura 2.5: Exemplo de traços de mobilidade numa tabela do Postgres (elaborado pela autora).

```
SELECT *
  FROM traces
  WHERE ST_DWITHIN(
    ST_MAKEPOINT(long, lat),
    ST_MAKEPOINT(-23.652, -46.776),
    95,
    true
  );

 bus_id |      register_date      |    long    |    lat
--------+-------------------------+------------+------------
    111 | 2016-01-01 04:00:34     | -23.652967 | -46.776522
    111 | 2016-01-01 04:00:36.697 |  -23.65232 | -46.776375
```

Figura 2.6: Consulta de coordenadas geográficas utilizando funções do PostGIS (elaborado pela autora).

Para garantir um maior desempenho, o Postgres pode paralelizar as consultas em múltiplas CPUs o que é recomendado, principalmente, para consultas em largos conjuntos de dados que retornam poucas linhas para o usuário.

2.5.3 Google BigQuery

O Google BigQuery é um serviço baseado no sistema de consultas e análise de dados interativo Dremel (MELNIK et al., 2010). O BigQuery é um serviço de análise de dados em grande escala que não requer infraestrutura, provisionando os recursos em segundo plano. Os dados podem ser analisados em lote ou podem prover de *streaming* (fluxos) de dados. As consultas a esses dados são feitas através de SQL (em conformidade com a ANSI) por meio de uma interface web do Google ou por *drivers* através de linguagens, como Python, Java, etc.

O BigQuery GIS analisa e processa dados geoespaciais com os tipos de dados geográficos e funções geográficas do SQL, incluindo pontos, linhas, polígonos e multipolígonos no formato WKT, como as do Postgres vistas na Subseção 2.5.2, e também dados no formato GeoJSON, como pode ser visto no MongoDB (Figura 2.3).

As funções geográficas do BigQuery são similares às do Postgres, como, por exemplo, a *ST_DWithin* que retorna verdadeiro caso a distância entre dois pontos seja menor ou igual a distância especificada. Na Figura 2.7, utiliza-se a função *ST_DWithin* para retornar todas as linhas da tabela de traços de modalidade cuja as coordenadas estejam a menos de 95 metros do ponto -23,652 e -46,776 segundo o sistema de coordenada de esfera. A consulta foi realizada pela interface web do BigQuery.

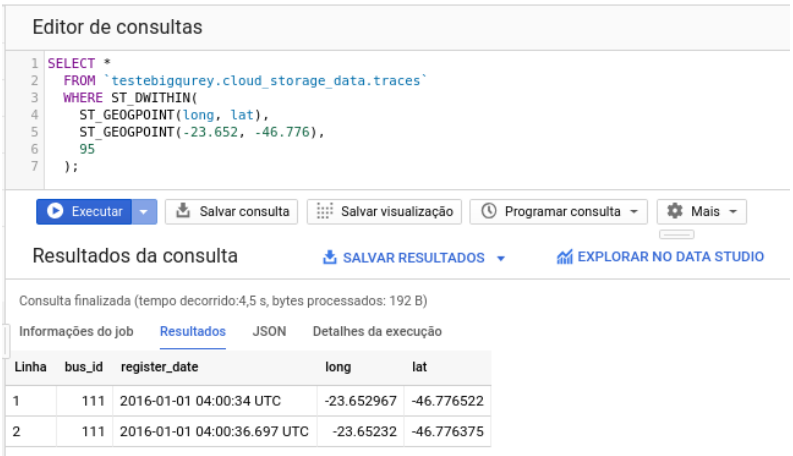


Figura 2.7: Consulta de coordenadas geográficas utilizando funções do BigQuery (elaborado pela autora).

O BigQuery pode ser escalável para lidar com petabytes de dados através do escalonamento

elástico, armazenamento gerenciado em colunas, e execuções em paralelo.

2.5.4 Apache Spark

Apache Spark é um mecanismo *open source* distribuído para análise de dados em larga escala em *batch* ou *streaming* (APACHE, 2019). A arquitetura do Spark é baseada em execução distribuída em *clusters*, como pode ser visto na Figura 2.8. Uma aplicação (*driver program*) inicializa o SparkContext que chama os funções e recursos do Spark que utiliza o Cluster Manager para se conectar-se aos nós. O Cluster Manager administra os nós *workers* do *cluster*. Cada nó *worker* executa as tarefas designadas e possui um cache para guardar os resultados e acelerar a leitura.

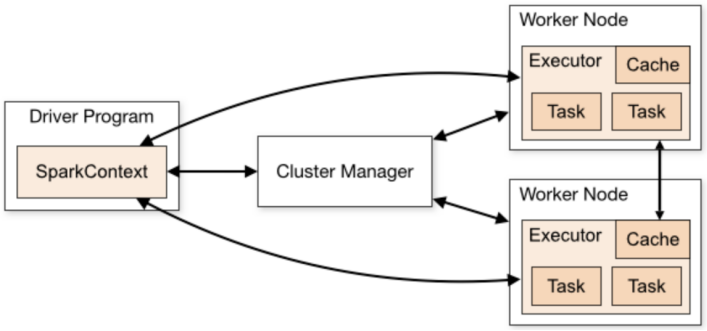


Figura 2.8: Componentes de um *cluster* Spark (APACHE, 2019)

A abstração principal do Spark é o *resilient distributed dataset* (RDD) que é uma coleção de objetos imutáveis particionados através de um conjunto de máquinas (ZAHARIA et al., 2010). Os RDDs podem ser armazenados na memória entre máquinas e reutilizado em várias operações paralelas do tipo MapReduce. Para exemplificar o funcionamento do RDD, pode-se seguir o exemplo do trabalho de Zaharia et al. (2016) visto na Figura 2.9. Na linha 1, a fonte de dados é definida e, internamente, define-se um RDD para representar o arquivo de texto que é um conjunto de linhas. Na linha 2, cria-se outro RDD internamente que está fazendo uma operação de transformação *filter* para filtrar cada linha do arquivo que comece com a palavra ERROR. Na linha 3, chama-se uma operação ação *count* que conta e retorna ao programa quantos elementos existem no RDD anterior. O Spark avalia os RDD de forma ”preguiçosa”para que formule um plano otimizado para realizar as computações requisitadas. O Apache Spark formula um grafo com todas as operações de transformação e ação para estabelecer um plano de execução eficiente, por exemplo, utilizando funções que podem ser paralelizadas (ex: *filter*) e de que forma distribuir, ler e processar os dados no *cluster*. Portanto para processar dados em larga escala de forma eficiente e rápida, o Spark utiliza seus planos de execução, computação distribuída, grafos, leituras em memórias nos nós *workers*, entre outras funcionalidades.

```
1   lines = spark.textFile(“hdfs://...”)
2   errors = lines.filter(  s => s.startsWith(“ERROR”))
3   print(“Totalerrors:“+errors.count())
```

Figura 2.9: Exemplo de código no Spark (ZAHARIA et al., 2016).

Para programar e acessar o Spark, é possível utilizar diferentes linguagens, como: Python, R, Scala e Java. No caso do Python, utiliza-se o pacote PySpark. O Spark pode ser utilizado para: contagem de palavras em arquivos, pesquisa de texto, processamento de dados em streaming, deteçcão de anomalias em *streaming* de dados, entre outras aplicações.

O Spark por si só não possui funções geográficas ou tipo de dados específicos para lidar com coordenadas geográficas, portanto são necessárias bibliotecas adicionais. Neste trabalho, o Spark foi utilizado na versão do Python (PySpark) em conjunto com as bibliotecas do Python que dão suporte a dados geográficos, como o shapely[7] e o geopandas[8]. Mais detalhes sobre esta implementação podem ser encontrados no Capítulo 4.

2.6 Conclusão do capítulo

Este capítulo apresentou os conceitos necessários para o entendimento da pesquisa: método de manipulação de dados, entendimento do tipo do dado que é caracterizado, como é o processo de caracterização de um *dataset* de mobilidade veicular e quais métricas podem ser utilizadas para isso. A caracterização do conjunto de dados de mobilidade pode ajudar no desenvolvimento de soluções de VANETs, como: teste e refinamento de protocolos, configuração de simulações de redes veiculares, ajuste nos parâmetros de modelos de mobilidade, estudo de viabilidade, dentre outros.

Foram exploradas ferramentas de dados que podem ajudar no processo de caracterização de traços reais de mobilidade. Para explorar, pré-processar e caracterizar os dados de mobilidade, a ferramenta Spark foi escolhida junto com bibliotecas de dados do Python. O Spark foi escolhido devido a disponibilidade de recursos para processar os dados de forma distribuída. Além disso, o Spark é escalável, consegue processar desde milhares de dados até *datasets* pequenos, podendo ser replicado tanto em ambiente local, como ambiente em nuvem.

[7] https://shapely.readthedocs.io/en/stable/manual.html
[8] https://geopandas.org/

Capítulo 3

TRABALHOS RELACIONADOS

Neste capítulo serão discutidos estudos que realizam algum tipo de caracterização, manipulam ou geram *datasets* de mobilidade de veículos. Os trabalhos deste capítulo são utilizados para a descoberta de métricas de caracterização, ferramentas e técnicas de manipulação de dados de mobilidade.

3.1 Geração e análise de um conjunto de dados de mobilidade urbana

Uppoor et al. (2014) ressaltam que a simulação é um mecanismo de teste de desempenho de redes veiculares, e que ela pode ser facilmente enviesada pelo *dataset* de mobilidade no qual é baseada, impactando na avaliação, por exemplo, de protocolos de roteamento. Para avaliar o impacto do realismo da representação da mobilidade veicular em simulações, Uppoor et al. (2014) geram um *dataset* sintético de tráfego de carros que cobre um janela de 24 horas e $400km^2$ com mais de 700 mil viagens da cidade de Colônia na Alemanha.

No trabalho de Uppoor et al. (2014), o conjunto de dados de tráfego de carros é gerado a partir de uma pesquisa do Institute of Transportation Systemsat the German Aerospace Center (ITS-DLR) sobre Origem e Destino da cidade de Colônia e a partir do simulador de tráfego SUMO (Simulation of Urban Mobility) que consegue simular o comportamento (ex: aceleração) de veículos para veículos individualmente ou de veículos para via. O OpenStreetMap (OSM)[1] é utilizado para fornecer a infraestrutura das vias e os mapas, possuindo alto nível de detalhes, como: vias principais, vias secundárias, sinalização, pontos de interesse, construções comerciais, entre outros. Durante o processo de geração, correções foram feitas no

[1] https://www.openstreetmap.org

conjunto de dados acerca do número de veículos e inconsistências das vias do OSM - cruzamentos, direção da ruas, acesso entre duas ou mais vias, informação inconsistente da sinalização.

O *dataset* gerado é descrito por Uppoor et al. (2014) através de métricas de mobilidade, que podem ser encontradas na Subseção 2.4.3. Um exemplo de métrica é a velocidade média dos veículos de acordo com as áreas (vias) da cidade, como pode ser visto na Figura 3.1.

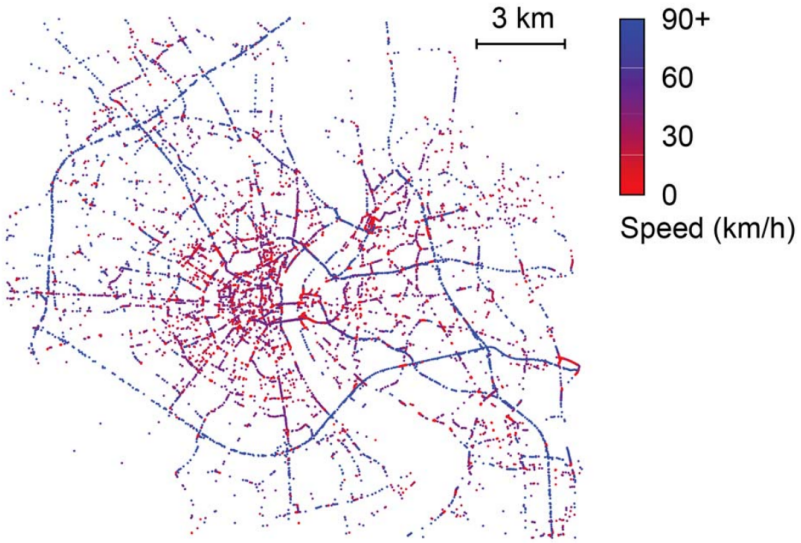


Figura 3.1: Velocidade média dos veículos em Colônia às 7:00 da manhã (UPPOOR et al., 2014).

Para validar o *dataset* gerado, Uppoor et al. (2014) comparam métricas de conectividade entre os veículos com outros conjuntos de dados de Zurique, Turin, e Colônia. Nesse caso, os autores compararam grupos (*clusters*) de veículos ao longo das horas para cada conjunto de dados, como pode ser observado na Figura 3.2.

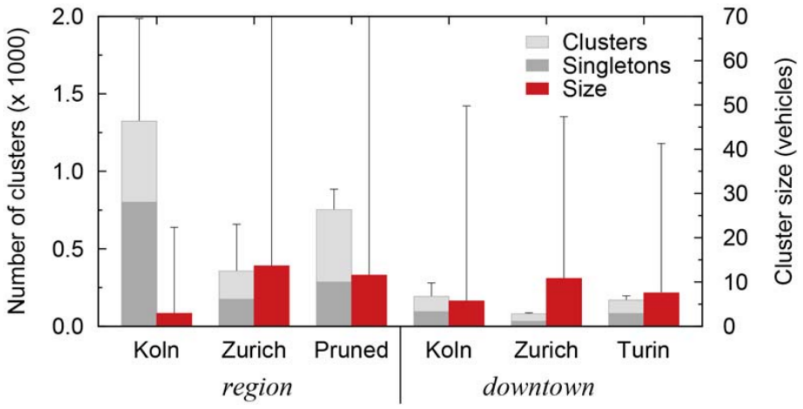


Figura 3.2: Média do número de *clusters* e do número de veículos por *cluster* (UPPOOR et al., 2014).

Para avaliar o impacto do realismo de diferentes *datasets* em redes veiculares, inclusive

o gerado, Uppoor et al. (2014) simulam uma rede veicular com disseminação epidêmica, extraindo métricas como o tempo que uma mensagem leva para atingir todos os nós da rede. Os autores ainda discutem como as características de cada conjunto de dados de mobilidade veicular influenciam nessas métricas de rede.

Esse trabalho contribui para identificação de métricas que podem ser aplicadas para caracterizar *datasets* de mobilidade veicular. Além disso, a pesquisa destaca de que modo características como velocidade, topologia de vias, janela de captura, número de veículos no conjunto de dados, podem afetar o comportamento da simulação de redes veiculares. Em relação ao tratamento de dados, Uppoor et al. (2014) demonstram como é necessário tratar a inconsistência de mapas e topologias, como foi o caso do OSM, e tratar os *datasets* fonte para configurar a simulação a fim de reduzir o viés na avaliação de redes veiculares.

3.2 Utilizando dados dos ônibus de São Paulo

Alguns trabalhos desta seção propõem a geração de dados e criação de modelo de mobilidade veicular a partir dos dados dos ônibus de São Paulo. Já outros trabalhos fazem uma análise mais técnica desses dados de mobilidade.

Santana, Kanashiro e Kon (2018) mencionam a dificuldade em encontrar rastros de mobilidade que correspondam ao comportamento de grandes cidades para o teste e experimentos com redes veiculares. Para contornar essa situação, os autores geram traços de mobilidade sintéticos a partir do simulador de cidade inteligentes InterSCSimulator [2].

O processo de geração consistiu em inserir dados de uma pesquisa de Origem e Destino de órgãos governamentais no InterSCSimulator juntamente com mapas e topologias provenientes do OSM. A saída da simulação foram traços de mobilidade sintéticos cobrindo cerca de $25km^2$ e 4 milhões de viagens de ônibus e carros. Algumas das métricas utilizadas para descrever o *dataset* gerado foram: duração das viagens ao longo das horas, distância percorrida pelos veículos ao longo das horas, número de veículos ativos em determinada hora do dia e a velocidade dos veículos ao longo das horas. Na Figura 3.3, a partir da observação dessas métricas, nota-se os picos de número de veículos, distância e duração de viagem estão próximos às 10 da manhã, enquanto as menores velocidades são encontradas no mesmo período. Isso pode acontecer devido ao alto tráfego e número de veículos nas vias.

Para validar os dados criados, eles foram inseridos no simulador de redes NS-3, simulando uma rede veicular com o protocolo Dedicated Short-Range Communications (DSRC) com um

[2] http://interscity.org/software/interscsimulator/

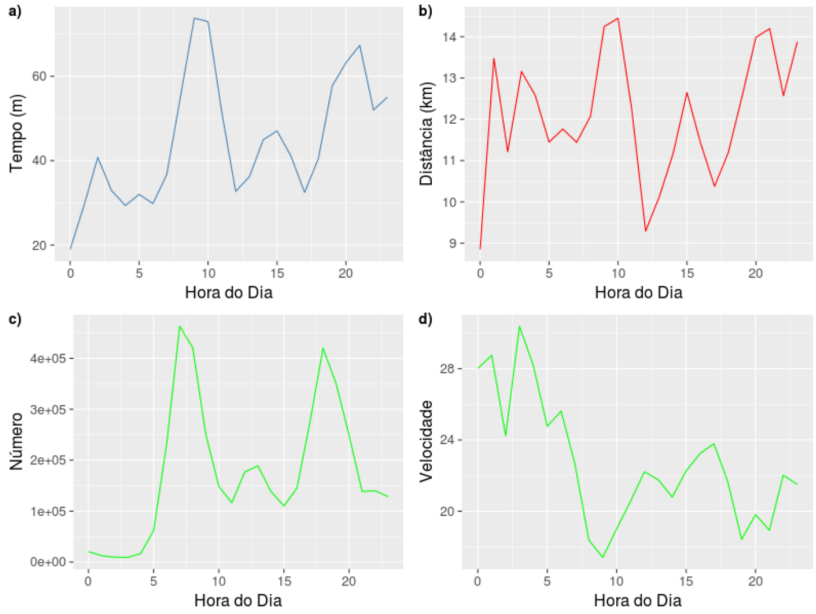


Figura 3.3: Análise das viagens de ônibus (SANTANA; KANASHIRO; KON, 2018).

raio de 300m para os carros.

Wen et al. (2018) processam dados de localização dos ônibus de São Paulo para melhorar o modelo de mobilidade do transporte público utilizado pelo InterSCSimulator. O modelo criado é validado por meio de análises comparativas entre os comportamentos observados dos dados reais e da simulação. Os dados utilizados para gerar o modelo foram os de GTFS (*General Transit Feed Specification*) - linhas de ônibus, itinerários, localização das paradas e o trajeto de uma parada para outra - e os de AVL (*Automatic Vehicle Location*) que representam o comportamento real do veículo - horário de início de circulação, frequência de saídas e velocidade média de deslocamento. O modelo baseou-se em viagens de 14.139 de ônibus em 7 dias.

Os trabalhos de Wen et al. (2018) e Santana, Kanashiro e Kon (2018) fornecem métricas diretamente relacionadas com o deslocamento de ônibus. Além disso, os dois trabalhos indicam as motivações da escolha da janela de captura dos dados para atingir o comportamento homogêneo e não-homogêneo desse tipo de veículo. Apesar de fazerem um processo de validação, ambos os estudos não apontam possíveis problemas da simulação e como ela afetaria os resultados, que é um ponto abordado por Uppoor et al. (2014). Esses estudos focaram na caracterização de dados oriundos de simulação enquanto esta dissertação propõe a caracterização de um *dataset* de traços reais de ônibus. Através desses trabalhos foi possível também entender que aspectos da mobilidade são levados em conta e validados quando uma simulação de VANETs ou geração de dados sintéticos são feitos.

Outros trabalhos que exploram diretamente o mesmo tipo de *dataset* de mobilidade vei-

cular dos ônibus de São Paulo são Yai (2016) e Pons, Monteiro e Speicys (2015). Yai (2016) tem como objetivo avaliar a qualidade do serviço de transporte público de São Paulo através de informações do GTFS e AVL, e de conjunto de métricas, como: distância entre as paradas, velocidade média do veículo, quantidades de linha por ônibus, tempo previsto de viagem, quantidade de linhas que passam por cada trecho da cidade, dentre outros. Esse trabalho traz descrições de arquivos e campos tanto dos arquivos GTFS quanto AVL. Essas descrições foram essenciais para entendimento e exploração do *dataset* fornecido pela SPTrans e que será caracterizado nesta pesquisa.

