UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

OBTENÇÃO DA DEMANDA DE PASSAGEIROS EM PONTOS DE PARADA DE ÔNIBUS UTILIZANDO CAPTAÇÃO DE SONDAS WI-FI PARA PLANEJAMENTO OPERACIONAL

SAMUEL SEIXAS MARTIN

ORIENTADORA: PROFA. FABIANA SERRA DE ARRUDA

COORIENTADOR: PROF. PASTOR WILLY GONZALES TACO

MONOGRAFIA DE PROJETO FINAL EM TRANSPORTES

BRASÍLIA, DF: DEZEMBRO DE 2025

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

OBTENÇÃO DA DEMANDA DE PASSAGEIROS EM PONTOS DE PARADA DE ÔNIBUS UTILIZANDO CAPTAÇÃO DE SONDAS WI-FI PARA PLANEJAMENTO OPERACIONAL

SAMUEL SEIXAS MARTIN

MONOGRAFIA DE PROJETO FINAL II SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL DA UNIVERSIDADE DE BRASÍLIA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM ENGENHARIA CIVIL.

APROVADA POR:

FABIANA SERRA DE ARRUDA, DOUTORA (Universidade de Brasília) (ORIENTADORA)

PASTOR WILLY GONZALES TACO, DOUTOR (Universidade de Brasília) (COORIENTADOR)

PAULO CÉSAR MARQUES DA SILVA, DOUTOR (Universidade de Brasília) (EXAMINADOR INTERNO)

YVES HENRIQUE MARQUES RIBEIRO, MESTRE (Universidade de Brasília), ESPECIALISTA EM PROJETOS E TRANSPORTE (Urbi Mobilidade) (EXAMINADOR EXTERNO)

DATA: BRASÍLIA/DF, 15 de DEZEMBRO de 2025

FICHA CATALOGRÁFICA

M379d	Martin, Samuel Seixas. Obtenção da demanda de passageiros em pontos de parada de ônibus utilizando captação de sondas Wi-Fi para planejamento operacional / Samuel Seixas Martin; Orientador: Fabiana Serra de Arruda; co-orientador Pastor Willy Gonzales Taco. Brasília, 2025. 119 f. Trabalho de Conclusão de Curso (Graduação – Engenharia Civil) Universidade de Brasília, 2025. 1. Demanda. 2. Pontos de ônibus. 3. Sensor. 4. Wi-Fi. I. Arruda, Fabiana Serra de, orient. II. Taco, Pastor Willy Gonzales, co-orient. III. Título.

REFERÊNCIA BIBLIOGRÁFICA

MARTIN, S. S. (2025). Obtenção da demanda de passageiros em pontos de parada de ônibus utilizando captação de sondas Wi-Fi para planejamento operacional. Monografia de Projeto Final em Engenharia Civil, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, DF, 119 p.

CESSÃO DE DIREITOS

AUTOR: Samuel Seixas Martin

TÍTULO DA MONOGRAFIA DE PROJETO FINAL: Obtenção da demanda de passageiros em pontos de parada de ônibus utilizando captação de sondas Wi-Fi para planejamento operacional.

GRAU / ANO: Bacharel em Engenharia Civil / 2025

É concedida à Universidade de Brasília a permissão para reproduzir cópias desta monografia de Projeto Final e para emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte desta monografia de Projeto Final pode ser reproduzida sem a autorização por escrito do autor.

_____

Nome do Aluno

RESUMO

A falta de dados contínuos de demanda em tempo real é um gargalo no Planejamento de Transportes, a fim de ajustar a oferta atualmente conhecida à demanda que varia no tempo e no espaço. Visando resolver este problema, o presente estudo propõe no âmbito de Sistemas Inteligentes de Transporte o desenvolvimento de um protótipo para captar dados de campo que possibilitem a estimativa do número de pessoas em pontos de parada de ônibus para fins de planejamento e operação do sistema. O protótipo é baseado no microprocessador ESP32 com tecnologia de captação passiva de pacotes WLAN públicos do tipo sondas Wi-Fi, dos quais podem ser aferidos os parâmetros de intensidade do sinal, endereço MAC, e rede procurada. Posteriormente, estes dados foram agrupados em atributos por janelas de um minuto, e alimentados em um modelo de Máquina de Vetores de Suporte para um aprendizado de máquina supervisionado correlacionando o número de dispositivos detectados com o número de passageiros, de forma a estimar a demanda no ponto de instalação do sensor. Foram realizadas cinco coletas de campo totalizando dez horas de dados, e nos resultados observaram-se uma grande quantidade de endereços MAC aleatorizados, que não se repetiam entre as coletas, e variabilidade significativa entre os histogramas de classes de intensidade de sinal para coletas diferentes. A impossibilidade de correlação entre os sinais fortes com os dados de referência por regressão linear evidenciou a utilidade do modelo em estimar a demanda no ponto de parada contanto que os passageiros se encontrem dentro do raio de alcance de aproximadamente seis metros. No entanto, o conjunto sensor se mostra viável em captar dados que viabilizem a operação de sistemas dinâmicos por meio do monitoramento em tempo real da demanda, de forma a assegurar respostas rápidas a eventos imprevistos, melhorando a qualidade do serviço prestado à população.

Palavras-Chave: Demanda. Pontos de ônibus. Sensor. Wi-Fi.

SUMÁRIO

1 INTRODUÇÃO 16
1.1 OBJETIVO 19
1.1.1 Objetivo Geral 19
1.1.2 Objetivos Específicos 19
1.2 JUSTIFICATIVA 20
1.3 ESTRUTURA DO PROJETO FINAL 21
2 REFERENCIAL TEÓRICO 22
2.1 IMPACTO DA TECNOLOGIA NA AVALIAÇÃO DA QUALIDADE DO SISTEMA E SATISFAÇÃO DO PASSAGEIRO 23
2.1.1 Sistemas Dinâmicos 23
2.1.2 Confiabilidade e Tempo de Espera 26
2.1.3 Indicadores para Estimativa da Demanda 29
2.2 TECNOLOGIAS PARA COLETA DE DADOS DA DEMANDA DE PASSAGEIROS 33
2.2.1 Sensores Baseados em Visão 35
2.2.2 Sensores Físicos Não-Visuais 37
2.2.3 Sensores de Rastreamento de Sinais 40
2.2.3.1 Ativos (dependem do usuário) 41
2.2.3.2 Passivos (dependem de sensores, não do usuário) 42
2.3 CARACTERIZAÇÃO DA TECNOLOGIA DE SENSORES PASSIVOS 44
2.4 MÉTODOS E MODELOS PARA ANÁLISE DE DADOS DOS SENSORES PASSIVOS WLAN 49
2.5 TÓPICOS CONCLUSIVOS 52
3 MÉTODO 54
3.1 ARQUITETURA DO SISTEMA 55
3.2 ESPECIFICAÇÃO E MONTAGEM DOS COMPONENTES DO EQUIPAMENTO SENSOR 57
3.3 PROGRAMAÇÃO DO PROTÓTIPO PARA COLETA DE DADOS 61
3.4 TESTES INICIAIS E LEVANTAMENTO DE CAMPO 64
3.5 MODELO DE MÁQUINA DE VETORES DE SUPORTE 66
3.6 ANÁLISE DOS DADOS COLETADOS 69
3.6.1 Recebimento e Armazenamento dos Dados Brutos 69
3.6.2 Verificação de Integridade e Compatibilização das Marcas Temporais 69
3.6.3 Agrupamento por Janela de Tempo e Extração dos Atributos 70
3.6.4 Otimização dos Hiperparâmetros do Modelo 72
3.7 TÓPICOS CONCLUSIVOS 73

4 RESULTADOS 74
4.1 TESTE INICIAL 74
4.1.1 Teste Piloto no Ponto de Ônibus 75
4.1.2 Discussão 77
4.2 APLICAÇÃO 77
4.2.1 Coleta dos Dados 78
4.2.2 Análise Estatística dos Dados 81
4.2.3 Comparação da Utilidade do Modelo SVM 84
4.2.4 Otimização dos Hiperparâmetros do Modelo SVM (*Grid Search*) 86
4.3 DISCUSSÃO 94
5 CONCLUSÃO 98
5.1 DISCUSSÃO DOS RESULTADOS 98
5.2 LIMITAÇÕES 100
5.3 TRABALHOS FUTUROS 101
REFERÊNCIAS BIBLIOGRÁFICAS 104
APÊNDICE A – Discussões Suplementares sobre o Método 108
APÊNDICE B – Lista de Atributos 111

LISTA DE TABELAS

Tabela 2.1: Dados primários coletados pelo conjunto sensor. 48
Tabela 2.2: Resumo dos métodos de análise de dados de trabalhos selecionados. 49
Tabela 3.1: Resumo dos componentes e equipamentos necessários para a coleta de dados. ... 59
Tabela 3.2: Exemplo de dados brutos provenientes do conjunto sensor. 64
Tabela 3.3: Exemplo de preenchimento do instrumento de coleta de dados de referência. 66
Tabela 3.4: Quantidade de atributos por tema. 71
Tabela 3.5: Hiperparâmetros do modelo SVM a serem otimizados dentro das opções consideradas..... 73
Tabela 3.6: Influência dos valores dos hiperparâmetros no sobreajuste ou subajuste do modelo. 73
Tabela 4.1: Relação das datas e locais de realização dos testes experimentais. 75
Tabela B.1: Estatísticas agregadas para cada janela de tempo. 111
Tabela B.2. Estatísticas de análise de sinal para cada janela de tempo. 112
Tabela B.3. Atributos de aleatorização de endereço MAC para cada janela de tempo. 113
Tabela B.4. Atributos de tempo entre sondas para cada janela de tempo. 114
Tabela B.5. Atributos de movimentação de dispositivos para cada janela de tempo. 114
Tabela B.6. Atributos de fabricante de dispositivo para cada janela de tempo. 115
Tabela B.7. Atributos de SSID para cada janela de tempo..... 116

LISTA DE FIGURAS

Figura 2.1: Principais objetivos de um Sistema Avançado de Gestão de Transporte Público (APTS). 23
Figura 2.2: Resumo visual dos principais benefícios da quantificação em tempo real da demanda citados pelos artigos mencionados. 25
Figura 2.3: Diferença entre os fatores principais de influência na QoS e os indicadores de confiabilidade. 28
Figura 2.4: Interação entre os sistemas ITS, APTS, e os indicadores medidos nesta pesquisa. 30
Figura 2.5: Fontes de falta de confiabilidade no serviço de TPC. 31
Figura 2.6: Esquema resumo do uso de dados para resolução de problemas de confiabilidade. 32
Figura 2.7: Principais causas, consequências, e soluções do agrupamento de ônibus (*bunching*). 33
Figura 2.8: Agrupamento temático dos diferentes tipos de sensores usados para contagem de passageiros. 35
Figura 3.1: Processo de aplicação do método, com entradas e resultados de cada etapa. 55
Figura 3.2: Arquitetura de comunicação entre os componentes e do fluxo de informações. 57
Figura 3.3: Gráfico de quadrantes ilustrando aproximadamente os parâmetros de detalhamento e aplicabilidade dos artigos mencionados. 58
Figura 3.4: Partes que compõem o dispositivo. 60
Figura 3.5: Fluxograma de captura de sondas Wi-Fi. 62
Figura 3.6: Vista frontal da parada de ônibus pesquisada. 65
Figura 3.7: Agrupamento dos atributos por tema. 71
Figura 3.8: Comparação dos processos de treino e teste do modelo. 72
Figura 4.1: Fluxograma do processo de análise de dados. 74
Figura 4.4: Gráfico comparativo do resultado de treino e teste na parada de ônibus com o sensor distante. 76
Figura 4.5: Gráfico comparativo do resultado de treino e teste na parada de ônibus com o sensor próximo. 77
Figura 4.6: Luminária modificada para incorporar o conjunto sensor em sua área interna. 78
Figura 4.7: Arranjo do equipamento dentro da luminária, com a armação removida para melhor visualização. 79

Figura 4.8: Visão geral do processo de coleta de dados, com o sensor em pleno funcionamento. 80
Figura 4.9: Comparação à noite do conjunto sensor com uma luminária normal. 80
Figura 4.10: Histograma de distribuição do RSSI das sondas captadas pelas cinco coletas de dados. 81
Figura 4.11: Gráfico UpSet que correlaciona as sondas captadas pelas cinco coletas de dados. 82
Figura 4.12: Histograma de distribuição das sondas por minuto com categorização dos RSSI. 83
Figura 4.13: Gráfico de correlação entre o número observado de pessoas, no eixo x, com o número de sinais fortes detectado pelo sensor, no eixo y. 84
Figura 4.13: Comparação dos dados de referência, em azul, com os sinais fortes detectados pelo sensor, em laranja. 85
Figura 4.15: Análise da previsão do modelo SVM, enquadrada como pertencendo ao grupo 1. 89
Figura 4.16: Análise da previsão do modelo SVM, enquadrada como pertencendo ao grupo 2. 90
Figura 4.17: Análise da previsão do modelo SVM, enquadrada como pertencendo ao grupo 3. 91
Figura 4.18: Análise da previsão do modelo SVM, enquadrada como pertencendo ao grupo do sobreajuste. 92
Figura 4.19: Análise da previsão do modelo SVM, enquadrada como pertencendo ao grupo dos insensíveis. 93
Figura 4.20: Vista lateral da parada de ônibus às 11h55 do dia 11 de novembro de 2025, em sua observação mais demandada das amostras do conjunto de dados. 94

LISTA DE ABREVIAÇÕES E SIGLAS

AFC: Cobrança Automática de Tarifas (*Automated Fare Collection*)

AP: Ponto de Acesso (*Access Point*)

APC: Contagem Automática de Passageiros (*Automatic Passenger Counter*)

APTS: Sistemas Avançados de Gestão de Transporte Público (*Advanced Public Transportation Systems*)

AVL: Localização Automática de Veículos (*Automatic Vehicle Location*)

BT: *Bluetooth*®

CAD: Despacho Assistido por Computador (*Computer-Aided Dispatch*)

CDR/UnB: Campus Darcy Ribeiro da Universidade de Brasília

CV: Visão Computacional (*Computer Vision*)

DF: Distrito Federal

DSRC: Comunicações Dedicadas de Curto Alcance (*Dedicated Short-Range Communications*)

GT: Dados de Referência (*Ground Truth*)

IC: Cartão de Circuito Integrado (*Integrated Circuit Card*)

ICC: Instituto Central de Ciências da Universidade de Brasília

IoT: Internet das Coisas (*Internet of Things*)

IPK: Índice de Passageiros por Quilômetro

ITS: Sistemas de Transporte Inteligentes (*Intelligent Transportation Systems*)

LSE: Estimador por Sensor de Luz (*Light Sensor Estimator*)

MaaS: Mobilidade como Serviço (*Mobility as a Service*)

MAC: Endereço de Controle de Acesso ao Meio (*Media Access Control*)

NTU: Associação Nacional das Empresas de Transportes Urbanos

OD: Origem-Destino

QoS: Qualidade de Serviço (*Quality of Service*)

RFID: Identificação por Radiofrequência (*Radio Frequency Identification*)

RSSI: Indicador de Intensidade do Sinal Recebido (*Received Signal Strength Indicator*)

SC: Cartão Inteligente (*Smart Card*)

SVM: Máquina de Vetores de Suporte (*Support Vector Machine*)

TPC: Transporte Público Coletivo

WLAN: Rede Local Sem Fio (*Wireless Local Area Network*)

Ora, àquele que é poderoso para fazer infinitamente mais do que tudo o que pedimos ou pensamos, conforme o seu poder que opera em nós, a ele seja a glória, na igreja e em Cristo Jesus, por todas as gerações, para todo o sempre. Amém!

(Efésios 3:20-21)

AGRADECIMENTOS

A Bíblia é imperativa ao ordenar que “em tudo dai graças” (1Ts 5:18). Este “tudo” se conecta com outro termo absoluto bem conhecido de Rm 8:28, “todas as coisas cooperam para o bem dos que amam a Deus.” Estes textos nos ajudam a entender que devemos confiar em Deus em todas as circunstâncias; tanto na vida fácil quanto na dificuldade, no sucesso e também na ruína. “Deus é o nosso refúgio e fortaleza, socorro bem presente nas tribulações” (Sl 46:1).

Agradeço ao meu Deus, criador e inteligente, que fez tudo do nada, de maneira ordenada, e de tal forma que podemos entender o mundo criado e extrair dele informações úteis.

Agradeço aos meus orientadores Fabiana e Pastor, que têm uma alegria contagiante e me ajudaram a ver mais longe por serem os ombros de gigantes nos quais eu pude estar de pé.

Agradeço à minha família que me ama, sempre me apoia, me deu a melhor educação que eu poderia receber, e me incentiva a sempre dar o meu melhor.

Agradeço à minha igreja e aos meus queridos irmãos, a quem tenho o privilégio de servir e “nos quais tenho todo o meu prazer” (Sl 16:3).

Agradeço ao Colégio Presbiteriano Mackenzie, no qual tive a honra de estudar, e cujo hino sempre nos lembra que “aprender é a lida que enobrece e conduz à vitória; ao progresso da Pátria querida, pela honra do nosso Brasil, empenhemos, colegas, a vida.” Agradeço também aos professores que me motivaram a buscar a excelência, às oportunidades de competir nas Olimpíadas do Conhecimento, que foram parte integrante do meu caráter, e às aulas de robótica, informática, e LEGO, que na minha infância despertaram meu interesse por tecnologia.

Agradeço à Universidade de Brasília, representada por todos os professores que tive. Alguns, por reforço positivo me inspiraram a amar o conhecimento e perseguir a carreira acadêmica; outros, por reforço negativo me lembraram que “através de muitas tribulações, nos importa entrar no Reino de Deus” (At 14:22). Seja por bem ou por mal, importa que esta casa de conhecimento me forjou na pessoa que hoje sou, certo de que não existem desafios impossíveis, afinal, eu consegui vencer “aquilo que parecia impossível.” Não existem impossíveis para um engenheiro, apenas limitações de tempo (“na sepultura, que é para onde você vai, não há obra”) e orçamento (com dinheiro infinito fazemos qualquer coisa); uma mão amiga também seria útil.

Finalmente, agradeço também a você, caro leitor, que chegou até aqui. Espero que este trabalho seja bênção na sua vida, e que você seja bênção na vida de muitas outras pessoas.

"A basic premise of this volume is that using transit service should be like eating potato chips – once you start, it's hard to stop. The fundamental question, though, is how do we design and approach such a service while knowing, because of the competition offered by cars, that an adequate transit design must provide superlative service in order to be good enough."

(Ceder, 2007)

PREFÁCIO

"Eu não acredito que esqueci de novo o aniversário da Abigail!", Davi rugiu. Então, às pressas, montou no seu cavalo e rumou veloz para a loja de Ezequiel, o confeiteiro favorito de sua esposa. Mas que surpresa ao chegar lá! Mil anos antes de Cristo, ainda não havia sido inventado o transporte público coletivo, então Aitofel, que morava nos arredores da cidade, não conseguiu um substituto para a sua mula que havia ficado doente. Será que ele empreitaria uma viagem a pé apesar de sua considerável idade? Como ele chegaria a tempo de atender o pedido especial de seu cliente?

O atual sistema de transporte público coletivo é algo que por vezes tomamos como garantido, mas frequentemente se esquece de que menos de dois séculos atrás pessoas gastavam horas do seu dia caminhando para satisfazer suas necessidades e desejos. Nunca foi tão fácil ir de um lado a outro da cidade de maneira tão *relativamente* rápida e econômica. O transporte é uma condição universal e absolutamente necessária para a manutenção da vida em sociedade; sempre foi e sempre será; ou, nas palavras de Ferraz *et al*. (2024): "o transporte urbano é tão importante para a qualidade de vida da população quanto os serviços de abastecimento de água, coleta de esgoto, fornecimento de energia elétrica, comunicação por telefone *etc*."

Os manuais costumeiros apontam três fatores que se destacam como responsáveis pela queda na demanda por Transporte Público Coletivo (TPC): (1) problemas na oferta, (2) má qualidade do serviço oferecido, e (3) limitações econômicas (Ceder, 2007; Ferraz *et al*., 2023).

Considerando a oferta no aspecto rigorosamente de demanda horária, o primeiro problema pode ser classificado como numérico quantitativo, enquanto os dois últimos podem estar mais relacionados com problemas intrinsecamente qualitativos e estruturais. Em relação a problemas de oferta, Ceder (2007) enfatiza que o problema se relaciona com a utilização eficiente dos recursos governamentais escassos; ter o ônibus certo, no lugar certo, na hora certa.

Em relação à Qualidade de Serviço (QoS) oferecido, Vuchic (2005) nota que as pessoas que têm a posse e habilitação para dirigir um automóvel e mesmo assim escolhem se deslocar pelo sistema de TPC rapidamente mudam o seu meio de transporte se este não for satisfatório. Então, se o problema de congestionamentos for o fator decisivo para a mudança do meio de transporte, este problema será agravado em uma visão global do sistema, embora localmente para o usuário possa haver alguma melhora. Isso acaba por gerar ainda mais insatisfação e suprimir o uso de sistemas de TPC e mobilidade, que é um efeito negativo de segunda ordem na oferta de um serviço de TPC insatisfatório (Vieira *et al*., 2020).

Quanto às limitações econômicas, popularmente se diz que “o dinheiro manda.” O conforto é diretamente proporcional ao custo do serviço, de forma que para uma operação ser economicamente viável ela pode demandar menor conforto e admitir maior lotação do ônibus. Em alguns casos, como em subúrbios nos quais o transporte público não é levado em consideração desde sua concepção, a implementação posterior desse sistema pode se tornar inviável (Vuchic, 2005).

Todos estes problemas afetam a parcela diária de tempo que os usuários gastam se deslocando entre seus destinos, o que tem impacto direto sobre a sua saúde e bem-estar. Quantificar essa impedância à realização de atividades produtivas é fundamental para compreender a complexa dinâmica entre os sistemas de TPC e os passageiros, o que deve ser feito não como um fim para si mesmo, mas como instrumento de direcionamento de recursos públicos para as áreas mais necessitadas (Litman, 2016; Pereira *et al*., 2021; Vieira *et al*., 2020).

Asakura e Hato (2001) exemplificam uma possibilidade de se adequar a oferta à demanda antes da era em que o *Machine Learning* estava disponível: em uma linha de ônibus na qual existiam problemas no planejamento da rota e do *headway*, os usuários solicitavam por telefone onde gostariam de embarcar e desembarcar dentre uma lista de opções disponíveis. Este embrião do *Mobility as a Service* (MaaS) aumentou de 7 para 30 passageiros diários, apenas com a melhora qualidade de serviço oferecido, sem alteração dos aspectos físicos da frota.

Os Sistemas de Transporte Inteligentes (ITS) também podem otimizar globalmente os atrasos nas linhas; Ceder (2007) argumenta que nunca há uma boa incerteza ou uma certeza ruim no TPC. É sempre importante reduzir as incertezas do sistema, o que só é possível com dados sobre a localização dos veículos, e contagem dos passageiros atendidos pelo serviço.

Computadores não têm olhos. Mas sistemas inteligentes requerem que os centros de controle "sintam" o ambiente da cidade. Uma das informações mais importantes a serem conhecidas é a relação entre a oferta e demanda de transporte, o que requer conhecê-las individualmente. Atualmente, é possível saber a localização precisa e a velocidade de cada ônibus do sistema. Contudo, falta uma ferramenta para conhecer a demanda de passageiros em tempo real, em qualquer lugar do sistema. Os sistemas dinâmicos atualmente usam informações defasadas temporalmente, e não verdadeiramente em tempo real. Isso se deve ao fato de a oferta ser monitorada em tempo real, mas a demanda surgir de estudos prévios, por exemplo de embarque e desembarque, e não de um sensor embarcado no veículo ou no ponto de parada que realmente poderia dizer a verdade a respeito deste dado (Oransirikul *et al*., 2016).

1 INTRODUÇÃO

O Transporte Público Coletivo (TPC) é a espinha dorsal de uma cidade, ligando as residências das pessoas a seus locais de trabalho e a tudo mais que elas precisem fazer. Vuchic (2005) afirma que este modo de transporte é inerentemente mais vantajoso para transportar altos volumes de passageiros, especialmente para motivos de trabalho, por serem concentrados espacial e temporalmente. Similarmente, Litman (2016) e Vieira *et al*. (2020) enfatizam que cidades com sistemas de TPC de alta qualidade veem um intenso fluxo de passageiros, o que influencia inclusive na dinâmica de trânsito da cidade e na qualidade de vida da população.

A demanda de passageiros por TPC não é constante, e varia no espaço e no tempo (Ferraz *et al*., 2023; Nunes, 2019). O conhecimento da variação dessa demanda no espaço e no tempo é fundamental para a definição adequada da oferta; os prestadores de serviço precisam de parâmetros de referência. O serviço de TPC pode ser modelado matematicamente segundo uma inequação na qual a oferta deve ser igual ou superior à demanda. Supondo que uma empresa gostaria de otimizar sua operação, a falta de oferta causa um excesso de lotação dos veículos que compromete a qualidade do transporte, e a oferta em excesso prejudica a eficiência operacional devido a aspectos econômicos. Dessa maneira, uma gestão baseada em dados é importante para a contínua avaliação e melhora da programação operacional, em busca da máxima eficiência operacional e econômica (Ceder, 2007; Ferraz *et al*., 2023; Vuchic, 2005).

O atendimento às demandas dos usuários também é importante como estímulo ao uso do transporte público como meio de transporte prioritário nos centros urbanos. É imperativo que seja oferecido aos passageiros um serviço de alta qualidade, pois o aumento da rejeição pelo TPC gera consequências negativas em cascata; passageiros em potencial, insatisfeitos com a falta de conforto ou demora no TPC que podem usar meios de transporte individuais sobrecarregam ainda mais o sistema e causam uma piora global, embora localmente vejam uma melhora (Vieira *et al*., 2020; Vuchic, 2005).