O trabalho de Pons, Monteiro e Speicys (2015) é um relatório técnico produzido pela SPTrans e pela empresa Scipopulis com outras instituições para avaliar a qualidade dos dados de rastreamento dos ônibus de São Paulo sob um ótica de técnicas de *big data*. Esse trabalho contribuiu para o entendimento do *dataset* da SPTrans, dinâmica dos ônibus da cidade, e que tipos de pré-processamento poderiam ser empregados para seguir com a análise dos dados. Dentre as técnicas encontradas no trabalho e utilizados nesta pesquisa, está o de usar as informações de GTFS (planejamento das rotas e paradas) para pré-processar e analisar a mobilidade dos ônibus.

3.3 Caracterizando traços de mobilidade

Os estudos desta seção têm como objetivo caracterizar traços de mobilidade focando no contexto de VANETs. Nenhum processo de validação foi realizado por esses estudos, pois nenhum deles realizou testes ou simulações.

Martins e Cunha (2018) consideram a comunicação de VANETs fortemente influenciada pelos padrões de deslocamento de veículos, rotina dos motoristas e diferentes períodos do dia. Com o objetivo de entender melhor os encontros entre veículos para auxiliar no desenvolvimento de protocolos de redes veiculares, os autores caracterizam a mobilidade veicular de traços reais dos táxis de São Francisco através de quatro métricas: quantidade de encontros repetidos, quantidade de encontros que se repetem na mesma localização e a razão entre a quantidade de encontros que se repetem sobre a quantidade total de encontros para cada veículo.

O conjunto de dados analisado por Martins e Cunha (2018) consiste em viagens de 500 táxis ao longo de 30 dias, com taxa de amostragem (granularidade) a cada 1 minuto. Pela granularidade ser baixa, os autores constatam que pode haver lacunas nas trajetórias de cada viagem, influenciando a análise de encontros. Portanto, os autores decidem usar a versão calibrada do *dataset* cujas lacunas já haviam sido preenchidas.

Na Figura 3.4, Martins e Cunha (2018) mostram o número de encontros entre veículos con-

siderando um raio de 100m para alguns dias da semana. Nota-se menor número de encontros no Domingo. Em relação às outras métricas, os autores concluem que entre os dias da semana houve maior número de encontros; a repetição dos encontros é maior durante a semana devido às distâncias mais curtas desempenhadas pelos táxis e, porque no final de semana eles se distribuem mais pela cidade, prolongando assim suas rotas aos Domingos, por exemplo.

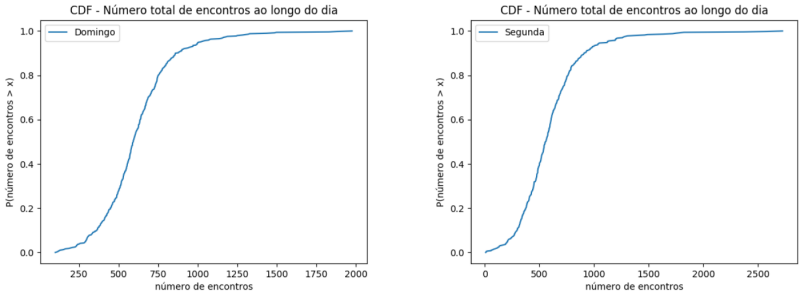
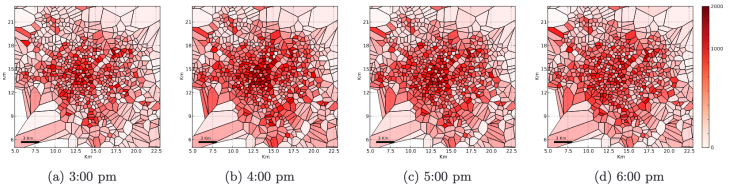


Figura 3.4: Total de encontros ao longo do dia para alguns dias da semana (MARTINS; CUNHA, 2018).

Uppoor e Fiore (2012) salientam que a dinâmica do tráfego afeta diretamente no desenho de soluções de redes veiculares. Para contribuir com o entendimento da dinâmica do tráfego em áreas metropolitanas, Uppoor e Fiore (2012) extraem características do conjunto de dados sintético da cidade de Colônia na Alemanha cobrindo um período de 24 horas e $400km^2$. A caracterização dos dados de deslocamento é dividida em macroscópica e microscópica. A primeira refere-se às métricas do comportamento dos veículos como um todo, como: densidade de veículos por área da cidade ao longo do dia e direção de destino ao longo das horas. Já a microscópica refere-se ao comportamento dos veículos individualmente ou interação entre eles, por exemplo: duração do contato entre veículos e tempo de permanência em cada área da cidade.

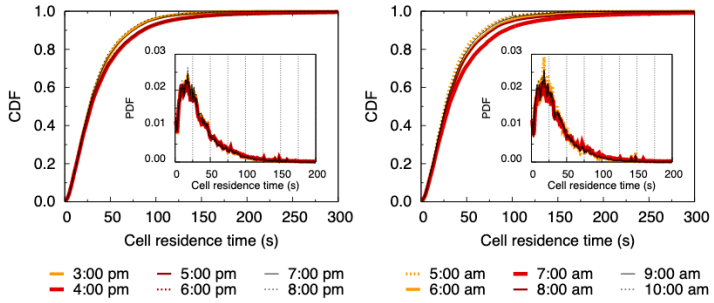
Na Figura 3.5, Uppoor e Fiore (2012) apresentam a distribuição geográfica da intensidade de veículos ao longo do período da tarde. Percebe-se que durante o meio da tarde, a maior concentração de veículos é na região central, enquanto que às 6 da tarde o tráfego começa a distribuir-se. Já na Figura 3.6, está representado o tempo de permanência dos veículos por região da cidade. Observa-se que a maior parte do carros permanecem menos de 50 segundos em cada célula do mapa.

Doering e Wolf (2015) destacam que entender as características especiais do transporte público, como os horários e as trajetórias definidas, é essencial para o desenho e avaliação de redes veiculares oportunísticas. Para tanto, os autores analisam e comparam dados de mobilidade do transporte público: a densidade de veículos, velocidade e intervalos de atualização de



(a) 3:00 pm (b) 4:00 pm (c) 5:00 pm (d) 6:00 pm

Figura 3.5: Evolução da intensidade de tráfego (UPPOOR; FIORE, 2012).



(a) Morning (b) Afternoon

Figura 3.6: Tempo de permanência dos veículos por região da cidade (UPPOOR; FIORE, 2012).

coordenadas. Os *datasets* utilizados foram os traços de mobilidade reais de ônibus de Seatle e Chicago nos Estados Unidos. O conjunto de dados de Seatle possui 765 ônibus e cobre um total de 16 dias com atualização de posição a cada 26-120 segundos. Já os traços de Chicago compreendem 1647 ônibus em 18 dias com uma taxa de amostragem a cada 20-40 segundos.

Ao comparar as métricas, Doering e Wolf (2015) explicam as diferenças entre as medições de Seatle e Chicago. A Figura 3.7 apresenta a distribuição de velocidades para cada cidade, ou seja, o número ônibus que possuem determinada velocidade. Os autores ressaltam que as velocidades são mais altas em Seatle devido a área de operação dos ônibus ser mais rural comparada a densa cidade como Chicago. Já na Figura 3.8, é mostrada a distância entre registros (pontos) consecutivos coletados para cada traço de mobilidade. Em Chicago, há maiores ocorrências de distâncias abaixo de 200 metros e, em Seatle, há mais veículos cujos pontos consecutivos do traço de mobilidade possuem mais de 450 metros. Essa diferença ocorre devido à taxa de amostragem ser maior em Chicago, ou seja, menor tempo de coleta entre os registros, enquanto em Seatle, esse tempo de coleta é maior, tendo maior lacuna de distância entre os registros de posição.

Entendendo as peculiaridades de cada *dataset*, e o porquê dos valores de métricas ocorrer, é possível identificar quais processos de pré-processamentos são necessários. Por exemplo, como preencher pontos de uma trajetória se a taxa de amostragem é muito baixa, ou seja, maior lacunas entre os registros de posição.

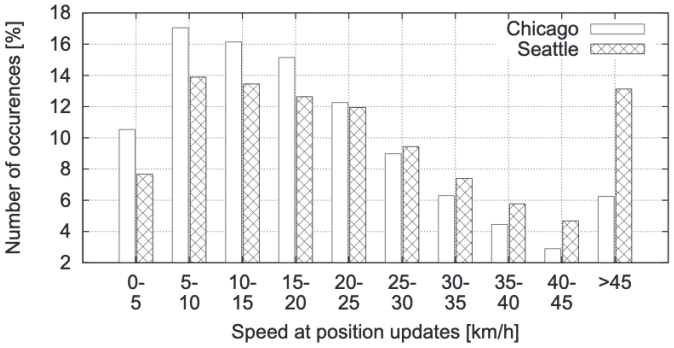


Figura 3.7: Distribuição de velocidades (DOERING; WOLF, 2015).

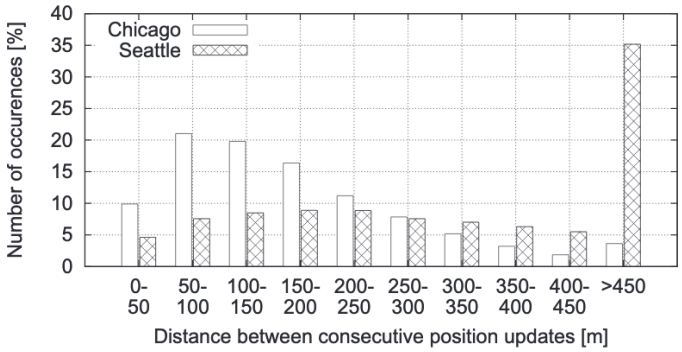


Figura 3.8: Distância entre atualizações de pontos consecutivos (DOERING; WOLF, 2015).

Polat e Soyturk (2016) consideram fundamental entender a mobilidade de veículos para o desenvolvimento de protocolos de roteamento para redes veiculares. Os autores apontam que é desafiador construir soluções para comunicações veiculares considerando a variabilidade na mobilidade de veículos - velocidades diferentes por conta de regiões diferentes, paradas por conta de semáforos, etc- e a variação da densidade de veículos dependendo da localização e do tempo. Para tentar contornar esses desafios, Polat e Soyturk (2016) propõem um método para análise da mobilidade veicular através de métricas espaço-temporais. Os autores dividem a cidade de Colônia em células, como um grafo, para entender métricas de mobilidade e conectividade dos *traces* analisados. As métricas que Polat e Soyturk (2016) extraem e que foram usadas como base nesta pesquisa são: densidade de veículo por região da cidade, velocidade média dos veículos, componentes mais conectados, média de distância entre veículos.

3.4 Síntese dos trabalhos relacionados

As principais motivações trazidas pelos trabalhos do porquê caracterizar ou analisar um conjunto de dados de mobilidade veicular são: desenvolver e avaliar soluções pra redes veicu-

lares de maneira acurada e realista, evitar viés nos testes ou simulações, conseguir incorporar a dinâmica real da movimentação de veículos em simulações de VANETs, e validar modelos ou *datasets* sintéticos gerados.

Todos os trabalhos elencados neste capítulo apresentaram métricas para caracterizar um *dataset* de mobilidade veicular e/ou validar dados gerados por simulação. Portanto, foi possível identificar métricas que podem ser utilizadas na caracterização de dados de mobilidade veicular. Além da caraterização dos conjuntos de dados através de métricas e gráficos de distribuição, a distribuição geográfica também foi utilizada para compreender os dados, como pode ser visto nas Figuras 3.5 e 3.1.

Capítulo 4

PRÉ-PROCESSAMENTO DO DATASET

Os sensores de localização, como o GPS ou AVL, inevitavelmente possuem erros, e ocasionalmente incluem ruídos devido a fatores como sinais de posicionamento ruins em ambientes urbanos (ZHENG; ZHOU, 2011). Portanto, traços reais de mobilidade captados por esses sensores não são totalmente acurados, sendo necessário pré-processar esses dados antes de realizar estudos, como a caracterização da mobilidade de veículos.

Zheng (2015) ressalta que os tipos de pré-processamento necessários para um *dataset* dependem do objetivo de cálculo. Por exemplo, se o objetivo é calcular quantos veículos estão próximos de outro veículo, talvez o dado não precise estar alinhado diretamente com a via (rua), basta que os pontos estejam o mais próximo possível da localização real do objeto, nesse caso filtros poderiam ser aplicados.

Zheng (2015) elenca uma série de pré-processamentos que podem ser aplicados a um *dataset* com informações espaciais (seção 2.4.1). Neste trabalho, alguns desses métodos de pré-processamento foram aplicados para o cálculo das métricas de caracterização coletadas nos trabalhos relacionados.

Este capítulo descreve quais foram os pré-processamentos necessários antes da caracterização do *dataset* através da extração de métricas. Na seções a seguir, são descritos quais foram os métodos utilizados, quais arquivos e campos utilizados, e apresentação de valores de redução dos dados.

4.1 Descrição do *dataset*

A cidade de São Paulo possui uma população de cerca de 12 milhões de pessoas segundo dados do IBGE (2019). O transporte coletivo nessa cidade é gerenciado pela empresa pública

de planejamento São Paulo Transporte S/A (SpTrans[1]) que supervisiona a operação das concessionárias que controlam os ônibus. Os bairros de São Paulo atendidos pelo transporte público são agrupados em 9 zonas e 96 bairros, como pode ser visto na Figura 4.1.

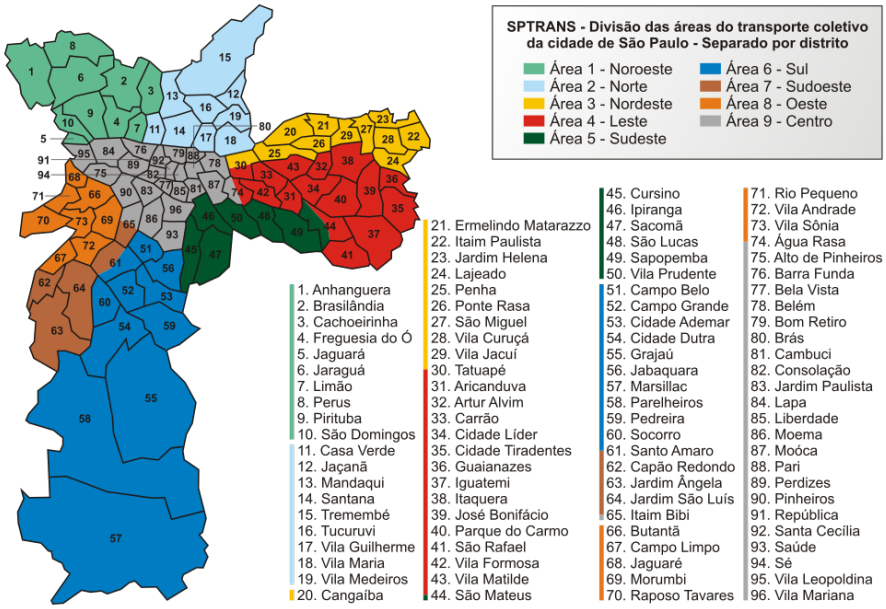


Figura 4.1: SPTrans - Áreas do sistema de transporte coletivo da cidade de São Paulo (SPTRANS, 2019).

Os traços reais de mobilidade dos ônibus de São Paulo podem ser obtidos em tempo real através da API Olho Vivo[2] da SPTrans. A API traz informações, como localização em tempo real do veículo, em que sentido está viajando, linha que está cobrindo, entre outros. Para ter um histórico desses traços é necessário capturar os dados pelo período de tempo desejado, ou solicitar uma janela de dados históricos para a SPTrans.

O *dataset* fornecido pela SPTrans cobre viagens de ônibus de 2009 a 2015 registradas por equipamentos AVL instalados nos ônibus. Para este trabalho foi selecionado a janela de tempo de outubro de 2015 das 06:00 às 22:59, pois é um mês que está fora da sazonalidade de férias, e não possui muitos feriados (comportamento atípicos). Este conjunto de dados possui viagens de mais de 14 mil ônibus e 2500 linhas. Os ônibus registram posições a cada 45 segundos, havendo um pequena parcela dos registros com maior intervalo de segundos. Os arquivos do *dataset* estão no formato CSV, e são dois arquivos por dia do mês. O arquivo com o nome "MO_1510X" representa os *traces* dos ônibus no dia X do mês de outubro. Cada arquivo desse possui cerca de 2,2 GB e 27 milhões de linhas totalizando, para os 31 dias do mês de outubro,

[1] http://sptrans.com.br
[2] http://olhovivo.sptrans.com.br/

68.2 GB. O arquivo "MO_1510X" possui 6 colunas com as seguintes informações:

- **dt_servidor:** data em que o registro chegou no servidor;
- **dt_avl:** data em que o registro foi coletado pelo equipamento AVL do ônibus;
- **longitude:** longitude do ônibus em determinado momento;
- **latitude:** latitude do ônibus em determinado momento;
- **id_avl:** identificador único do aparelho de AVL de cada ônibus;
- **line_id:** identificador único da linha que o ônibus está associado em determinado momento.

O arquivo auxiliar "AL_1510X" em formato CSV contém as seguintes informações para cada dia X do mês de outubro: número da linha, complemento, identificador da linha e sentido da viagem. O número da linha junto ao complemento é a informação mostrada ao passageiro, e cada número de linha está associado a identificadores de linha de ônibus do arquivo "MO_1510X". Cada arquivo desse possui 45KB e 2784 linhas.

4.2 Ferramentas para o pré-processamento e análise dos dados

Os dados não processados e processados foram armazenados no serviço Amazon Simple Storage Service (S3) que é um serviço de armazenamento de objetos da nuvem Amazon Web Services (AWS), podendo receber dados de qualquer tipo.

Para analisar e pré-processar grandes volumes de dados foi utilizado o PySpark, que é uma abstração em Python para acessar o Spark, em clusters de máquinas virtuais do Amazon Elastic MapReduce (EMR) que auxilia no gerenciamento e instalação do Spark e outras ferramentas de *big data* também na nuvem AWS. O Spark oferece funcionalidades para carregar, filtrar, e transformar dados de forma distribuída, sendo eficiente para grandes volumes de dados. Junto ao PySpark, foi utilizado o pandas[3], que é uma biblioteca do Python para estruturas e análise de dados. Tanto o Spark quanto o pandas, além de oferecer funcionalidades para transformar os dados (filtros, mapeamentos, reduções, amostragem), fornecem métodos para análise estatística (contagem, soma, média, mediana, quantis) que também foram utilizados para caracterização dos dados.

[3]https://pandas.pydata.org/

Em conjunto com o Spark, bibliotecas do Python foram utilizadas em variadas fases da caracterização. Para a manipulação de dados geoespaciais e polígonos foram utilizados bibliotecas shapely[4] e o geopandas[5]. Já para visualização dos dados, e das métricas coletadas, os gráficos foram gerados com a biblioteca matplotlib[6]. Algumas informações de geolocalização foram plotadas com a biblioteca folium[7] que plota as localizações em mapas interativos utilizando camadas do OpenStreetMaps (OSM).

Em relação às informações de mapas, o mapa e os *shapes* (polígono e geometrias) da cidade de São Paulo foram extraídas do OSM já processados pelo governo do estado de São Paulo[8].

Para a execução de todos os códigos criados, e interação com o cluster de Spark no EMR foi utilizado o Jupyter Notebook[9] que é um ambiente web interativo que une o código e texto, permitindo que se execute pedaços de um *script*, veja os resultados, e também adicione anotações.

4.3 Exploração inicial, filtros de valores nulos, hora e elementos duplicados

Após a obtenção dos dados, foi aplicado um filtro de valores nulos em ambos os arquivos de *traces* (MO) e auxiliar (AL) para filtrar registros que tivessem valores nulos ou vazios em alguma das colunas, porém nenhum valor nulo ou vazio foi encontrado.

Após verificação dos elementos nulos, foi realizado estudo dos tipos de arquivos, quais colunas estavam disponíveis e que valores elas poderiam assumir. Os trabalhos de Pons, Monteiro e Speicys (2015) e Yai (2016) auxiliaram na descoberta do que cada coluna representava, e qual era a função de cada arquivo. O arquivo MO_[data].csv é o arquivo que contém os traços de mobilidade dos ônibus, e o arquivo auxiliar AL_[data].csv apresenta dados sobre as linhas dos ônibus.

Em relação aos valores que cada coluna pode assumir no arquivo de traços de mobilidade, verificou-se que:

- as datas do servidor e do equipamento AVL possuem discrepâncias. A média de diferença entre as duas datas considerando todos os dias do mês permaneceu entre 40 e 163 segundos, sendo que o quantil de 75% de todos os dias permaneceu entre 5 e 28 segundos. Isso

[4] https://shapely.readthedocs.io/en/stable/manual.html
[5] https://geopandas.org/
[6] https://matplotlib.org/
[7] https://python-visualization.github.io/folium/quickstart.html
[8] http://datageo.ambiente.sp.gov.br/
[9] https://jupyter.org/

mostra que a diferença entre as colunas de datas para 75% dos dados é menor que 28 segundos. Entretanto, há momentos em que a diferença pode chegar a 6 horas de diferença, ou seja, servidor recebeu os dados 6 horas depois, ou relógio estava fora de sincronia;

- existem mais de 2500 linhas por dia para a coluna *line_id*;
- existem mais de 14600 ônibus nos registros, pela contagem de *id_avl* (identificador do equipamento AVL);
- um ônibus pode estar atribuído a mais de uma linha por dia, identificado pela contagem do par *id_avl* e *line_id*.

Como o interesse deste trabalho é a caracterização de *dataset* de mobilidade com foco em VANETs, optou-se por analisar períodos de maior atividade dos ônibus quando pode haver maior probabilidade de conexão entre eles. Portanto, foram filtrados registros da madrugada entre às 23:00 e 5:00 da manhã, e mantidos registros do arquivo MO_[data].csv cuja hora de detecção do equipamento AVL foi entre 6:00 e 22:59.

Outro filtro aplicado junto com o de hora, foi o filtro de elementos duplicados. Foram eliminados registros duplicados com base nas colunas *dt_avl*, *id_avl*, *line_id*, *latitude* e *longitude*, ou seja, ônibus que registraram a mesma posição mais de uma vez num determinado momento.

Após a aplicação de todos esses filtros o número de registros em cada arquivo foi reduzido de 27 milhões para valores na casa de 19 milhões, cerca de 30% de redução do número de linhas. A redução também ocorreu no tamanho dos arquivos, pois foi utilizado o formato de dado colunar PARQUET. O tamanho dos arquivos foram de 2.2GB no formato CSV para 540MB em média com este formato e após os filtros aplicados.

A coluna de data do servidor também foi eliminada, pois não foi utilizada para análise dos dados.

4.4 Filtrando dados fora da cidade de São Paulo

Para auxiliar na análise e pré-processamento dos dados, uma coluna foi adicionada ao conjunto de dados para identificar a que bairro um registro pertence. Para realizar tal processo, o arquivo do *shape* da cidade de São Paulo foi utilizado junto com as bibliotecas do python geopandas e shapely. Um arquivo *shape* é um conjunto de polígonos que descrevem a geometria de determinada região, ou seja, conjunto de localizações que delimitam uma área. Nesse caso, o *shape* da cidade de São Paulo possui os limites de cada um dos 96 bairros.

A Figura 4.2 demonstra como é feito o processamento de identificação de bairro. Nas linhas 13-16 do código, o Spark chama a função *get_region* para cada linha do arquivo, passando como parâmetro os campos de *latitude* e *longitude* do traço de mobilidade. Nas linhas 1-6, a biblioteca shapely utiliza o método *within* para verificar se o ponto (localização) passado como parâmetro está dentro de um dos 96 bairros de São Paulo. Se o ponto estiver dentro de um bairro, imediatamente o código retorna o nome do bairro, caso o ponto não pertença a nenhum bairro a função retorna o valor "None".

```
1   def get_region(row, sp):
2       point = Point((float(row[0]), float(row[1])))
3       for i in range(96):
4           if point.within(sp.loc[i, "geometry"]):
5               return sp.loc[i, "Nome"]
6       return "None"
7
8   def get_region_udf(sp):
9       return udf(lambda x: get_region(x, sp))
10
11  traces = spark.read.parquet("MO_1510")
12
13  traces.withColumn("region", get_region_udf(
14      sc.broadcast(sp_shape).value)
15      (struct(traces["longitude"],
16      traces["latitude"]))
17  )
```

Figura 4.2: Código em Spark que identifica o bairro dos traços de mobilidade (elaborado pela autora).

Após a identificação do bairro de cada traço de mobilidade, eliminam-se todos os traços fora da região da cidade de São Paulo, ou seja, cujo bairro seja "None" retornado pela função. Em média, das 19 milhões de linhas restantes dos pré-processamento anteriores, apenas 150 mil linhas foram removidas por esse pré-processamento, menos de 1% dos registros de cada arquivo.

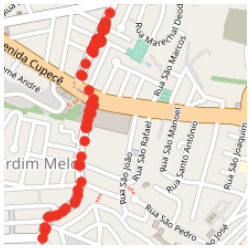
4.5 Visualização de traços de mobilidade no mapa e tempo de atualização de posições

Para entender a distribuição dos pontos de localização pelo mapa de São Paulo foram plotados na Figura 4.3 os traços de mobilidade de um ônibus cumprindo uma linha durante o dia 1/10/2015 dentro da região da cidade de São Paulo.

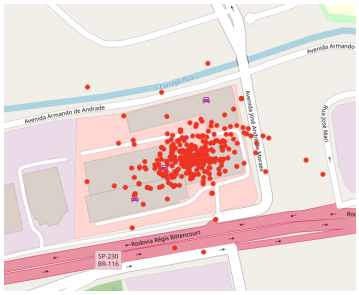
Na Figura 4.4, há três situações de pontos fora da região de São Paulo. A primeira delas (a) é quando um ônibus trafega por outras regiões produzindo traços de mobilidade nessas

Figura 4.3: Traços de mobilidade de um ônibus no dia 1/10/2015 (elaborado pela autora).

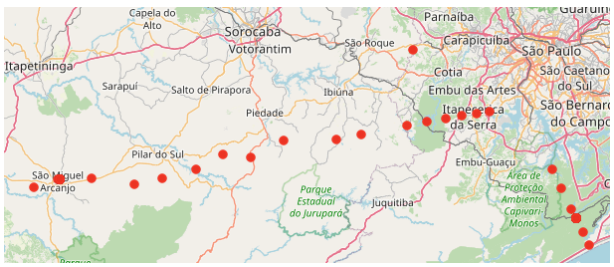
áreas, nesse caso em Diadema. No momento (b), há uma situação em que o ônibus registrou diversos pontos num terminal ou estacionamento em Taboão da Serra, podendo indicar um ônibus intermunicipal. Esse tipo de situação também ocorre em ônibus que estão dentro de São Paulo, emitindo registros dentro dos terminais durante vários segundos. Já no momento (c), o equipamento AVL registrou posicionamento do ônibus, porém fora de vias trafegáveis, indicando um possível erro do equipamento e de sinal.



(a) Traços de mobilidade



(b) Pontos no estacionamento



(c) Erro de equipamento

Figura 4.4: Situações de localizações fora da região da cidade de São Paulo (elaborado pela autora).

Em relação ao tempo de atualização das posições consecutivas de um ônibus, verificou-se que a média de atualização em todos os dias da semana estão próximas a 45 segundos, sendo

que os quantis de 25%, 50%, e 75% são 45 segundos, revelando que grande parcela dos dados possui tempo de atualização nesse intervalo de tempo.

4.6 Explorando os arquivos GTFS e auxiliar AL

Após os filtros das seções anteriores, outro mecanismo que pode ser utilizado para pré-processar dados de ônibus ou de transporte público são dados de GTFS. O GTFS[10] (General Transit Feed Specification) é um padrão que define um formato comum para os horários e informações geográficas de transporte público. As agências de transporte podem ter APIs ou mecanismos para compartilhar esse tipo de informação publicamente, por exemplo, com desenvolvedores. No caso de São Paulo e de outras cidades no mundo há diversos portais para obter o conjunto de arquivos GTFS, como o TransitFeeds[11], Governo Aberto[12], SPtrans[13], entre outros.

O GTFS possui um conjunto de arquivos que descrevem os seguintes aspectos do transporte público: paradas, trajetos, viagens e outros dados relativos a horário. Essas informações podem servir de referência para algum tipo de processamento. Por exemplo, no relatório técnico produzido pela SPTrans em conjunto com empresa Scipopulis (PONS; MONTEIRO; SPEICYS, 2015), utiliza-se informações de paradas e *shapes* dos ônibus para identificar a velocidade e o cumprimento dos horários previstos, além de extração de outras métricas como velocidade. Já em Weng et al. (2016), os autores propõem um modelo de cálculo de velocidade de ônibus utilizando paradas de ônibus dos arquivos GTFS e mais de 20 mil ônibus da cidade de Beijing.

Esta pesquisa utilizou dados do GTFS de São Paulo para a eliminar ônibus parados e eliminar traços de mobilidade com possibilidade de serem ruídos. Num primeiro momento, foram explorados o arquivo auxiliar AL_[data].csv e quais arquivos do GTFS poderiam auxiliar no pré-processamento.

Foram comparados os arquivos do GTFS com os campos do arquivo AL. Foram explorados 4 arquivos do GTFS:

- **routes:** conjunto de trajetos (grupo de viagens) do transporte público que o ônibus desempenhará e é mostrado ao passageiro como um único serviço. Por exemplo: 1015-10. Em relação ao arquivo auxiliar AL, 1015 seria o número da linha e 10 a coluna de complemento;

[10] https://developers.google.com/transit/gtfs/reference
[11] http://transitfeeds.com/
[12] http://catalogo.governoaberto.sp.gov.br/dataset/gtfs-dados-operacionais-dos-onibus-metropolitanos
[13] http://www.sptrans.com.br/desenvolvedores/

- **trips:** representa as viagens de cada trajeto. Uma viagem é um sequência de paradas que acontece em um certo período. Este arquivo possui o *route_id*, *trip_id*, *direction_id*, e *shape_id*. O *trip_id* é composto do *route_id* mais o *direction_id* que pode assumir valores 0 ou 1. Por exemplo: 1015-10-0 é um *trip_id* que indica que a rota é 1015-10 com direção 0. Em relação ao arquivo auxiliar AL, o *trip_id* pode ser formado unindo as colunas número da linha, complemento e direção;
- **stops:** paradas onde os ônibus pegam e deixam passageiros;
- **shapes:** regras para mapear os caminhos de viagem de ônibus. Um *shape* possui um conjunto de geolocalizações (latitude e longitude) que possuem um campo de sequência mostrando qual a sequência da viagem do ônibus. *Shapes* são conjuntos de localizações que representam as localizações das vias onde trafegam ônibus.

O GTFS possui outros arquivos que não foram citados aqui, mas todas as descrições e *schema* dos dados são descritos na documentação[14].

A Figura 4.5 demonstra como são alinhados os *shapes* e as paradas no mapa. A figura (a) mostra que o *shape* (em verde) e as paradas (vermelho) estão alinhadas numa visão macro. Na figura (b) quando detalhes desse alinhamento são mostrados, percebe-se que as paradas podem estar entre duas localizações do *shape*.

Antes de fazer o cruzamento do arquivo auxiliar AL com os do GTFS para descobrir o *shape* que cada ônibus está seguindo, o campo de direção do arquivo AL assume dois valores 1 ou 2, já o campo direção do *trips* pode ser 1 ou 0, então uma análise precisou ser feita para descobrir quais valores desses eram correlatos. Foi analisada a linha 2435-10 do arquivo AL. Para descobrir a direção do ônibus, foi escolhido um ônibus do arquivo de traços de mobilidade que estava desempenhando essa linha, e foram plotados traços de mobilidade desse veículo, sobrepondo com os dois *shapes* candidatos (com as duas direções). Na Figura 4.6, os pontos vermelhos representam os *traces* do ônibus, em verde e azul indicam as localizações dos *shapes*. Há indicações em ambas as figuras, onde cada caminho começa e onde termina. Constatou-se que o *trace* do ônibus começa próximo a localização onde começa o *shape* 58695 com direção 0 na figura (b), e se encaminha para finalizar também na mesma direção onde *shape* termina. Portanto, conclui-se que o valor 1 no campo direção do arquivo AL é igual ao valor 0 dos arquivos de GTFS, e o valor 2 do AL é igual a direção 1 do GTFS. Esses valores indicam ida e volta da direção do ônibus.

O próximo passo utilizando os arquivos de GTFS e auxiliar AL foi encontrar qual o *shape_id*

[14]https://developers.google.com/transit/gtfs/reference#file_requirements

(a) *shape* em verde e paradas em vermelho

(b) Paradas entre os pontos do *shape*

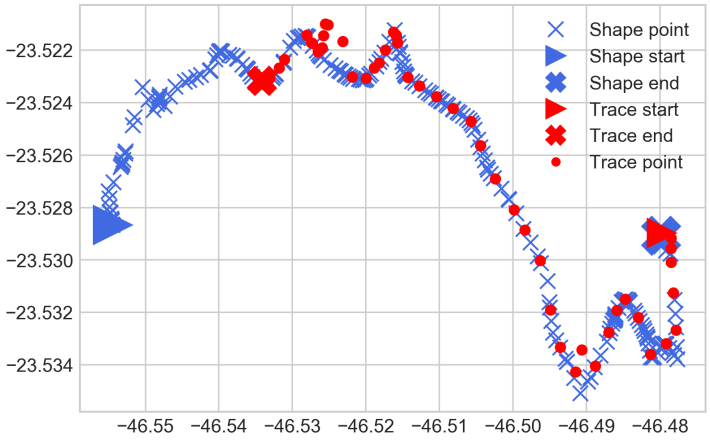
Figura 4.5: Alinhamento de *shapes* e paradas do GTFS de São Paulo (elaborado pela autora).

de cada traço de mobilidade. Primeiro, cruzou-se o arquivo *trips* com o auxiliar AL para descobrir qual é o *shape_id* de cada linha de ônibus para cada dia do mês. As colunas número da linha, complemento e direção do arquivo auxiliar AL foram concatenadas, o que resulta no *trip_id*, e então foi feito a operação de *join* com arquivo de *trips* a partir do campo *trip_id*. No final da operação, os registros do arquivo AL que não obtiveram associação com nenhum *shape_id* do arquivo de *trips* foram excluídos.

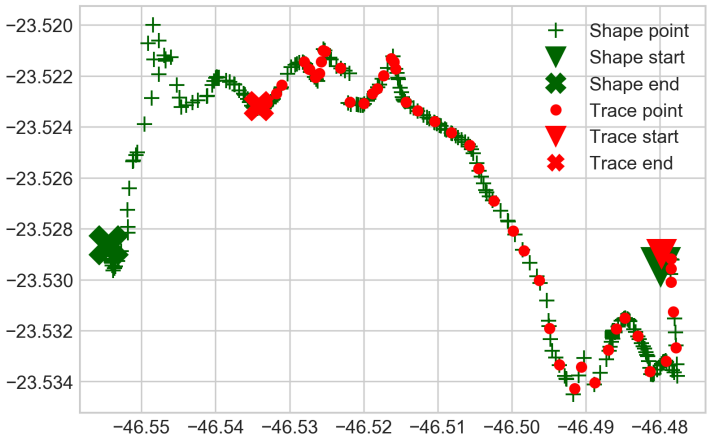
O último passo foi cruzar o arquivo AL com o arquivo de traços de mobilidade (MO) para descobrir qual o *shape_id* de cada registro. Todos registros do arquivo MO que não tiveram nenhuma associação com *shapes* (*shape_id*) foram excluídos do *dataset*.

4.7 *Map Matching*

Para eliminar possíveis ruídos e descartar ônibus que não estão se movendo - por exemplo estão estacionados na garagem- foi realizado o processo de *map matching* dos traços de



(a) Shape 53709 e linha 2435-10 direção 1



(b) Shape 58695 e linha 2435-10 direção 0

Figura 4.6: Identificando direção do arquivo auxiliar AL (elaborado pela autora).

cada ônibus com o *shape* do GTFS mais próximo. Outros trabalhos já realizaram processos semelhantes.

O trabalho de Weng et al. (2016) realizou o processo de *map matching* mapeando os *traces* às paradas mais próximas, ou estimando posições quando um ônibus localizava-se entre duas paradas. Já no relatório técnico da SPTrans produzido por Pons, Monteiro e Speicys (2015) em conjunto com a empresa Scipopulis, interpola-se a posição dos ônibus até o *shape* ou parada mais próxima, e utiliza-se as distâncias planejadas já conhecidas para realizar o cálculo das métricas. Zheng (2015) ainda ressalta que existem diferentes maneiras como fazer o *map matching*: algoritmos geométricos que é o caso deste trabalho, topológicos, probabilísticos, dentre outros. O trabalho de Domingues, Silva e Loureiro (2018) realiza esse processo com informações do OpenStreetMaps, realizando algoritmos topológicos, considerando as conexões entre as vias.

A seção 4.6 descreveu o processo para identificar o *shape_id* de cada registro de ônibus do

arquivo MO. Apesar de saber qual *shape* o ônibus seguirá, um *shape* possui um conjunto de geolocalizações (sequências) que descrevem o trajeto do ônibus ao longo do mapa. Portanto, para saber a que sequência do *shape* o registro de posição do ônibus está mais próximo foi realizado o processo de *map matching*.

O *map matching* foi executado através do Spark, do pandas e da biblioteca do python haversine[15] que calcula a distância em metros entre 2 pontos. Para cada registro (traço) do arquivo MO, o Spark chama a função que recebe como parâmetro a coordenada do *trace* latitude e longitude, e também o *shape_id*. Como mostra a Figura 4.7, a função calcula a distância haversine (linha 4), e compara se a distância atual é menor do que a distância da sequência anterior, se sim, nas linhas 7-10 armazena-se as informações do *shape_sequence* mais próximo. Ao final, na linha 12, retorna-se qual a sequência do *shape* mais próxima da localização (traço) do ônibus naquele momento.

```
1  for _,shape in candidate_shapes.iterrows():
2          shape_coord = (shape["shape_pt_lon"], shape["shape_pt_lat"])
3
4          distance = haversine(shape_coord,trace_coord)
5
6          if distance <= min_distance:
7              min_distance = distance
8              min_shape_sequence = shape["shape_pt_sequence"]
9              min_shape_coord_lat = shape["shape_pt_lat"]
10             min_shape_coord_lon = shape["shape_pt_lon"]
11
12 return (min_shape_sequence,min_distance,min_shape_coord_lat,min_shape_coord_lon)
```

Figura 4.7: Código em Spark que identifica a sequência do *shape* mais próximo do *trace* (elaborado pela autora).

Ao final do processo, foi identificado qual a sequência do *shape* mais próxima de cada traço do ônibus, e qual a distância esses dois pontos estão um do outro. Foi feita uma análise sobre as distâncias encontradas. Foram encontradas distâncias superiores a 30km, podendo indicar ruído ou ônibus fora de rota. O estudo dos quantis revelou que para dias úteis, 75% dos dados estão a menos de 56 metros do *shape* mais próximo. Já aos finais de semana e feriado, o valor do quantil de 75% fica entre 1383 e 2100 metros. Essa análise revela que grande parcela dos dados estão próximos aos *shapes* encontrados, portanto próximo as vias cerca de 55 metros em dias de semana, e um pouco mais distantes em finais de semana.

Constatou-se que em média, em um dia de semana, um ônibus associado a uma linha atinge entre 98 e 143 *shape_sequences* diferentes. Essa análise revela que o algoritmo conseguiu identificar grande diversidade de *shapes* para os traços de mobilidade de um ônibus associado a uma linha, mas ainda sim existem ônibus que podem estar na garagem parados, seguiram por rotas não planejadas, ou o aparelho AVL está produzindo ruídos, o que pode justificar poucos *shapes*

[15]https://pypi.org/project/haversine/

associados. Analisando quantis do dados, encontra-se uma parcela de veículos que atingiram menos de 4 sequências de *shape* num dia.

Com base na análise dos quantis, foram filtrados traços de ônibus associado a uma linha que possuíssem menos de 4 *shapes_sequence* distintos associados em determinado dia. Esse filtro pode eliminar ônibus que não estejam se movimentando por estarem na garagem, eliminar dados que podem conter ruído por erros de detecção, ou que não estavam seguindo rota planejada. Outro filtro aplicado foi o de distância mínima encontrada, foram eliminados todos os registros cuja distância para o *shape* mais próximo fosse superior a 2100 metros. Esse valor foi identificado como maior valor de corte encontrado no quantil de 75% dos finais de semana, ou seja, de todos os dias do mês 75% dos traços de ônibus estão até 2100 metros do *shape* mais próximo.