No Brasil, a Associação Nacional das Empresas de Transportes Urbanos (NTU) informa que esse serviço de transporte é usado por cerca de um terço da população brasileira, com fontes indicando uma peculiar estagnação da demanda por este modo de transporte nas últimas décadas, mesmo com o crescimento da população (NTU, 2025, Pereira *et al*., 2021). Em especial, Pereira *et al*. (2021) apontam uma estagnação com leve tendência de queda nos indicadores de TPC desde 1999, tanto no número de passageiros equivalentes transportados por veículo por dia, quanto no Índice de Passageiros por Quilômetro (IPK). O IPK mede a eficiência

econômica de um sistema de transporte público, calculando a média de passageiros transportados a cada quilômetro percorrido pelo veículo; quanto maior este número, maior a rotatividade ao longo de uma viagem. Em 1993, o IPK era de aproximadamente 2,5, e foi caindo ano após ano até estagnar em 2001 por volta de 1,6 e desse valor não sair até onde se tem registro. Essa tendência também foi observada no número de passageiros equivalentes transportados, com quase 50% de queda somente na última década (NTU, 2024). No entanto, como o IPK é função da densidade populacional, ele pode sugerir outras interpretações sobre a forma como as cidades brasileiras têm crescido.

Nesse sentido, os Sistemas de Transporte Inteligentes (ITS: *Intelligent Transportation Systems*) são relevantes em aspectos globais e locais para conhecer a demanda em função do tempo, inclusive em tempo real. A coleta de dados para fins de planejamento e operação possibilita aos operadores do sistema detectar situações anormais e otimizar a operação da linha, permitindo amenizar os efeitos de agrupamento de ônibus (*bunching*) e reduzir o tempo de espera para os usuários do sistema, efetivamente melhorando a Qualidade de Serviço (QoS: *Quality of Service*) prestado (Minea, 2022; Roriz Junior *et al*., 2022).

Diversas tecnologias no âmbito dos ITS vêm surgindo com o objetivo de auxiliar a quantificação da demanda em cada parte do sistema, uma tarefa que é primordial para o estabelecimento de sistemas dinâmicos, que permitem flexibilizar a oferta de acordo com as necessidades da demanda. Sendo os dados a base de qualquer sistema dinâmico, a informação que pode ser extraída destes dados fundamenta as escolhas da gestão e uso deste sistema, respectivamente por parte dos operadores do sistema e dos usuários (Madsen *et al*., 2022).

Em países emergentes, as políticas públicas tipicamente não são eficazes e a aquisição de dados sobre transporte é deficiente, de modo que é de suma importância melhorar a atratividade do transporte público. Contudo, as políticas públicas devem ser embasadas por dados confiáveis, o que faz com que a coleta e a análise de dados sejam fundamentais para a avaliação de um sistema de transporte público existente, seja para introduzir novos serviços ou comparar estratégias que possam melhorar a qualidade oferecida ao usuário. O sistema de transporte urbano é complexo, e a falta de dados históricos fidedignos que possam auxiliar na tomada de decisões atrapalha a construção de modelos contextualizados para o local de estudo (Pereira *et al*., 2021; Vieira *et al*., 2020; Vuchic, 2005).

Tradicionalmente, as coletas em campo são feitas por meio de pesquisas de lotação ao longo da linha e movimento nas paradas (sobe-desce) embarcadas, ou por meio da pesquisa Origem-

Destino (OD) domiciliar. Contudo, sabe-se ser esta operação custosa e sujeita a limitações estatísticas. O fator humano introduz erros intrínsecos e a definição de uma amostra representativa torna-se um desafio logístico complexo, oferecendo apenas um recorte estático da realidade (Ceder, 2007; Vuchic, 2005; Xu *et al*., 2025).

Sendo assim, torna-se imperativo substituir ou complementar esses métodos por ferramentas de sensoriamento remoto. Tais tecnologias permitem transicionar da dependência de amostras limitadas para a possibilidade de trabalhar com o universo de dados (ou Big Data), capturando o fluxo de passageiros de forma contínua e massiva. Isso não apenas desonera os usuários, cujo custo das pesquisas manuais acaba absorvido na tarifa, mas também aumenta a resolução temporal da informação para o planejamento.

O sensoriamento remoto é uma tendência atualmente, permitindo a captação e manipulação de dados reais e atuais de maneira econômica, algo que, se não impossível, seria economicamente inviável no passado. Essa tecnologia é caracterizada pela aquisição de informações sobre um objeto ou fenômeno sem fazer contato físico direto com ele. Sensores são a base e a ponta da lança de qualquer sistema dinâmico, e a ausência de um robusto sistema de captação de dados inviabiliza qualquer tentativa de compreensão da realidade da demanda. (Nunes, 2019)

Atualmente existem estudos sobre a contagem de passageiros usando sensores Wi-Fi, mas a maioria voltados à pesquisa de lotação ao longo da linha e de movimento nas paradas (sobe-desce). Contudo, essas pesquisas, embora úteis tanto para o planejamento das linhas quanto para operação macroscópica do sistema dinâmico, não fornecem a preciosa informação da demanda não atendida, uma vez que a pesquisa sobe-desce apenas informa o número de passageiros atendidos, e não os que ficaram esperando na parada por falta de espaço disponível no veículo. No que se refere especificamente ao comportamento da demanda no ponto de parada, apenas Oransirikul *et al*. (2016) investigaram esse aspecto; os sensores podem revelar informações preciosas sobre a demanda não atendida.

Nos estudos encontrados sobre o tema, a prática comum é a coleta de sondas WLAN (*Wireless Local Area Network*: Rede Local Sem Fio), na qual os dispositivos habilitados, como telefones celulares, carregados pelos passageiros transmitem solicitações públicas para descobrir dispositivos próximos, que são coletadas passivamente por um sensor que está atento a estes sinais. Essa abordagem oferece uma estimativa não intrusiva do número de passageiros em uma pequena área, o que é adequado para a estimativa de demanda em um ponto de ônibus (Madsen

et al., 2022; Minea, 2022; Nunes, 2019; Oransirikul *et al*., 2016; Roriz Junior *et al*., 2022; Santos e Almeida, 2021; Vieira *et al*., 2020; Xu *et al*., 2025).

Todos os modelos apresentados têm o mesmo objetivo: como diferenciar de maneira assertiva os passageiros de não passageiros, principalmente os transeuntes que estão próximos ao sensor? Este problema pode ser generalizado com base na analogia do sinal e do ruído, e como separar um do outro: o objetivo é descartar todo o ruído e manter todo o sinal em sua forma mais pura. No entanto, os sinais não vêm todos da mesma fonte, sendo difíceis de serem identificados todos da mesma forma, e até mesmo os ruídos têm diferentes fontes e comportamentos. Os ruídos estáticos, por exemplo, são aqueles provenientes de por exemplo apartamentos ou escritórios próximos, com dispositivos fixos que estão no raio de influência do sensor, e que podem ser removidos com base no filtro por um tempo de permanência longo sem alteração no RSSI. Por outro lado, os ruídos dinâmicos são os mais difíceis de serem filtrados, pois são aqueles gerados por transeuntes, sejam eles pedestres com dispositivos, carros com conexão Wi-Fi, ou ônibus cheios de passageiros com dispositivos, que emitem sondas ao passar pelo raio de alcance do sensor ou que podem até passar por dentro da parada de ônibus, mas que não estão efetivamente esperando o transporte e que dessa forma não deveriam ser contados.

A operação com dados reais e atuais é um imperativo categórico para qualquer sistema dinâmico de transportes inteligente. Nesse sentido, este trabalho tenciona propor uma alternativa prática, economicamente viável, e que respeita a privacidade dos passageiros, para a estimativa de demanda em pontos de acesso de transporte público, particularmente pontos e estações de ônibus, para fins de auxílio à operação desse modo.

1.1 OBJETIVO

1.1.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver um protótipo de dispositivo que permita a estimativa de demanda de passageiros em pontos de parada de ônibus utilizando a captação passiva de sondas Wi-Fi.

1.1.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

- Definir parâmetros que possam caracterizar a demanda nos pontos de parada de ônibus;
- Aplicar a técnica de percepção de sondas WLAN para estimar demanda em pontos de parada de ônibus;
- Avaliar a sensibilidade do modelo de Máquina de Vetores de Suporte com os dados de campo.

1.2 JUSTIFICATIVA

A técnica de contagem e estimativa de demanda em acessos de transporte público coletivo urbano por ônibus usando sensores passivos WLAN se mostra factível e pronta para aplicação imediata. Ela possibilita a coleta de indicadores que, de outro modo, seriam inviáveis de obter manualmente, como a individualização das medidas do tempo de espera por usuário em pontos de parada.

De acordo com Asakura e Hato (2001), Ceder (2007), e Murdan *et al*. (2020), o conhecimento do número de passageiros em cada ponto de acesso ao sistema de transporte público permite informar aos usuários a ocupação dos veículos da linha, o que proporciona uma melhor experiência de jornada do usuário. Isso pode ser feito tanto por telões informativos em pontos de parada, quanto por aplicativos de telefone celular, que estão cada dia mais presentes no cotidiano das pessoas.

Além do monitoramento, operação, e gestão em tempo real, a análise posterior e histórica do comportamento de viagem dos usuários do transporte público é importante para o planejamento e aprimoramento do sistema, por exemplo localizando pontos de demanda não atendida satisfatoriamente (*hotspots*), bem como ajustando a rota e número de veículos envolvidos, ou até mesmo basear estudos para relocação de garagens de veículos (Asakura e Hato, 2001; Madsen *et al*., 2022; Murdan *et al*., 2020).

Finalmente, notou-se uma dificuldade em replicar o equipamento de coleta da maioria dos trabalhos estudados, por sua falta de detalhes na descrição do equipamento, como por exemplo em Madsen *et al*. (2022), Vieira *et al*. (2020), e Xu *et al*. (2025).

Neste trabalho, avaliou-se a usabilidade de um sistema de sensor passivo que não usa imagem, mas que se baseia nas informações públicas emitidas por dispositivos dos passageiros, para a estimativa da demanda atual de passageiros, compreendida como o volume de passageiros em pontos de acesso ao sistema de transporte coletivo. Isso se justifica pela elevada acurácia do

sistema em conjunto com sua escalabilidade e baixo custo operacional e de infraestrutura (Madsen *et al*., 2022; Oransirikul *et al*., 2016; Vieira *et al*., 2020).

1.3 ESTRUTURA DO PROJETO FINAL

Este Projeto Final se estrutura em cinco capítulos. O Capítulo 1 introduz o tema, delimitando os objetivos e a justificativa da pesquisa. O Capítulo 2 apresenta o Referencial Teórico, discutindo a relevância e o estado da arte das tecnologias de contagem de passageiros em tempo real. O Capítulo 3 detalha o Método proposto, descrevendo a arquitetura do sistema e os procedimentos de coleta. O Capítulo 4 exibe e analisa os resultados obtidos a partir da aplicação experimental do método. Por fim, o Capítulo 5 apresenta as conclusões, limitações e sugestões para trabalhos futuros, seguido pelas Referências e Apêndices.

2 REFERENCIAL TEÓRICO

Sabe-se que as informações extraídas de dados são a base de sistemas dinâmicos de transporte. Mas como isso é feito atualmente, e qual o estado da arte disponível para implementação?

Sensores são a ponta da lança de qualquer sistema dinâmico. Nos transportes, não é possível aferir a realidade da demanda sem um robusto sistema de captação de dados que esteja instalado em pontos estratégicos e que forneça informações em tempo real (Nunes, 2019). Para que se implemente uma verdadeira gestão baseada em dados, não é possível continuar a usar somente o planejamento fruto de pesquisas estáticas, como as de contagem volumétrica ou ocupação visual de veículos de Transporte Público.

Quando Vuchic (2005) listou os dados usados para a previsão da demanda, ainda não era factível a contagem automatizada e confiável de passageiros em um ponto qualquer da rota. Antes do advento dos computadores, e mesmo atualmente, os sistemas dinâmicos usam informações defasadas temporalmente, e não verdadeiramente em tempo real, o que causa situações como superlotações em alguns pontos de parada e ônibus que passam vazios em outros. Isso se deve ao fato de a oferta ser monitorada em tempo real, mas a demanda surgir de estudos prévios, por exemplo de embarque e desembarque, e não de um sensor embarcado no veículo ou no ponto de parada que realmente poderia dizer a verdade a respeito deste dado tão importante. Contudo, sabe-se hoje que tal sistema é não apenas possível, mas também economicamente viável, conforme se vê em Vieira *et al*. (2020), Roriz Junior *et al.* (2022), e González *et al*. (2021).

Ademais, Asakura e Hato (2001) discutem a possibilidade de acompanhar em tempo real a demanda verdadeira e ainda não atendida ao longo de todo o sistema por meio de sensores localizados estrategicamente em paradas-chave de TPC, de forma que não seja necessário monitorar cada uma das paradas da rede. A ideia é que amostras, ou pessoas-chave, permitem estimar a distribuição da população ao longo da rede, o que seria economicamente mais vantajoso do que contar todo e cada passageiro do sistema para que se tenha uma estimativa razoavelmente precisa dessa demanda. Essa técnica poderia fornecer dados para auxiliar os gestores a balancear adequadamente a oferta de veículos e a demanda de passageiros, bem como otimizar os custos e reduzir o tempo de espera.

2.1 IMPACTO DA TECNOLOGIA NA AVALIAÇÃO DA QUALIDADE DO SISTEMA E SATISFAÇÃO DO PASSAGEIRO

2.1.1 Sistemas Dinâmicos

Sistemas considerados dinâmicos são aqueles que permitem ajustes nas programações horárias, o que se traduz em potenciais reduções de custos operacionais pela redução da ineficiência do uso dos veículos, e melhorias na qualidade do serviço oferecido ao reduzir o tempo de espera de passageiros (Minea, 2022; Oransirikul *et al*., 2016) Roriz Junior *et al*, 2022; Vieira *et al*., 2020; Xu *et al*., 2025). Asakura e Hato (2001), e Murdan *et al.* (2020) por exemplo trazem uma proposta de implementação de veículos autônomos capazes de operar em pelotões e assim ajustar suas capacidades de oferta em ambas as direções de uma linha. Contudo, sem um conhecimento íntimo da demanda tal modulação da oferta é ineficiente.

No propósito de resolver estes problemas surgiram os Sistemas Inteligentes de Transporte (ITS), que têm aplicações em várias áreas nos Transportes; dentro do contexto de TPC, destaca-se o Sistema Avançado de Gestão de Transporte Público (APTS). Segundo Ceder (2007), conforme resumido na Figura 2.1, os principais objetivos deste sistema são: (1) melhorar a gestão da operação por meio do desenvolvimento e validação de modelos de planejamento de TPC; (2) melhorar a eficiência do uso de recursos por meio da melhor correspondência entre oferta e demanda; (3) melhorar a Qualidade do Serviço (QoS) e os níveis de serviço por meio do aumento da confiabilidade resultante de melhor controle da operação; (4) reduzir os custos de coleta, processamento e análise de dados; e (5) integrar as informações de gestão referente aos veículos (operações) e ao comportamento dos passageiros (demanda).

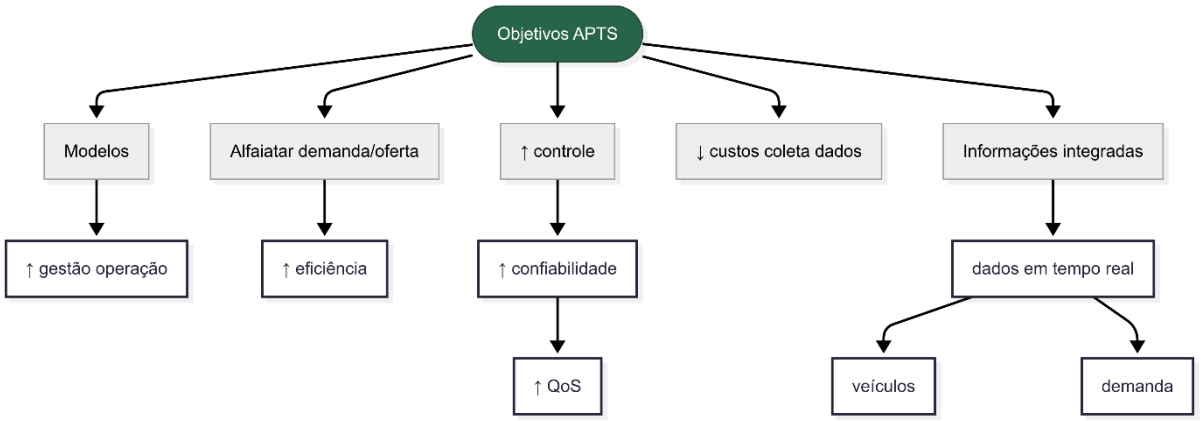


Figura 1: Principais objetivos de um Sistema Avançado de Gestão de Transporte Público (APTS).

Mesmo atualmente, as agências de trânsito raramente possuem dados suficientes sobre as operações de serviço de modo geral, quanto mais em tempo real, visto que o atual modelo de coleta manual de dados utilizado para este procedimento é de alto custo financeiro devido ao alto volume de dados necessário. Este é um dos motivos por que estas pesquisas são feitas somente a cada muitos anos, resultando na falta de informações detalhadas sobre a demanda de passageiros e as características do serviço em nível de linha, que são essenciais para implantar um serviço de transporte público eficiente e alinhado à demanda. (Madsen *et al*., 2022, Paradeda *et al*., 2019)

Ceder (2007) e Madsen *et al*. (2022) afirmam que os operadores de empresas de ônibus estão interessados em ter conhecimento do número de passageiros dentro de seus ônibus, preferencialmente com custos mínimos de implantação e de forma automatizada, mantendo alta precisão. Esses dados de densidade de usuários por pontos de parada são úteis para a compreensão da distribuição espaço-temporal da demanda, permitindo o ajuste dinâmico da oferta de TPC dentro de sistemas de ônibus responsivos à demanda e uma melhor operação das linhas em comboio. Quando essas informações estão disponíveis em tempo quase real, com um atraso de apenas alguns minutos, as operadoras de ônibus podem alocar veículos adicionais a linhas específicas, inclusive usando estratégias conhecidas como *deadheading*, otimizando a capacidade de responder proativamente às variações observadas.

Este conjunto de ferramentas e princípios dos ITS aplicados ao TPC tem a capacidade de remediar problemas de confiabilidade, ao permitirem o uso de mensurações do tempo de espera e picos inesperados de demanda tanto para fins diagnósticos a fim de corrigir problemas detectáveis quanto para a tomada de decisões tático-operacionais e de planejamento, inclusive iterativamente por meio da análise do impacto de intervenções no sistema. Os dados coletados permitem identificar gargalos na rede, otimizar a programação de veículos e a frequência de linhas, bem como disponibilizar informações aos usuários que reduzam a incerteza durante a espera. Essas medidas contribuem para um ciclo de melhoria contínua da qualidade do serviço e, em última instância, para a promoção de um transporte público mais atrativo e sustentável, o que pode inclusive aumentar a demanda por TPC (Ceder, 2007; Madsen *et al*., 2022; Minea, 2022; Oransirikul *et al*., 2016; Roriz Junior *et al*., 2022; Vieira *et al*., 2020; Xu *et al*., 2025).

Em serviços de TPC com intervalos curtos, é razoável supor que os passageiros chegam aleatoriamente, ou seja, sem planejar o horário; algo que não é válido para *headways* longos, uma vez que o tempo de espera se torna incômodo, de forma que se assume que os usuários se planejam para chegar ao ponto de embarque pouco antes da chegada do veículo. Contudo, sabe-

se que um sistema de informações com dados integrados aos veículos pode potencialmente influenciar o horário de chegada dos passageiros, fazendo-os esperar menos tempo na parada (Ceder, 2007). Isso pode ser visto atualmente com serviços como o Google Maps ou Moovit, que informam ao usuário o horário de chegada dos veículos de Transporte Público.

Por essa razão, a obtenção de dados operacionais a um custo mais baixo, usando subsistemas do APTS como o AVL (Localização Automática de Veículos), APC (Contagem Automática de Passageiros), CAD (Despacho Assistido por Computador) e AFC (Cobrança Automática de Tarifas), pode auxiliar os prestadores de TPC, incluindo as agências governamentais, em seu processo de tomada de decisão. Estas tecnologias têm o potencial de aumentar a eficiência operacional, melhorar a programação horária, e aprimorar a confiabilidade do serviço para os passageiros, mesmo tendo um alto investimento inicial, de cerca de US$ 6.000 e um custo anual de US$ 600 (Ceder, 2007; Murdan *et al*., 2020).

Em resumo, a literatura disponível sobre o tema de quantificação em tempo real da demanda em pontos de parada de ônibus afirma que esta tecnologia produz consequências positivas em seis áreas temáticas principais, que podem ser divididas em subcategorias interrelacionadas, conforme representado pelo mapa conceitual da Figura 2.2.

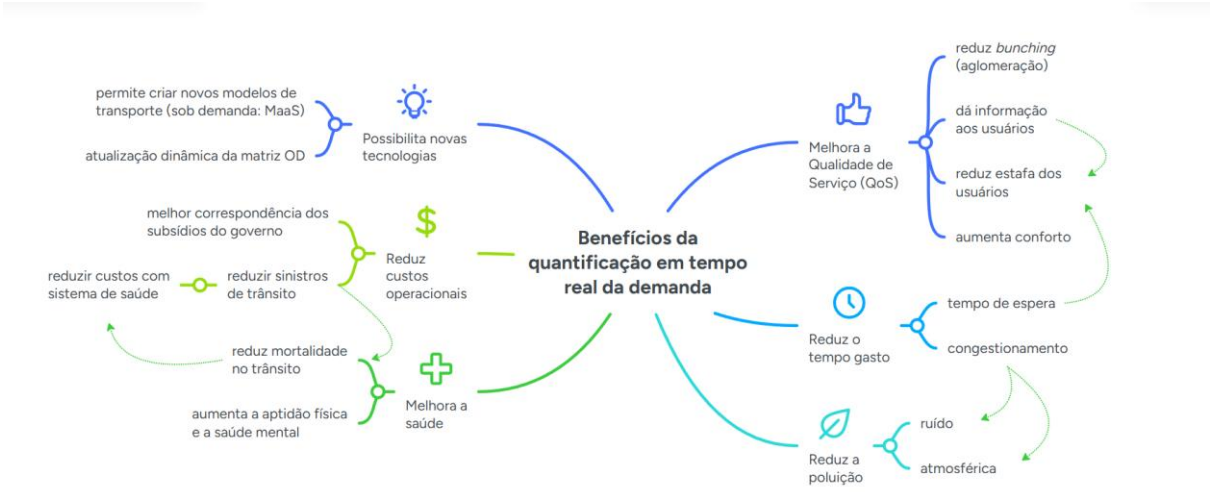


Figura 2: Resumo visual dos principais benefícios da quantificação em tempo real da demanda citados pelos artigos mencionados.

Primeiro, o sistema dinâmico permite a criação de novas tecnologias, tanto no sentido de novos serviços, como o MaaS (Asakura e Hato, 2001; Vieira *et al*., 2020), quanto a geração de uma

matriz OD dinâmica, que permite a identificação de situações anormais ou levar em conta passageiros menos frequentes (Xu *et al*., 2025).

Segundo, ele reduz custos operacionais, por meio de três áreas principais; planejamento e gestão baseados em dados (Vieira *et al*., 2020; Xu *et al*., 2025), melhor correspondência dos subsídios do governo com os passageiros transportados (Madsen *et al*., 2022), e a redução dos custos financeiros materiais e hospitalares com sinistros de trânsito (Litman, 2016; Vieira *et al*., 2020). Esta última subcategoria, de redução de sinistros, está intimamente ligada com o terceiro tema, de melhora da saúde pública, uma vez que, pela Pirâmide de Bird, menos sinistros equivalem a menos mortes (Litman, 2016; Vieira *et al*., 2020); a quantificação da demanda em tempo real, por suas consequências positivas, também aumenta a aptidão física e saúde mental dos usuários (Litman, 2016).

Além disso, ela melhora a QoS percebida pelos usuários, diminuindo os impactos causados pelo agrupamento de ônibus conforme descrito por Oransirikul *et al*. (2016); dá informações em tempo real sobre a ocupação da linha aos usuários, o que os permite tomar decisões informadas com base em suas preferências (Asakura e Hato, 2001; Oransirikul *et al*., 2016); o que também ajuda a reduzir a estafa (*stress*) dos usuários (Minea, 2022).

Ademais, essa medida atualizada reduz o quinto aspecto, que é o tempo gasto pelos passageiros, tanto na espera (Roriz Junior *et al*., 2022; Vieira *et al*., 2020), quanto em congestionamentos durante a operação (Minea, 2022; Vieira *et al*., 2020). Finalmente, o sexto aspecto é beneficiado pela redução da poluição sonora, ou do ruído (Minea, 2022), que traz benefício para os seres humanos, e da atmosférica (Litman, 2016; Minea, 2022), o que também traz benefício ao meio ambiente.

2.1.2 Confiabilidade e Tempo de Espera

Em se tratando de passageiros, não é incomum se ouvir reclamações dos usuários a respeito da qualidade do serviço de TPC oferecido (Ceder, 2007). Especificamente em relação à confiabilidade do serviço prestado, os efeitos mais sentidos pelos passageiros são externados em frustrações quanto à falta de pontualidade, falhas na integração com outras linhas, superlotação (conforto precário), e agrupamento de ônibus, sem contar com problemas operacionais que eles não podem ver, como problemas de manutenção ou falta de técnicos, que causam problemas de frota, entre outros (Ceder, 2007; Ferraz *et al*., 2023).