4.8 Filtro por número de registros

Após todos os pré-processamentos anteriores, foi realizado um estudo para identificar quantos traços um ônibus associado a uma linha produzia por dia. Constatou-se que em média, em dias úteis, o conjunto ônibus/linha produz de 551 a 585 registros, já aos finais de semana e feriado este valor fica entre 586 e 776. Portanto, há maior média de registros por ônibus e linha aos finais de semana. Para o quantil de 75%, os cortes chegam a valores de 758 a 766 para dias úteis, e para finais de semana de 856 a 1212. Portanto, tanto a média quanto o corte dos quantis aponta um maior número de registros por linha e ônibus aos finais de semana, o que pode ser explicado por menos ônibus estarem rodando pelas vias, e os ônibus ficarem mais tempo associados a uma única linha.

Com análise dos valores mínimos de traços produzidos por um ônibus associado a uma linha, foram encontrados veículos que produziam menos de 10 traços de mobilidade por dia. Os traços de mobilidade dessa associação de ônibus e linha que produziram menos de 10 *traces* por dia foram eliminados, pois foram considerados não relevantes para aplicação de VANETs por conta da possível duração curta em espaço de tempo e metros.

Após os filtros, os dados foram reduzidos da casa de 18 milhões para 16 milhões, tendo uma média de 18,14% na redução dos dados em cada arquivo de traço de mobilidade.

4.9 Cálculo da velocidade escalar instantânea

A velocidade média de ônibus foi utilizada para filtrar os dados e também como uma das métricas discutidas no Capítulo 5.

Para o cálculo da velocidade média de um veículo associado a uma linha num determinado dia, foram considerados as formulações a seguir.

Dado um ônibus associado a uma linha em determinado dia, o conjunto de localizações registradas para esse par consiste de um conjunto de coordenadas geoespaciais (x,y) e um *timestamp* tal que cada registro pode ser representado por $p = (x,y,t)$. Já o conjunto desses pontos é ordenado pelo *timestamp* t, composto por n pontos e representado por $P = \{p_1, p_2, ..., p_n\}$.

Dado um ônibus associado a uma linha em determinado dia, a velocidade escalar instantânea V_i de cada um dos registros desse ônibus ordenados pela data do AVL (*dt_avl*), pode ser calculada pela fórmula 4.1. A distância d é calculada por um função em Python usando a biblioteca haversine entre um ponto p_i e seu ponto adjacente anterior p_{i-1}, a distância é retornada em metros. A variação de tempo calculada em segundos se dá pela diferença entre a data de um registro t_i e a data do registro anterior t_{i-1}. Por fim, divide-se a distância pela variação de tempo, e obtém-se a velocidade escalar em m/s, para convertê-la para km/h, multiplica-se a divisão por 3,6.

$$V_i = \frac{d(p_i, p_{i-1})}{t_i - t_{i-1}} x 3,6 \tag{4.1}$$

A velocidade média final da associação de um ônibus a uma linha em determinado dia é representada pela fórmula 4.2. A velocidade média final M é resultante da soma de todas as velocidades instantâneas calculadas divido pelo número de registros n.

$$M(V) = \frac{1}{n} \sum_{1}^{n} V_i \tag{4.2}$$

As fórmulas foram baseadas nos trabalhos Campos, Moraes e Silva (2010), Weng et al. (2016) e Silva (2010) que também utilizam velocidades instantâneas e uma agregação final de velocidade média como forma de cálculo.

A análise das velocidades calculadas aponta que a média, em todos os dias de outubro, foi na casa de 13 km/h. O estudo do quantil de 75% mostra que as velocidades permanecem entre 19 e 20 km/h, ou seja, 75% do dados tem velocidades menores que esses valores. O estudo dos extremos, revelam que há velocidades superiores a 1000 km/h, podendo indicar algum tipo

de ruído na geolocalização, ou informação insuficiente para calcular a velocidade de forma apropriada (distância ou tempo). Portanto, é necessário um filtro de velocidade para eliminar velocidades anômalas.

4.10 Filtro de velocidade

Após o cálculo de velocidade instantânea dos registros, foram eliminados registros cuja velocidade fosse superior a 80 km/h e inferior a 0,1km/h - nivelamento feito nos trabalhos de Campos, Moraes e Silva (2010), Weng et al. (2016), Silva (2010), e como referência de velocidade extremamente alta no relatório técnico da SPTrans (PONS; MONTEIRO; SPEICYS, 2015).

4.11 Resumo do pré-processamento

O pré-processamento é uma das etapas requeridas dependendo do tipo de cálculo que se faz com um *dataset*. No caso dos traços de mobilidade de ônibus da SPTrans, a captura dos dados é feito com um equipamento AVL que inevitavelmente é afetado por aspectos físicos, como concentração de prédios, falhas de hardware, dentre outros, que impactam em erros e ruídos no registro do posicionamento.

Neste trabalho utilizou-se técnicas de pré-processamento que foram extraídas de outros trabalhos ligados a caracterização e manipulação de *datasets* de mobilidade ou de VANETs. Os mecanismos escolhidos foram requisitados para que fossem calculadas as métricas apresentadas no Capítulo 5.

Os métodos utilizados para o pré-processamento deste *dataset* foram:

- Eliminação de valores nulos;
- Filtro por janela de tempo de interesse;
- Filtro de elementos duplicados;
- Filtro de dados fora da região da cidade de interesse;
- *Map Matching* que permitiu a eliminação de ônibus parados, fora das rotas conhecidas ou com possíveis ruídos de detecção;
- Filtro pelo número de registros;
- Filtro de velocidade.

Durante todas as fases de pré-processamento, foram extraídos valores estatísticos como média, desvio padrão, contagem, e valores dos quantis de 25%, 50% e 75% a fim de entender o comportamento dos dados em cada momento, e em cada dia do mês de outubro. As médias e os valores de quantis foram fatores de corte, filtros e nivelamento em processos como, filtro de velocidade, *map matching*, e número de registros. Além de decisório, esse estudo pode auxiliar no discernimento dos comportamento padrões e exclusivos, neste caso, entre dias da semana, finais de semana e feriado.

Ao final do pré-processamento houve uma redução de 68,2GB de dados para 9,9GB. Como próximo passo, prosseguiu-se para o cálculo das métricas que caracterizam o *dataset* no Capítulo 5.

Capítulo 5

CARACTERIZAÇÃO DE UM DATASET DE MOBILIDADE DE ÔNIBUS DA CIDADE SÃO PAULO

A caracterização de *datasets* de mobilidade veicular a partir de traços de mobilidade tem um grande impacto na área de VANETs, pois estudos desse tipo contribuem para que pesquisadores avaliem a viabilidade de redes veiculares, gerem modelos de mobilidade, conduzam simulações baseados em parâmetros observados no conjunto de dados, e façam ajustes nos testes de protocolos de roteamento.

Existem *datasets* públicos de traços de mobilidade que são amplamente explorados na área de VANETs, como: táxis de São Francisco, ônibus e táxis de Beijing e Shangai, táxis de Roma, dentre outros. São Paulo é uma cidade relevante no território nacional e no cenário internacional, sendo um dos centros financeiro e tecnológico, possui cerca de 12 milhões de pessoas e um vasto transporte público para atendê-las. Entretanto, São Paulo ainda não é muito explorada dentro da área de redes veiculares, apesar de possuir *datasets* públicos de mobilidade de diferentes modais, como os ônibus e metrô. A cidade de São Paulo pode ser relevante para o estudo de mobilidade e VANETs, principalmente em relação aos ônibus contando com um contingente com mais de 14 mil veículos e 2500 linhas, e também é um cenário pouco explorado por outros pesquisadores fora do Brasil.

Para demonstrar aspectos da mobilidade dos ônibus de São Paulo com foco na área de VANETs, este capítulo apresenta a caracterização de um *dataset* de mobilidade de ônibus da cidade.

O primeiro passo para a caracterização é coletar os dados. A empresa SPTrans que controla o transporte público de São Paulo forneceu conjunto de traços de mobilidade de ônibus de outubro de 2015. Após passar por fases de pré-processamento, como visto no Capítulo 4, o

dataset se tornou adequado para extração de características.

A caracterização do *dataset* de mobilidade veicular deste trabalho se dá pela extração de métricas de mobilidade e conectividade identificadas ou derivadas dos trabalhos relacionados do Capítulo 3. As características do *dataset* são apresentadas com recursos gráficos e através de explicações para os valores obtidos sob uma visão da dinâmica da cidade de São Paulo e sobre o funcionamento planejado do transporte público.

Este capítulo organiza-se da seguinte maneira: são apresentadas as características gerais do *dataset*, descrição como cada métrica foi calculada e discussão dos resultados obtidos; e por fim discute como as métricas obtidas podem influenciar o cenário de VANETs.

5.1 Características gerais do dataset

O *dataset* de mobilidade de ônibus da cidade de São Paulo foi fornecido pela SPTrans. Foram analisados os traços de mobilidade do mês de outubro de 2015, levando em conta traços de mobilidade registrados das 6:00 da manhã às 22:59. Esse conjunto de dados possui mais de 14 mil ônibus e 2500 linhas. O registro da posição de um ônibus ocorre a cada 45 segundos, havendo alguns intervalos de captura maiores e menores.

Para caracterizar o *dataset*, serão extraídas métricas de conectividade e mobilidade. As métricas escolhidas foram selecionadas a partir do comum uso nos trabalhos relacionados, por serem parâmetros configuráveis em alguns simuladores de tráfego e redes para VANETs, serem aspectos de ajuste em modelagens, e também poderem avaliar a conectividade da rede para possíveis aplicações de redes veiculares.

5.2 Número de ônibus ativos

A métrica de número de ônibus ativos corresponde ao número de veículos que estão trafegando, independente da linha de ônibus que este veículo está associado. Este tipo de métrica pode ser encontrado diretamente ou indiretamente nos trabalhos Uppoor et al. (2014), Uppoor e Fiore (2012), Doering e Wolf (2015) e Santana, Kanashiro e Kon (2018).

5.2.1 Cálculo da métrica

Para executar o cálculo da métrica de ônibus ativos, realiza-se a contagem de diferentes identificadores únicos de ônibus (*id_avl*). Por exemplo, para saber quantos ônibus estavam

ativos no dia 1/10/2015, basta contar quantos (*id_avl*) distintos há nesse arquivo, como mostra a linha 2 da Figura 5.1. Se a contagem for para algum agrupamento, por exemplo, hora ou região, primeiro realiza-se a ação com o operador *group_by* do Spark para agrupar os dados pelo atributo especificado. Posteriormente, é computado a contagem de (*id_avl*) distintos para cada um desses grupos. A linha 3 da figura mostra o agrupamento por hora, ou seja, contanto o número de ônibus ativos em cada hora do dia 1/10/2015. Já a linha 4 da figura mostra como calcular os ônibus por região em cada horário do dia 1/10/2015. A coluna *hour_avl* representa a hora inteira da data do registro, por exemplo 22:50, *hour_avl* é igual a 22.

```
1  traces = spark.read.parquet("s3://mobility-traces-sp/MO_15101/")
2  n_vehicles_day = traces.select(F.countDistinct("id_avl").alias("number_buses"))
3  n_vehicles_hour_day = traces.groupby("hour_avl").agg(F.countDistinct("id_avl").alias("number_buses"))
4  n_vehicles_hour_region_day = traces.groupby("hour_avl","region").agg(F.countDistinct("id_avl").alias("number_buses"))
```

Figura 5.1: Código em PySpark para calcular ônibus ativos (elaborado pela autora).

5.2.2 Número de ônibus ativos por dia

Esta métrica está relacionada a quantos ônibus estão ativos por cada dia do mês de outubro de 2015. Na Figura 5.2 é possível observar a quantidade de ônibus ativos para cada dia. Nos dias de trabalho (dias úteis), a quantidade chega a valores próximos a 14 mil ônibus ativos. Já aos sábados (dias 3, 10, 17, 24 e 31) este valor chega a valores pouco acima de 10 mil ônibus. Aos domingos (dias 4, 11, 18, e 25) e feriado (dia 12) estes valores não chegam a 8000 ônibus.

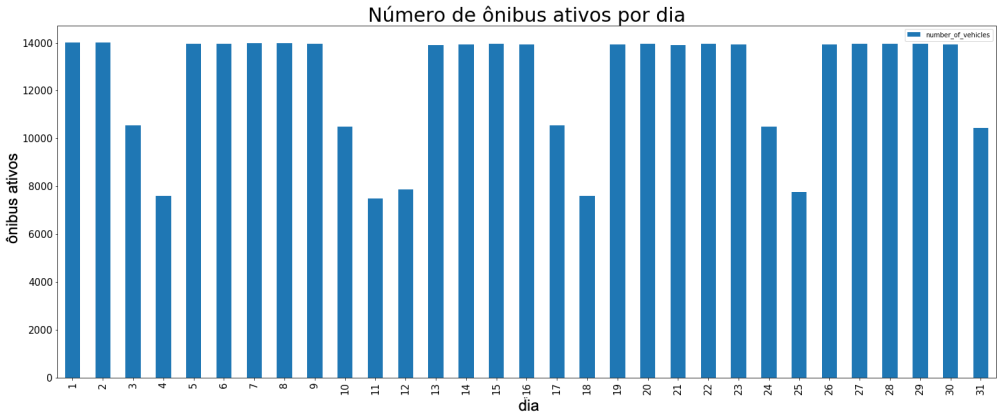


Figura 5.2: Número de ônibus ativos por dia do mês de outubro de 2015 (elaborado pela autora).

Esse comportamento de mais ônibus em dias úteis, ocorre principalmente para atender a demanda colaboradores deslocando-se para seus trabalhos, jovens para escola, e outras atividades comerciais que ocorram em dias úteis. No dia de feriado (12) o comportamento foi similar aos domingos.

5.2.3 Número de ônibus ativos ao longo do dia

O número de ônibus ativos ao longo do dia descreve quantos ônibus estão trafegando pela cidade ao longo da horas em cada dia do mês de outubro de 2015. Para demonstrar o comportamento desta métrica, foram selecionados 6 dias do mês - dia 5/10 uma segunda-feira, dia 20/10 terça-feira, 1/10 quinta-feira, 17/10 sábado, dia 4/10 domingo e 12/10 feriado - que estão disponíveis nos gráficos da Figura 5.3.

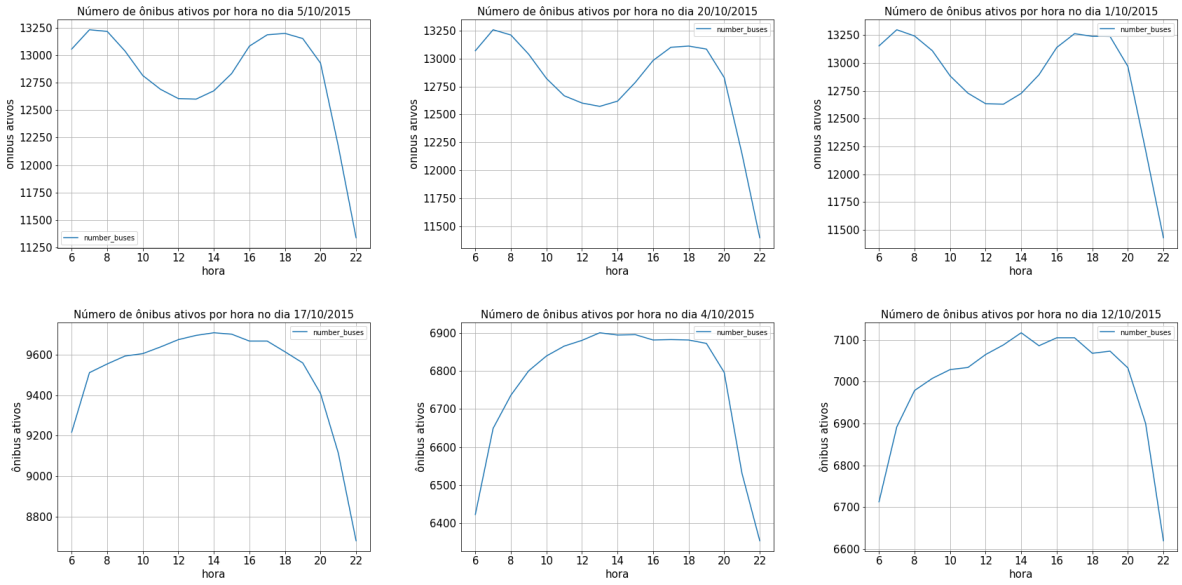


Figura 5.3: Número de ônibus ativos por hora (elaborado pela autora).

Nota-se que em dias úteis (5, 20 e 1), o número de ônibus ativos atinge seu primeiro pico às 7:00 da manhã com aproximadamente 13250 veículos. A partir das 08:00 esse valor começa a decrescer até as 13:00, quando encontra-se um valor próximo a 12600 ônibus. O número de ônibus ativos volta a crescer a partir das 14:00 até às 18:00 quando atinge seu último pico do dia de aproximadamente 13100 ônibus ativos. A partir das 18:00, a atividade dos veículos decresce, e a partir das 20:00 o valor decai ainda mais rapidamente chegando a valores próximos a 11250 ônibus às 22:00.

O comportamento do gráfico nos dias úteis acompanha o horário comercial. Entre às 6 e 8 da manhã acontece o primeiro pico do dia, pois há mais ônibus trafegando para atender a demanda de deslocamento inicial das pessoas de casa para o trabalho, para a escola, dentre outros locais. À medida que o dia passa, há menos ônibus ativos principalmente na parte da manhã até o começo da tarde, pois a demanda também decresce, pois há pessoas trabalhando e estudando que podem não precisar de deslocamento imediato.

O segundo pico volta a aparecer entre às 16 e 18 horas, quando o horário comercial chega

próximo ao final. Uma motivação poderia ser de que os ônibus começam a voltar a ativa principalmente a partir das 16:00 para atender a demanda crescente de pessoas que começam a finalizar o expediente de trabalho e escola por volta das 17:00 e 18:00 e estão voltando para suas residências. Já a partir das 20:00 a demanda decresce, pois a maior parte das pessoas já se deslocaram para suas casas, e não há a mesma quantidade de estabelecimentos comerciais abertos para que pessoas desenvolvam atividades nesse horário.

Aos finais de semana, há menor variação da quantidade de ônibus ativos ao longo do dia, principalmente, entre o período das 8:00 às 20:00. Entretanto, os picos de valores não são tão altos quanto os dias de semana. No sábado 17/10, o número de ônibus ao longo do dia permanece acima de 9400, enquanto aos domingos 6700 ônibus. O número de ônibus volta a cair drasticamente a partir das 20 às 22 horas seguindo o comportamento dos dias úteis para esse intervalo de tempo.

Aos finais de semana o número de ônibus ativos é menor do que nos dias úteis, pois não há demanda nesse dia para escolas, e tantos colaborares indo para o trabalho. Ainda há um grande contingente de ônibus andando principalmente aos sábados, pois parte do comércio está aberto até às 18:00, ainda há pessoas que trabalham de sábado, dentre outros motivos. Já aos domingos, esse valor cai ainda mais, pois não há comércio aberto (exceto supermercados, serviços essenciais, algumas exceções de lojas), ainda menos pessoas trabalhando em comparação com dias úteis ou sábados.

O feriado do dia 12/10 (segunda-feira) acompanha os valores dos picos próximos aos finais de semana com valores constantemente acima dos 7000 ônibus entre às 8 e 20 horas. Num feriado, não há tanta demanda de pessoas trabalhando, nem de deslocamento comercial ou escolas, pois a maioria deles interrompem as atividades.

5.2.4 Número de ônibus ativos por bairro ao longo do dia

A métrica de número de ônibus ativos por bairro ao longo do dia demonstra quantos veículos estavam em cada bairro em certo horário do dia.

Na Figura 5.4, estão presentes as quantidades de veículos por bairro para os horários 7 da manhã, 13 horas, 18 e 22 horas do dia 1/10/2015, uma quinta-feira. Às 7 da manhã nota-se que os bairros com maior concentração de ônibus estão na região central e na sul: Pinheiros (90), Jardim Paulista (83), Moema (86), Itaim Bibi (65), Santo Amaro (61), Jardim São Luís (64), República (91), Barra Funda (76), dentre outros. Já às 13 horas, nota-se que o centro e parte do sul ainda estão concentrados, mas com valores menores que às 7 da manhã. Às 18 horas, a

concentração retorna para os bairros do centro e do sul. Às 22 horas, o centro e o sul não estão com a maiores concentrações, há poucos bairros ainda com intervalo máximo de ônibus ativos. Os bairros que ficam altamente concentrados entre 710 e 1667 ônibus ativos durante todos os horários mostrados são: Jardim São Luís (64), Santo Amaro (61), Tatuapé (30), e Itaquera (38).