A percepção de confiabilidade por parte dos passageiros está relacionada ao nível de certeza dos usuários de que o veículo de TPC vai passar pela origem de desejo do passageiro e chegar ao destino pretendido, ambos no horário previsto ainda que com alguma tolerância, e com alta probabilidade de se adequar à expectativa de seus atributos, como conforto e segurança. O alinhamento das expectativas dos passageiros é fator importante na percepção de atrasos; por exemplo, a percepção de atraso é muito menor em uma linha que frequentemente atrasa do que em uma linha pontual. Nesse sentido, a medida objetiva de confiabilidade de um serviço pode ser definida como a variabilidade dos atributos de um serviço, especificamente no sentido de que quanto menor a variabilidade de um atributo, mais confiável ele é. Estes atributos influenciam as decisões dos atores envolvidos, e podem ser relacionados aos passageiros, prestadores de serviço ou a fatores exógenos (Ceder, 2007).

Ferraz *et al*. (2023) apontam doze fatores que influenciam a qualidade do serviço prestado para os usuários, dos quais quatro podem ser interpretados como atributos de confiabilidade: tempo de viagem, lotação, e o que o autor chama de "confiabilidade" em si, que será tratada separadamente em termos de pontualidade, e efetividade. Dentre esses, considera-se o tempo de viagem (que é a soma dos tempos de caminhada até a origem, de espera no ponto inicial, no interior do veículo, e do destino para completar a viagem), como uma consequência, e não uma causa, do problema de confiabilidade, embora a falta de planejamento do usuário possa aumentar consideravelmente o tempo de espera. Conforme ilustra a Figura 2.3, apenas os outros três atributos podem ser considerados indicadores de confiabilidade; a (1) pontualidade, (2) lotação, e (3) efetividade. Vale destacar que o que o autor chama de "frequência de atendimento" não deve ser considerado um indicador de confiabilidade por si só porque ele é um parâmetro de planejamento do TPC, enquanto o indicador de confiabilidade se propõe a medir as discrepâncias entre o planejado e executado, de forma que uma frequência de atendimento mal planejada, mas executada à risca, não traz problemas *de confiabilidade* propriamente ditos, porque não há atrasos, embora de fato exista um problema mais profundo. Por isso, preferiu-se trabalhar em termos do quesito pontualidade.



Figura 3: Diferença entre os fatores principais de influência na QoS e os indicadores de confiabilidade.

A (1) pontualidade é um nome alternativo para a discrepância no tempo de espera. Essa denominação faz referência ao inverso do fenômeno, na qual uma maior pontualidade implica em menor tempo de espera, o qual afeta intrinsecamente o tempo de espera dos usuários, chegando eles aleatoriamente ou não. Os atrasos na chegada, e consequentemente na saída, podem surgir por várias causas, desde problemas de manutenção, congestionamentos na via, situações problemáticas com passageiros, descuidos do condutor, entre outros. Quanto à (2) lotação, ela diz respeito à quantidade absoluta de passageiros dentro do veículo, ou proporcional à lotação máxima deste; o problema reside em uma densidade de passageiros próxima ou maior do que o limite aceitável, o que gera atrasos no embarque e desembarque, além do desconforto (entendido como perda de qualidade do serviço, com todas as suas consequências de potencialmente perder passageiros para outro modo de transporte). A maximização do conforto dos passageiros instituiria que todos viajassem sentados, mas essa medida aumentaria o custo de transporte; a presença de usuários em pé é aceitável, desde que não excessiva. Um problema grave de lotação é o fenômeno do agrupamento de veículos, que será abordado nos próximos parágrafos. Finalmente, a (3) efetividade da programação operacional diz respeito à capacidade do prestador de serviço de superar falhas oriundas de problemas de manutenção, sinistros de veículos, ou simplesmente de situações extraordinárias que ocorram; às vezes, um congestionamento pode "travar" os ônibus em uma rua da qual eles não podem sair por horas, o que pode, inclusive, impedir a realização das próximas viagens programadas, haja vista que é necessário que haja veículos suficientes na garagem, e equipes de plantão (Ferraz *et al.*, 2023).

Para uma avaliação consistente da confiabilidade do sistema, é fundamental o estabelecimento de indicadores objetivos. Atributos de confiabilidade são dados quantificáveis ou qualidades observáveis do serviço de TPC que podem variar; caracterizam-se por serem diretamente

observáveis, e podem ser descritos estatisticamente (por exemplo, tempos de espera reais) ou percebidos diretamente pelos usuários e pela agência (por exemplo, a sensação de um ônibus lotado ou de falta de segurança). Já indicadores de confiabilidade são as métricas derivadas dos atributos, que podem inclusive ser iguais aos próprios atributos, usadas para avaliar e diagnosticar o desempenho do serviço em relação a padrões definidos. Eles fornecem maneiras objetivas e mensuráveis de avaliar a confiabilidade real do serviço (por exemplo, "o tempo médio de espera é de 5 minutos" ou "85% dos ônibus estão no horário"). Vale ressaltar que nem todo indicador de confiabilidade necessariamente configura problemas de confiabilidade, mas servem como ferramentas de diagnóstico (Ceder, 2007).

2.1.3 Indicadores para Estimativa da Demanda

O tempo de espera em pontos de parada de TPC é um dos indicadores mais críticos da qualidade do serviço e da satisfação do usuário (Oransirikul *et al*., 2016; Wu *et al*., 2022). Uma vez que é possível medir individualmente o tempo de espera nos locais de acesso ao TPC, ele pode ser ainda mais caracterizado por suas medidas de média, valores máximos, desvio padrão, entre outros. Os métodos específicos de mensuração dos dados de demanda serão tratados na próxima subseção, destacando-se a medição individual destes valores com os sensores passivos.

A Figura 2.4 apresenta os dois indicadores tidos como chave para a estimativa e caracterização da demanda em pontos de acesso ao transporte público. São eles: o atributo Tempo de Espera (medido de forma direta), e o indicador Demanda Imprevisível de Passageiros, que pode ser derivado da demanda histórica para determinado ponto de coleta, e que influencia a distribuição de passageiros dentro de um ônibus e entre ônibus consecutivos (Mandal *et al*., 2023; Wu *et al*., 2022). Ambos os indicadores são relacionados tanto ao prestador de serviço quanto ao passageiro, uma vez que afetam a QoS percebida por este, e é um dado útil para a reprogramação da operação por aquele. Ceder (2007) e Ferraz *et al*. (2023) abordam diversos outros indicadores, mas eles são mais indicados para a medição da demanda dentro dos veículos, ou de outros parâmetros operacionais que não são possíveis de mensurar sem o cruzamento de dados dos veículos, ficando fora do escopo deste projeto.

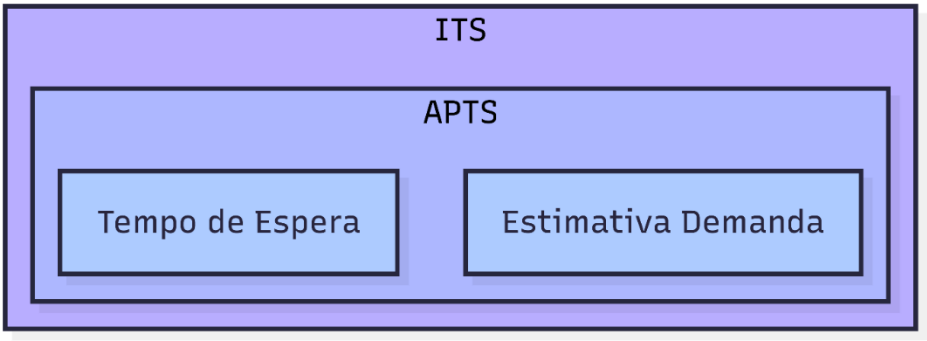


Figura 4: Interação entre os sistemas ITS, APTS, e os indicadores medidos nesta pesquisa.

Qualitativamente, a experiência de espera influencia diretamente a fidelidade ao sistema, especialmente no quesito negativo, produzindo mudanças no modo de transporte utilizado. Vale destacar a dificuldade de medição do tempo de espera, uma vez que, devido ao viés da percepção do usuário, a precisão dos dados reportados por eles é gravemente afetada. Dessa maneira, não é possível aferir essa medida por entrevistas aos usuários, de maneira que na coleta tradicional é necessário que o pesquisador controle a chegada e saída de cada usuário monitorado. No entanto, é o tempo de espera subjetivo o principal fator na percepção da qualidade do serviço, e não o objetivo. Conforme demonstrado em uma pesquisa no Reino Unido, a confiabilidade de prestação do serviço foi apontada como um fator duas vezes mais importante do que o aumento da frequência dos ônibus. Isso se deve ao fato de a percepção de conforto do usuário ser não linear e, portanto, difícil de quantificar; uma tentativa consiste no uso da razão de possibilidades para simular de maneira objetiva o sentimento de desconforto do usuário. A minimização e a previsibilidade do tempo de espera são, portanto, objetivos centrais para operadores e planejadores de TPC (Ceder, 2007).

As variações nos tempos de espera dos passageiros podem ter causas crônicas, que são diagnosticáveis com antecedência e tratáveis com mudanças de planejamento e ajustes operacionais, ou causas intrinsecamente imprevisíveis e que requerem mudanças na programação operacional em tempo real (de preferência baseados em planos de contingência), como, por exemplo, sinistros de trânsito ou pneus furados. Fontes de falta de confiabilidade no serviço são caracterizadas por elevada variância, independente da causa.

Muitos são os fatores que influenciam a variação do horário de chegada do TPC no horário previsto. Conforme ilustrado na Figura 2.5, Ceder (2007) traz cinco problemas principais que degradam a confiabilidade de serviço prestado: (1) incertezas no ambiente operacional, seja por

sinistros, engarrafamentos, ou outras externalidades diversas, (2) falta de dados adequados para um planejamento eficiente de operações, tal como medidas individualizadas de tempos de embarque/desembarque em cada estação, que poderiam indicar infraestrutura inadequada em paradas específicas, (3) projeto inadequado para a prestação do serviço, tal como escolha de vias que estão mais sujeitas a congestionamentos ou *headways* que não atendem à demanda existente, (4) monitoramento e controle de serviço inadequados, para gestão de eventos inesperados, e (5) falhas na execução dos cronogramas, como atrasos no início, saídas antecipadas, velocidades acima do esperado, e não cumprimento dos pontos de controle de horário.

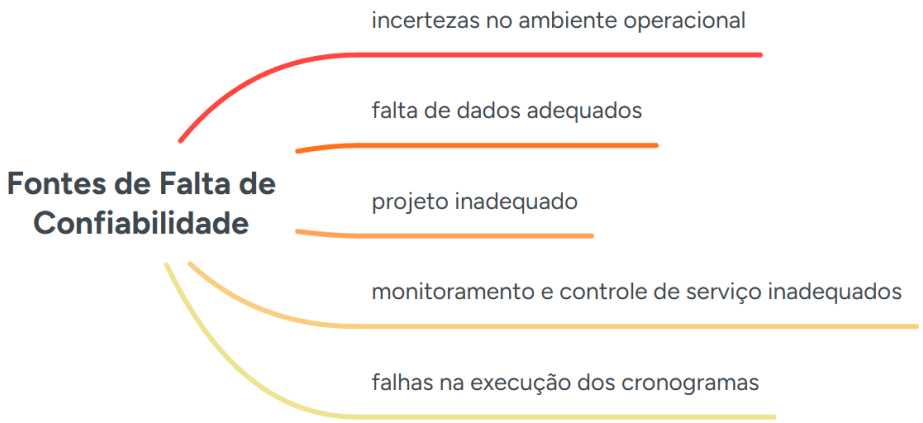


Figura 5: Fontes de falta de confiabilidade no serviço de TPC.

Considerando os pontos positivos de sistemas dinâmicos e os negativos dos atuais problemas de confiabilidade, propõe-se na Figura 2.6 um esquema explicitando como esses dois paradigmas podem ser combinados para melhorar o Transporte Público. Essa solução seria baseada em 2 pilares centrais; a captação e distribuição de dados operacionais e de demanda, e uma gestão operacional baseada no uso desses dados. A captação e distribuição de dados operacionais e de demanda ajuda a resolver os problemas 1 e 2 por meio do monitoramento permanente, o que reduz as áreas cinzentas da operação, bem como dá suporte a tomadas de decisão tática, operacional, ou estratégica. Além disso, a gestão operacional baseada no uso desses dados ameniza os problemas 3, 4, e 5, permitindo a modelagem matemática e otimização contínua dos projetos, bem como a elaboração de planos de contingência antecipados para problemas que podem acontecer, e redução dos pontos de falha nos controles de horário.

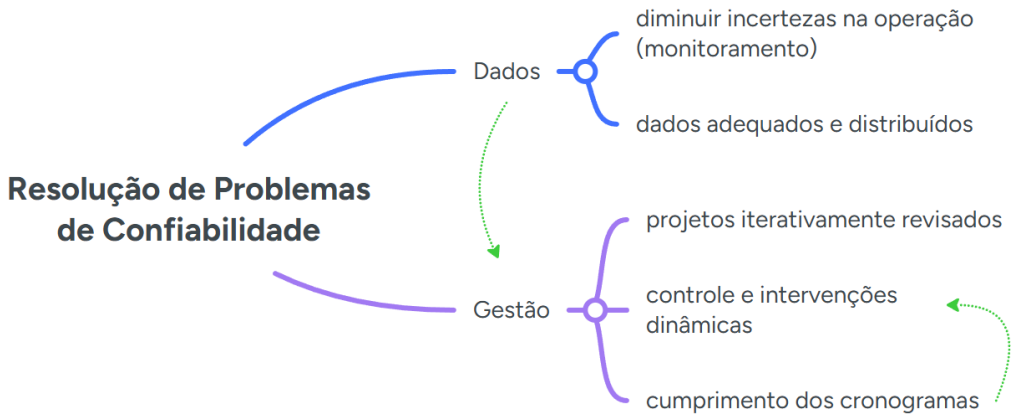


Figura 6: Esquema resumo do uso de dados para resolução de problemas de confiabilidade.

Nessa perspectiva, é importante discutir as causas, efeitos, e soluções do fenômeno de agrupamento de ônibus, conforme ilustrado na Figura 2.7. São três as possíveis causas do fenômeno de aglomeração: (1) veículo atrasado que se encontra com o anterior que está no horário, (2) veículo despachado antecipadamente e/ou em alta velocidade que se encontra com seu predecessor que está no horário, e (3) excesso inesperado de passageiros, que atrasa o veículo da frente e acelera o veículo de trás. Cada uma dessas fontes, ou alguma combinação delas, pode levar à criação de dois ou mais ônibus chegando em fila em um ponto de acesso ao TPC. Como tipicamente acontece em fenômenos de aglomeração, as consequências são (1) desregulação dos horários, com alterações nos *headways*, e passageiros esperando consideravelmente após a passagem de uma aglomeração de ônibus, e (2) carga desequilibrada entre os veículos, causando problemas de desconforto no veículo mais lotado. Para solucionar este problema, basta manter um *headway* regular entre os ônibus, mesmo em situações em que há desregulação em um ônibus da rota, por meio do controle da marcha mais rápida ou devagar conforme a circunstância, buscando manter o intervalo programado entre veículos consecutivos ao longo da linha (Ceder, 2007; Ferraz *et al*., 2023).

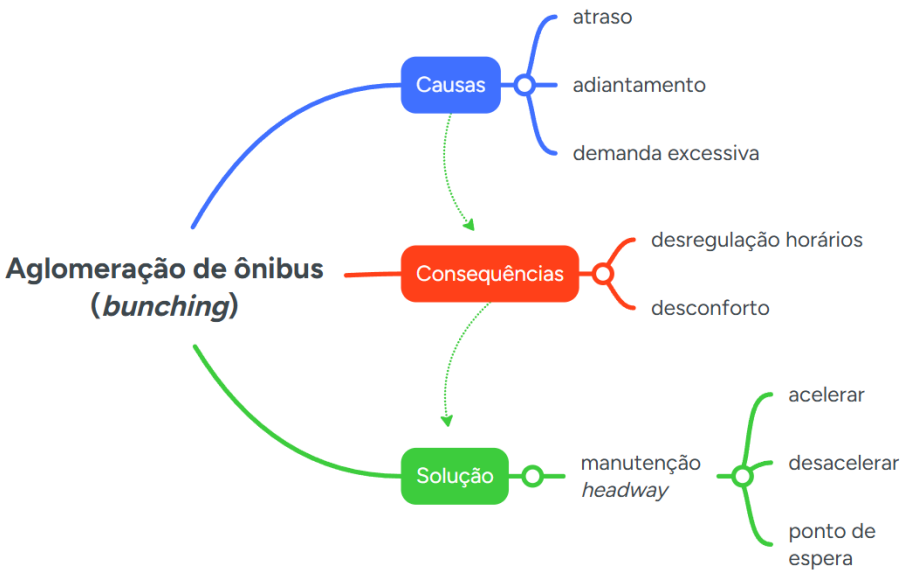


Figura 7: Principais causas, consequências, e soluções do agrupamento de ônibus (*bunching*).

Entretanto, não é o próprio operador do ônibus que irá tomar esta decisão de acelerar, desacelerar, ou esperar pela sincronização dos *headways* em relação ao veículo da sua frente e o de trás. Esta decisão precisa ser tomada pelo centro de controle operacional ou algum operador externo que tenha a visão completa do sistema, para coordenar os veículos (Ceder, 2007). Mas como ele tomará esta decisão sem o conhecimento da localização exata dos ônibus? Ademais, como o controlador saberá que há um excesso de passageiros em uma ou mais paradas sem o monitoramento destas? É neste ensejo que as tecnologias para coleta de dados de demanda de passageiro se inserem.

2.2 TECNOLOGIAS PARA COLETA DE DADOS DA DEMANDA DE PASSAGEIROS

Atualmente, os aplicativos de telefone celular já informam aos passageiros quanto às melhores rotas, horários, e atualizações no trajeto, mas ainda não conseguem informar quanto à lotação atual do veículo. Isso se deve ao fato de os ônibus serem atualmente equipados com um *kit* transmissor/receptor GPS, mas falta uma integração de informações com sistemas de bilhetagem ou sensores que permitem a estimativa de ocupação dos ônibus (Madsen *et al.*, 2022).

De acordo com a literatura disponível sobre as tecnologias de contagem de passageiros, a quantificação da demanda traz três vantagens principais ao ser feita em pontos de acesso ao sistema de transporte coletivo, tais como os pontos de parada. Primeiro, é possível oferecer um serviço com menor incerteza, o que permite uma oferta dinâmica mais próxima à demanda verdadeira, inclusive com uma matriz OD dinâmica (González *et al*., 2021; Madsen *et al*., 2022; Oransirikul *et al*., 2016; Xu *et al*., 2025). Com isso, é possível se obter um melhor aproveitamento operacional dos ônibus, o que se traduz em um menor investimento rendendo melhores resultados à população (Roriz Junior *et al*. 2022). Por último, mas de grande importância, estão a detecção e resposta a eventos imprevistos e passageiros não-frequentes, que podem eventualmente ser omitidos nas pesquisas normais (Madsen *et al*., 2022; Minea, 2022; Xu *et al*., 2025).

Existem diversas abordagens possíveis para a contagem automatizada de pessoas em uma área, sejam elas para a estimativa de público em um *show*, ou a estimativa de número de passageiros no transporte público. Em todos os casos, os sensores devem atender a dois requisitos principais; um, ter uma elevada precisão na contagem, mesmo considerando um ambiente pequeno que pode estar denso (como por exemplo um ônibus), e dois, ter suficiente precisão temporal, quase em tempo real, devido à dinâmica de passageiros entrando e saindo dos ônibus (Madsen *et al*., 2022; Nunes, 2019).

Em especial, três categorias de sensores são apontadas como principais, em função da forma mediante a qual os dados são captados, conforme ilustrado na Figura 2.8, que inclui também algumas tecnologias possivelmente aplicáveis para este trabalho. As categorias principais são as baseadas em vídeo (visão computacional), sem dispositivo e sem imagem, ou com dispositivo e sem imagem, que serão explicadas com maiores detalhes a seguir.

Na literatura encontrada sobre o tema, a maioria das referências trata da contagem de passageiros já dentro de um veículo ou com uso de dados de bilhetagem. Apenas Oransirikul *et al*. (2016) registram um experimento parecido com a proposta deste trabalho; sensoriar estações ou pontos de parada de ônibus para a coleta de dados de demanda. No entanto, Minea (2022) também desenvolveu pesquisas a respeito da contagem de pessoas em diversos contextos dentro de estações de trem usando a contagem de dispositivos com sensores estáticos, afixados não nos veículos. Portanto, um dos objetivos desta pesquisa bibliográfica é avaliar a aplicabilidade no presente estudo das tecnologias usadas para fins semelhantes.

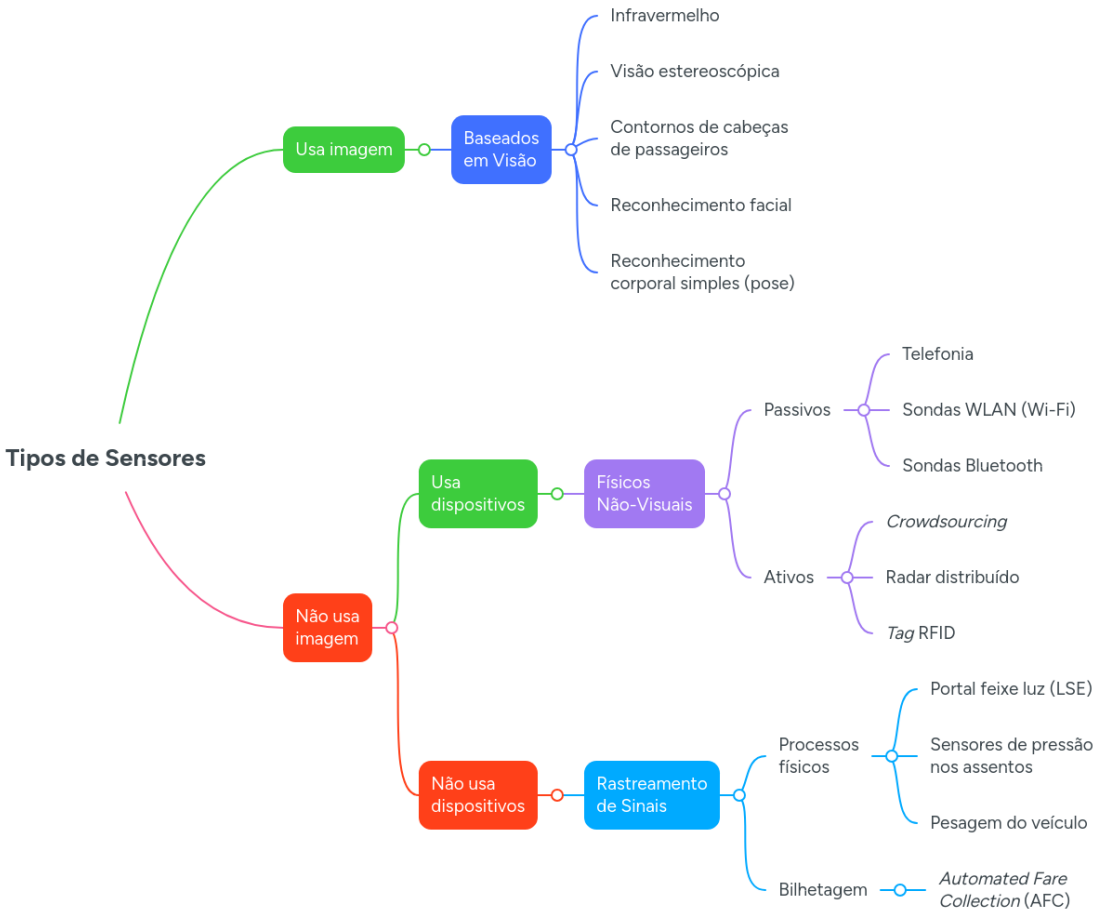


Figura 8: Agrupamento temático dos diferentes tipos de sensores usados para contagem de passageiros.

2.2.1 Sensores Baseados em Visão

A Visão Computacional, ou, como é conhecida em inglês, *Computer Vision*, é um tipo de sensor com princípio de funcionamento fácil de se entender, uma vez que simula a forma pela qual os seres humanos usualmente percebem o mundo. O sensor emula computacionalmente o processo de visão humana, atribuindo classes a objetos com características reconhecíveis. Esta é uma alternativa possível e que oferece excelentes resultados, mas que possui um ainda alto custo envolvido na compra de equipamentos, mesmo com a popularização desta tecnologia (Chen *et al.*, 2021; Hsu *et al.*, 2020; Madsen *et al.*, 2022; Xu *et al.*, 2025).

Há duas principais limitações comuns destes sensores: a necessidade de uma linha de visada para o reconhecimento dos atributos (feições/*features*) presentes na imagem, e o que a literatura cita como “preocupações com a privacidade”. No quesito linha de visada, é necessária uma

visão clara e desobstruída do objeto a ser detectada; fatores que podem atrapalhar são a meteorologia, a limpeza das lentes, manutenção do equipamento, e a oclusão de outros objetos que passam na frente do objeto-alvo a ser detectado (Madsen *et al*., 2022; Martin e Arruda, 2024). No que se refere à privacidade, Magro e Fortes (2021) consideram intensamente inconstitucional o uso de reconhecimento facial até para fins legítimos, afirmando que a privacidade e a proteção de dados pessoais devem atuar como uma barreira contra a vigilância, impedindo que esta os anule. No entanto, a visão computacional ainda se localiza em uma área cinzenta, uma vez que esta tecnologia usa atributos dos objetos ou pessoas para a identificação que não são exatamente informações pessoais, mas que podem ser derivadas das imagens.

Dentro da categoria de reconhecimento corporal simples, ou de reconhecimento de pose, Martin e Arruda (2024) desenvolveram uma ferramenta para contagem de fluxo de pessoas ao redor de faixas de pedestre usando uma única câmera, que também poderia ser usada para contar passageiros dentro de uma região-fonte delimitada pela parada de ônibus, por exemplo. De forma mais aprofundada, Tsiktsiris *et al*. (2025) trazem uma abordagem multissensorial com uma estrutura completa de sensores e câmeras na cabine, integrando análise visual e de áudio por meio de algoritmos de aprendizado de máquina profundos, visando aumentar a segurança, monitorar a presença de passageiros, e personalizar a experiência de viagem dos passageiros, ao mesmo tempo em que otimizam a carga de trabalho de agentes de segurança humanos, com corroboração dos resultados experimentais.

Na categoria de reconhecimento facial, Zhang *et al*. (2025) acompanharam passageiros ao longo de suas rotas para reconstruir matrizes OD, usando reconhecimento facial, feições dos passageiros (como roupas e proporções corporais) e probabilidades históricas de desembarques, alcançando resultados reportados de redução de erros quase pela metade. Contudo, conforme anteriormente citado por Magro e Fortes (2021), esta tecnologia possivelmente encontraria desafios na aplicação em território nacional.

Em relação ao uso do reconhecimento de contornos de cabeças de passageiros, esta era uma técnica muito usada no passado, quando se usavam uma ou mais câmeras instaladas acima das portas para calcular o fluxo de passageiros por meio da subtração entre os que embarcam e desembarcam. Nessa abordagem, Yu *et al*. (2007) analisaram as imagens de uma única câmera para reconhecer os contornos das cabeças dos passageiros. Atualmente, no entanto, com o avanço das pesquisas em Visão Computacional, pesquisas sobre a estimativa direta do número de passageiros em um ônibus agora são viáveis, como demonstrado por Hsu *et al*. (2020), que usaram câmeras frontais e traseiras instaladas dentro da cabine do ônibus para estimar o número

de passageiros em tempo real, usando técnicas complementares de contagem de cabeças onde elas são visivelmente distintas, e estimativa de densidade de pessoas. Em ambos os estudos, os resultados demonstraram alta precisão nos testes, mas também apresentaram problemas de susceptibilidade com relação a certas combinações de cor do cabelo e luz, cor de fundo, e alta exposição à luz.

No campo da visão estereoscópica, Bell *et al*. (2023) usaram um sistema de visão com duas câmeras frontais operando em conjunto em vez de apenas uma. Essa configuração alternativa, denominada câmera estereoscópica, oferece uma melhor percepção de profundidade do que a tradicional câmera única, permitindo medições mais precisas de distâncias que, neste artigo, foram usadas para verificar o distanciamento social implementado alguns anos atrás. Os autores também usaram um programa de Visão Computacional para agregar dados de ocupação dos assentos e pose dos passageiros. Contudo, esta solução não foi testada em um ambiente com alta densidade de passageiros.

Também é possível usar a tecnologia do infravermelho, abaixo do espectro visível da cor vermelha, como demonstrado por Lin e Zhao (2025), que usaram um tipo de "Visão Computacional com pixels que sentem calor", em um sistema de monitoramento e posicionamento de ocupação humana que utiliza um sensor térmico infravermelho e diversas técnicas de processamento de imagem. Eles enfatizam eficácia na detecção e rastreamento da presença humana em diferentes condições internas, incluindo iluminação padrão e escuridão total, sem comprometer a privacidade. "alcançamos uma segmentação robusta de regiões térmicas, mesmo quando vários indivíduos se aproximavam, aumentando assim a precisão e a confiabilidade da detecção humana." Contudo, eles enfatizam como problemas os altos níveis de ruído ambiental, especialmente em regiões quentes como é o caso do Brasil.

2.2.2 Sensores Físicos Não-Visuais

"Não ser baseado em dispositivo" significa que as contagens de passageiro são feitas de forma direta, geralmente por medições físicas de "pegadas" que os seres humanos deixam ao passar pelo sensor, para contrastar o próximo caso, em que a contagem é feita indiretamente pelos dispositivos eletrônicos que os passageiros levam consigo. Estes "eventos", termo técnico da ciência da computação quando há algo que aciona um processo, podem ser físicos ou relacionados à bilhetagem, por isso os sensores podem ser subdivididos em dois outros

conjuntos. Os físicos dependem de vários tipos de sensores localizados no ônibus (Madsen *et al.*, 2022).

Uma tecnologia bem conhecida desta categoria é a do portal, na qual um aparato montado nas portas do veículo conta o número de pessoas que passam por ele, seja por meio de feixes que são interrompidos, ou simples sensores infravermelhos de presença. Esta é uma tecnologia que usa dados físicos sentidos pelo sensor no ambiente que está, mas, diferentemente do sensor criado por Lin e Zhao (2025), os de portal não conseguem discernir as formas ou número de objetos na cena; apenas contar valores intermitentes que passam da tolerância especificada. Nesse sentido, Madsen *et al*. (2022) usaram sensores de luz infravermelha localizados nas portas de ônibus, denominados LSE (Estimador por Sensor de Luz) ou portais, os quais funcionam como um sensor de presença que detecta pessoas passando pelas portas, que, quando acumuladas, podem fornecer informações sobre o número atual de passageiros no ônibus. De acordo com os autores, o sistema de portais fornece alta precisão e, por essa razão, é uma das abordagens mais comuns usadas para contar passageiros atualmente, mas o sensor de presença tem a desvantagem de que as portas devem ser usadas estritamente para entrada ou saída, uma vez que se uma porta puder ser usada para ambos, o sensor de presença não consegue identificar se os passageiros estão entrando ou saindo ao passar pela área do sensor. Murdan *et al*. (2020), por meio de seu sistema BUSsit, mostrou que é possível contornar essa limitação por meio do emprego de um par de sensores infravermelhos de feixe, que, pela ordem que o passageiro interrompe os feixes, permite determinar se a pessoa está entrando ou saindo. Mesmo assim, os sensores estão sujeitos a danos acidentais causados pela passagem de passageiros, e possuem erros que vão se acumulando dentro de cada contagem.

Outra tecnologia física sem dispositivo é a de sensores de pressão nos assentos. Murdan *et al*. (2020), além de usarem os sensores de feixe para validar as entradas e saídas, também se destacaram pelo uso de sensores de pressão em cada assento, não para medir o peso de cada ocupante, mas somente para verificar a ocupação dos assentos do ônibus. Com base no dado de entrada e saída georreferenciado com GPS, e na ocupação do ônibus validada também pelo sensor pressiométrico, foi possível construir um modelo de previsão de demanda. Os autores também instalaram uma câmera, mas a usaram apenas para a contagem manual de forma a se obter a *Ground Truth* (GT), que é a coleção das categorias verdadeiras dos dados reais, para comparação com a previsão do modelo; em trabalhos futuros a aplicação de visão computacional seria mais uma camada de validação interna do modelo que eles poderiam usar. Contudo, uma grave desvantagem deste tipo de sensor é que ele não consegue contar quem está

em pé, desconsiderando-os. Além disso, “Sensores são integrados aos assentos do ônibus para detectar se os passageiros os ocupam ou não. Essa abordagem proporciona alta precisão na estimativa de passageiros em tempo real, mas a desvantagem é que um sensor deve ser instalado em cada assento” (Madsen *et al*., 2022).

Similarmente à alternativa anterior, que media a pressão individual dos assentos, Spaeth *et al.* (2020) apontam que o uso de sensores de pressão com alta precisão nos pneus e suspensões a ar do ônibus de forma a correlacionar com o peso do veículo carregado é um método viável para detectar mudanças na carga útil global do veículo, uma vez que dependendo do volume de passageiros, a massa total e o alcance do ônibus variam em cerca de 30%. Comparado à alternativa anterior, esta já consegue medir a quantidade de passageiros em pé, o que é uma vantagem por permitir uma medição mais precisa da realidade. Com essa informação, a operação pode ser otimizada.

Já em relação aos processos de bilhetagem, argumenta-se que os dados obtidos a partir de um sistema eletrônico ou, em inglês, *Automated Fare Collection* (AFC), são a forma mais precisa de estimativa de passageiros em tempo real, assumindo que haja informações tanto de entrada quanto de saída. Contudo, para isso é necessário que todos os passageiros precisem fazer o *check-in* e o *check-out* com seus cartões, o que não é conveniente para todos os modos. Para a maioria dos sistemas de metrô isso é factível, mas, em linhas de ônibus, que requerem apenas que o usuário passe o cartão na entrada, estando ele livre para desembarcar sem processamento adicional, é muito difícil prever os locais de desembarque dos usuários, o que impede a elaboração de uma matriz OD com este método (Chen *et al.*, 2021; Madsen *et al.*, 2022; Nunes, 2019; Xu *et al.*, 2025).

O AFC objetiva descrever o movimento completo do passageiro individualmente, referenciado por horário, de forma que seja possível descrever o fluxo de passageiros com a elaboração de uma matriz OD ou projetar cenários futuros com dados históricos. Estas informações podem ser acessadas *online*, e inclusive de maneira dinâmica. No entanto, se essas informações estiverem disponíveis apenas *offline*, a previsão do fluxo futuro de passageiros é necessária, o que torna essa abordagem vulnerável a picos no número de passageiros devido a eventos inesperados.

Os dados de bilhetagem também podem ser cruzados com outras bases de dados para obter resultados mais contextualizados. Wu *et al.* (2022) investigaram as características de transferência em tempo real dos modos metrô-ônibus e ônibus-metrô em diferentes estações de

metrô durante vários períodos, a partir de dados de Cartão Inteligente (SC: *Smart Card*) do transporte público, dados de rastreamento de GPS e dados de coordenadas estáticas de estações de metrô e pontos de ônibus. Isso permitiu avaliar a transferência de passageiros entre os dois modos de transporte apresentados. No metrô, os passageiros precisavam validar o cartão na entrada e saída, mas no ônibus somente na entrada. Portanto, cruzaram dados de ida e volta para obter o carregamento nos ônibus. O cartão usado, tipo BRB Mobilidade ou *Oyster Card*, pode receber várias nomenclaturas: IC (Cartão de Circuito Integrado),, entre outros.

Também é possível aplicar análises estatísticas robustas para modelar a demanda com dados já coletados, inclusive para identificar e responder a eventos imprevistos. Sun *et al*. (2021) estudaram a previsão do fluxo de passageiros em um modelo de duas etapas, com dados extraídos de um sistema AFC. Primeiro, os dados originais foram decompostos em três componentes: tendência, periodicidade e resíduo; em seguida, se fizeram previsões para cada componente separadamente, as quais foram somadas para se obter o resultado final; contudo, este modelo foi apenas testado em duas estações de grande porte, em condições regulares. Como trabalho futuro, os autores se propõem a estudar como o fluxo de passageiros se comportaria durante alguns eventos especiais ou incidentes não-previstos, como condições climáticas extremas, nas quais há perturbação anormal da demanda. Como conclusão, destaca-se que a decomposição da demanda em 3 partes complementares pode fornecer uma visão das características internas do fluxo de passageiros, e permite uma previsão globalmente mais rápida, e potencialmente mais precisa para cada componente, reduzindo os erros cumulativos durante a previsão das etapas seguintes.

2.2.3 Sensores de Rastreamento de Sinais

Ser “baseado em dispositivo” significa que o número de passageiros é medido indiretamente por meio da detecção de dispositivos que eles carregam consigo. Dentro desta categoria, também é possível classificar os sensores em intrusivos, que podem incomodar o usuário ou os arredores do ambiente onde ele for instalado, como por exemplo, um sensor WLAN que emula uma rede Wi-Fi, o qual pode perturbar os pontos de acesso legítimos já existentes, ou não-intrusivos, que não fornecem um ponto de acesso para conexões Wi-Fi, mas apenas recebem as sondas públicas e as armazenam (Madsen *et al*., 2022; Minea, 2022).

2.2.3.1 Ativos (dependem do usuário)

Abordagens baseadas em dispositivos ativos significam que os dispositivos ativamente fornecem informações do ambiente à sua volta a um sistema centralizado (Madsen *et al*., 2022). Conforme a Figura 2.8, eles podem ser categorizados entre sistemas de *crowdsourcing*, de radar distribuído, e de *tags* (etiquetas) RFID (Identificação por Radiofrequência).

Sistemas de *crowdsourcing* foram desenvolvidos para que os próprios usuários informem o número de passageiros em seus veículos, por meio de formulários especializados. Essa abordagem sempre requer que uma pessoa participante do sistema de *crowdsourcing* esteja presente no ônibus para obter informações sobre a contagem de passageiros. Além disso, para ter informações atualizadas quase em tempo real, os usuários devem continuamente relatar a contagem de passageiros ao sistema (Madsen *et al*., 2022). Em um caso concreto, Mandal *et al.* (2023) usaram voluntários equipados com um aplicativo em seus telefones que, simultaneamente, extraía dados dos sensores e permitia aos voluntários responderem um formulário, que funciona como os Dados de Referência (GT: *Ground Truth*) para uma análise de *Machine Learning* posterior usando os dados coletados pelos sensores. Isso permite verificar se é possível usar os dados dos sensores para se chegar na GT do formulário.

Essa é uma abordagem mais participativa em relação à próxima, na qual o aplicativo no celular do usuário funciona também ativamente como um sensor, mas sem a necessidade de interação com o usuário. Danielis *et al*. (2017) usaram uma malha de dispositivos para coletar informações sobre a presença de outros dispositivos, funcionando como uma rede descentralizada de radares distribuídos em uma área. As informações são coletadas, centralizadas e analisadas para descobrir se os dispositivos estão próximos de muitos ou poucos outros dispositivos. Com base nisso, a densidade de pessoas pode ser estimada em diferentes locais. Essa abordagem tem uma fraqueza semelhante à anterior, pois a presença de dispositivos especializados é necessária para estimar o número de pessoas (Madsen *et al*., 2022).

Já Oberli *et al.* (2010) coletaram estatísticas de reconhecimento de etiquetas RFID transportadas por voluntários em suas mochilas, carteiras, bolsos e mãos através de uma maquete de porta de ônibus equipada com quatro antenas de leitura de etiquetas. As etiquetas foram incorporadas em cartões plásticos laminados semelhantes a cartões de crédito. Nesta configuração, a porcentagem média de reconhecimento entre os quatro modos de transporte dos cartões RFID é de 91%, destacando-se como aspectos negativos no desempenho a quebra da linha de visão entre a etiqueta e a antena, o padrão de radiação das antenas, e o posicionamento delas ao redor do portal; para otimização, elas devem ser montadas em cada lado do portal na altura da cintura,

e adicionar mais de duas antenas não melhora o desempenho medido em mais de 2%. As porcentagens de reconhecimento são influenciadas principalmente pela posição da antena e pelo padrão de radiação, bem como pelas condições da linha de visão entre a etiqueta e as antenas. É interessante notar como que a maioria das sugestões que os autores fizeram foram colocadas em prática hoje em dia; aplicações para medição e previsão de demanda em tempo real, síntese de matrizes OD e operação responsiva à demanda de transporte público, conforme Murdan *et al.* (2020), Wu *et al.* (2022), Xu *et al.* (2025).

2.2.3.2 Passivos (dependem de sensores, não do usuário)

São abordagens que coletam informações oportunisticamente sobre dispositivos, com base na comunicação já gerada pelos dispositivos em outros contextos (Madsen *et al*., 2022). Por exemplo, sondas de Wi-Fi de um telefone procurando pontos de acesso (AP: *Access Point*), ou similarmente com Bluetooth (BT).

O autor Minea (2022) se preocupou com a correlação entre as medições físicas do Indicador de Intensidade do Sinal Recebido (RSSI: *Received Signal Strength Indicator*) pela fonte receptora, e a distância até o dispositivo do usuário, usando aparelhos e obstáculos diversos. Notou-se que cada celular tem uma potência de transmissão de sinal e curva de atenuação diferente, de modo que sua constatação de que uma linha de visão desobstruída entre o dispositivo do usuário e o sensor passivo é um fator importante para a manutenção da RSSI ajuda este trabalho ao direcionar a colocação do sensor na parte superior interna dos pontos de parada, maximizando a linha de visão do equipamento.

Anteriormente, para a contagem automatizada de pessoas por meio da conexão de seus dispositivos às torres de telefonia celular usavam-se informações sobre quais dispositivos estão conectados a quais torres de celular, juntamente com o conhecimento da localização das torres de celular, para se estimar o tráfego dos dispositivos em uma grade de células. Contudo, esse sistema é altamente dependente da definição da estrutura da grade em relação à localização das torres de celular, o que influencia diretamente a precisão, em termos de atribuição dos dispositivos à célula correta. No entanto, mesmo que o modelo acerte a qual célula da grade o dispositivo pertence, a célula em si continua sendo de uma baixa resolução espacial, o que pode não ser aplicável em alguns cenários, como o atual, no qual objetiva-se diferenciar pessoas que estão esperando em uma parada de ônibus, o que requer uma precisão de no máximo poucos metros. Além disso, a abordagem também não é capaz de distinguir se um dispositivo é

carregado por uma pessoa em um ônibus ou caminhando na rua (Asakura e Hato, 2001; Madsen *et al*., 2022).

Porém, Molina *et al*. (2022), avançando no estado da arte em termos de desenvolvimento tecnológico, testou a aplicabilidade da tecnologia DSRC (*Dedicated Short-Range Communications*; Comunicações Dedicadas de Curto Alcance), que é baseada em Wi-Fi IEEE 802.11p e nas redes celulares.. Ela permite que os veículos se comuniquem entre si e com outros usuários da estrada diretamente, sem a necessidade de comprar ou usar equipamentos adicionais, uma vez que usa os telefones celulares habilitados sem envolver infraestrutura celular, para "ver ao redor das esquinas" ou através de obstáculos, o que garante uma comunicação direta e rápida quando os veículos estão próximos entre si, e uma comunicação de longo alcance robusta via rede de celular para garantir que os avisos cheguem aos carros mesmo em situações mais complexas. Para este trabalho, isso significaria poder detectar dispositivos telefônicos inclusive em locais de alto ruído ou com barreiras físicas espessas, como abrigos em concreto. Destaca-se também uma característica fundamental dessa tecnologia, que é a sua capacidade de rapidamente estabelecer comunicação entre os dispositivos, sem a necessidade do tempo de configuração usual que as conexões de internet sem fio normalmente exigem (denominado em inglês *4-way handshake*), que é essencial para veículos que estão em movimento. Neste trabalho, essa tecnologia permitiria menores tempos de detecção dos dispositivos que adentram a área da parada.

Outra abordagem possível para sensores passivos baseados em dispositivos consiste na coleta de sondas WLAN. Este acrônimo é o termo genérico para qualquer rede local que usa ondas de rádio (em vez de cabos) para conectar dispositivos, sendo o Wi-Fi a implementação mais popular e dominante de uma WLAN, referindo-se a produtos que seguem os padrões da família IEEE 802.11 para redes locais sem fio. Em suma, toda rede Wi-Fi é uma WLAN, mas nem toda WLAN é uma rede Wi-Fi. Esta abordagem explora um recurso da tecnologia Wi-Fi na qual os dispositivos habilitados periodicamente transmitem sondas (*probes*) de conexão para detectar redes (APs) próximas. Estas sondas são públicas e contém o endereço MAC (*Media Access Control*) do dispositivo, que é o identificador universal de cada dispositivo existente no mundo. Estas sondas são então coletadas e registradas por um sensor simples e, com base nessas informações, é possível contar os dispositivos nas proximidades do ponto de acesso (Madsen *et al*., 2022; Vieira *et al*., 2020). Uma abordagem mais profunda deste método será tratada na próxima seção.

Nos seguintes trabalhos analisados, esta abordagem baseada em sondas WLAN oferece uma estimativa não intrusiva de passageiros em uma pequena área, o que é adequado para o presente trabalho. No entanto, a abordagem está sujeita a ruído ou sinais de dispositivos nas proximidades que não devem ser contados. Além disso, a frequência das sondas emissoras não é fixa, o que torna difícil dizer quando todos os dispositivos nas proximidades foram ouvidos. Outra fraqueza, comum a todas as abordagens baseadas em dispositivos, é a necessidade de correlação entre o número de dispositivos e o número de pessoas, uma vez que a medida direta é feita para os dispositivos na área, e não para o número de pessoas, o que requer uma correlação das chances de um passageiro carregar vários dispositivos ou não carregar nenhum dispositivo, um parâmetro que varia inclusive localmente (Madsen *et al*., 2022; Oransirikul *et al*., 2016; Roriz Junior *et al*. 2022).

2.3 CARACTERIZAÇÃO DA TECNOLOGIA DE SENSORES PASSIVOS

Portanto, a seguir discutir-se-ão características, vantagens, e desvantagens deste tipo de sensor. A tecnologia Wi-Fi, padronizada pela norma IEEE 802.11™, revolucionou a comunicação e o acesso à informação globalmente, estabelecendo-se atualmente como a forma mais popular de conectividade sem fio. Desde sua introdução em 1997 e o marco comercial em 1999 com a Apple, a evolução contínua dos padrões IEEE 802.11 tem proporcionado maiores taxas de transmissão de dados, alcance e segurança na conexão, tudo isso enquanto mantém a compatibilidade retroativa. Atualmente universalmente presente em dispositivos móveis como smartphones, tablets e na crescente Internet das Coisas (IoT), a tecnologia Wi-Fi permite que aparelhos móveis constantemente busquem pontos de acesso (APs), transmitindo requisições de sonda que contêm endereços MAC, que são globalmente únicos. Ao mesmo tempo que essa tecnologia possibilita a identificação mútua entre dispositivos que já se conectaram anteriormente, ela também pode ser usada para identificar e rastrear uma pessoa específica em um ambiente controlado (IEEE Standards Association, 2023; Vieira *et al*., 2020).

O endereço MAC é um identificador único atribuído a cada interface de rede para comunicações em um segmento de rede. Funcionando como um "endereço de hardware", ele é gravado diretamente na placa de rede do dispositivo pelo fabricante. Um endereço MAC possui 48 bits (6 bytes), geralmente representado por seis grupos de dois caracteres hexadecimais (por exemplo; *00:1A:2B:3C:4D:5E*). A primeira metade, com 3 bytes, identifica o fabricante, enquanto a segunda metade é um identificador único atribuído pelo fabricante. Seu

funcionamento é fundamental em redes locais, onde é utilizado para identificar dispositivos. Quando um dispositivo envia dados pela rede, ele inclui o endereço MAC de destino e seu próprio endereço MAC no cabeçalho do pacote. Os switches de rede utilizam esses endereços para encaminhar o tráfego apenas para a porta correta, evitando que os dados sejam transmitidos para todos os dispositivos na rede (Teca e Natkaniec, 2023).

Todo dispositivo eletrônico habilitado com Wi-Fi tem um endereço MAC físico referente a seu *hardware*. Contudo, as boas práticas de telecomunicações mais atuais não recomendam a disseminação desse endereço em transmissões públicas, devido ao risco de este endereço ser clonado por um agente malicioso, e também por permitir o rastreamento deste dispositivo, seja por lojas que querem analisar a influência do marketing no perfil de consumo, seja de novo por agentes maliciosos; tudo isso está contra o melhor interesse do usuário. Por isso, os dispositivos modernos (iOS 14+ e Android 10+) usam um endereço MAC administrado localmente (aleatorizado) para transmitir suas requisições publicamente. O normativo IEEE 802 especifica que é necessário que essa informação seja codificada no próprio endereço MAC, atribuindo o valor 1 ao segundo bit menos significativo do primeiro byte (conhecido como bit U/L ou Universal/Local) para indicar que aquele endereço é "Administrado Localmente" e não um endereço físico gravado de fábrica (OUI), que recebe valor 0. Isso inviabiliza a análise de distribuição de fabricantes de dispositivos, uma informação potencialmente útil para o modelo de análise. Contudo, não somente esses endereços virtuais impossibilitam a análise de fabricante; eles também têm uma janela de persistência limitada. Isso significa que o dispositivo que os administra também os altera periodicamente, tornando esses endereços efêmeros e novamente incapazes de serem rastreados, o que é valioso para o usuário, que não tem seus dados expostos, mas para os pesquisadores essa medida é agridoce; por um lado é mais fácil estar em conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD), mas por outro informações valiosas são perdidas.

Em estudos de estimativa de demanda usando sensores passivos, a prática comum é a coleta de pacotes WLAN do tipo sonda, na qual os dispositivos transmitem requisições contendo o endereço MAC do dispositivo, como explicado anteriormente, para descobrir APs tanto conhecidos como desconhecidos nas proximidades. Para tal, os dispositivos especificam o nome do AP no parâmetro SSID, que identifica a rede que está sendo procurada; caso contrário, a transmissão será denominada pública (*null probe request*). As sondas são coletadas passivamente por um sensor que está atento aos canais de transmissão pública de dados, que também pode medir a intensidade do sinal recebido, de forma que é possível contar os

dispositivos nas proximidades e medir a distância radial de cada dispositivo. Isso permite uma coleta anonimizada, não intrusiva, de larga escala, a baixo custo, e com potencial de integração com sistemas dinâmicos, algo impossível para coletas manuais (Madsen *et al*., 2022; Oransirikul *et al*., 2016; Roriz Junior *et al*., 2022; Santos e Almeida, 2021; Vieira *et al*., 2020).

O método aqui discutido de coleta de sondas só é possível devido à tecnologia de Varredura Ativa (*Active Scanning*) implementada nos dispositivos habilitados com Wi-Fi. Ela possui duas estratégias principais de descobertas de rede; a emissão de sondas direcionadas, contendo o SSID de redes conhecidas, e as sondas nulas, também chamadas de "coringa". Embora a primeira permita a conexão rápida a redes conhecidas, inclusive ocultas, ela está cada vez mais em desuso devido às preocupações de privacidade, pois ela (1) expõe publicamente todas as redes que o usuário tem salvas, e (2) permite o rastreamento do dispositivo emissor mesmo que ele esteja com a anonimização de endereço MAC ativa; basta rastrear os MACs que emitem a mesma SSID. Por outro lado, as requisições coringa (*wildcards*) têm a lógica inversa, pedindo para que os APs públicos se identifiquem, em vez do dispositivo chamar cada uma de suas redes conhecidas; embora este processo seja menos direto, ele agrega valor em privacidade e evita rastreamentos.