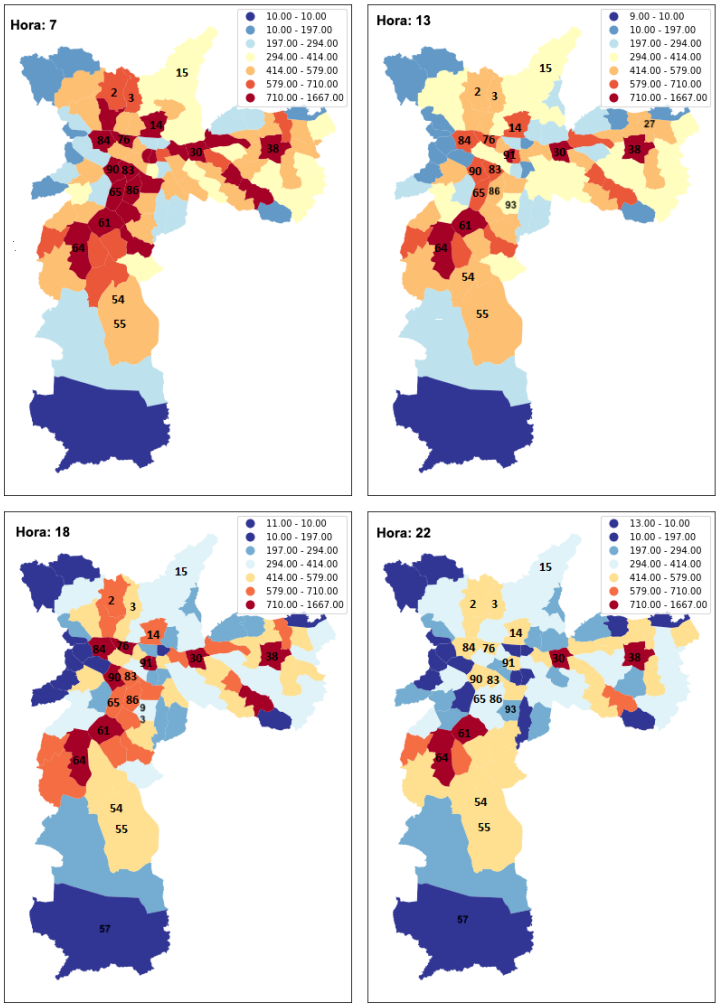


Figura 5.4: Número de ônibus ativos por bairro ao longo dia 1/10/2015 (elaborado pela autora).

Os bairros 65 (Itaim Bibi) e 86 (Moema) da Figura 5.4 ilustram o que acontece durante o dia com os bairros centrais. Durante o início da manhã, possuem o maior intervalo de ônibus ativos de 710-1667, já às 13 horas essa concentração cai no Itaim Bibi para 579-710 e em Moema para 414-579. Os números voltam a aumentar às 18 horas em Moema para o intervalo 579-710. No final da noite às 22 horas, a concentração de ônibus dos dois bairros cai para o intervalo de 294-414 ônibus ativos.

Já em bairros das extremidades como Tremembé, Marsillac (57) ou bairros próximos ao bairro 38 (Itaquera) - Vila Jacuí, São Miguel, José Bonifácio, etc. - o número de ônibus permanece nos mesmos intervalos durante o dia todo (entre 294 a 414 ônibus) com pouca variação.

A alta concentração de ônibus nas regiões centrais e sul, podem ser justificados por serem eixos econômicos e comercial de São Paulo, e também pela presença de terminais de ônibus e de outros tipos de transporte público. Por exemplo, Itaim Bibi e Moema são centros econômicos onde há muitas empresas. Já República e Santo Amaro são destinos de muitos colaboradores que trabalham nesses bairros, e são locais onde há vários modais como ônibus, metrô, e trem, portanto também sendo zona de troca para outras regiões. Já no comportamento da noite, o número de ônibus ativos decresce às 22 horas, como pode ser observado na subseção 5.2.3, por esse motivo a maior parte dos bairros estão menos concentrados comparados às outras horas do dia.

A Figura 5.5 mostra os ônibus ativos no final de semana 4/10/2015 (domingo). Diferentemente dos dias úteis, a maior concentração de ônibus está na região sul e sudoeste, em bairros como: Capão Redondo (62), Jardim São Luís (64), Jardim Ângela (63), Campo Grande (52), Cidade Dutra (54), Grajaú (55), dentre outros. Há grande concentração também em bairros da região Leste: Sapobempa (49) e São Mateus (44). A região Nordeste possui um bairro que fica no maior intervalo de ônibus ativos em todos os horários apresentados, Penha (25). Já da região Noroeste, a Brasilândia (2) fica sempre nos maiores intervalos considerando todos os horários apresentados.

O comportamento do domingo revela que os ônibus estão mais concentrados em áreas de extremidades e fora da região central, não havendo tanta variação na concentração dos bairros entre as 7 e 13 horas. O contingente de ônibus pode estar atendendo demandas específicas de cada bairro ou proximidades, podendo os pontos de interesse da população dessas áreas ser uma área de estudo. Vale ressaltar que os limites dos intervalos do mapa (quantis) são menores que os dias de semana, pois há um número menor de ônibus trafegando pela cidade, quase metade dos dias de semana.

5.3 Velocidade média de ônibus

A velocidade média dos ônibus foi calculada previamente como um dos filtros na etapa de pré-processamento dos dados. O método de cálculo pode ser encontrado Seção 4.9.

A velocidade de veículos é um fator que impacta na mobilidade de VANETs. Por exemplo, se os nós de uma rede estão em alta velocidade o tempo de contato entre os veículos é mínimo. A velocidade é um parâmetro configurável em simuladores de mobilidade, como o SUMO[1], e também pode ser levada em conta na criação e modelagem de protocolos e estudo de viabilidade

[1] https://sumo.dlr.de/docs/index.html

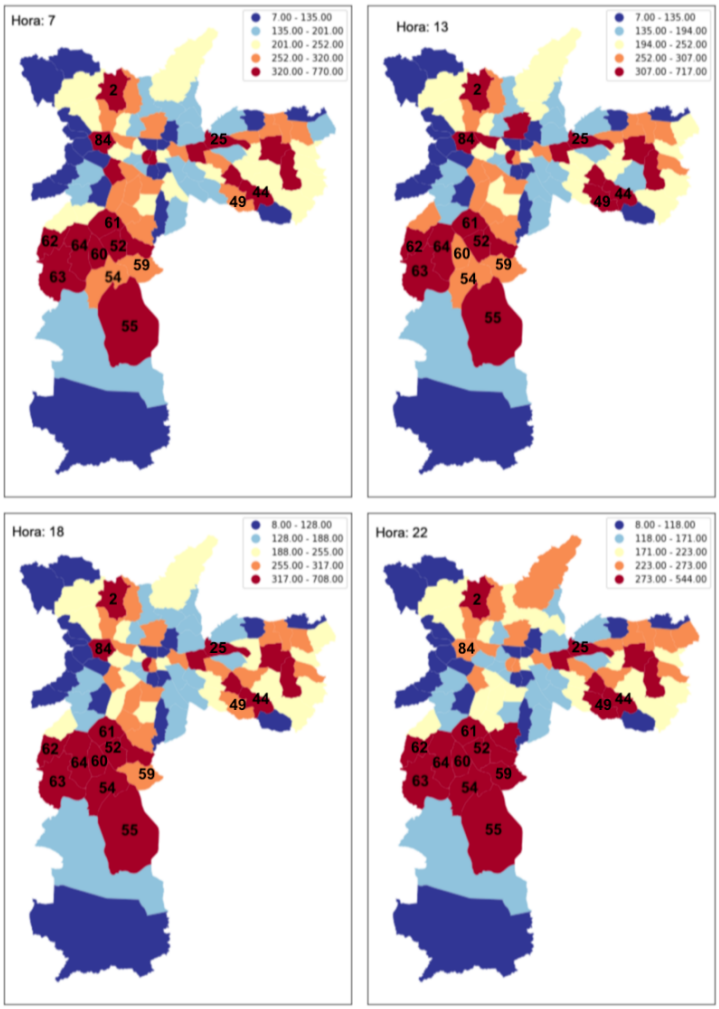


Figura 5.5: Número de ônibus ativos por região ao longo dia 4/10/2015 (elaborado pela autora).

de VANETs.

A velocidade média final de um ônibus associado a uma linha em determinado dia foi calculada a partir das velocidades instantâneas dos traços de mobilidades que é baseada na variação de distância e tempo de pontos adjacentes dos traços de mobilidade ordenados pelo tempo de registro. A partir dessas velocidades instantâneas é possível agregar os dados e produzir as métricas de velocidade desta seção.

5.3.1 Velocidade média dos ônibus por dia do mês

A velocidade média dos ônibus por dia do mês corresponde a média das velocidades médias finais dos ônibus em um dia.

Para calcular essa métrica, computou-se a velocidade média final para cada par ônibus e linha de um dia a partir da média das velocidades instantâneas. Posteriormente, calculou-se a

média de todas as velocidades médias finais daquele dia.

As velocidades médias de ônibus por dia estão presentes na Figura 5.6. Verifica-se pela imagem, que em dias úteis e sábados a velocidade está próxima a 14km/h, já em domingos (dias 4, 11, 18, 25) e feriado (dia 12), a velocidade ultrapassa 14km/h chegando a casa dos 15km/h. Em dias úteis e sábados a média fica entre 13,3km/h e 13,9km/h. Já aos domingos e feriado, a velocidade média fica entre 14,1km/h e 14,6km/h. O desvio padrão em todos os dias assume valores entre 3,9km/h e 5,9km/h.

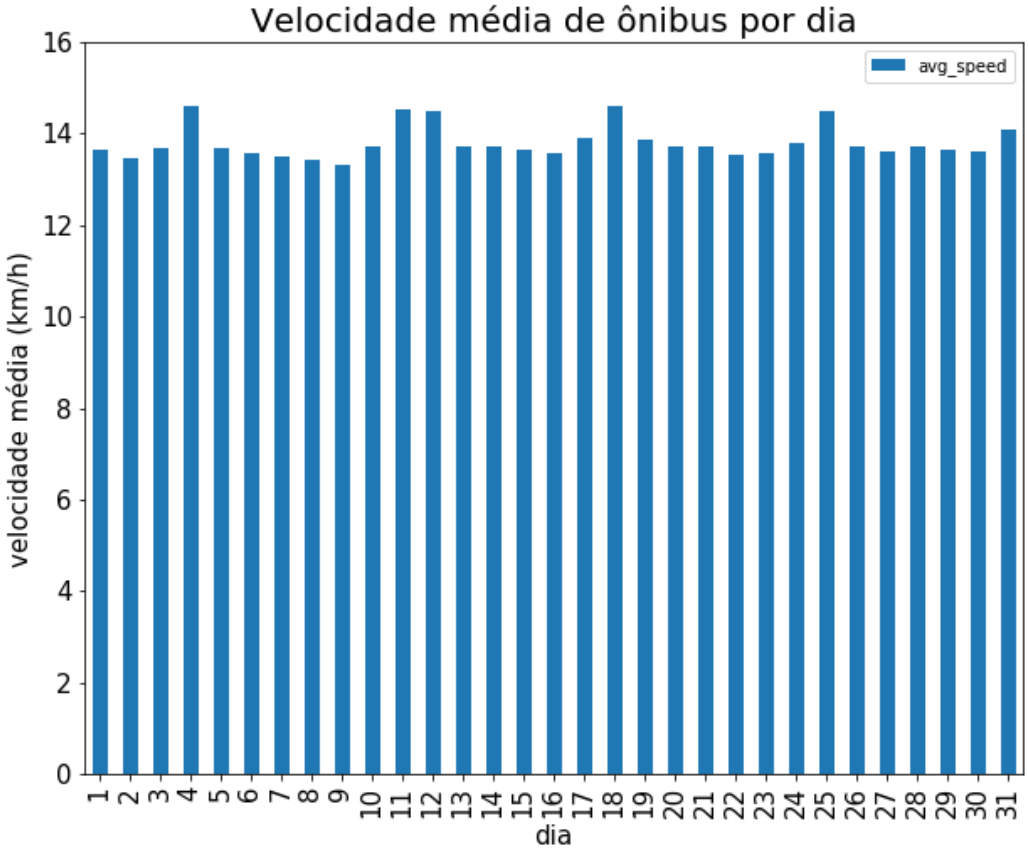


Figura 5.6: Velocidade média dos ônibus por dia do mês (elaborado pela autora).

A Figura 5.7 apresenta o gráfico ECDF (Empirical cumulative distribution function - função empírica de distribuição ou Distribuição acumulativa da frequência) para todas as velocidades média ônibus/linha registradas no dia 1/10 e no dia 4/10. Nos dias de semana, temos mais de 80% dos ônibus com velocidade menor que 15km/h. Já aos finais de semana, como no dia 4/10, 80% dos dados tem velocidade menor do que 20km/h. Esse gráfico reafirma o gráfico anterior, mostrando que finais de semana tendem a ter velocidade média de ônibus maior que os dias de semana. Em ambos os gráficos, a maior parte das velocidade médias finais registradas por

veículos está próxima a 15km/h.

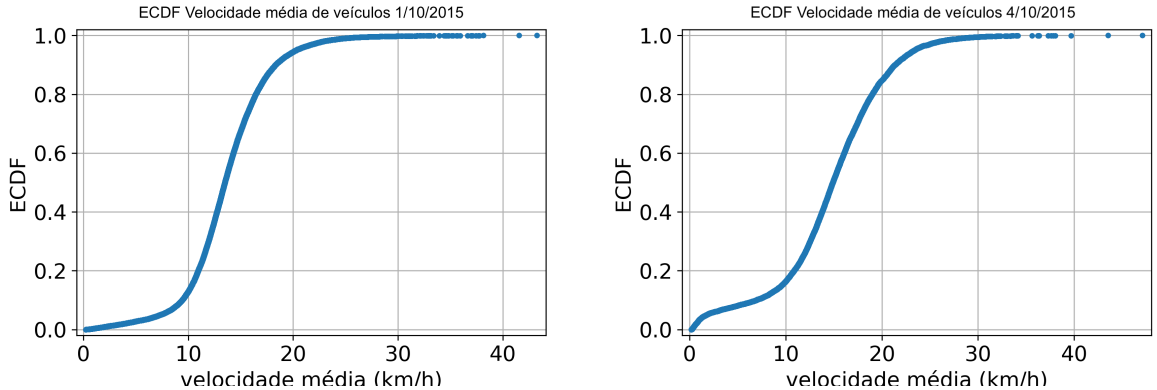


Figura 5.7: ECDF Velocidade média de ônibus (elaborado pela autora).

A razão para velocidades maiores aos finais de semana deve-se ao menor número de ônibus nas vias e também de veículos pessoais, pois há menos pessoas trabalhando, indo a escola, e deslocando-se para lugares rotineiros. Portanto, essa métrica pode ser impactada diretamente pelo número de veículos trafegando, enquanto em dias úteis há cerca de 13 mil ônibus ativos, aos finais de semana há cerca de 7000 mil ônibus.

5.3.2 Velocidade média dos ônibus por hora

A velocidade média dos ônibus ao longo das horas fornece uma visão mais granular do que acontece com a velocidade durante o dia.

Para calcular esta métrica, os traços de mobilidade com velocidade escalar instantânea já computada foram agrupados por *id_avl*, *line_id* e *hour_avl*, como pode ser visto nas linhas 3 e 4 do código da Figura 5.8. Portanto, calculou-se a média das velocidades instantâneas para o par ônibus/linha dentro de cada hora. Posteriormente, a média de todas as velocidades finais dentro daquele grupo de hora resultam na velocidade média de ônibus por hora (linhas 6-7).

```
1  traces = spark.read.parquet("s3://mobility-traces-sp/MO_15101/")
2
3  df_speed_hour_per_vehicle = traces.groupby("id_avl","line_id","hour_avl")\
4      .agg(F.avg("speed").alias("avg_speed_vehicle")))
5
6  df_speed_hour_day = df_speed_hour_per_vehicle.groupby("hour_avl")\
7  .agg(F.avg("avg_speed_vehicle").alias("avg_speed"))
```

Figura 5.8: Código em Spark para calcular a velocidade média de veículos por hora (elaborado pela autora).

A Figura 5.9 mostra as velocidades média de veículos por hora para os dias 5/10 (segunda-feira), 20/10 (terça-feira), 1/10 (quinta-feira), 17/10 (sábado), 4/10 (domingo), e 12/10 (feriado).

Os dias úteis apresentam comportamento semelhante. Às 6 da manhã observa-se que os ônibus atingem a maior velocidade média por dia acima de 13,5km/h. Às 8 da manhã esse valor cai para valores entre 13km/h e 13,5km/h. A partir das 8 da manhã, esse valor permanece caindo até ao 12:00 quando a velocidade é próxima a 11,5km/h. A partir do 12:00 a velocidade volta a subir, e às 16:00 a velocidade média chega a 12km/h. Às 18 horas, a velocidade decai para valores abaixo de 11,5km/h. A partir das 19:00, a velocidade média volta a subir para 12,5km/h.

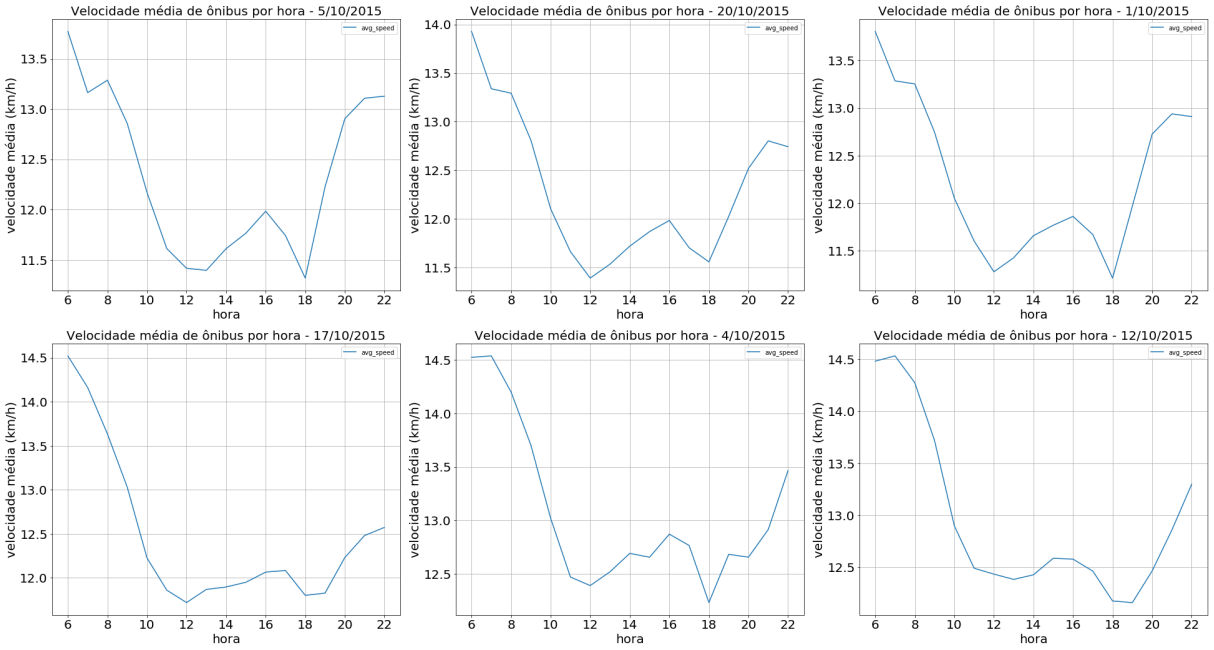


Figura 5.9: Velocidade média de veículos por hora (elaborado pela autora).

Nos dias úteis esse comportamento ocorre, pois no início da manhã às 6:00, apesar de haver mais de 13 mil ônibus ativos trafegando, não há tantos veículos pessoais andando para criar congestionamentos. Além disso, até às 9:00 a cidade de São Paulo possui faixas exclusivas de ônibus. A partir das 09:00, há mais veículos pessoais e comerciais trafegando em São Paulo, pois o horário comercial já começou, o que cria congestionamentos e diminui a velocidade dos ônibus. A partir do 12:00 até às 16:00, a velocidade volta a subir, o que pode ser explicado por menos ônibus trafegando, como visto na subseção 5.2.3, e menos veículos pessoais ou comerciais, pois este horário abrange horário de almoço até às 14 horas, e também horário de expediente e escolar, onde as pessoas estão paradas em algum lugar da cidade. Ainda sim, esse valor alto da tarde ainda não é alto como o da manhã, pois podem existir mais veículos de variados tipos (pessoais, ambulâncias, ônibus, comerciais, etc.) trafegando. Já a partir das 16:00 até às 18:00 a velocidade decresce, pois mais ônibus passam a rodar, e são horários finais do período comercial, a população está retornando para suas residências. A partir das 19:00, a velocidade aumenta novamente, pois o deslocamento da população para suas residências já está no final ou longe das regiões centrais, assim como o número de veículos comerciais e ônibus

diminuem.