Assim como as consideráveis vantagens apresentadas, a abordagem também está sujeita alguns pontos de atenção, que serão apresentados a seguir: (1) ruído, (2) alternativas para canais de transmissão, (3) choque de sondas, e (4) correlação entre dispositivos e pessoas (Madsen *et al*., 2022; Minea, 2022; Oransirikul *et al*., 2016; Paradeda *et al*., 2019; Roriz Junior *et al*., 2022; Santos e Almeida, 2021; Vieira *et al*., 2020).

Imersos em um mar de sinais, os sensores devem ser capazes de discernir os sinais verdadeiros dos ruídos ou sinais de dispositivos próximos que não devem ser contados; por exemplo, nos arredores de um ponto de ônibus pode haver cafeterias com roteadores ligados, que emitem sinais que são captados pelo sensor e podem confundi-lo a considerá-los como dispositivos no ponto de parada.

Além disso, a frequência do canal de emissão das sondas emissoras não é fixa; isto é, existem vários canais com frequências de onda diferentes que o padrão Wi-Fi estabelece para conexão, e o sensor precisa alternar entre elas para não deixar de incluir um dispositivo que emite sondas por aquela frequência, o que faz com que seja impossível garantir que todos os dispositivos foram ouvidos.

Há também a possibilidade de dois ou mais dispositivos emitirem sondas ao mesmo tempo na mesma frequência, o que causaria uma colisão que potencialmente poderia gerar interferência destrutiva impedindo uma ou ambas as sondas de chegarem no receptor; não há problema de uma chegar enquanto a outra está sendo processada porque o microcontrolador possui um *buffer* que pode armazenar alguns pacotes enquanto o processador está ocupado.

Outra fraqueza da abordagem baseada em dispositivos sem a correlação com a imagem é que o dado bruto efetivamente coletado pelo sensor é o número de dispositivos detectáveis no ambiente, e não o número de pessoas, que é estimado baseado em correlações. A estimativa é feita por meio de um mapeamento entre o número de dispositivos e o número de pessoas, levando em consideração as probabilidades de um passageiro carregar zero ou mais dispositivos, podendo carregar vários ou não carregar nenhum. Especialmente no caso de pessoas mais velhas, notou-se uma maior dificuldade em estimar sua presença por sensores passivos, o que pode ter variadas causas, mas o efeito principal é que há um menor número de dispositivos detectáveis. Além disso, conforme se discutiu anteriormente, o usuário, embora tenha um dispositivo perfeitamente detectável, pode simplesmente escolher desligar o Wi-Fi de seu dispositivo por preferência pessoal (economizar bateria *etc*.), o que o impediria de ser detectado.

Mesmo com estes pontos negativos, o sensor ainda é considerado viável uma vez que os modelos de correção de dados levam em conta esses fatores, embora o erro aumente em função da diminuição da proporção de pessoas que tenham dispositivos detectáveis consigo. Madsen *et al*. (2022) advogam a favor de usar outro tipo de sensor, como por exemplo físico, para correção dos dados, considerando os diferentes tipos de erro intrínseco de cada sensor, o que poderia contribuir à capacidade de autodepuração do sistema. Especificamente, aliviaria os falsos negativos, que se configuram por dispositivos que as pessoas carregam e que emitem sinais, podendo ser detectados, mas, por alguma falha entre a transmissão e recepção do sinal, não o foram.

Também por isso, existem discussões a respeito do local e número de sensores para uma melhora na captação de dados. Madsen *et al*. (2022) afirmam que escolheram a localização do sensor WLAN pela necessidade de uma conexão de energia local, fornecendo apenas alguns locais de montagem dentro do ônibus; ao lado do motorista ou sob o teto perto do centro do ônibus, tendo esta última sendo considerada a melhor opção. Adicionalmente, apenas um sensor foi usado por ônibus, por critérios declarados de redução de custo, baseado na suposição de que o alcance de captação do sensor “cobria facilmente todo o veículo”, com múltiplos sensores

tendo pouco efeito adicional, embora se discuta o balanço entre a melhora real da efetividade com o uso de antenas direcionais e o esforço adicional para montá-las e calibrá-las. Minea (2022) destaca a complexidade computacional de usar mais de um sensor, particularmente influindo nos custos de implantação e operação do sistema, uma vez também que existem influências ambientais e variação nas antenas dispositivos que interferem na captação dos sinais. No entanto, o autor conclui que os sensores passivos são excelentes no que se propõe, usando um método não-intrusivo e anônimo para a coleta de dados de fluxo de passageiros e demanda.

A fim de se trabalhar com clareza técnica, ressalta-se uma questão de nomenclatura importante para o entendimento do funcionamento do conjunto sensor, composto por partes diferentes que se integram formando uma unidade de processamento que recebe dados do ambiente e os transforma em grandezas mensuráveis. Dessa forma, as principais partes são: (1) o transdutor, ou a antena, que é de fato o sensor, pois transforma a grandeza física das ondas Wi-Fi em sinais elétricos que podem ser interpretados pelo (2) o chip ESP32, que funcionalmente é também descrito como um microcontrolador. O conjunto desses dois componentes, que são ligados entre si via circuito integrado, é conhecido como a placa de desenvolvimento (*devkit*), também conhecido como módulo ou kit de desenvolvimento. Contudo, à placa são ligados outros componentes auxiliares, como o cartão SD em seu módulo, e a bateria, de modo que, a este último e maior conjunto que agrupa todos os componentes dá-se o nome de conjunto sensor ou nó de sensoriamento; no contexto específico de monitoramento de sondas Wi-Fi, pode-se usar também o termo *sniffer* (“farejador”), monitor de tráfego, ou interface de rádio do ESP32 operando em modo promíscuo.

Na Tabela 2.1 apresentam-se os dados primários que podem ser coletados pelo conjunto sensor, em consonância com os estudos de Santos e Almeida (2021)**,** Roriz Junior *et al*. (2022), Vieira *et al*. (2020), Oransirikul *et al*. (2016), e Madsen *et al*. (2022), e conforme discutido anteriormente. Enfatiza-se para o último item a discussão do início desta subseção a respeito da presença ou não da procura de uma rede específica, que traz este parâmetro adicional.

Tabela 1: Dados primários coletados pelo conjunto sensor.

Tipo de Sensor	Dados primários coletados
Monitor de tráfego	Marca temporal (*timestamp*)
	Endereço MAC do transmissor
	Intensidade do sinal (RSSI)
	SSID procurado/transmissão pública

2.4 MÉTODOS E MODELOS PARA ANÁLISE DE DADOS DOS SENSORES PASSIVOS WLAN

Uma vez que o tema deste trabalho diz respeito ao uso de sensores passivos WLAN, na Tabela 2.2 a seguir analisar-se-ão os métodos empregados nos artigos selecionados, em especial os modelos para a previsão da demanda real a partir dos dados brutos dos sensores.

Tabela 2: Resumo dos métodos de análise de dados de trabalhos selecionados.

Estudo	Método	Observações
(Lin e Zhao, 2025)	Interpolação	Não aplicável aos dados do sensor WLAN
(Madsen *et al*., 2022)	Filtragem simples com fusão de informações	Parâmetros de tempo e RSSI para classificação binária (dentro/fora do ônibus), combinado com dados do sensor infravermelho
(Oransirikul *et al*., 2016)	Filtragem simples	Parâmetros de tempo, número de sondas, e RSSI para classificação binária (dentro/fora do ônibus)
(Roriz Junior *et al*., 2022)	Máquina de Vetores de Suporte (SVM)	Classificação binária (dentro/fora do ônibus)
(Santos e Almeida, 2021)	Filtragem simples	Parâmetros para tempo, número de sondas, RSSI
(Vieira *et al*., 2020)	Filtro Kalman	Usa dados históricos de contagem manual para ajuste do modelo
(Xu *et al*., 2025)	*Fuzzy C-means Clustering* (FCM Clustering)	Classificação binária (dentro/fora do ônibus) baseado em atributos extraídos do conjunto de dados

Roriz Junior *et al.* (2022) usaram os dados coletados pelo sensor passivo e do sensor GPS para estimar o número de passageiros dentro do ônibus, usando um método de aprendizado de máquina (*Machine Learning*) denominado Máquina de Vetores de Suporte (SVM: *Support Vector Machine*) que determinava se a sonda Wi-Fi era oriunda de uma fonte interna ao ônibus, correspondente a um passageiro, ou externa, oriunda de ruído ambiental para o sensor. De forma geral, o sensor apresentou comportamento aderente às curvas de embarque e desembarque, com exceção dos pontos inicial e final da linha, algo compreensível devido à latência das requisições de sondagem WLAN. No entanto, destaca-se o zelo por maximizar tanto a precisão do modelo, reduzindo os falsos positivos, caracterizados pela errônea identificação de dispositivos fora do veículo como estando dentro dele, quanto o *recall*, reduzindo os falsos negativos, que são dispositivos dentro do veículo com o Wi-Fi ligado mas que por algum motivo não foram reconhecidos, conforme a definição de Paradeda *et al*. (2019). No entanto, assim como em

outros artigos semelhantes, os autores não deixaram claro como garantir que os endereços MAC fornecidos para os dados de treino eram de fato os dentro do ônibus tendo em vista a aleatorização dos endereços MAC atualmente em prática pelos fabricantes de dispositivos eletrônicos. Supôs-se que eles utilizaram dispositivos habilitados com Wi-Fi que fossem antigos e ainda que ainda não tivessem a referida tecnologia de aleatorização, ou que eles também fabricaram dispositivos auxiliares que pudessem informar o endereço MAC aleatório a cada momento, uma vez que dispositivos *vanilla* (comercializados sem modificação) normalmente não o fazem.

Vieira *et al*. (2020) usaram o cruzamento de informações do sensor GPS e do sensor de sondas Wi-Fi para derivar dos dados brutos o número corrigido de passageiros dentro do ônibus usando um modelo chamado Filtro Kalman. Ele é baseado em programação matemática iterativa para uma estimativa rápida de uma variável aleatória em situações com presença de ruído no sensor, por meio da ponderação continuamente otimizada entre previsões anteriores e medições com ruído do sensor, dando mais peso ao dado com menor incerteza. Este modelo é muito usado em processos dinâmicos, e os autores observaram que seu uso reduziu os erros em até 300% quando comparado ao uso somente dos dados brutos.

A análise dos trabalhos anteriores destaca como a pesquisa nacional sobre o tema, embora esparsa, não está qualitativamente defasada em relação à atualidade e relevância do tema, especialmente as pesquisas conduzidas por Roriz Junior *et al*. (2022), e Vieira *et al*. (2020).

Oransirikul *et al.* (2016) apresentam uma perspectiva robusta sobre sensores passivos WLAN fixamente posicionados em pontos de parada de ônibus, seja terminais ou pontos intermediários na linha. Os autores conseguiram correlacionar a RSSI com a distância do dispositivo, e usaram esse parâmetro como um dos critérios de filtragem dos dados. A preocupação era eliminar dispositivos fixos, como em um escritório, bem como de transeuntes, que não estariam propriamente no ponto de ônibus; foi para isso arbitrado um tempo mínimo de um minuto no qual o dispositivo deveria ser sentido no ambiente, e um máximo de trinta minutos, bem como a emissão de no mínimo duas sondas para que um dispositivo fosse contado como no ponto de parada. Durante o experimento, um valor arbitrário de -79 dB foi usado para a filtragem, mas a análise experimental mostrou que um valor de -72 dB de RSSI teria sido melhor. Além disso, introduziu-se um adiantamento de um minuto entre a captura dos dados e o registro de sua emissão, para compensar o atraso gerado para o atendimento às condições do filtro.

Embora tenham discutido apenas estratégias de processamento de dados e não detalhes da implementação do equipamento, Madsen *et al*. (2022) propuseram em seu trabalho uma técnica de fusão de informações de sensores físicos e passivos para obter uma estimativa precisa do número de passageiros a bordo do ônibus, otimizando os dados brutos dos sensores por modelagem paramétrica, e, de maneira ainda mais impressionante, sem a necessidade de calibração prévia, o que é uma imensa vantagem. De forma astuta, perceberam as características de tempo de resposta e do erro experimental de cada sensor, sendo o sensor físico melhor para perceber o comportamento instantâneo da demanda, e o sensor WLAN melhor para corrigir distorções na contagem de passageiros durante o percurso. Após a fusão de informações, percebe-se uma impressionante correção nas estimativas dos dois sensores. Com isso, os autores aproveitaram as potencialidades de cada sensor para compor de maneira ponderada as medidas que são feitas por cada um deles, inclusive usando o sensor físico para corrigir a medida inferida pelo sensor passivo do número de dispositivos por pessoa.

Ao olhar para os dados de sonda individuais, os modelos desagregados têm a desvantagem de serem passíveis de serem avaliados somente em laboratório ou em testes controlados, uma vez que, devido à aleatorização dos endereços MAC, não é possível biunivocamente garantir que as sondas recebidas pelo sensor vêm de algum dispositivo específico, algo que seria necessário para estabelecer os dados de referência (GT). Ou seja, seria possível comparar somente o resultado final do modelo, e não seus dados de entrada, conforme se supôs em Roriz Junior *et al.* (2022).

No entanto, apesar das vantagens discutidas acima, um parâmetro de atenção continua sendo o de correlação entre o número de dispositivos e número de passageiros na filtragem por sondas; ou seja, quantos dispositivos detectáveis cada pessoa carrega. Pesquisas diferentes encontraram valores diferentes, dependendo de onde a pesquisa foi realizada; Madsen *et al.* (2022) assumiram 0,63 dispositivos por pessoa baseado em estatísticas populacionais da Alemanha, e Minea (2022), na Romênia, encontrou valores de até 0,49 baseado em medidas em campo. Oransirikul *et al*. (2016) encontraram entre 0,4 e 0,2 dispositivos por pessoa quase dez anos atrás no Japão. Estes valores estão intrinsecamente relacionados ao local e ano de aplicação da pesquisa (Madsen *et al.,* 2022).

A fim de quantificar tanto a proximidade do modelo à realidade e, por isso, estabelecer um critério de comparação entre alternativas de configuração do modelo, estabeleceu-se o uso do coeficiente de determinação, representado pelo termo R^2. O ajuste do modelo aos dados pode também ser avaliado visualmente pela proximidade relativa entre as curvas real e prevista, mas

essa forma analítica mede o ajuste do modelo em relação à variância dos dados, por meio de uma proporção modificada entre o erro das previsões do modelo e o erro de uma previsão constante e igual à média do conjunto. Esse coeficiente pode ser aplicado globalmente no conjunto de previsão, sendo conhecido como R^2 total, ou subdividido para cada etapa do processo, gerando um indicador para a etapa de treino, e outro para o teste. Este último é particularmente importante para verificar se o modelo conseguiu identificar padrões dos dados, uma vez que que ele foi gerado aplicando o modelo em dados reais nunca vistos por ele e comparando suas predições com a referência do GT.

Olhar apenas para o R^2 total pode ser enganoso. Isso se deve ao fato de que em geral cerca de 80% dos dados são usados para fins de treinamento, de modo que é possível que um modelo que consiga explicar com perfeição todos eles mas que falhe miseravelmente nos dados reais recebera uma nota muito melhor do que aquele modelo que possui um desempenho semelhante tanto no treino quanto no teste. Por isso, o que importa não são os valores absolutos principalmente do R^2 total, mas, sim, a comparação entre o R^2 de treino e o R^2 de teste. Entende-se que, por ordem de importância, o coeficiente de determinação de teste vem antes do coeficiente de treino, sendo o total calculado por uma média ponderada, mesmo que não-linearmente, entre eles.

É importante interpretar essa informação de discrepância entre as medidas de treino e teste de uma maneira contextualizada, porque poder-se-ia usar simplesmente a diferença entre os dois valores, mas isso não permitiria dizer apenas usando esse indicador se o modelo teve um bom ou mau desempenho, pois um modelo poderia ter uma pequena diferença, o que é bom, mas ter um desempenho próximo de zero, o que é pior do que ter uma diferença ligeiramente maior com resultados melhores.

2.5 TÓPICOS CONCLUSIVOS

Ceder (2007) aponta a dinâmica de causa e efeito da aplicação de ITS no TPC; aumentar a correspondência entre oferta e demanda gera um aumento na confiabilidade do sistema. Isso só pode ser feito com o conhecimento íntimo da demanda, de maneira similar ao sistema de monitoramento e controle da oferta, o que minimiza os atrasos e superlotação dos veículos, que é o efeito benéfico principal que pode ser percebido pela população. Um sensor que possa atender parcial ou totalmente os requisitos apresentados é um forte candidato para ser implementado na transformação dos sistemas de TPC existentes em sistemas ITS, aumentando

sua qualidade de serviço prestado e deixando-os mais atraentes ao público, o que, em troca, pode inclusive aumentar a demanda por TPC, o que traz uma leva de benefícios, conforme sistematizados na Figura 2.2.

Para fazer isso, conforme apresentado na comparação anterior entre os sensores, os sistemas Wi-Fi são fortes candidatos por permitirem a coleta contínua, anonimizada, e a baixo custo dos dados publicamente disponibilizados no ambiente, embora eles precisem de tratamento antes de se tornarem dados de demanda, pois sua natureza é indireta ao dado primário da demanda.

3 MÉTODO

Para a operação com dados reais e atuais não é possível continuar a usar somente o planejamento fruto de uma pesquisa estática. Nesta seção, é discutida a aplicação modificada da tecnologia APC no ponto de parada de ônibus para fins de planejamento de transportes.

Este capítulo descreve os procedimentos usados para o desenvolvimento e validação do sistema de estimativa de densidade de pessoas. Utilizou-se uma abordagem experimental e quantitativa baseada em sensoriamento passivo de redes sem fio usando sinais públicos de radiofrequência, em consonância com os estudos anteriormente apresentados, especialmente os de Madsen *et al*. (2022), Minea (2022), Oransirikul *et al*. (2016), Roriz Junior *et al*. (2022), e Santos e Almeida (2021).

O tipo de sensor passivo WLAN por Wi-Fi foi escolhido dentre as alternativas por três motivos principais, a saber: (1) anonimização da coleta, (2) baixo custo de implementação, e (3) automatização. Este sensor trabalha com dados públicos enviados voluntariamente pelos dispositivos usados pelos passageiros, com vários trabalhos já tendo sido feitos corroborando sua funcionalidade e preservação da privacidade do usuário. Além disso, ele não necessita de equipamentos específicos, patenteados, ou caros, sendo possível sua montagem com partes eletrônicas genéricas que podem ser encontradas comercialmente e instruções na internet a respeito de sua montagem, bem como parâmetros bem-definidos de programação de suas funcionalidades, dirigindo a coleta e processamento de dados. Isso permite a coleta ininterrupta de dados para fins de operação e planejamento de transportes, mesmo que uma calibração inicial seja necessária para assegurar a fidelidade dos dados de saída.

No entanto, o sensor apenas capta os dados; um modelo é necessário para interpretá-los. Por isso, é realizada também uma breve comparação entre possibilidades de método de análise a fim de escolher um método que seja capaz de demonstrar que os dados são significativos para a descrição da demanda, mesmo que de forma aproximada.

Conforme disposto na Figura 3.1, a aplicação do método é feita de modo sequencial e dividida em quatro etapas principais, a saber: (1) especificação e montagem dos componentes do equipamento sensor, (2) programação do protótipo para coleta de dados, (3) testes iniciais e levantamento de campo, e (4) análise dos dados coletados. Adicionalmente, o método é suplementado por discussões metodológicas da arquitetura do sistema e do modelo usado na análise de dados.

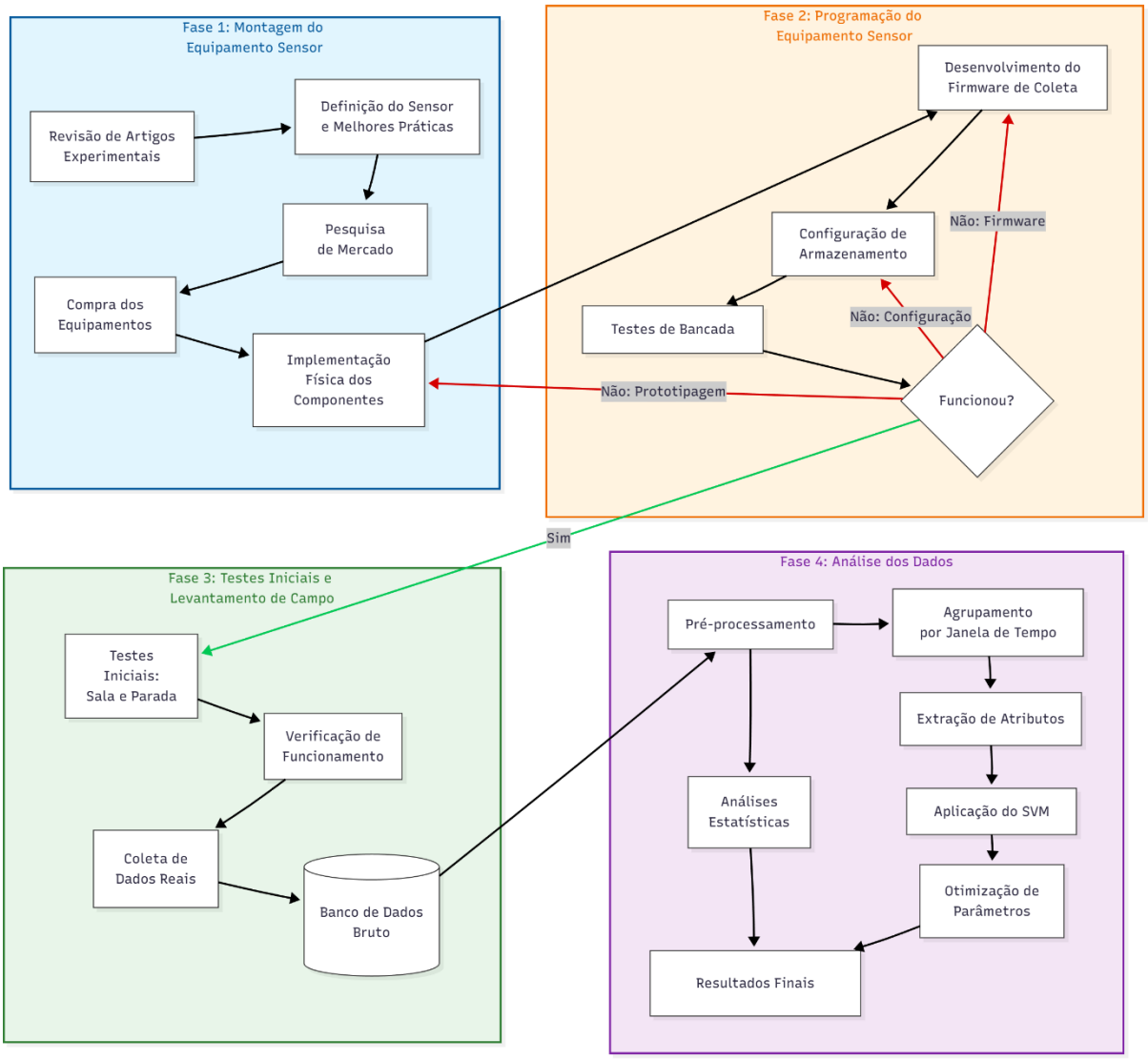


Figura 9: Processo de aplicação do método, com entradas e resultados de cada etapa.

3.1 ARQUITETURA DO SISTEMA

A Figura 3.2 evidencia a arquitetura de comunicação entre os componentes do equipamento sensor e as etapas posteriores de tratamento de dados, aos moldes de camadas ITS. Esta é uma visão conceitual do processo, que é descrito em maiores detalhes nas subseções correspondentes.

O sistema inicia seu processo por meio da inserção do conjunto sensor no ambiente físico. Naturalmente, as pessoas que carregam dispositivos móveis consigo são como estações-base de rádio, pois seus eletrônicos habilitados com comunicação Wi-Fi estão constantemente procurando pontos de acesso Wi-Fi, emitindo sondas (pacotes do tipo *Probe Request*).

Entrando na camada de sensoriamento, o sensor possui uma antena que pode ser regulada para receber quaisquer sinais Wi-Fi. Uma vez funcionando, ela capta e transfere todos os sinais recebidos para o microprocessador ESP32, que verifica se o pacote recebido é de fato do tipo sonda. Em caso positivo, ele é armazenado no cartão SD; senão, ele é descartado.

Uma camada importante é a de comunicação externa, composta de dois procedimentos distintos; o de sincronização de horário e a de comunicação de emergência. A sincronização de horário é feita apenas sob demanda do pesquisador, e somente no momento em que o dispositivo é ligado à energia. Já a comunicação de emergência é um procedimento que pode ser acionado a qualquer hora que o sensor perceber que há um erro em seu processo de coleta de dados, e avisa o pesquisador via mensagem de texto.

Após a finalização da coleta de dados em campo, o pesquisador então leva o sensor para seu computador a fim de extrair os dados do cartão SD do equipamento. Depois, as marcas temporais registradas pelo conjunto sensor devem ser compatibilizadas com o horário real da coleta. Caso haja mais de uma coleta no mesmo local a fim de visualizar o comportamento ao longo do tempo, os dados devem ser juntados em um arquivo único, que é então pré-processado ao ser inserido em uma base de dados para melhor eficiência de processamento.

Em seguida, o programa de processamento e análise de dados entra em cena, realizando as análises estatísticas dos dados brutos e elaborando os gráficos correspondentes, para depois organizar os dados em janelas temporais e atributos definidos previamente à análise. Caso já haja hiperparâmetros ótimos definidos por análises prévias daquela parada, a base de dados de atributos é repassada diretamente para alimentar o modelo SVM; caso contrário, é realizada a etapa opcional de otimização dos hiperparâmetros do modelo (*grid search*). Após o treinamento do modelo com os dados de treino fornecido pelo banco de dados dos atributos, o modelo é aplicado aos dados de teste, que quantificam a precisão do modelo.