Nos dias de final de semana e feriado, o comportamento é similar. A velocidade média pico encontrada é maior que nos dias úteis, chegando a 14,5 km/h às 6:00 da manhã, o que pode ser explicado por menos ônibus (quase metade dos dias úteis) ativos e menos veículos pessoais. Já a partir das 8:00 até ao 12:00 a velocidade cai e fica em valores abaixo de 12,5km/h no domingo e no feriado, e no sábado abaixo de 12km/h. Isso ocorre, pois nesses horários o contingente de ônibus aumenta, de veículos pessoais e comerciais também. É preciso apontar, que o sábado possui uma velocidade menor no período da manhã do que o domingo e o feriado, pois é um dia que parte do comércio ainda abre até às 18:00, então existe certa atividade comercial e também trabalhadores se deslocando. Já no período da tarde, a velocidade volta a subir e permanece constante entre 12 km/h e 13km/h até às 17:00. Esse comportamento acompanha o número de ônibus com pouca variação a tarde - acima de 9600 para sábados, e 6400 para domingo e feriado. Às 18:00, a velocidade média cai valores próximos a 12km/h, isso decorre do retorno para residência no final de período comercial (sábado), passeios (domingos e feriado), etc. A partir das 19:00, a velocidade média aumenta, pois há menos veículos e ônibus nas ruas por conta da demanda diurna já ter cessado.

5.3.3 Velocidade média dos ônibus por bairro

A velocidade média dos ônibus por bairro descreve a velocidade média dos veículos de cada bairro em determinado dia do mês. Para fazer este cálculo, computou-se a velocidade média de cada ônibus/linha dentro de um bairro, como mostra a linha 3-4 da Figura 5.10. Já a velocidade média por bairro foi calculada pelo agrupamento da computação anterior por bairro (campo *region*), como mostra a linha 6-7 da figura.

```
1  traces = spark.read.parquet("s3://mobility-traces-sp/MO_15101/")
2
3  df_speed_region_per_vehicle = traces.groupby("id_avl","line_id","region")\
4      .agg(F.avg("speed").alias("avg_speed"))
5
6  df_speed_region_day = df_speed_region_per_vehicle.groupby("region")\
7      .agg(F.avg("avg_speed").alias("avg_speed_region"))
```

Figura 5.10: Código em Spark para calcular a velocidade média por bairro (elaborado pela autora).

Na Figura 5.11, no dia 1/10 (dia útil), a maior parte dos bairros possuem velocidade inferior a 18km/h. Os bairros, centrais e do centro-sul possuem velocidades médias abaixo de 16km/h. Esse comportamento ocorre, pois ao longo do dia essas são áreas mais densas de veículos pessoais, comerciais e também ônibus, por serem centros comerciais, financeiros e tecnológicos da cidade de São Paulo. Já a velocidade média em bairros de extremidade, como Parelheiros,

Marsillac e Brazilândia, permanece próxima aos 12km/h e abaixo. Fatores físicos que podem justificar esse comportamento são bairros com vias com mais desvios entre as paradas, e ruas menos largas que as do centro, exigindo menor velocidade dos ônibus para seguir o caminho.

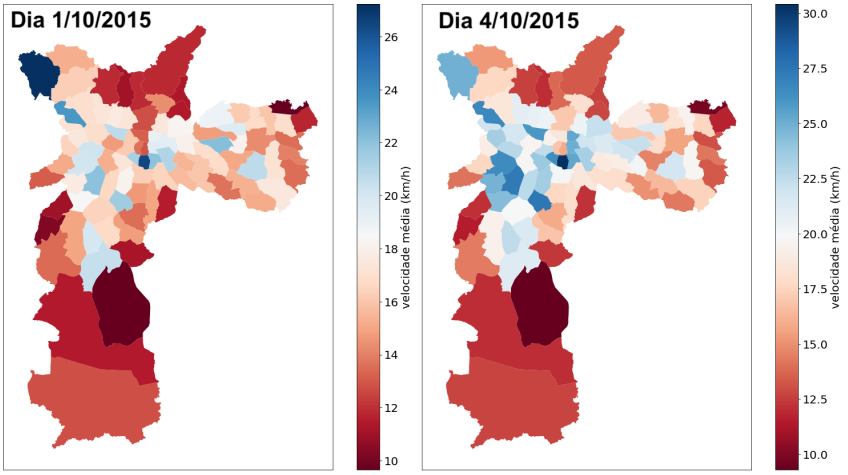


Figura 5.11: Velocidade média de ônibus por bairro (elaborado pela autora).

Já aos finais de semana, como no dia (4/10), as extremidades permanecem com comportamento similar aos dias úteis. Entretanto, os bairros centrais estão em azul indicando velocidades média mais altas de 20 a 25km/h. Aos finais de semana, a cidade de São Paulo tem menos ônibus circulando e veículos pessoais trafegando, o que pode aumentar a velocidade média. Além disso, as regiões centrais aos finais de semana não são as mais concentradas com ônibus ativos. Vale ressaltar a diferença entre os intervalos, enquanto no dia 1/10, a velocidade média do maior intervalo é 26km/h, no dia 4/10 esse valor é 30km/h.

5.4 Conectividade entre os ônibus

As métricas de conectividade entre os ônibus estão relacionadas a oportunidades de comunicação (encontros) que podem haver entre ônibus. Neste trabalho, uma oportunidade de conexão entre dois ônibus ocorre quando em determinado momento eles estão a menos de 100 metros um do outro, adotando a referência do protocolo IEEE 802.11p onde a comunicação entre dois veículos ocorre se estiverem no máximo a 100 metros um do outro. Esse é um valor de referência usado como raio de comunicação em VANETs (UPPOOR; FIORE, 2012) (ALVARENGA et al., 2014) (MARTINS; CUNHA, 2018) (UPPOOR et al., 2014) (POLAT; SOYTURK, 2016) (CAMPOS; MORAES; SILVA, 2010).

As métricas de conectividade demonstram as interações que o ônibus fazem com sua vizinhança em determinado momento. Veículos com alto grau de conectividade e mais ativos na rede po-

dem atuar como canal de troca de informações numa rede veicular. As métricas dessa seção proveem dos trabalhos de Martins e Cunha (2018), Alvarenga et al. (2014), Polat e Soyturk (2016) e Uppoor et al. (2014).

5.4.1 Modelagem dos traços de mobilidade em grafo temporal

Para determinar o momento de conexão entre dois ônibus, os traços de mobilidade foram modelados num grafo temporal baseado nos trabalhos de Alvarenga et al. (2014), Martins e Cunha (2018) e Polat e Soyturk (2016).

Alvarenga et al. (2014) e Martins e Cunha (2018) modelam os traços de mobilidade de Roma, Helsinque e São Francisco em um grafo $G(t) = (V, E)$. G é um grafo não direcionado em um tempo t, V um conjunto de veículos V_i e E um conjunto de arestas E_{ij}. A aresta E_{ij} só existe durante o tempo t entre o veículo V_i e V_j se $i \neq j$ (veículos diferentes). Nesses dois trabalhos, t pode ser definido como uma janela de tempo que agrupa os registros por um intervalo definido. Nesse caso, os dois estudos utilizaram uma janela de 15 minutos, portanto o grafo G representa o conjunto de todos os registros de veículos V que possuam traços de mobilidade dentro daquela janela de tempo. Por exemplo, existe um grafo que agrupa registros entre as 10:00 e 10:14:59. As arestas só existem entre dois veículos se eles estão a 100 metros um do outro no momento t.

Polat e Soyturk (2016) utilizam um conceito de grafo similar aos trabalhos do parágrafo anterior com tempo de agregação de 200 segundos, porém adiciona a componente geográfica. Este estudo divide o mapa da área geográfica de Colônia na Alemanha em células de 250x250m, então o grafo $G(t) = (V, E)$ existe para o tempo t e na célula C. A aresta entre o veículo V_i e V_j só existe se a distância entre os dois vértices (veículos) for menor do que o limite definido.

Essas técnicas de modelagem auxiliam na redução do escopo de busca para os veículos candidatos que podem estar se comunicando, além de diminuir a complexidade para lidar com variáveis de temporais e espaciais ao mesmo tempo.

Nesta pesquisa, foi utilizado o conceito de grafo temporal para modelar os traços de mobilidade. Um grafo $G(t) = (V, E)$ existe no momento t em um bairro da cidade de São Paulo. t agrupa os dados a cada 1 minuto. Por exemplo, existirá um grafo para o intervalo 10:00 e 10:00:59 para o bairro do Itaim Bibi. V representa o conjunto de veículos V_i que possuem traços de mobilidade naquele bairro em determinado momento t. E representa um conjunto de arestas. Uma aresta é uma oportunidade de conexão entre dois veículos no momento t em um bairro. A aresta só existe entre o veículo V_i e V_j, se $i \neq j$, ou seja, veículos diferentes, e se a distância geográfica entre esses dois vértices é menor ou igual a 100 metros.

Para definir a qual grafo o traço de mobilidade pertence, foi criada uma coluna *graph_id* que é a concatenação da hora inteira, do minuto da data de registro do equipamento AVL, e da região onde o traço foi registrado. Por exemplo, se o traço foi registrado às 10:48:35 no bairro da Barra Funda, então o grafo ao qual ele pertence é o *10-48-BarraFunda*.

5.4.2 Determinando as oportunidades de conexão

Para determinar a oportunidade de conectividade entre os veículos, os traços de mobilidade foram cruzados (com eles mesmos) a partir do campo *graph_id* (operação de *outer join*). Em seguida, foram eliminados do conjunto produzido, todos os registros cujo os identificadores dos dois veículos da conexão fossem iguais, ou seja, um ônibus se conectando a ele mesmo. O próximo passo foi computar a distância em metros entre os dois veículos de cada linha da tabela. Foram eliminados todos os registros cuja distância calculada fosse maior do que 100 metros.

Para facilitar a computação de métricas de conectividade, foi adicionada uma coluna *connection_id*, que é a concatenação dos identificadores dos dois ônibus de forma ordenada. Por exemplo, se há uma aresta entre o ônibus com o *id_avl* 35501 e o ônibus 12101 dentro de um grafo, então o *connection_id* dessa conexão (ou aresta) é ”12101-35501”.

Outro pré-processamento feito para o cálculo das métricas de conectividade foi dentro de cada grafo, eliminou-se elementos duplicados baseando-se nas colunas *id_avl_1*, e *id_avl_2*. Na vida real, os ônibus podem se comunicar mais de uma vez por minuto, porém para fins de contagem, é considerado um contato por minuto, ou seja, uma oportunidade conexão ocorre em determinada hora, minuto e localização. Portanto, só foi considerado o primeiro contato de um ônibus com o outro dentro de um grafo.

5.4.3 Número total de oportunidades de conexão

Uma oportunidade conexão representa uma aresta que existe entre dois ônibus para cada grafo num momento t em uma região, sendo t intervalos de 1 minuto ao longo do dia.

O grafo do dia 1/10 (quinta) possui cerca de 8197926 de vértices ao longo do dia. Já no dia 4/10 (domingo) houve 3512699 vértices. No total, ocorreram 25602814 arestas em dia de semana (1/10), ou seja, houveram cerca de 25 milhões de oportunidades de conexão entre ônibus ao longo do dia, descontando a reincidência de conexões foram 5 milhões oportunidades distintas. Já no dia 4/10, foram 8 milhões (8013742) de conexões totais ao longo do dia, e 1 milhão (1132836) de conexões únicas.

A Figura 5.12 mostra o número total de oportunidades de conexão entre os ônibus ao longo das horas para os dias 1/10 e 4/10. Nos dias de semana, como o dia 1/10, o número de conexões ao longo do dia tendem a permanecer constante, nesse caso entre 1,4 e 1,5 milhão por hora. Esse valor sobe a partir das 18:00, por conta do maior número de ônibus na ruas até às 19:00, e após esse horário ocorrem mais contatos em terminais e paradas. Já aos finais de semana, como no dia 4, o número de conexões varia ao longo do dia, crescendo rapidamente após às 14:00. O número de conexões por hora no dia 4/10 é menor do que nos dias de semana, pois há menos ônibus ativos trafegando, cerca de metade dos dias da semana.

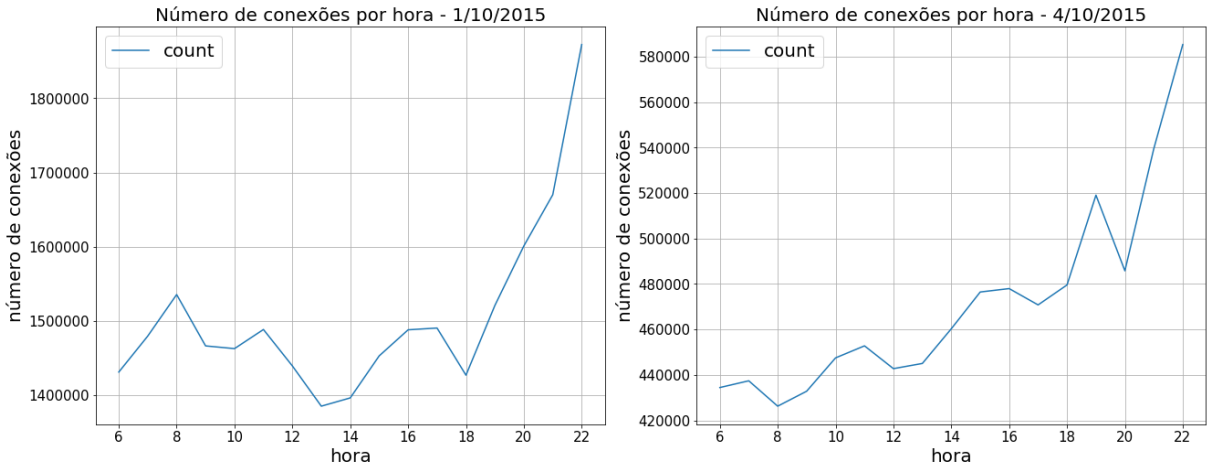


Figura 5.12: Número de conexões ao longo das horas (elaborado pela autora).

A Figura 5.13 mostra quais bairros tiveram mais oportunidades de conexão ao longo do dia. Tanto no dia 1/10 quanto no dia 4/10, os bairros onde há maior número de conexões são bairros de centros comerciais e onde há terminais de ônibus e outros modais, como: Grajaú (59), Santo Amaro (72), Capão Redondo (55), Jardim São Luís (65), Itaquera (32). O dia 1/10, um dia de semana, possui outros bairros com mais conexões: Pirituba (92), Lapa (87), Brás (10), São Mateus (42), e Santana (14).

Esses resultados revelam que o maior número de oportunidades conexão ocorrem na parte da noite e nas regiões onde há terminais de ônibus e centros comerciais.

5.4.4 Grau médio de conectividade dos ônibus

O grau de conectividade de um veículo representa a quantidade de oportunidades de comunicação que o mesmo tem com sua vizinhança em determinado grafo, ou seja, a quantidade de arestas que incidem sobre o vértice (veículo).

Para calcular essa métrica, foi computado o grau dos vértices de cada grafo do momento t, totalizando 8197926 vértices no dia 1/10 e 3512699 vértices ao longo do dia 4/10. Em seguida,

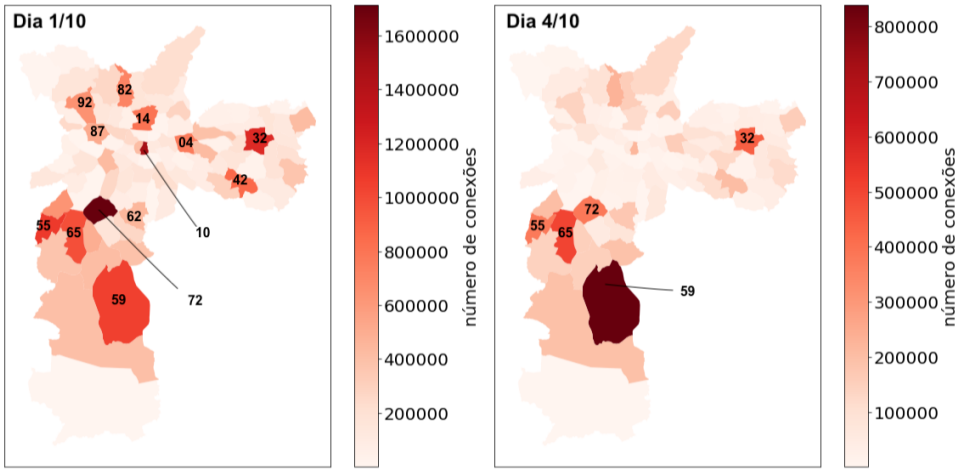


Figura 5.13: Número de conexões por região (elaborado pela autora).

calculou-se o grau médio de conectividade dos ônibus a partir da média do grau de todos vértices de grafos que ocorrem no momento t.

A Figura 5.14 aponta qual o grau médio de conectividade de ônibus ao longo do dia. Para dias úteis, observa-se que ao longo das horas até às 19:20 o grau de conectividade permanece constantemente em valores entre 5,5 e 6,5. Já a partir das 19:20 esse valor começa a subir, e às 22:40 o grau de conectividade chega a 11.

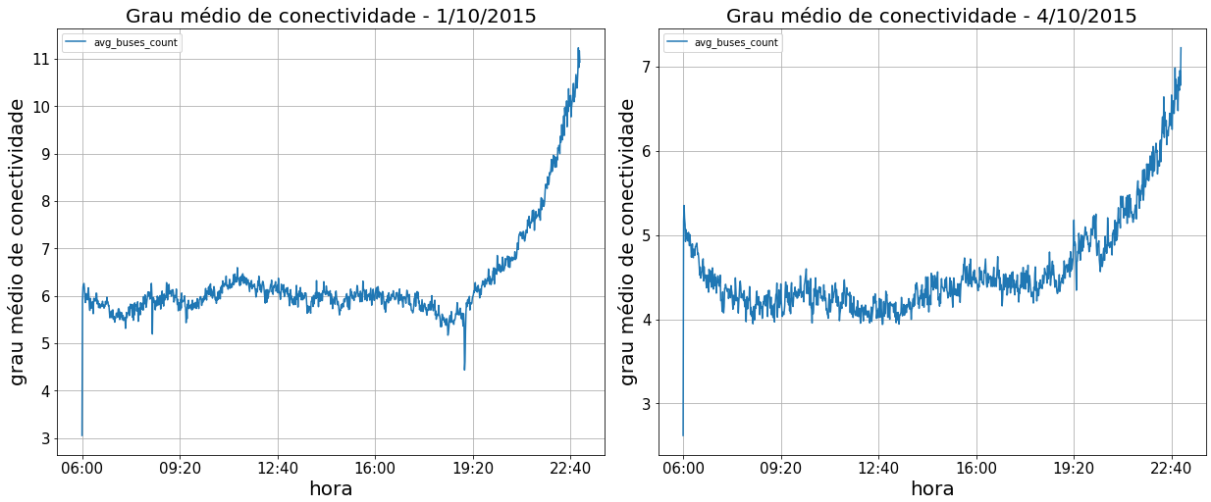


Figura 5.14: Grau médio de conectividade dos ônibus (elaborado pela autora).

Através da observação dos quantis do dia 1/10, encontrou-se os seguintes graus de conectividade: 25% com grau 2, 50% com grau 3, e 75% com grau 7. Esses resultados indicam que ao longo do dia os 75% vértices tem até grau 7 de conectividade, o que acompanha o valor constante encontrado no gráfico da Figura 5.14, e também o gráfico de distribuição ECDF da Figura 5.15. Investigando o quantil acima de 75% verificou-se que há uma parcela considerável dos vértices com grau de conectividade entre 7 e 21 (1475055 vértices no dia 1/10, o grau 20 no

Bairro	Número de vértices incidentes no dia 1/10
SANTO AMARO	131948
JARDIM SAO LUIS	87644
TATUAPE	68982
ITAQUERA	67701
CAPAO REDONDO	59642
GRAJAU	56224
PARELHEIROS	40148
LAPA	39494
PINHEIROS	39172

Tabela 5.1: Número de vértices incidentes por bairro

gráfico ECDF aponta que 90% dos dados possuem 20 conexões ou menos), o que pode representar possíveis candidatos para serem disseminadores centrais de informações para toda rede. Explorando essa parcela dos vértices com grau entre 7 e 21, verificou-se que a maior parte deles ocorre em Santo Amaro, Tatuapé, Itaquera, dentre outros bairros mostrados na Tabela 5.1. Em relação ao horário em que essas conexões ocorrem, verificou-se que a maior parte delas ocorreu às 20, 19, 21, 8 e 7 da manhã.

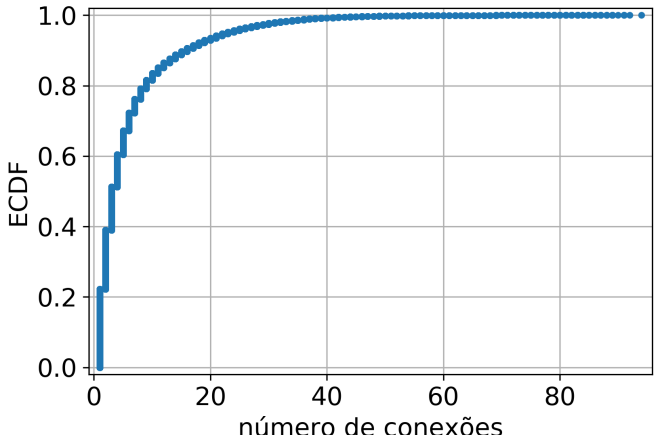


Figura 5.15: ECDF Grau de conectividade dos ônibus (elaborado pela autora).

A justificativa para aumento do grau de conectividade a partir das 19:20, pode estar associada aos ônibus estarem mais próximos em direção a terminais ou estacionamentos, que ficam situados em bairros como os da Tabela 5.1.

Já aos finais de semana, como no dia 4/10, o comportamento é similar aos dias de semana tendo ao longo do dia grau de conectividade entre 4 e 5. O grau de conectividade mais baixo que os dias de semana deve-se ao menor contingente de ônibus trafegando. Enquanto em dias de semana há cerca de 13 mil ônibus, aos domingos há por volta de 6900 ônibus.

Em relação aos vértices com grau entre 7 e 21 para os finais de semana, os bairros com maior concentração de vértices (veículos) com esse grau são os mesmos do dia 1/10. Para os finais de semana, esse grau de conectividade (7-21) ocorrem, principalmente, em horários a partir das 18:00. Esse comportamento contribui para a média do grau de conectividade aumentar após as 19:20, momento em que ocorrem mais vértices (veículos) com graus mais altos.

Polat e Soyturk (2016) e Alvarenga et al. (2014) mencionam que a densidade dos locais onde os ônibus estão afeta diretamente o grau de conectividade dos vértices, o que pode estar acontecendo a partir das 19:00, quando os ônibus podem estar mais concentrados próximos a terminais, e paradas, saindo ou chegando lá.

Para investigar melhor o aumento do grau de conectividade dos ônibus após às 19, etapas de pré-processamento do Capítulo 4 foram reaplicados em cima dos dados fazendo novas considerações a fim de acompanhar possíveis mudanças no grau de conectividade. Nesse novo pré-processamento, os traços de mobilidade foram particionados em intervalos de 15 minutos. Dentro dessa partição, para cada ônibus associado a uma linha foi feita a contagem de sequências de *shapes* (*shape_sequence*) naquele intervalo de 15 minutos. Se nesse intervalo, o ônibus não tivesse pelo menos 3 *shapes*, os traços do ônibus durante aquele período seriam eliminados indicando que o ônibus estaria parado, por exemplo, num terminal ou alguma parada. Os outros pré-processamentos, como o filtro de velocidade e distância máxima do traço para o *shape* mais próximo foram mantidos e também aplicados nesse novo processo.

Através desse novo pré-processamento, foi identificado a queda do grau de conectividade entre os ônibus a partir das 19:20 nos dias de semana, como pode ser visto na Figura 5.16. Já para finais de semana e feriado, este valor permanece constante o dia todo. Outra diferença que vale lembrar são os limites dos gráficos que também mudaram, indo para grau médio de conectividade de 4,5 a 6 durante o dia em dias de semana, e de 3 a 3,5 aos finais de semana. Esse comportamento revela que parte das oportunidades de conexões entre os ônibus pode ocorrer quando estão se encaminhando para terminais ou paradas, sendo momentos oportunos de troca de mensagens.

5.4.5 Oportunidades de conexão repetidas ao longo do dia

A métrica de oportunidades de conexão repetidas ao longo do dia diz respeito a quantas vezes o encontro entre dois veículos se repetiu considerando todos os grafos de um dia. É importante identificar os elementos que mais se conectam durante o dia, pois eles podem ser disseminadores de mensagens para grupos de veículos que não se conectam tão constantemente, ou seja, podem fazer com que pacotes de uma rede VANETs atinjam o maior número de nós

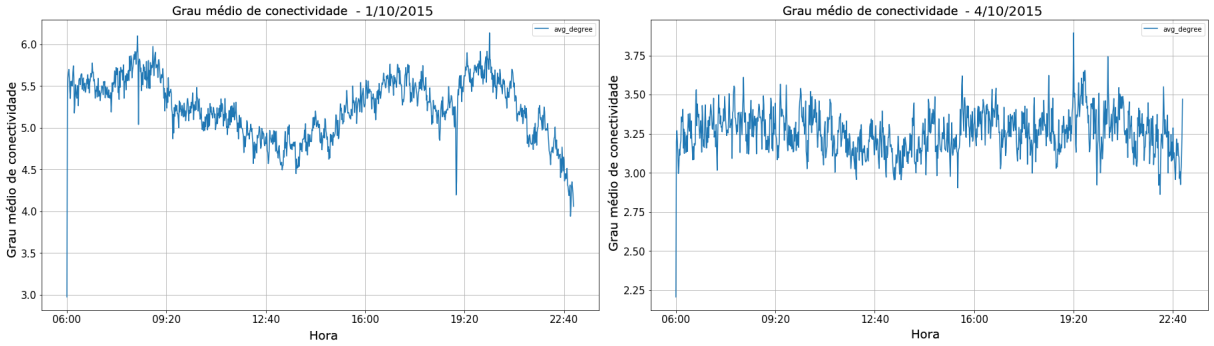


Figura 5.16: Grau médio de conectividade dos ônibus com novo filtro de 15 minutos (elaborado pela autora).

possível.

A Figura 5.17 demonstra o histograma de conexões repetidas para o dia 1/10 (quinta), 4/10 (domingo) , 20/10 (terça-feira) e 12/10 (feriado). Ambos os grafos revelam que a maior parte dos encontros ocorrem menos de 200 vezes por dia. Ao analisar os dados dos quantis presentes na Tabela 5.2, percebe-se que no corte de 75% para todos os dias, o valor é 5, com exceção do dia 17/10 que é 6. Esse valor aponta que pelo menos 75% das conexões se repetem até 5 vezes por dia. Há valores extremos, como conexões que se repetem 800 vezes, ou até 1000, o que pode indicar a permanência de conjuntos de ônibus em terminais durante alguns minutos por dia. Vale ressaltar que o processo de pré-processamento do conjunto de dados deste trabalho, apenas elimina ônibus que não se deslocam durante o todo dia, ou seja, ônibus parado o tempo todo. Portanto, há momentos do dia em que o ônibus se desloca, e em outros momentos encontra ônibus em terminais e paradas.

Dia	Média	Desvio padrão	min	max	25%	50%	75%
01-Out	5,1	10,74	1	858	1	2	5
04-Out	7,07	26,74	1	1001	1	2	5
12-Out	7,45	29,1	1	987	1	2	6
17-Out	6,42	20,39	1	994	1	2	5
20-Out	5,23	11,34	1	850	1	2	5

Tabela 5.2: Quantis das conexões repetidas ao longo do dia

Se o mesmo filtro (ônibus parado por 15 minutos) for aplicado nesse contexto, encontramos os resultados presentes na Figura 5.18. Nesse caso, o limite de conexões repetidas cai para 200 e 400 nos gráficos, porém ainda mantendo o padrão da maior parte das conexões se repetirem poucas vezes. Vale ressaltar que o comportamento de repetições ocorrendo mais vezes aos finais de semana e feriado (limites maiores dos gráfico) ainda acontece.

Na Figura 5.19 foram explorados o momento de encontro de ônibus e plotados no mapa

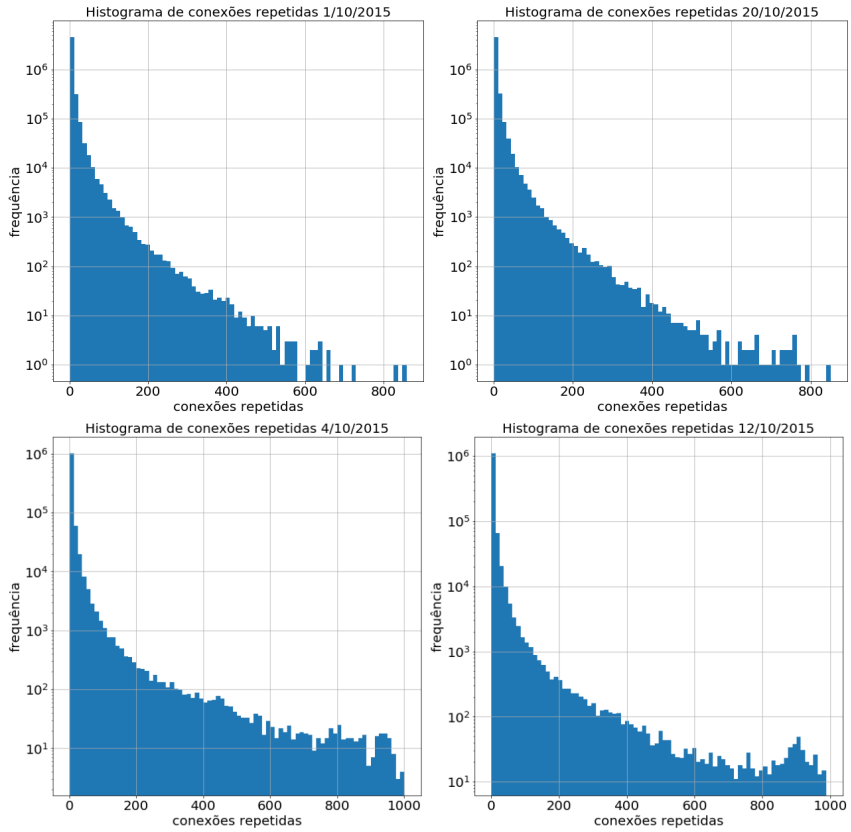


Figura 5.17: Histograma para conexões repetidas (elaborado pela autora).

para entender o local de encontro e o horário de encontro. Os pontos em vermelho e verde são traços de mobilidade de dois ônibus distintos em variadas horas do dia. No mapa A, mostra-se o encontro dos ônibus nas vias, e também as conexões que ocorrem dentro do terminal no bairro do Itaquera. Neste caso, esses encontros aconteceram em várias partes do dia dentro de horários das 8 , 7, 16, 20, e 14 horas. O mapa A ressalta que há encontros que ocorrem nas vias, mas também dentro dos terminais de parada, que podem servir para troca de informações entre vários veículos ao mesmo tempo. No mapa B, são encontros que ocorrem em lugares próximos em horários distintos no bairro de São Lucas, nesse caso os três encontros aconteceram às 17:14, 13:50, e 9:27. Essa situação ressalta o planejamento e rotina do transporte público mesmo em períodos distintos do dia é possível haver conexões repetidas. No mapa C, mais encontros em horários variados entre dois ônibus no bairro Belém.

A Figura 5.20 demonstra mais comportamentos que podem ocorrer nas conexões entre ônibus. No mapa A, dois ônibus se encontram em bairros distintos e em horários distintos, às 10:31 e às 12:15 em Socorro, e às 13:21 em Cidade Dutra. O mapa B, mostra um comportamento muito comum nos encontros, os ônibus se conectam por um período de minutos. Nesse caso, esses três encontros ocorreram às 9:17, 9:18, e 9:20.

Em média em dias de semana, um ônibus tem contato com 720 ônibus diferentes ao longo

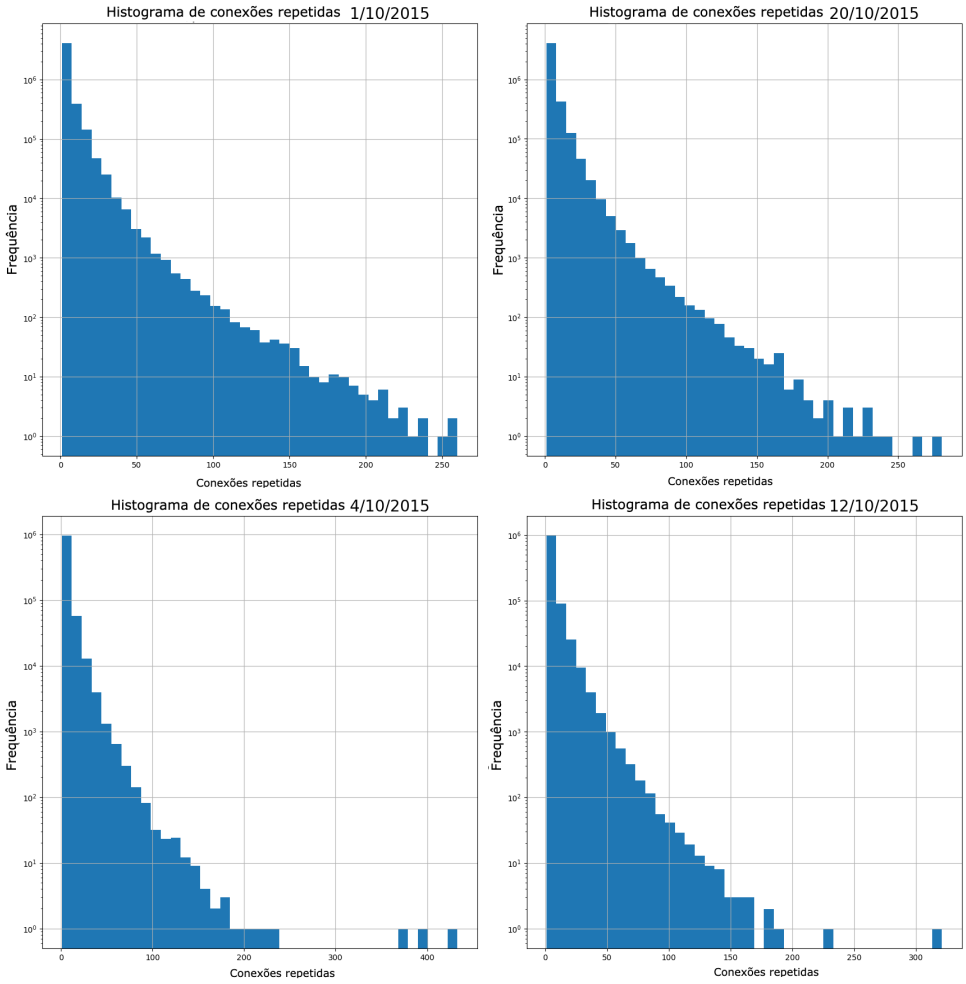


Figura 5.18: Histograma para conexões repetidas com novo filtro de 15 minutos (elaborado pela autora).

do dia, já aos finais de semana esse número cai para 300.

5.5 Resumo da caracterização

Para caracterizar o *dataset* de traços de mobilidade de ônibus fornecido pela SPTrans, foram utilizadas 3 métricas:

- **Número de ônibus ativos:** número de ônibus trafegando pela cidade. Foram usados 3 tipos de agregação - ônibus ativos ao longo das horas, por dia do mês de outubro e por bairro/hora;
- **Velocidade média dos ônibus**: média da velocidade média final de todas as velocidade escalares instantâneas dos pares ônibus linha. Foram utilizados como agregação de velocidade média: dia do mês, horas e bairros;

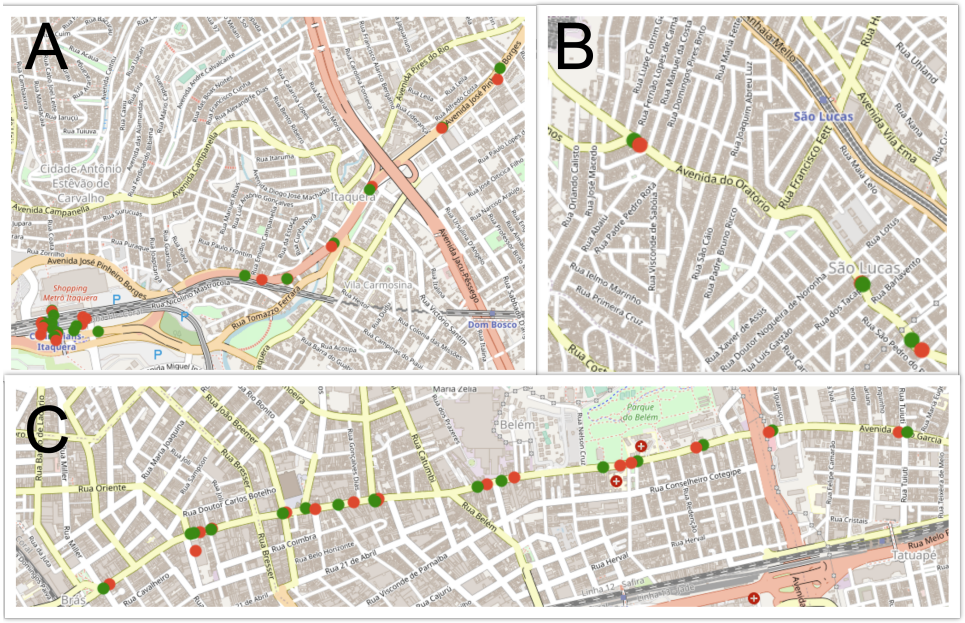


Figura 5.19: Exemplos de encontros entre ônibus (elaborado pela autora).

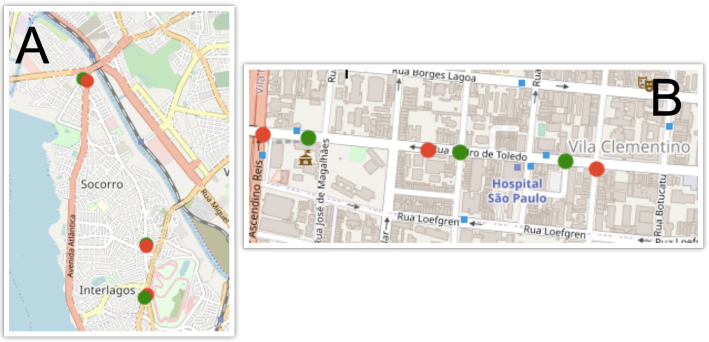


Figura 5.20: Encontros entre veículos no mapa (elaborado pela autora).

- **Conectividade entre os ônibus:** demonstra as oportunidades de interação que os ônibus tem com sua vizinhança em determinado momento. As métricas que compõem essa métrica são: número total de oportunidades de conexão que os ônibus fazem (número de arestas) ao longo das horas e por bairro, grau médio de conectividade dos ônibus (grau médio dos vértices) ao longo das horas, conexões repetidas ao longo do dia.

As métricas revelam que grande parte do número de ônibus ativos estão concentrados no início da manhã e no final da tarde em dias úteis, permanecendo na casa de 13 mil ônibus ao longo do dia, e decrescendo a partir das 22:00. Já aos finais de semana, o número de ônibus ativos fica entre 6600 e 6900 o dia todo. Esses comportamentos influenciam na velocidade

média dos ônibus que no início da manhã possuem as maiores velocidades média, mas decaem ao longo da tarde pelo número de veículos pessoais e outros ônibus na via.

O número de ônibus ativos também influencia na métrica de conectividade de ônibus a partir do momento que o maior número de conexões no período do dia ocorre junto com o pico do horário de mais ônibus ativos às 8:00 da manhã. Além disso, a densidade de ônibus ativos próximos a terminais e paradas tende a aumentar o grau de conectividade e o número de conexões a partir das 22:00 e no início da manhã.

Em relação ao grau de conectividade ao longo do dia, ele permanece constante na casa de 6 veículos conectados. Essa métrica acompanha o número de conexões que fica constante durante o dia e sobe a partir das 19:00.

O maior número de conexões e grau de conectividade ocorrem em bairros que possuem terminais de ônibus. Observa-se pelos mapas plotados, que é comum o encontro de ônibus nas vias quando estão fazendo o caminho para os terminais.

Em relação às conexões repetidas, grande parte ocorre apenas uma vez, mas o estudo dos quantis releva que o corte na parcela de 75% tem valor 5, ou seja, pelo menos 75% das conexões observadas acontecem até 5 vezes num dia. Conexões repetidas com valores muito maiores que 5 tendem a ocorrer nas vias próximas aos terminais.