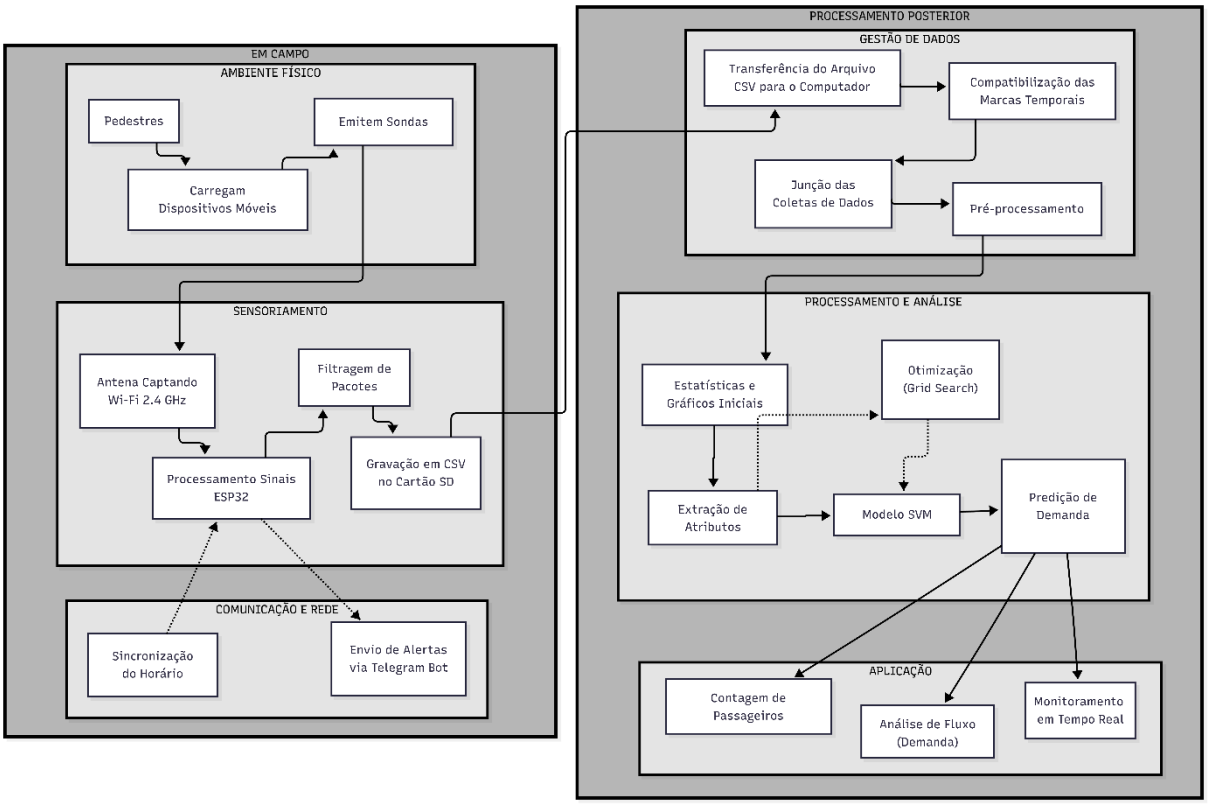


Figura 10: Arquitetura de comunicação entre os componentes e do fluxo de informações.

A sub-etapa de validação da predição da demanda é essencial para identificar possíveis limitações no modelo, que pode então ser aplicado para contagem instantânea de passageiros, análise de demanda horária, ou até mesmo evoluir para um monitoramento em tempo real, que, embora não seja um objetivo do presente trabalho, é uma aplicação validada pela literatura.

Todo equipamento eletrônico é constituído de uma parte física (*hardware*), e uma parte programada (*software*), que são complementares, integradas, e funcionam ao mesmo tempo. Dessa maneira, o conjunto sensor é uma simbiose entre os dois, mas, para o propósito de clareza metodológica, são discutidos de forma separada.

3.2 ESPECIFICAÇÃO E MONTAGEM DOS COMPONENTES DO EQUIPAMENTO SENSOR

Discute-se nesta subseção a parte de *hardware*, ou o equipamento físico em si. Levou-se em conta para a construção do equipamento os trabalhos representados na Figura 3.3, especialmente os de Santos e Almeida (2021), Roriz Junior *et al*. (2022), e Oransirikul *et al*.

(2016), sendo mais semelhante a este último. O trabalho de Lin e Zhao (2025) também foi levado em conta pela semelhança na implementação de *hardware*, embora os sensores utilizados sejam diferentes, situação inversa à de Xu *et al*. (2025), que tem semelhança no sensor, mas não dá detalhes quanto ao equipamento usado.

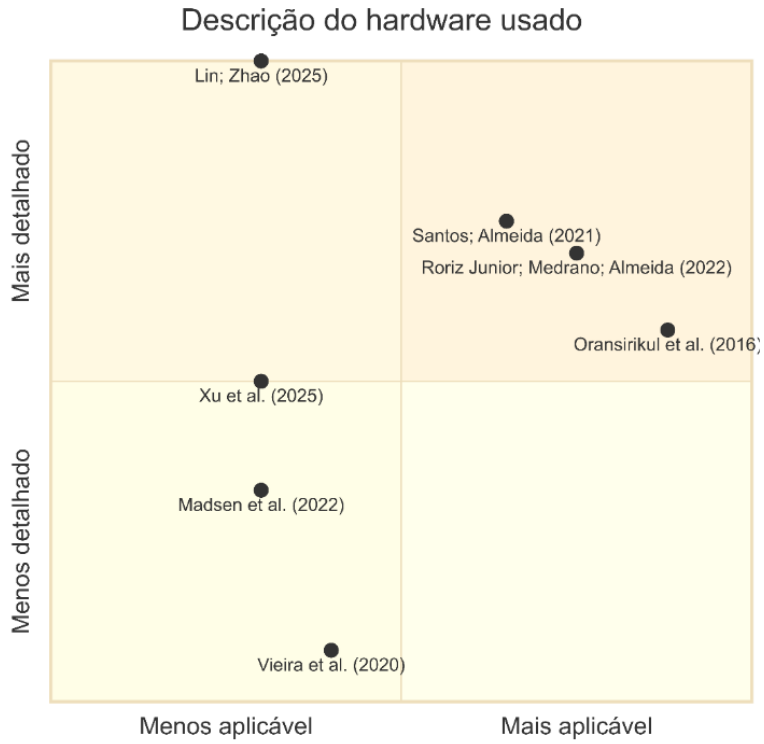


Figura 113: Gráfico de quadrantes ilustrando aproximadamente os parâmetros de detalhamento e aplicabilidade dos artigos mencionados**.**

Enfatiza-se que a avaliação de aplicabilidade levou em conta as informações disponíveis sobre o aparato experimental, de modo que os artigos menos detalhados não forneceram informações suficientes sobre o equipamento usado de forma a avaliar a semelhança com o que se propõe neste trabalho. Métodos pouco descritivos tornam a pesquisa desenvolvida não replicável.

Em teoria, apenas a placa de desenvolvimento com sua antena e conexões, e um cartão de memória, já seriam o suficiente para os propósitos aqui discutidos, uma vez que este módulo é um microcontrolador completo, com uma antena que capta sinais tanto das tecnologias Wi-Fi e Bluetooth (BT), e um processador *dual-core* mais potente que a maioria dos Arduinos. Ele pode, portanto, ser programado diretamente para operar em modo monitor e coletar endereços MAC em seu cartão de memória, procedendo à análise posterior destes dados. Outra opção é o uso do módulo trabalhando junto com o microcontrolador Raspberry Pi, no qual cada um

desempenha uma tarefa específica, aproveitando suas fortalezas; a placa ESP32 funcionaria como coletor de dados por sua capacidade nativa e eficiência como monitor Wi-Fi, e a placa Raspberry Pi serviria como processador central, recebendo os dados coletados pelo monitor Wi-Fi via comunicação serial e processando ou enviando os dados à internet, uma vez que o chip ESP32 não permite a coleta e transmissão de dados Wi-Fi simultaneamente (Roriz Junior *et al*., 2022; Santos e Almeida, 2021).

Diferentemente dos trabalhos que usaram microcontroladores como o Arduino ou o Raspberry Pi, este trabalho se propôs a ser ainda mais econômico e usar somente o sensor ESP32. Para operações permanentes, seria necessário um ponto de energia e uma conexão de internet para transmissão dos dados. Contudo, para um teste-piloto, o sensor poderá ser alimentado por um carregador portátil (*power bank*), e ter as informações coletadas gravadas na memória do dispositivo, para coleta e posterior análise. A relação completa dos componentes necessários para a montagem do equipamento sensor encontra-se descrita na Tabela 3.1, e a fotografia deles juntos, na Figura 3.4. O custo total aproximado para a realização da pesquisa foi de US$50, ou R$ 260, valores referentes ao mês de outubro de 2025.

Tabela 3: Resumo dos componentes e equipamentos necessários para a coleta de dados.

ID	Componentes	Quantidade	Custo (jul-out/25)
A	Módulo Wi-Fi/Bluetooth ESP32 Devkit	1	US$ 15,26
B	Módulo cartão micro SD + soldagem dos pinos	1	US$ 4,49
C	Cartão micro SD 32 GB Sandisk Ultra com adaptador para cartão SD	1	US$ 7,18
D	Carregador Portátil *Power Bank* Trust 4400 mAh	1	US$ 10,30
E	Luminária Arandela Tartaruga LED 20 W 6500K Avant	1	US$ 4,33
F	Cabos *jumper* fêmea-fêmea	1	US$ 2,01
G	Cabo micro-USB	1	US$ 4,60
-	**Custo total do conjunto sensor**	**-**	**US$ 48,18**

Cotação média do dólar USD entre julho e outubro de 2025: R$ 5,43

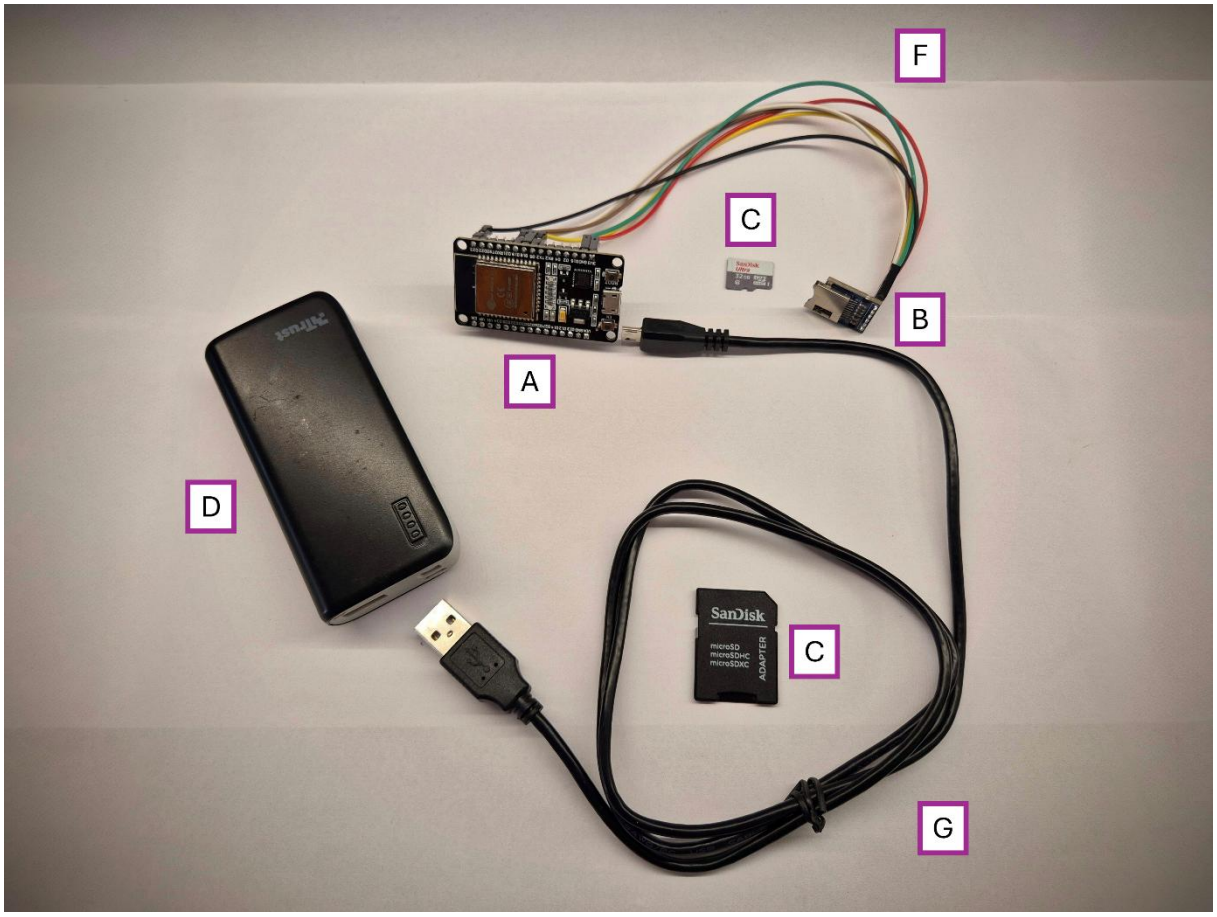


Figura 12: Partes que compõem o dispositivo.

Após a aquisição dos componentes do dispositivo sensor, foi necessário montá-los e testar sua operação em um ambiente controlado, para que se verifique sua funcionalidade. O raciocínio que baseia esta escolha é de identificar problemas à medida que eles forem surgindo, para ir validando a montagem em etapas. Por exemplo, a alimentação elétrica da placa foi a primeira coisa a ser testada, a fim de garantir que não havia mau funcionamento; depois, foi testada a conexão da placa com o computador para o descarregamento de um programa simples, a fim de ver se os circuitos do componente estavam funcionando como deveriam; depois, um programa simples de detecção de sondas para ver se ela consegue detectar um celular próximo; depois, um teste de captura de sondas de múltiplos aparelhos próximos; para então testar a capacidade do tempo de operação contínuo do equipamento, que foi arbitrado em no mínimo absoluto 2 horas, com alguma folga para a montagem.

Em seguida, testou-se o armazenamento dos dados no cartão SD, que na realidade é da tecnologia micro SD, mas, onde esse nível de refinamento não for necessário, será tratado

genericamente como cartão SD. Uma observação importante é que o módulo cartão micro SD foi vendido para os autores desmontado, de modo que se contratou um técnico para efetuar a soldagem dos pinos macho-macho nos respectivos buracos. Com isso, foi possível conectá-lo ao módulo de desenvolvimento via cabos elétricos do tipo *jumper* nos pinos correspondentes do circuito.

3.3 PROGRAMAÇÃO DO PROTÓTIPO PARA COLETA DE DADOS

Nota-se que esta etapa também inclui a programação que permite a coleta de dados, desde os primeiros até os últimos testes. Para tal, foram utilizados os programas Visual Studio Code, da Microsoft, para a escrita de fato do programa a ser usado, e o Arduino IDE, para compilar o programa em C++ e enviar ao chip ESP32. Mesmo que o microcontrolador não seja propriamente da família Arduino, é por meio deste programa que o código é processado e transmitido à máquina.

Ressalta-se que os resultados gerados pela IA foram revisados em sua totalidade antes de serem colocados em prática, e que o autor tem total responsabilidade pelo resultado final. Na Figura 3.5 encontra-se descrito visualmente o processo cíclico de captura das sondas Wi-Fi por parte do equipamento sensor atuando em campo, após os procedimentos iniciais de sincronização de horário, descritos a seguir. Na Figura, os retângulos indicam ações; quando verdes, indicam o início ou fim de uma repetição do processo cíclico, enquanto que a cor azul indica a manipulação de um arquivo do cartão SD, e vermelho indica o protocolo de envio de uma mensagem de emergência. Os losangos laranjas indicam etapas decisórias ou de verificação, na qual o programa computacional precisa verificar os valores recebidos contra as ordens que ele recebeu, e seguir no caminho correspondente (sim ou não).

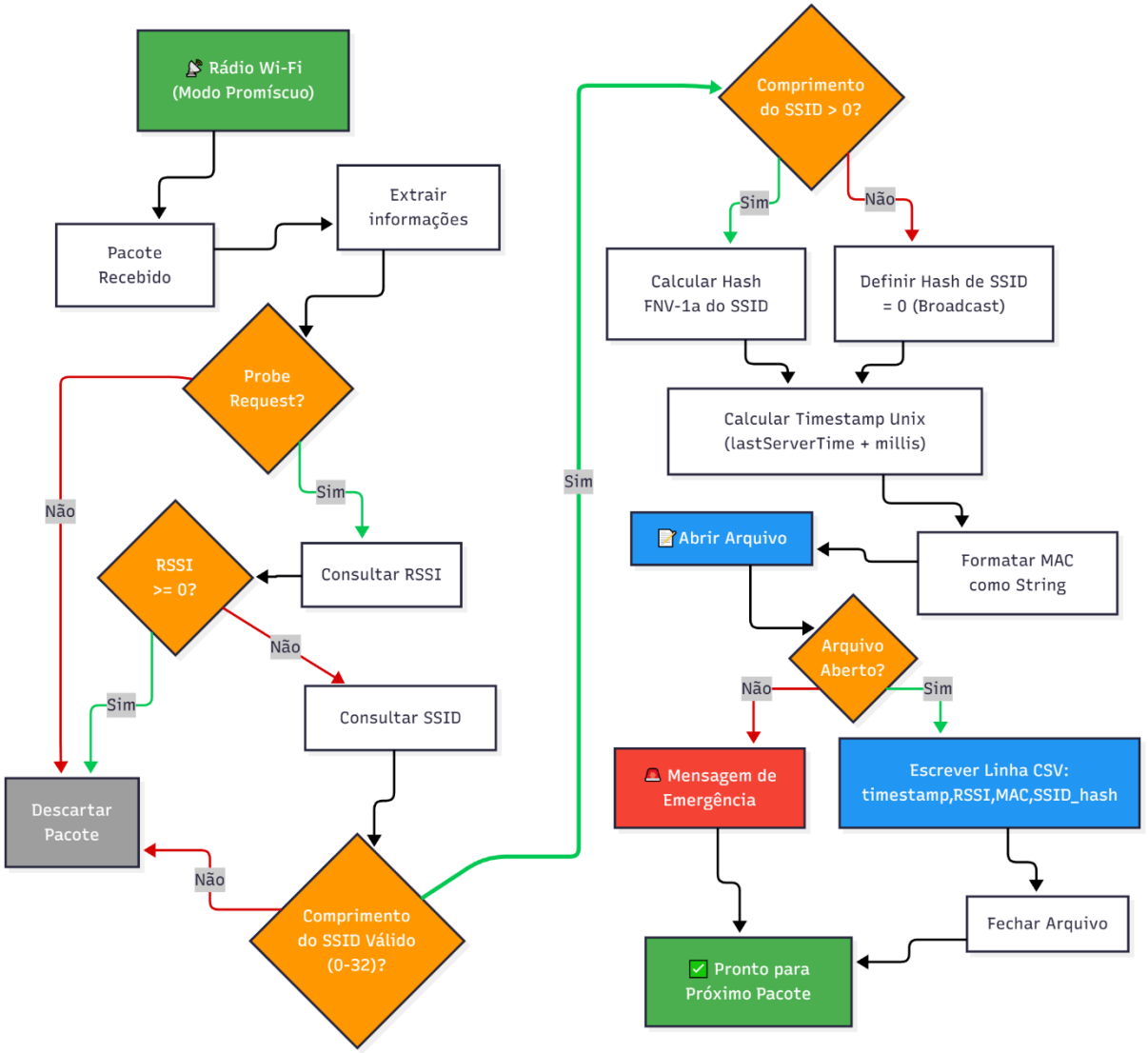


Figura 13: Fluxograma de captura de sondas Wi-Fi.

Antes de cada coleta de dados, é importante realizar a sincronização inicial do relógio interno do sensor, para que os dados sejam gravados com um horário de referência facilmente identificável, que pode ser pré-especificado ou o horário daquele instante, ao se conectar a um servidor de horário pela internet. De qualquer forma, este horário é escolhido pelo operador e pode ser ajustado posteriormente. Diferente dos valores transientes dentro do módulo sensor, que são apagados após a desconexão elétrica, este valor é registrado permanentemente na memória do componente, sendo apenas apagado caso uma nova versão do programa seja sobrescrita na memória. É importante descrever que o programa foi elaborado com uma função de verificação de integridade antes de o início da coleta de dados, de forma que, ao ser ligado à energia, o equipamento tenta se conectar a uma rede Wi-Fi preestabelecida (especialmente o

telefone celular do pesquisador, atuando como roteador), para enviar uma mensagem ao pesquisador por meio de uma API do Telegram Messenger dizendo que foi feita uma verificação inicial da antena e do cartão SD, e que o equipamento está pronto para coletar os dados. Então, o sensor se desconecta da rede do pesquisador e muda o funcionamento de sua antena para o modo promíscuo, de captura de pacotes Wi-Fi que obedeçam às especificações de quadro da normativa IEEE 802.11. Além disso, o sensor foi orientado a manter sua antena fixa no canal 1, evitando possíveis perdas causadas pela alternância entre canais, mas potencialmente não captando pacotes enviados apenas por outros canais, algo que se julgou um risco menor, uma vez que requisições do tipo sonda são emitidas para diversos canais.

A programação do equipamento sensor tem uma lógica cíclica e orientada a eventos para a captura de dados, no qual o sensor fica passivamente "escutando" o seu ambiente, esperando a detecção de pacotes Wi-Fi. O programa funciona de maneira assíncrona e não bloqueante, de forma que apenas no momento em que a antena capta um pacote válido a pilha de protocolos Wi-Fi do microprocessador ESP32 invoca uma função de retorno (*callback*) e ocupa a CPU, sendo, portanto, a aquisição de dados por parte do sensor acionada pela atividade do transmissor. Isso reduz a chance de perder uma sonda, pois a outra alternativa seria especificar uma frequência de amostragem da antena, que, por menor que fosse, poderia perder requisições.

Para cada pacote capturado, valida-se a correspondência do cabeçalho com o tipo de requisição 0x40 (*Probe Request*). Caso a condição não seja satisfeita, o pacote é imediatamente descartado, evitando o consumo desnecessário de memória, porque não existe motivo factível para salvar esse dado se ele legalmente não pode ser usado. Por isso, como não é um impacto computacional muito grande fazer essa etapa de validação, optou-se por integrar esse passo no próprio sensor, e não salvar todas as sondas para analisá-las depois. Agora que se verificou que o pacote de fato é uma sonda, continua-se o processo consultando a informação de RSSI, que somente é positiva para reflexões de sinal originadas dentro do próprio sensor, e mantendo somente os valores negativos, o que garante que não haja sinais indevidos na base de dados. Depois, verifica-se o comprimento do SSID, que deve ter no máximo 32 bytes; caso haja uma requisição com uma informação maior que o limite, ela provavelmente está corrompida e deve ser descartada. Essa informação que pode ter um tamanho variável é transformada e compactada em um número de tamanho fixo de apenas 4 bytes, por meio do algoritmo *hash* FNV-1a, que é explicado em maiores detalhes no Apêndice A. Caso o SSID venha não-especificado, a ele é atribuído um valor numérico de 0, o que preserva esta informação pois o algoritmo *hash* é incapaz de produzir este valor. Em seguida, é calculada a marca temporal para ser registrada no

arquivo interno ao somar o horário de referência gravado no equipamento com o seu relógio interno que informa a quantidade de milissegundos (*millis*) desde o momento em que foi energizado; toda vez que é desenergizado, este contador é zerado. Finalmente, o endereço MAC é transformado de um número para texto, usando uma forma padrão de 6 bytes. Para salvar essas informações, é feita uma verificação de abertura de arquivo, para salvar as informações recebidas e deixar o sensor novamente pronto para receber outro pacote Wi-Fi. Caso o equipamento não consiga abrir o arquivo, significa que algo deu errado internamente; neste caso, um módulo de emergência é acionado, que desliga o modo promíscuo e tenta se conectar a uma rede conhecida, para, por meio do Telegram, enviar uma mensagem ao pesquisador dizendo que há um problema com o sensor.

Após todas estas verificações, o dispositivo escreve as propriedades especificadas em um arquivo CSV dentro do cartão SD que pode posteriormente ser facilmente transformado em uma base de dados ou até em uma tabela para uma visualização mais limpa. A Tabela 3.2 apresenta um exemplo fictício de dados brutos provenientes do conjunto sensor, que contém os títulos na primeira linha. Na coluna "time" estão as marcas de tempo em formato Unix, na coluna "RSSI" estão os valores medidos pelo sensor, que são sempre negativos[1], na coluna "MAC" estão os endereços conforme o sensor os recebeu, e na coluna "SSID_hash" estão os valores do *hash* FNV-1a das requisições que continham um SSID especificado a procurar, e 0 caso contrário.

Tabela 4: Exemplo de dados brutos provenientes do conjunto sensor.

time	RSSI	MAC	SSID_hash
1762870085	-94	aa:bb:cc:dd:ee:ff	1526184382
1762967448	-49	a1:b2:c3:d4:e5:f6	400712832
1763070586	-66	11:22:33:44:55:66	0

3.4 TESTES INICIAIS E LEVANTAMENTO DE CAMPO

O pré-teste é feito em ambiente controlado, potencialmente dentro de um ambiente doméstico ou de escritório, e deve durar cerca de 3 semanas.

[1] O RSSI na verdade não é um dado primário, como o endereço MAC, que o conjunto sensor recebe. Quando um pacote é recebido, ele possui uma intensidade de sinal própria, que é comparada em base logarítmica com a potência de 1 mW para gerar o indicador de RSSI (Minea, 2022; Roriz Junior *et al*., 2022). Como as potências de sinal transmitidas pelo meio aéreo são fisicamente atenuadas, é muito difícil equipamentos comerciais comuns chegarem a potências similares, ficando eles quase sempre na faixa inferior a 1 mW e portanto resultando em uma fração com numerador menor que denominador, que ao ser avaliada pelo logaritmo recebe um sinal negativo.