A dinâmica da cidade de São Paulo afeta diretamente no resultado das métricas. Por exemplo, o horário comercial das 09:00 até às 18:00 aumenta o número de veículos pessoais e comerciais na rua reduzindo a velocidade dos ônibus. A demanda de ônibus acompanha a rotina de trabalho e de deslocamento dos residentes, com maior concentração de ônibus na regiões centrais e sul onde ficam grande parte dos centros comerciais, financeiro e tecnológico. As faixas de ônibus exclusivas afetam a velocidade de veículos no período da manhã, permitindo que os ônibus circulem mais rápido até às 09:00.

Aspectos físicos que podem afetar a mobilidade e conectividade são as estruturas das vias públicas e prédios. Geralmente, as regiões centrais possuem ruas mais largas e com menos desvios, impactando em velocidades maiores para os ônibus. Já em bairros na extremidade do mapa possuem mais desvios, vias mais irregulares, o que impacta em velocidades menores. Na região central da cidade onde há maior concentração de prédios, o número de conexões pode ser reduzido por conta das construções que prejudicam no sinal de captura, além de outros tipos de sinais provenientes de diferentes antenas nos prédios.

A rotina dos ônibus de trafegar pela cidade e em determinadas horas do dia passar por terminais, permite um maior número de conexões para os ônibus. Este trabalho avaliou também

ônibus que trafegam pela cidade e em determinados momentos do dia passam por terminais.

5.6 Mobilidade dos ônibus de São Paulo e o cenário de VANETs

A caracterização do *dataset* de traços de mobilidade de ônibus fornecido pela SPTrans permitiu identificar alguns aspectos dos dados que podem ser aplicados e/ou influenciar redes veiculares no cenário de São Paulo.

O número maior de conexões nos bairros de terminais (Santo Amaro, Grajaú, São Mateus, Lapa, etc.) e zonas próximas permite que esses lugares sejam zonas de troca de informação entre ônibus que normalmente não se encontram. Além disso, podem ser zona de descarga de informações dos ônibus para múltiplos nós ao mesmo tempo. Os terminais ou paradas podem servir como estações centrais caso sejam requisitadas na rede.

Bairros de extremidade podem ter ônibus com rotas mais específicas que não tem contato com os centrais. É preciso investigar quais são os nós que conseguiriam transmitir informações das regiões centrais para as extremidades da cidade.

As métricas encontradas permitem criar uma perspectiva para aplicação de protocolos de roteamento a partir da observação da variação de tráfego, número de conexões e velocidade, e também dos padrões (por região, por horário, dia da semana ou final de semana). As métricas de conectividade permitem ter uma definição de panorama da probabilidade de retransmitir os dados, e estudo de lugares e nós (ônibus) que permitam uma melhor cobertura da rede.

As zonais centrais e sul, aliadas de regiões de terminais, são regiões onde há maior densidade de ônibus e maior número de ônibus ativos. Para estudos de viabilidade, testes ou simulação, os traços de mobilidade que incidem sobre essas regiões podem ser começo de estudos para VANETs, onde há maior probabilidade de conexão.

As velocidades médias encontradas entre 13km/h e 20km/h para os ônibus se assemelham a de outros *datasets* de ônibus amplamente utilizados em VANETs, Chicago e Seatle (DOERING; WOLF, 2015). A velocidade contribui como uma parâmetro que pode ser configurado em simuladores de tráfego, e também pode ser levado em conta em modelagens.

Pontos de atenção em relação ao *dataset* deste trabalho para aplicação em VANETs são: zonas periféricas apresentam comportamento diferente das zonas centrais e sul; sazonalidades do tráfego e do número de ônibus ativos; fora da região dos terminais o grau de conectividade dos ônibus gira em torno de valores de 1 a 5; há veículos que mesmo longe de terminais pos-

suem conectividade (entre 7 e 21 ônibus), podendo ser nós centrais de conectividade da rede e disseminadores de informações; terminais podem servir como *hubs* centrais caso informação do ônibus precisa ser descarregada ou trocadas com maior número de ônibus e/ou outros modais ao mesmo tempo; as conexões da noite representam conexões que ocorrem mais frequentemente em terminais, portanto pode-se começar o estudo conexões com dados diurnos.

Apesar do *dataset* de ônibus seguir padrões de comportamento, principalmente devido à natureza do planejamento do transporte público, ele oferece diferentes cenários que podem ser testados com VANETs. Cenários de regiões com mais ou menos conexões, bairros com maior número de ônibus ativos, bairros com maior probabilidade de disrupção do sinal, dentre outros. Esses diferentes cenários permitem que se escolha parte dado ou utilize-o integralmente com base no objetivo do protocolo, dos testes com redes veiculares, ou até mesmo *benchmark* com outras bases.

Capítulo 6

Conclusão

Para caracterizar o *dataset* de mobilidade de ônibus da cidade de São Paulo, o primeiro passo foi uma pesquisa bibliográfica para obter métricas que poderiam caracterizar traços de mobilidade. Foram encontradas métricas relacionadas a mobilidade dos veículos, como velocidade, e de conectividade com foco em VANETs, como o grau de conectividade de veículos ao longo do dia.

Antes de iniciar o processo de caracterização, o *dataset* precisou passar por etapas de pré-processamento para: formatar os dados, eliminar dados desnecessários, e filtrar ruídos. Após a conclusão das etapas de pré-processamento, foram extraídas as métricas de conectividade e mobilidade dos traços de mobilidade dos ônibus de São Paulo.

Através da caracterização do *dataset* de mobilidade dos ônibus de São Paulo, foi possível compreender o comportamento dos veículos ao longo do mês de outubro de 2015, encontrando comportamentos que se repetem em certos intervalos. Além disso, foram obtidos valores e dados estatísticos que podem ser utilizados para o desenvolvimento de modelos de mobilidade, e também para validação de simulações e teste em VANETs. Os traços de mobilidade analisados compreendem registros de geolocalização de mais de 14 mil ônibus entre às 6:00 e 22:59 durante o mês de outubro de 2015.

A métrica de ônibus ativos revela que em dias úteis há cerca de 13 mil ônibus trafegando pela cidade. Já aos finais de semana e feriados, há cerca de 7 mil ônibus, pouco mais da metade dos dias de semana. O maior número de ativos ocorre entre às 6:00 e 8:00, e entre às 17 e 18 num dia de semana. Enquanto aos finais de semana o número é constante ao longo do dia. As áreas com maior concentração de ônibus são os bairros centrais e do sul da cidade em dias de semana. Aos finais de semana, os bairros com mais veículos ativos são os bairros onde há terminais de ônibus, como Santo Amaro, Grajaú, e São Mateus.

A métrica de velocidade média dos ônibus aponta que em dias de semana os ônibus tendem a ter velocidade próximas a 13km/h, já aos finais de semana esse valor sobe para 14,5km/h.

A métrica de conectividade revela que os ônibus tem grau médio de conectividade 6 durante o dia. Em relação a repetição das conexões, 75% se repetem até 5 vezes por dia. Verificou-se que esses encontros podem ocorrer tanto durante o tráfego do ônibus pelas vias, quanto em terminais. As conexões podem ocorrer também no mesmo local em horários distintos, e podem ocorrer em lugares distintos da cidade.

O valor das métricas, principalmente, quando visualizadas ao longo do tempo, indicam a presença de comportamentos que se repetem ao longo dos dias, e outros que são específicos de certos dias da semana. As variações ao longo do dia ocorrem principalmente devido à dinâmica da cidade de São Paulo, como congestionamentos em dias de semana no começo da manhã, faixas exclusivas para ônibus, localidade dos centros comerciais e financeiros, estrutura das vias e do bairro que pode tornar a conexão entre os ônibus difícil de ocorrer.

Em relação às redes veiculares, os valores encontrados podem servir como parâmetro para condução de simulação e de testes. Por exemplo, as zonas centrais e sul da cidade, aliadas as regiões de terminais, são regiões onde há maior número de ônibus ativos, e onde ocorreu grande parte das conexões. Para testes, os traços de mobilidade que incidem sobre essas regiões podem ser começos do estudo, onde há maior probabilidade de conexão.

Apesar do *dataset* de ônibus ter padrões de comportamento, principalmente devido à natureza do planejamento do transporte público, ele oferece diferentes cenários que podem ser testados em VANETs. Cenários de regiões mais e menos concentradas, bairros com maior e menor probabilidade de disrupção da comunicação. Esses contextos permitem que se escolha parte do dado ou utilize-os de forma integral com base no objetivo do protocolo, ou dos testes.

Como contribuições desta pesquisa:

- foi possível identificar conjuntos de métricas de mobilidade e conectividade na literatura;
- foram extraídas características do *dataset* através das métricas. As características foram quantificadas e disponibilizadas com recursos gráficos e de maneira descritiva;
- esta pesquisa descreveu todo o processo de pré-processamento e cálculo da métricas o que pode ser aplicado para outros conjuntos de dados de mesma natureza;
- disponibilização de todos os códigos de pré-processamento e cálculos das métricas disponíveis no Github no link ⟨https://github.com/caroljunq/sptrans-data-analysis⟩;
- disponibilização de parte do *dataset* sem nenhum processamento, e parte pré-processado

pronto para extração das métricas através da plataforma Kaggle no link ⟨https://www.kaggle.com/caroljunq/sao-paulo-bus-mobility-traces-oct-2015⟩.

O primeiro ponto de melhora no trabalho poderia ser o método de pré-processamento. Para a fase de *map matching*, além de considerações geométricas, o processo pode ser feito através do OpenStreetMap garantindo um mapeamento que considera aspectos de topologia, assim melhorando a exatidão na detecção e eliminação de ruídos. Outro ponto de melhora é separar as conexões entre conexões de terminal e conexões de via, para identificar qual a proporção dos tipos de conexões ocorrem durante o dia.

6.1 Trabalhos futuros

Como trabalhos futuros propõe-se:

- aplicar os processos deste trabalho para outros traços de mobilidade para comparação de resultados;
- aplicar o *dataset* produzido em simulações de VANETs ou criação de modelo de mobilidade;
- identificar redes de ônibus através de clusterização hierárquica, e medir o tempo para pacotes chegaram a maior parte dos grupos da rede através de protocolos epidêmicos;
- utilizar o *dataset* produzido para avaliar diferentes protocolos de roteamento para que os pesquisadores tenham uma noção do impacto de desempenho da mobilidade dos ônibus de São Paulo numa rede veicular;
- realizar o estudo do aspecto social para incorporar à caracterização dados sobre o comportamento de veículos pessoais, demandas de outros modais de transporte, dando mais contexto e explicação da ocorrência dos valores das métricas coletadas.

REFERÊNCIAS

AL-SULTAN, S.; AL-DOORI, M. M.; AL-BAYATTI, A. H.; ZEDAN, H. A comprehensive survey on vehicular Ad Hoc network. *Journal of Network and Computer Applications*, Elsevier, v. 37, n. 1, p. 380–392, 2013. ISSN 10958592. Disponível em: ⟨http://dx.doi.org/10.1016/j.jnca.2013.02.036⟩.

ALVARENGA, D.; CUNHA, F. D. D.; VIANA, A. C.; MINI, R. A. F.; LOUREIRO, A. A. F. Classificando Comportamentos Sociais em Redes Veiculares. In: SBC. *XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. Florianópolis, Brazil, 2014. Disponível em: ⟨https://hal.inria.fr/hal-01302236⟩.

APACHE. *Apache Spark™ - Unified Analytics Engine for Big Data*. 2019. Disponível em: ⟨https://spark.apache.org/⟩.

AYYASH, M.; ALSBOU, Y.; ANAN, M. Introduction to Mobile Ad-Hoc and Vehicular Networks. In: ____. *Wireless Sensor and Mobile Ad-Hoc Networks Vehicular and Space Applications*. New York, Estados Unidos: Springer, 2015. cap. 1, p. 33–46. ISBN 9781493924684.

CAMPOS, C. A. V.; MORAES, L. F. M. de; SILVA, R. F. Caracterização da mobilidade veicular e o seu impacto nas redes veiculares tolerantes a atrasos e desconexões. In: *Proceedings of the 28nd Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuidos*. [S.l.: s.n.], 2010. p. 585–598.

CHODOROW, K. *MongoDB: The Definitive Guide*. Sebastopol, CA - Estados Unidos: O'Reilly Media, Inc., 2013. ISBN 1449344682, 9781449344689.

DOERING, M.; WOLF, L. Opportunistic vehicular networking: Large-scale bus movement traces as base for network analysis. In: *2015 International Conference on High Performance Computing Simulation (HPCS)*. [S.l.: s.n.], 2015. p. 671–678.

DOMINGUES, A. C. S. A.; SILVA, F. A.; LOUREIRO, A. A. F. Space and time matter: An analysis about route selection in mobility traces. In: *2018 IEEE Symposium on Computers and Communications (ISCC)*. [S.l.: s.n.], 2018. p. 00958–00963. ISSN 1530-1346.

HARRI, J.; FILALI, F.; BONNET, C. Mobility models for vehicular ad hoc networks: a survey and taxonomy. *IEEE Communications Surveys Tutorials*, v. 11, n. 4, p. 19–41, Fourth 2009. ISSN 1553-877X.

HARTENSTEIN, H.; LABERTEAUX, K. *VANET: Vehicular Applications and Inter-Networking Technologies*. 1. ed. Torquay, UK: John Wiley & Sons Ltd., 2010. 1–435 p. ISBN 978-0-470-74056-9.

IBGE. *IBGE — Cidades@ — São Paulo — São Paulo — Panorama*. 2019. Disponível em: ⟨https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama⟩.

LAKATOS EVA MARIA; MARCONI, M. d. A. *Fundamentos de metodologia científica*. 5. ed. São Paulo, Brasil: Editora Atlas, 2003.

MARTINS, K. S.; CUNHA, F. D. Explorando dados urbanos: um estudo usando viagens de taxi da cidade de são francisco. In: *Proceedings of the 2nd Workshop on Urban Computing*. [S.l.: s.n.], 2018. (CoUrb 2018).

MELNIK, S.; GUBAREV, A.; LONG, J. J.; ROMER, G.; SHIVAKUMAR, S.; TOLTON, M.; VASSILAKIS, T. Dremel: Interactive analysis of web-scale datasets. *Proc. VLDB Endow.*, VLDB Endowment, v. 3, n. 1–2, p. 330–339, set. 2010. ISSN 2150-8097. Disponível em: ⟨https://doi.org/10.14778/1920841.1920886⟩.

MOUSTAFA, H.; ZHANG, Y. *Vehicular Networks: Techniques, Standards, and Applications*. 1. ed. Boston, MA, USA: Auerbach Publications, 2009. 450 p. ISBN 9781420051841.

NEWSON, P.; KRUMM, J. Hidden markov map matching through noise and sparseness. In: . [s.n.], 2009. p. 336–343. Disponível em: ⟨https://www.microsoft.com/en-us/research/publication/hidden-markov-map-matching-noise-sparseness/⟩.

POLAT, B. K.; SOYTURK, M. An alternative approach to mobility analysis in vehicular ad hoc networks. In: *2016 IEEE Symposium on Computers and Communication (ISCC)*. [S.l.: s.n.], 2016. p. 244–249.

PONS, I.; MONTEIRO, J.; SPEICYS, R. *Big Data para análise de métricas de qualidade de transporte: metodologia e aplicação*. São Paulo, SP, 2015. 90 p. Disponível em: ⟨https://scipopulis.com/docs/RELAT\%C3\%93RIO_TECNICO_ANTP-SPTRANS-03-PRE_IMPRESS\%C3\%83O-04.pdf⟩.

POSTGIS. *18. Geography — Introduction to PostGIS*. 2019. Disponível em: ⟨http://postgis.net/workshops/postgis-intro/geography.html⟩.

POSTGIS. *9. Geometries*. 2019. Disponível em: ⟨http://postgis.net/workshops/postgis-intro/geometries.html⟩.

SANTANA, E. F. Z.; KANASHIRO, L.; KON, F. Geração de rastros de mobilidade para experimentos em redes veiculares. In: *Proceedings of the 2nd Workshop on Urban Computing*. [S.l.: s.n.], 2018. (CoUrb 2018), p. 1–10.

SILVA, F. A.; CELES, C.; BOUKERCHE, A.; RUIZ, L. B.; LOUREIRO, A. A. Filling the gaps of vehicular mobility traces. In: *Proceedings of the 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. New York, NY, USA: ACM, 2015. (MSWiM '15), p. 47–54. ISBN 978-1-4503-3762-5. Disponível em: ⟨http://doi.acm.org/10.1145/2811587.2811612⟩.

SILVA, R. F. *Caracterização da mobilidade veicular e o seu impacto nas redes veiculares tolerantes a atrasos e desconexões*. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, São Paulo, SP, 1 2010.

SINGH, S.; AGRAWAL, S. Vanet routing protocols: Issues and challenges. In: *2014 Recent Advances in Engineering and Computational Sciences (RAECS)*. [S.l.: s.n.], 2014. p. 1–5.

SPTRANS. *SPTrans*. 2019. Disponível em: ⟨http://www.sptrans.com.br/⟩.

The PostgreSQL Global Development Group. *PostgreSQL 11.3 Documentation*. [S.l.], 2019. 2626 p. Disponível em: ⟨https://www.postgresql.org/files/documentation/pdf/11/postgresql-11-A4.pdf⟩.

UPPOOR, S.; FIORE, M. Insights on metropolitan-scale vehicular mobility from a networking perspective. In: *Proceedings of the 4th ACM International Workshop on Hot Topics in Planet-scale Measurement*. New York, NY, USA: ACM, 2012. (HotPlanet '12), p. 39–44. ISBN 978-1-4503-1318-6. Disponível em: ⟨http://doi.acm.org/10.1145/2307836.2307848⟩.

UPPOOR, S.; TRULLOLS-CRUCES, O.; FIORE, M.; BARCELO-ORDINAS, J. M. Generation and analysis of a large-scale urban vehicular mobility dataset. *IEEE Transactions on Mobile Computing*, v. 13, n. 5, p. 1061–1075, May 2014.

WEN, M.; ROSA, T. d. O.; SOUZA, M. C.; ALEIXO, R. P.; ALVES, C.; Sá, L.; SANTANA, E. F. Z.; KON, F. Criação de modelo para simulação de movimentação de Ônibus a partir de dados reais. In: *Proceedings of the 1st Brazilian Workshop on Smart Cities*. [S.l.: s.n.], 2018. (WBCI 2018), p. 1–10.

WENG, J.; WANG, C.; HUANG, H.; WANG, Y.; ZHANG, L. Real-time bus travel speed estimation model based on bus gps data. *Advances in Mechanical Engineering*, v. 8, 11 2016.

WU, Y.; ZHU, Y.; LI, B. Trajectory improves data delivery in vehicular networks. In: IEEE. *2011 Proceedings IEEE INFOCOM*. [S.l.], 2011. p. 2183–2191.

XU, R.; LI, Y.; CHEN, S. On the opportunistic topology of taxi networks in urban mobility environment. *IEEE Transactions on Big Data*, p. 1–1, 2018. ISSN 2332-7790.

YAI, A. K. *Análise e visualização dedados do transportepúblico de ônibus dacidade de São Paulo*. Dissertação (Mestrado) — Universidade de São Paulo, São Paulo, SP, 1 2016.

ZAHARIA, M.; CHOWDHURY, M.; FRANKLIN, M. J.; SHENKER, S.; STOICA, I. Spark: Cluster computing with working sets. In: *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*. Berkeley, CA, USA: USENIX Association, 2010. (HotCloud'10), p. 10–10. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=1863103.1863113⟩.

ZAHARIA, M.; XIN, R.; WENDELL, P.; DAS, T.; ARMBRUST, M.; DAVE, A.; MENG, X.; ROSEN, J.; VENKATARAMAN, S.; FRANKLIN, M.; GHODSI, A.; GONZALEZ, J.; SHENKER, S.; STOICA, I. Apache spark: A unified engine for big data processing. *Communications of the ACM*, v. 59, p. 56–65, 11 2016.

ZHENG, L.; XIA, D.; ZHAO, X.; LIU, W. Mining trip attractive areas using large-scale taxi trajectory data. In: *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*. [S.l.: s.n.], 2017. p. 1217–1222.

ZHENG, Y. Trajectory data mining: An overview. *ACM Transaction on Intelligent Systems and Technology*, September 2015.

ZHENG, Y.; ZHOU, X. *Computing with Spatial Trajectories*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2011. ISBN 1461416280, 9781461416289.