Para fins de teste do equipamento, optou-se por realizar um levantamento compreensivo de campo na parada de ônibus entre a Faculdade de Tecnologia e o ICC Norte, do Campus Darcy Ribeiro da Universidade de Brasília (CDR/UnB), conforme ilustrada na Figura 3.6. Sabe-se de trabalhos anteriores do grupo de pesquisa que esta é uma das paradas de ônibus mais movimentadas dentro do CDR/UnB, em conjunto com as paradas do ICC Sul, ICC Norte, e da Faculdade de Medicina (Universidade de Brasília, 2023). A respeito do posicionamento do sensor no campo, primeiro foi escolhido um ponto de ônibus, por conveniência ou por expectativa de melhor resultado, depois será buscada a melhor posição para o sensor, de modo a maximizar a centralidade na área de espera dos passageiros, ao mesmo tempo maximizando a linha de visada do sensor e evitando deixá-lo visível aos transeuntes.



Figura 14: Vista frontal da parada de ônibus pesquisada.

Ademais, inicialmente foi planejada uma coleta de dados ininterrupta ao longo de um dia, de manhã cedo à noite, com o posicionamento do sensor idealmente de madrugada antes das 5h, para que se pudesse coletar dados até por volta das 20h, quando se projeta o fim do pico de demanda noturno. Contudo, isso somente seria possível se o autor obtivesse acesso às câmeras de segurança que observassem o ponto de parada, algo que não foi possível dentro do cronograma de desenvolvimento do projeto final. Por isso, alternou-se a estratégia para a coleta por períodos de duas horas, acompanhados pela coleta visual com registro manual dos dados de referência (GT).

Finalmente, procurou-se coletar os dados de demanda preferencialmente nas terças e quintas-feiras. Isso foi escolhido primeiro para observar a representatividade dos dados em um dia normal de pesquisa de transportes (terça, quarta, ou quinta-feira), e também para que houvesse tempo na quarta-feira de descarregar os dados para um computador e recarregar a bateria. Contudo, com a alternativa de estratégia, este último motivo não precisava mais ser observado, de modo que inclusive se aumentou o número de coletas, inicialmente quatro levantamentos ao longo de duas semanas, totalizando quatro períodos de análise, para cinco períodos de análise de duas horas cada, o que se acredita ser suficiente para que os dados sejam representativos.

Utilizou-se a ficha de registro de dados exemplificada na Tabela 3.3 como instrumento de coleta para os dados de referência (GT). Sua versão inicial contava apenas com as colunas de registro de horário com precisão de 1 minuto, presença ou ausência de passagem de ônibus, e número instantâneo de passageiros no ponto de parada. Contudo, à medida que o instrumento foi aplicado, observaram-se dois comportamentos anômalos que, pelo bem da indagação científica de representatividade, mereciam ser registrados; (1) alguns acompanhantes de passageiros ficavam por certo tempo no ponto de parada mas depois iam embora, o que por vezes coincidia com a chegada de outro passageiro, de modo que, embora o número total de pessoas permanecesse o mesmo, para o sensor seria inconcebível a detecção de dois dispositivos diferentes serem negligenciados, e (2) algumas pessoas que estavam sentadas na parada por motivos quaisquer (esperando carona, ou outro ônibus, ou só passando o tempo), e que permaneciam mesmo após a passagem do ônibus, que normalmente "zera" o número de passageiros.

Tabela 5: Exemplo de preenchimento do instrumento de coleta de dados de referência.

Hora	Ônibus?	Número	Unique
10h37	0	6	6
10h40	1	6	6
10h41	0	2	7
10h42	0	4	9

3.5 MODELO DE MÁQUINA DE VETORES DE SUPORTE

Interrompe-se brevemente a cronologia da explicação do método para melhor justificar a escolha do modelo de Máquina de Vetores de Suporte (SVM) para o presente trabalho. Baseado na busca bibliográfica da Tabela 2.2, elencam-se cinco motivos principais para a escolha do modelo SVM conforme descrito por Roriz Junior *et al.* (2022).

Um, a maioria dos trabalhos analisados usa métodos de filtragem simples, na qual se estabelece um valor de corte (*threshold*) para classificação binária de sondas referentes a dispositivos que se encontram dentro ou fora do ônibus. Esse limite é baseado em uma correlação de RSSI com distância, tal qual feita por Madsen *et al*. (2022), Minea (2022), e Minea *et al*. (2020), que, por melhor que seja para contar passageiros dentro do espaço físico do ônibus, no qual há controle de acesso, não serve para o presente caso, no qual transeuntes podem passar livremente por baixo do sensor, sem compromisso de serem passageiros que estão esperando na parada de ônibus. Isso limita gravemente a aplicabilidade de usar somente o dado de intensidade do sinal para separar passageiros e não passageiros.

Dois, ao usar esta técnica de filtragem com valores-limite, ainda há o problema da correlação entre sondas válidas e número de pessoas. Os métodos estudados fazem essa conexão usando dados secundários de natureza estatística e amostral da quantidade de dispositivos em uma determinada população local, como apresentado a seguir, o que adiciona mais uma camada de incerteza ao dado.

Três, embora a técnica de FCM Clustering (*Fuzzy C-means Clustering*) utilizada por Xu *et al*. (2025) seja muito relevante para categorização e agrupamento de padrões, ela não foi usada neste trabalho por ser incapaz de fazer regressão; isto é, mapear dados de entrada a uma saída numérica contínua. Além disso, ela se enquadra na categoria de aprendizado de máquina não supervisionado, que não suporta o fornecimento e treinamento com dados de referência (GT), e é do tipo *fuzzy* (difuso), que permite que um dado possa pertencer a mais de uma categoria ao mesmo tempo, perdendo o sentido físico do dado e não sendo o mais adequado para cumprir com os objetivos do trabalho.

Quatro, quanto às técnicas de Filtro Kalman e interpolação usadas por Vieira *et al*. (2020) e Lin e Zhao (2025), elas não são métodos de predição como o SVM o é, sendo apenas ferramentas de suavização e compatibilização de dados, seja de uma única fonte, para a interpolação, ou vindo de fontes diversas, para o Filtro Kalman.

Cinco, esta escolha se torna ainda mais justificada quando comparada com modelos de previsão semelhantes; no trabalho, os autores compararam o modelo utilizado com três outros, e o que teve melhor resultado foi o deles, tanto para acerto de corretos positivos quanto para redução de falsos negativos. Por uma diferença irrisória um modelo otimizado para evitar falsos positivos teve um desempenho melhor nesse aspecto e na atribuição correta de negativos;

contudo, foi muito pior na atribuição correta de positivos e teve o pior resultado de falsos negativos, o que não justifica sua adoção.

Todos os modelos apresentados têm o mesmo objetivo: diferenciar de maneira assertiva os passageiros de não passageiros, principalmente os transeuntes que estão próximos ao sensor. Este problema pode ser generalizado com base na analogia do sinal e do ruído, e como separar um do outro: o objetivo é descartar todo o ruído e manter todo o sinal em sua forma mais pura. No entanto, os sinais não vêm todos da mesma fonte, sendo difíceis de serem identificados todos da mesma forma, e até mesmo os ruídos têm diferentes fontes e comportamentos. Os ruídos estáticos, por exemplo, são aqueles provenientes de por exemplo apartamentos ou escritórios próximos, com dispositivos fixos que estão no raio de influência do sensor, e que podem ser removidos baseado no filtro por um tempo de permanência longo sem alteração no RSSI. Por outro lado, os ruídos dinâmicos são os mais difíceis de serem filtrados, pois são aqueles gerados por transeuntes, sejam eles pedestres com dispositivos, carros com conexão Wi-Fi, ou ônibus cheios de passageiros com dispositivos, que emitem sondas ao passar pelo raio de alcance do sensor ou que podem até passar por dentro da parada de ônibus, mas que não estão efetivamente esperando o transporte e que dessa forma não deveriam ser contados.

É importante também enfatizar uma diferença técnica que diz respeito a dados de entrada e resultados; o modelo SVM faz predições, e não previsões. Predições acontecem quando se têm dados em mão e, usando-os, se fazem asserções ou estimativas sobre o estado atual das coisas, especialmente em fenômenos intangíveis ou que não podem ser diretamente observáveis, enquanto que previsões são baseadas em asserções sobre o futuro levando em consideração dados históricos ou o que aconteceu no passado, o que não precisa de dados de entrada.

Alguns modelos conseguem interpretar padrões de temporalidade para estabelecer correlações adicionais entre os dados. Um ponto negativo do modelo SVM é que ele desconsidera a temporalidade em suas análises, considerando cada pacote recebido na sequência de capturas como um evento independente dos demais, quando na realidade isso não é completamente verdadeiro.

Além disso, os dados podem ser trabalhados individualmente (sonda-a-sonda), ou de maneira agregada, na qual previamente se extraem atributos relevantes para a interpretação do modelo. A fim de possibilitar o treino da Máquina de maneira supervisionada, será necessário agregar os dados temporalmente em janelas de um minuto, em detrimento ao tratamento individual dos dados de sonda.

Nesse sentido, a modelagem proposta fundamenta-se em uma Máquina de Vetores de Suporte que opera sobre uma base de dados agregada em janelas de um minuto, que é feita utilizando atributos meticulosamente projetados, discutidos na subseção 3.6.3 e no Apêndice B, que visam extrair os padrões de atividade dos dispositivos detectados pelo equipamento sensor.

3.6 ANÁLISE DOS DADOS COLETADOS

Após a coleta de dados em campo, é possível realizar a análise destes dados, que é dividida em cinco sub-etapas, a saber: (1) recebimento e armazenamento dos dados brutos, (2) verificação de integridade e compatibilização das marcas temporais, (3) agrupamento por janela de tempo e extração dos atributos, e (4) otimização dos hiperparâmetros do modelo (*grid search*).

3.6.1 Recebimento e Armazenamento dos Dados Brutos

Após o término da coleta de dados em campo, o sensor é recolhido pelo autor desta pesquisa para uma etapa de pré-processamento, na qual os dados gravados na memória interna do aparelho são extraídos por meio da desconexão de seu cartão micro SD e sua conexão em um adaptador para cartão SD que por sua vez se conecta em um computador. Também é possível realizar a transmissão de dados via API do Telegram, mas é uma forma que o autor ainda não conseguiu usar de maneira tão eficiente; ela serve mais como opção alternativa. Os dados brutos são então baixados para a memória interna de um computador, salvos na nuvem em duas vias, e inspecionados visualmente se eles parecem íntegros, pelo tamanho do arquivo e por amostragem visual. Em seguida, os dados são carregados em uma base de dados baseada em SQL para serem permanentemente armazenados. Os dados brutos não precisam ser filtrados por tipo de requisição (*probe requests*), porque o próprio sensor já armazena somente este tipo de requisição, algo que é a favor da LGPD, conforme discutido anteriormente na seção 3.2.

3.6.2 Verificação de Integridade e Compatibilização das Marcas Temporais

Há uma dificuldade na sincronização das marcas temporais registradas pelo equipamento e os horários registrados no GT pelo observador. Isso se deve ao fato de o monitor de tráfego ser ligado por uma bateria, que não fica permanentemente ligada, e por ele não conseguir simultaneamente se conectar à internet e fazer seu trabalho de monitorar sondas Wi-Fi. Por isso,

escolheu-se definir previamente um horário genérico esperado para começar os trabalhos, por exemplo, quinta-feira 11 de novembro de 2025 às 17h, e registrar três horários diferentes entre si na ficha de monitoramento de campo: o valor de horário guardado no sensor, o valor de horário do relógio do observador, e o horário real UTC-3 (fuso horário local) na hora de início da coleta, que seria usado para posterior compatibilização dos horários de forma a entrelaçar os dados provenientes da análise do sensor com os dados do GT.

3.6.3 Agrupamento por Janela de Tempo e Extração dos Atributos

Depois, os dados brutos passam pelo programa de extração de atributos. Mais especificamente, o programa analisa os dados brutos minuto a minuto e calcula para cada atributo previamente especificado o seu valor para aquela janela de tempo.

A Engenharia de Atributos (*feature engineering*) se refere à estruturação das principais informações que podem ser extraídas dos dados de modo a caracterizar algum comportamento observável para o modelo escolhido, funcionando como a ponte que conecta a ciência de dados com suas áreas de aplicação. Dessa maneira, dos quatro dados de entrada que o sensor pode captar diretamente, conforme representado na Tabela 2.1, pode-se fazer categorizações, descrições estatísticas, correlações com bases externas, proporções, entre outras manipulações dos dados brutos de forma a fornecer ao modelo o maior número de informações potencialmente úteis. Para tal, foram estabelecidas três categorias de intensidade de sinal, de forma a tentar descrever a distância do sinal emitido. Baseado nos estudos de Roriz Junior *et al.* (2022), e Minea (2022), e conforme os dados empíricos analisados, os sinais Wi-Fi foram classificados em três categorias delimitadas pelos valores de RSSI:

- Forte: RSSI > -65 dBm (distância próxima, até aproximadamente 6 metros)
- Médio: -80 dBm < RSSI ≤ -65 dBm (distância moderada, entre 6 e 20 metros)
- Fraco: RSSI ≤ -80 dBm (distância longa, maior que 20 metros)

Dessa maneira, conforme sintetizado pela Figura 3.7 e Tabela 3.4, chegou-se ao resultado de 68 atributos selecionados, respectivamente agrupados por tema. Uma lista detalhada de cada um dos atributos está disponível no Apêndice B.

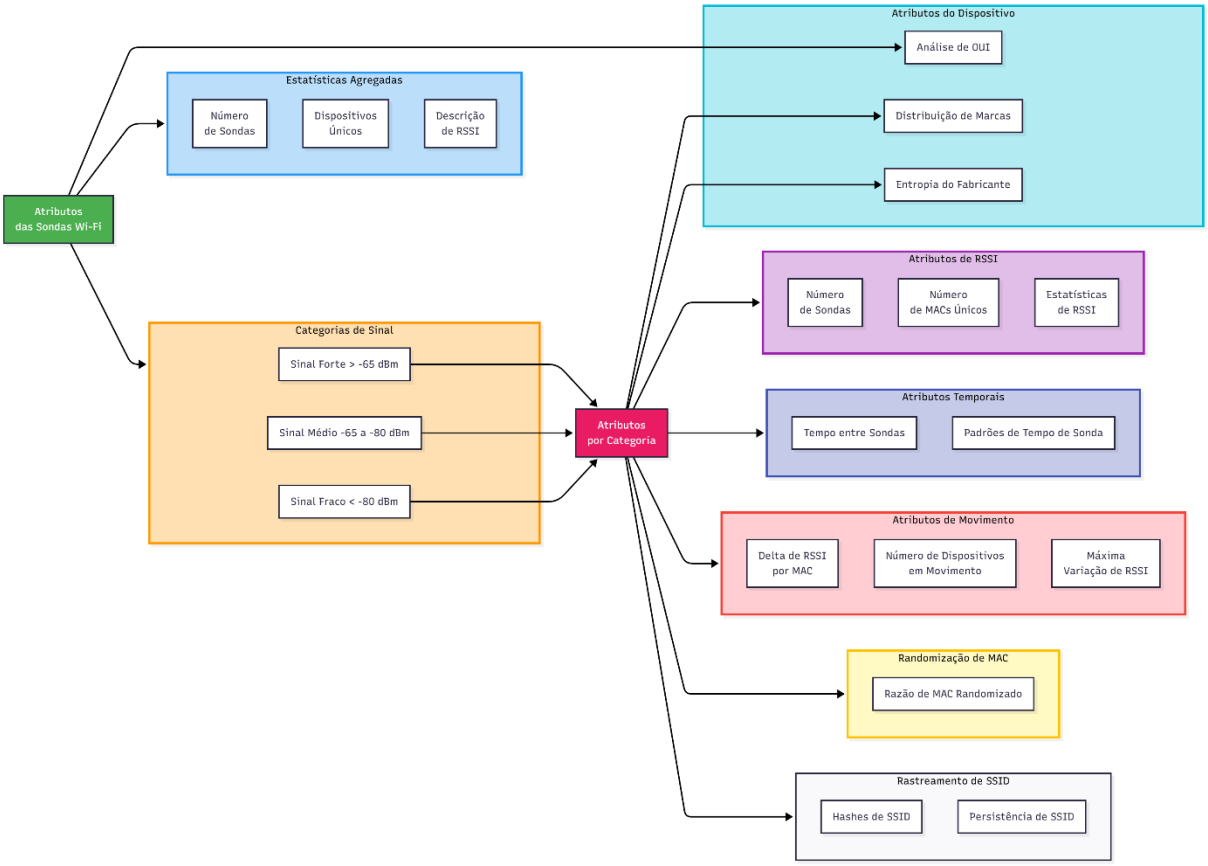


Figura 15: Agrupamento dos atributos por tema.

Tabela 6: Quantidade de atributos por tema.

Tema	Atributos
Estatísticas Agregadas	6
Estatísticas de RSSI	18
Aleatorização de MAC	3
Atributos Temporais	12
Atributos de Movimento	9
Atributos do Dispositivo (OUI)	8
Atributos de SSID	12
Total	68

3.6.4 Otimização dos Hiperparâmetros do Modelo

Nessa etapa, conhecida em inglês como *grid search*, é realizada uma busca dos valores dos hiperparâmetros do modelo que resultam em um melhor resultado do classificador SVM, que é extremamente sensível a cada caso. Para isso, será feita a análise da combinação de todos os valores arbitrários definidos para os hiperparâmetros, o que resulta em algo da ordem de centenas de combinações, que são então comparadas em relação a critérios pré-estabelecidos como o teste do coeficiente de determinação R^2, discutido no referencial teórico. Genericamente, essa técnica permite fazer uma comparação quantitativa entre modelos diferentes para indicar quando o ajuste está inadequado, ou, neste caso específico, como a escolha dos hiperparâmetros influencia na capacidade do modelo em explicar os dados.

A Figura 3.8 ilustra o processo de aplicação do modelo, no qual ele é primeiro treinado com ambos os dados de campo e os dados de referência a fim de "aprender" a transformar a presença dos dispositivos ativos em presença de passageiros, por meio do ajuste matemático de seus parâmetros internos. Para tal, existem hiperparâmetros do modelo que podem ser ajustados para influenciar o aprendizado de máquina interno do modelo. Em seguida, na etapa de teste o modelo recebe apenas os dados de campo e deve então transformá-los em dados de demanda, aos moldes dos dados de referência nos quais ele foi treinado.

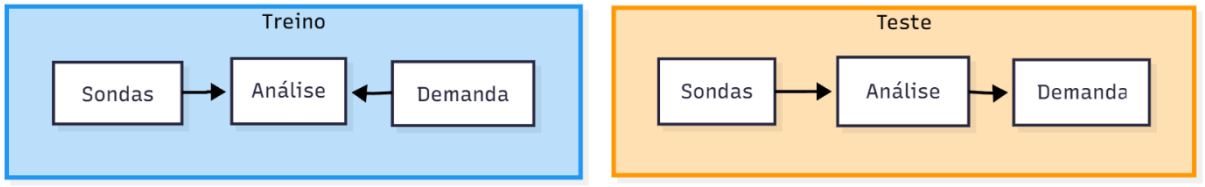


Figura 16: Comparação dos processos de treino e teste do modelo.

Para fazer isso, o SVM tenta encontrar um hiperplano que separe suas classes. A forma exata e a rigidez desse hiperplano dependem fundamentalmente de dois hiperparâmetros: o C e o Gama, mas também depende do Kernel escolhido e do hiperparâmetro Epsilon. Por isso, usando os valores nos intervalos listados na Tabela 3.5, chegou-se a um total de 180 combinações. Embora a multiplicação dos cinco fatores de C pelos quatro de Epsilon pelos três de Kernel pelos quatro de Gama resulte em 240 combinações, deve-se considerar que o kernel linear só admite o Gama configurado como "*scale*".

Tabela 7: Hiperparâmetros do modelo SVM a serem otimizados dentro das opções consideradas.

Hiperparâmetro	Intervalo de valores
C	[100, 10, 1, 0.1, 0.01]
Epsilon	[0.05, 0.1, 0.2, 0.5]
Kernel	["rbf", "linear", "poly"]
Gama	["scale", 0.001, 0.01, 0.1]

Resumiu-se na Tabela 3.6 a influência dos valores de cada um dos hiperparâmetros no sobreajuste ou subajuste do modelo, no qual uma seta apontando para cima indica que um maior valor daquele hiperparâmetro aumenta o ajuste especificado, seja ele sobreajuste ou subajuste, e uma seta apontando para baixo indica que um maior valor daquele hiperparâmetro aumenta o ajuste especificado. Uma discussão mais aprofundada sobre o tema pode ser encontrada no Apêndice A.

Tabela 8: Influência dos valores dos hiperparâmetros no sobreajuste ou subajuste do modelo.

Hiperparâmetro	Sobreajuste	Subajuste
C (Regularização)	↑	↓
Gama	↑	↓
Epsilon	↓	↑

3.7 TÓPICOS CONCLUSIVOS

Este trabalho adota o uso da tecnologia de sensor passivo WLAN por Wi-Fi como a mais viável para os propósitos desta pesquisa, considerando a discussão do referencial teórico. Em consonância com o exposto, recorda-se o funcionamento deste sensor, que coleta sondas Wi-Fi e registra como parâmetros: (1) o horário exato de recebimento delas, bem como (2) o endereço MAC do dispositivo transmissor, seja ele o MAC físico ou o atribuído localmente (aleatorizado), (3) a intensidade relativa de sinal entre os dispositivos, e (4) o SSID procurado, caso haja. Estes dados têm então seus atributos extraídos após serem agregados temporalmente com janelas de 1 minuto, e são analisados posteriormente utilizando o modelo SVM para o estabelecimento da curva de demanda estimada.

O modelo selecionado é capaz de cumprir com a proposição inicial de medição instantânea da demanda. Contudo, o uso de dados agregados para minimizar os efeitos da interferência dos transeuntes não permite a medição do tempo de espera individualizado dos passageiros. Este é um *tradeoff* do método, conforme discutido na seção 3.5.

4 RESULTADOS

A avaliação da capacidade do conjunto sensor em traduzir dados ambientais em demanda quantificável se baseia na comparação entre o resultado predito pelo sensor com a verdade medida em campo tendo os mesmos dados de entrada, que são as sondas emitidas pelos dispositivos no ambiente. Neste capítulo são discutidos os resultados das etapas de aplicação do método, que vão evoluindo em complexidade, na esperança de melhor descrever a realidade do ponto de parada mediante a generalização dos dados de entrada.

É crucial enfatizar a importância do processo de captação de dados, pois dele depende toda a análise. Simultaneamente à captação dos dados pelo sensor nos ambientes pesquisados, foi feita a descrição do número real de pessoas por minuto no ambiente, de modo que fosse possível estabelecer critérios de comparação (*benchmark*) entre a predição e a realidade, o que permite avaliar tanto a correlação global do modelo com a realidade, quanto os melhores hiperparâmetros de ajuste do modelo. A Figura 4.1 mostra uma visão geral do processo de análise de dados, conforme discutido na seção anterior.

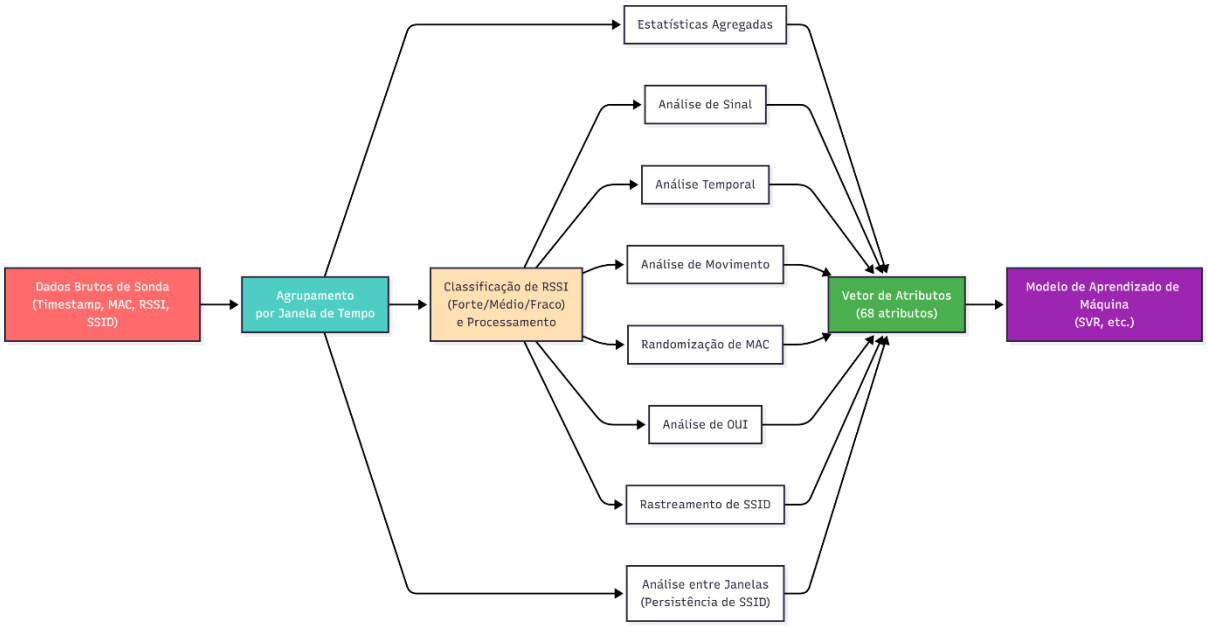


Figura 17: Fluxograma do processo de análise de dados.

4.1 TESTE INICIAL

Entre as datas de 15 de outubro e 6 de novembro de 2025, foram realizados quatro testes controlados para avaliar o desempenho inicial do sensor, principalmente a sensibilidade à

distância entre o sensor e os dispositivos dos passageiros simulados. O local escolhido para o teste foi a parada de ônibus entre a FT e o Instituto Central de Ciências (ICC) Norte, localizada no CDR/UnB, nas datas representadas na Tabela 4.1.

Tabela 9: Relação das datas e locais de realização dos testes experimentais.

		PROXIMIDADE DO SENSOR	
		Distante	**Próximo**
LOCAL	**Ponto de ônibus**	28/10/25	4/11/25

4.1.1 Teste Piloto no Ponto de Ônibus

Na parada de ônibus, experimentou-se a colocação do sensor em dois pontos diferentes. Inicialmente, no dia 28 de outubro de 2025, colocou-se o sensor em um assento da parada propriamente dita. Todos os testes da parada foram não-controlados, já que o número de pessoas que passaram não foi previamente acordado, seguindo a chegada natural dos passageiros na parada. Os resultados deste experimento se encontram na Figura 4.4. Algo que se verificou no local é que, ao contrário do esperado, os passageiros não se assentavam nos bancos próximos ao pesquisador, preferindo ficar em pé esperando o ônibus em uma fila, algo que foi posteriormente constatado também quando o pesquisador observava de longe o comportamento dos frequentadores da parada de ônibus.

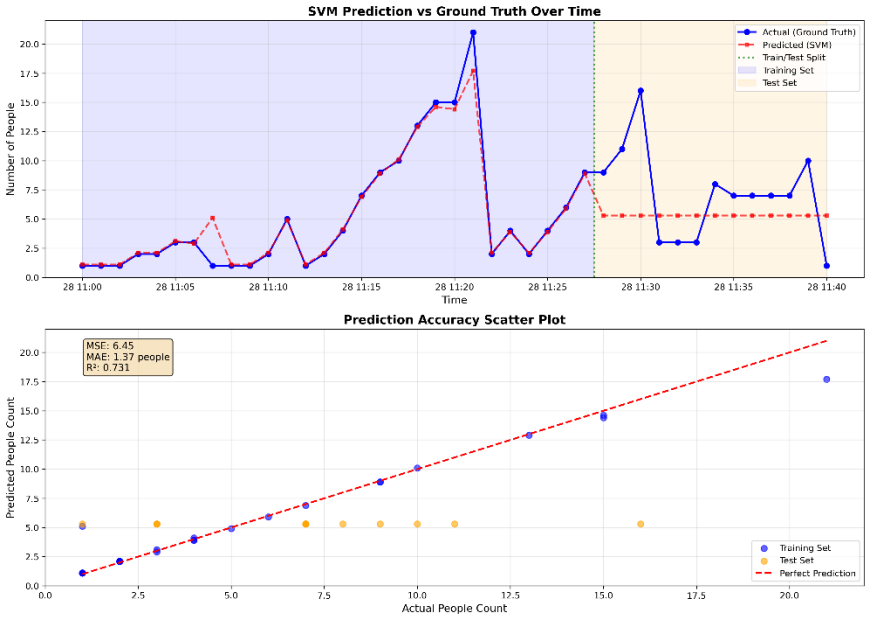


Figura 18: Gráfico comparativo do resultado de treino e teste na parada de ônibus com o sensor distante.

Verificou-se após análise dos resultados que o modelo não conseguiu depreender dos dados de entrada parâmetros internos úteis para prever a demanda, tendo gerado como resultado uma linha reta e constante. Mais uma vez, supôs-se que a distância teve participação na má qualidade destes dados, de modo que resolveu-se experimentar a colocação do sensor em um lugar mais próximo em relação aos dispositivos que seriam percebidos. Assim, no dia 4 de novembro de 2025, após afixar o sensor pendurado na estrutura de toldo na qual os estudantes fazem uma fila esperando o ônibus passar, obtiveram-se os resultados da Figura 4.5.

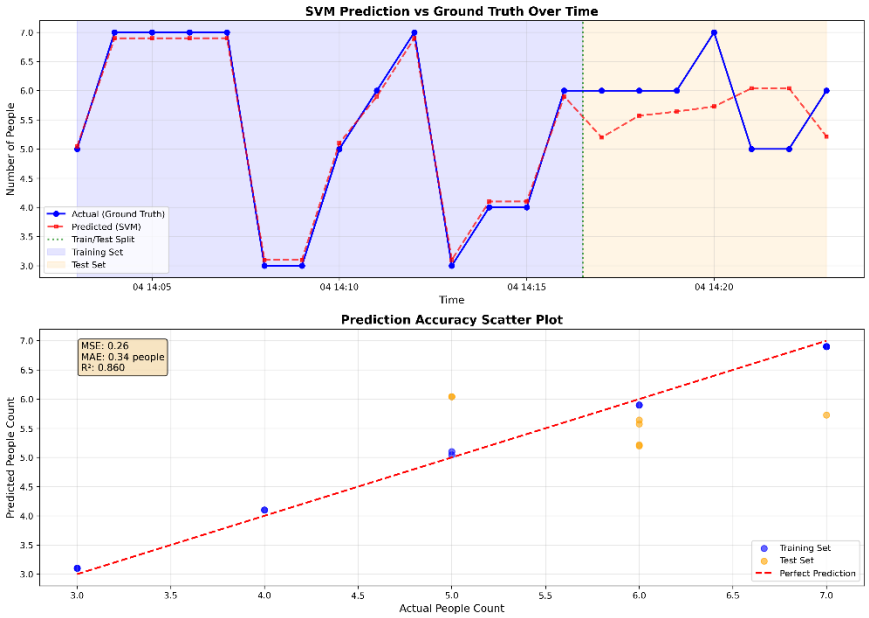


Figura 19: Gráfico comparativo do resultado de treino e teste na parada de ônibus com o sensor próximo.

4.1.2 Discussão

A análise comparativa dos cenários mostra que, conforme se antecipava, a proximidade do sensor ao público-alvo, e, consequentemente, aos dispositivos que serão de fato contados, é um fator-chave para o sucesso no treino do modelo e sua capacidade preditiva. Isto corrobora as observações de Oberli *et al.* (2010) e Minea (2022), que destacaram a importância da manutenção de uma linha de visão desobstruída entre o equipamento e o público-alvo para obter dados de melhor qualidade.

Embora os resultados não sejam perfeitamente conclusivos nesse sentido para o lado positivo da afirmação, verifica-se especialmente pelo experimento no ponto de ônibus que a maior distância entre o sensor e o ponto de demanda foi um fator de influência negativa para a habilidade do sensor em estimar a demanda com sensibilidade.

4.2 APLICAÇÃO

Após a realização dos testes iniciais, aplicaram-se os ensinamentos aprendidos na coleta real de dados, em um período agregado de cerca de dez horas, distribuídas ao longo de três dias.

4.2.1 Coleta dos Dados

Tendo escolhido como lugar de aplicação do sensor para coleta de dados a parada de ônibus entre a FT e o ICC Norte, e pelos testes iniciais estabelecendo o toldo na qual os passageiros esperam o ônibus como ponto ótimo de colocação do sensor, observou-se que o melhor local para o afixar seria pendurado na armação da estrutura, como evidenciado nas fotos seguintes, fabricou-se a estrutura evidenciada na Figura 4.6 para fácil colocação e remoção do conjunto sensor embarcado na luminária.



Figura 20: Luminária modificada para incorporar o conjunto sensor em sua área interna.

Dessa maneira, descreve-se a seguir o procedimento realizado antes de todas as coletas de dados para preparação do equipamento. Removeu-se da mochila do autor a luminária e o estojo do equipamento sensor, em seguida conectando o cabo de energia no módulo de desenvolvimento, que já estava com suas conexões ao módulo SD e cartão SD efetivas, e colocando-se todo esse equipamento dentro da luminária, conforme se evidencia na Figura 4.7, que está sem a armação que se encaixa na estrutura da parada de ônibus apenas para melhor visualização. Ligou-se o ponto de acesso móvel (acesso pessoal) do telefone, para que o ESP32 que já estava esperando esse AP se conectasse a ele, recebendo-se então no telefone celular a mensagem de confirmação de que todas as verificações haviam sido feitas. Ao fim da coleta, para desmontá-lo, realizou-se o procedimento inverso. Graças ao bom Deus durante as cinco realizações da coleta de campo o conjunto sensor não teve nenhum erro.

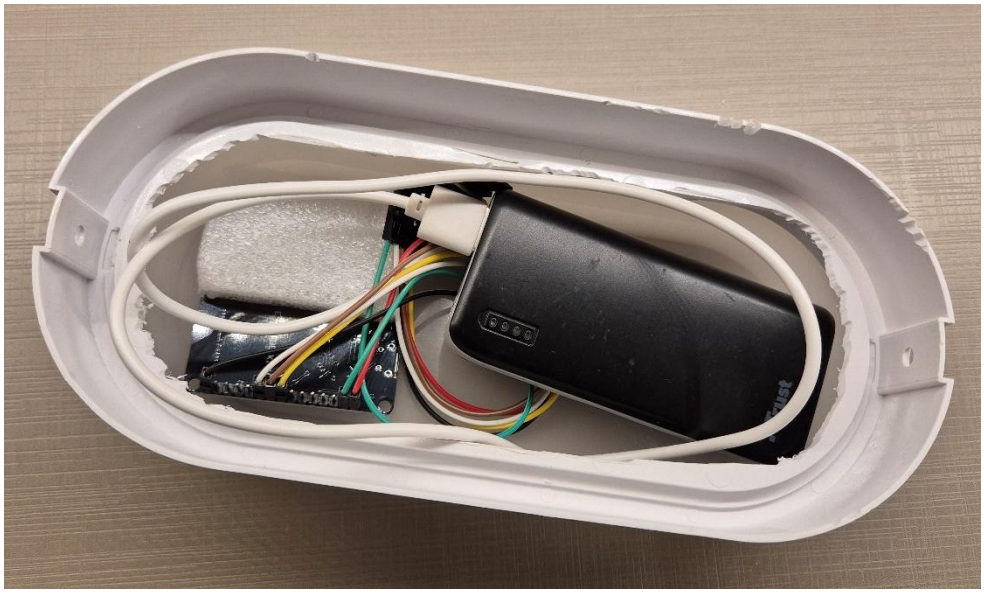


Figura 21: Arranjo do equipamento dentro da luminária, com a armação removida para melhor visualização.

A Prefeitura da UnB (PRC) foi informada e autorizou a coleta de dados, mediante processo SEI 23106.108478/2025-61, público para consulta. A coleta de dados reais de campo foi feita entre os dias 11 e 13 de setembro de 2025, sendo cada coleta realizada em um horário específico. Na terça-feira dia 11 de setembro, foram realizadas duas coletas; a primeira das 11h às 13h, e a segunda das 17h às 19h. Na quarta-feira dia 12 de setembro, foi realizada apenas uma coleta das 13h45 às 15h45. Finalmente, na quinta-feira dia 13 de setembro, foram realizadas duas coletas; a primeira das 10h30 às 12h30, e a segunda das 16h45 às 18h45. Houve alguma variabilidade no início da coleta, principalmente para um horário depois do proposto, devido à necessidade de esperar um ônibus passar e levar todos os passageiros que estavam na parada, a fim de que a instalação do sensor na parada não levantasse suspeita por parte dos usuários, e eles não se sentissem incomodados com um aparelho cheio de luzes e fios coloridos logo acima de suas cabeças. Mas, uma vez instalado, o equipamento ficou *realmente imperceptível*, camuflando-se quase que perfeitamente com o ambiente e principalmente com as outras luminárias operantes do mesmo modelo, conforme evidenciado pela Figura 4.8. A Figura 4.9 ilustra o único caso em que poderia se perceber alguma diferença, pois a presente implementação do conjunto não possui uma fonte de luz semelhante à luminária original; contudo, o usuário que percebesse essa discrepância ainda poderia pensar que se tratasse apenas de uma lâmpada defeituosa.



Figura 22: Visão geral do processo de coleta de dados, com o sensor em pleno funcionamento.



Figura 23: Comparação à noite do conjunto sensor com uma luminária normal.

4.2.2 Análise Estatística dos Dados

Cada coleta de dados produziu um arquivo com todas as sondas capturadas pelo equipamento sensor naquele período. Seria possível analisar cada arquivo separadamente, criando inclusive modelos diferentes para cada um deles. Contudo, isso seria contrário ao objetivo do trabalho, que é estabelecer um modelo único que pode ser usado em qualquer hora do dia em um mesmo local; isto é, admite-se que pode haver coeficientes diferentes para ajuste do modelo dependendo do local de aplicação. Os dados das cinco coletas foram então juntados no computador, resultando em um arquivo que possui 131.729 linhas de dados e uma linha de título ocupando 5,01 MB de armazenamento, com cada linha do arquivo representando uma sonda captada pelo equipamento sensor. Destas, 57.215 eram duplicatas por terem os mesmos valores de marca temporal e endereço MAC, sendo portanto removidas da análise, que ficou com 74.514 sondas distintas entre si. A Figura 4.10 representa a distribuição por histograma dos valores de RSSI percebidos pelo sensor ao longo de todas as dez horas de coleta. Nota-se uma distribuição com cauda mais alongada à direita, indicando uma forte prevalência das intensidades próximas a −90 dBm, bem como alguns picos localizados de intensidade de sinal, que podem indicar dispositivos estáticos ou que enviaram muitas sondas em um curto período de tempo. Conforme os estudos de Minea (2022) e Roriz Junior *et al.* (2022), quanto mais próximo de 0 dBm, mais forte o sinal e consequentemente mais próximo do sensor está o dispositivo.

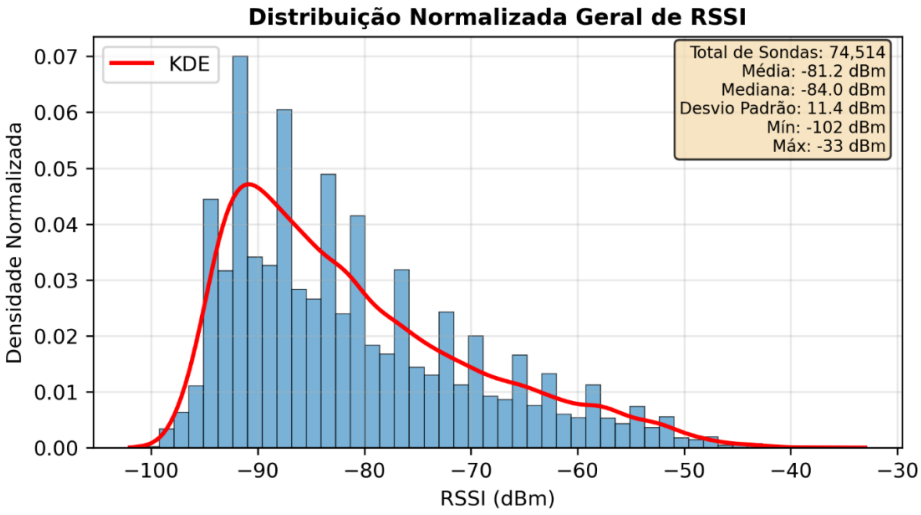


Figura 24: Histograma de distribuição do RSSI das sondas captadas pelas cinco coletas de dados.

Já a Figura 4.11 representa a correlação entre as aparições de endereços MAC entre as coletas de dados. Por contraste visual vê-se imediatamente que a maioria das sondas coletadas são únicas para o período, e que há combinações relevantes entre endereços coletados entre as coletas do mesmo dia. Contudo, também há combinações entre coletas aparentemente sem relação, que devem ser melhor investigadas, mas que podem simplesmente se tratar de endereços conflitantes gerados pela aleatorização de MACs, o que é improvável, ou que a mesma pessoa tenha aparecido na parada de ônibus para se locomover nos horários em que a pesquisa estava sendo feita, o que é mais plausível.

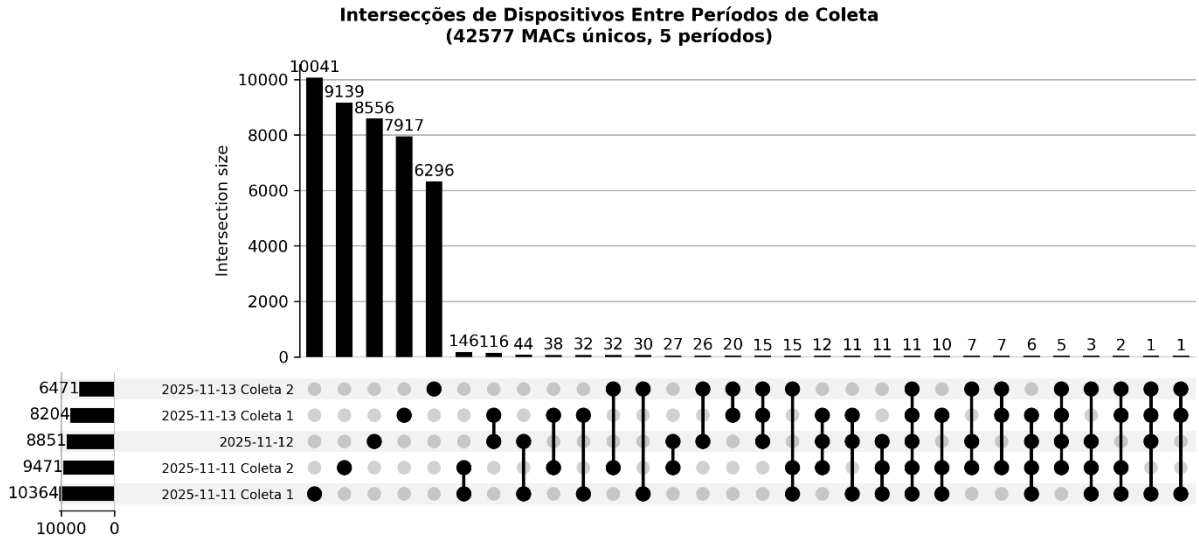


Figura 25: Gráfico UpSet que correlaciona as sondas captadas pelas cinco coletas de dados.

Em seguida, a Figura 4.12 representa uma descrição visual da quantidade de sondas recebidas pelo sensor por minuto, para cada um dos períodos de coleta. A cor mais escura representa o sinal mais forte (mais próximo), enquanto que a cor mais clara representa o sinal mais fraco (mais distante). Pelas escalas dos gráficos percebe-se também que há variabilidade entre as amostras, mesmo quando comparadas em horários parecidos entre dias diferentes. O sensor era sempre colocado no mesmo local e na mesma posição, de modo que não se conseguiu chegar a uma conclusão sobre o acontecido.

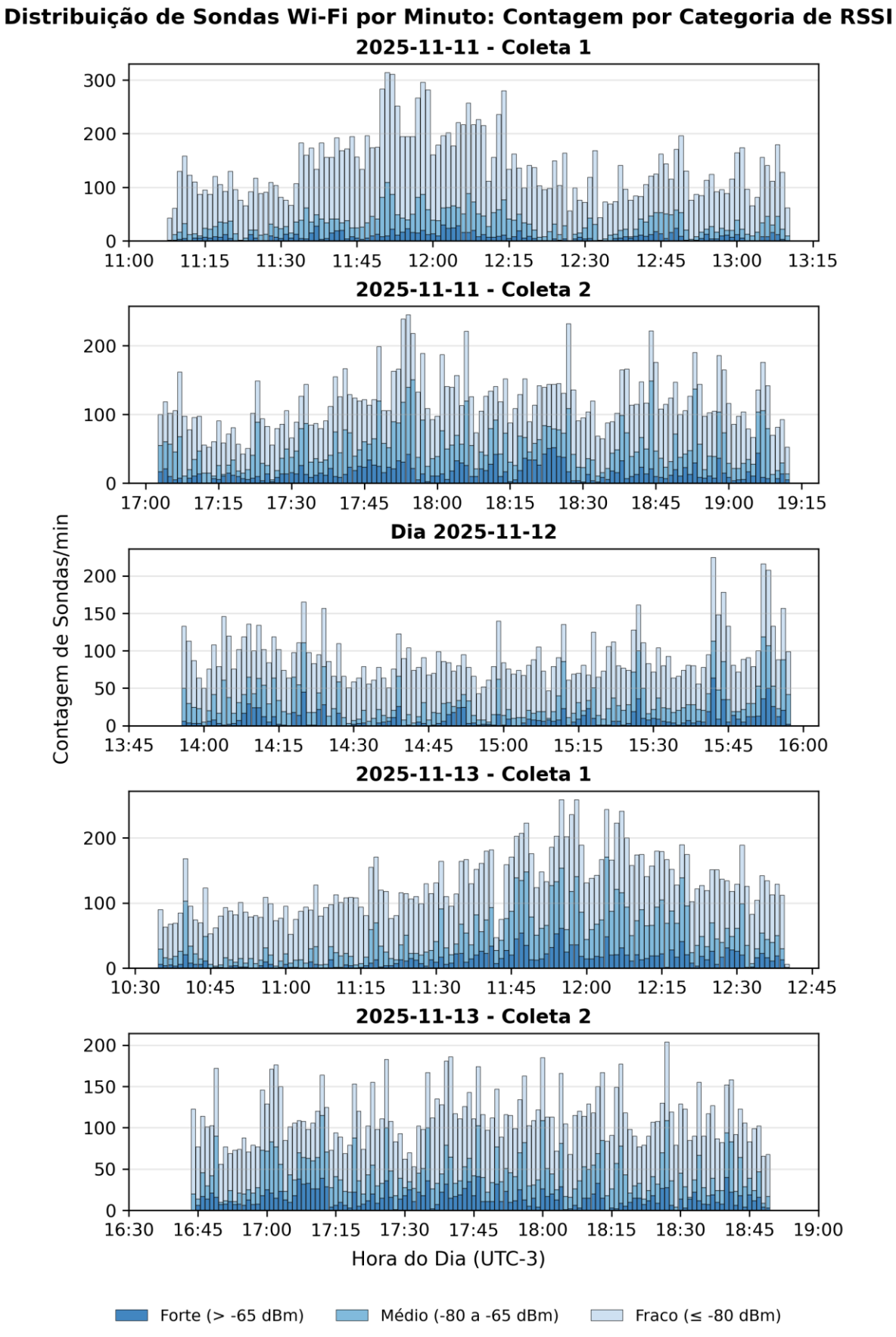


Figura 26: Histograma de distribuição das sondas por minuto com categorização dos RSSI.

4.2.3 Comparação da Utilidade do Modelo SVM

Inicialmente, realizou-se a análise dos dados com cada coleta de dados sendo analisada de forma separada das outras, com os valores de hiperparâmetros padrão do modelo; isto é, C = 10.0, Epsilon = 0.1, Gama = “scale”, e Kernel = “rbf”. Contudo, após análise detalhada dos dados de referência, percebeu-se que não faria sentido a realização de cinco análises diferentes, com cinco modelos diferentes, já que o sensor deveria com um modelo único explicar todas as coletas. Com isso, surgiram os primeiros resultados de R^2 total próximo a 0,64 e agrupamento das previsões ao redor dos dados de referência, seguindo aproximadamente as tendências do gráfico. Mas, a fim de dirimir qualquer dúvida se o modelo realmente era útil, criou-se um gráfico dos sinais fortes, conforme utilizado por outros trabalhos, para ver se ele seria aplicável também a este trabalho, apesar de suas diferenças intrínsecas de aplicação. O gráfico de correlação direta entre o número de sinais fortes detectados pelo sensor e o número real de pessoas na parada naquele instante está na Figura 4.13, e os gráficos de contagem de passageiros e sinais fortes em função do tempo estão na Figura 4.14.

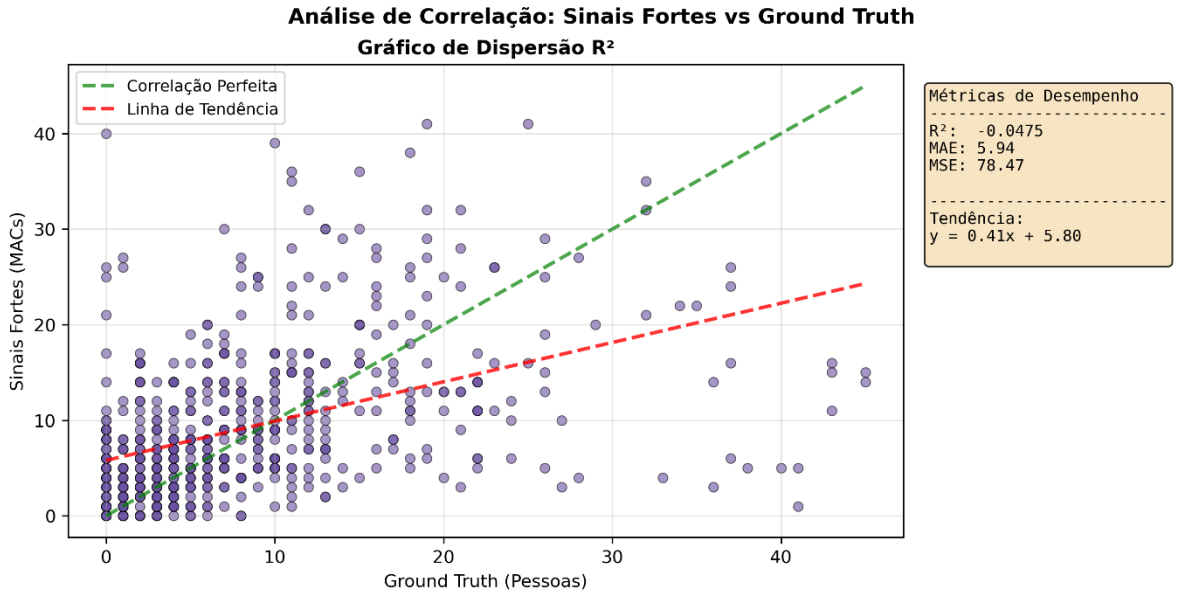


Figura 27: Gráfico de correlação entre o número observado de pessoas, no eixo x, com o número de sinais fortes detectado pelo sensor, no eixo y.

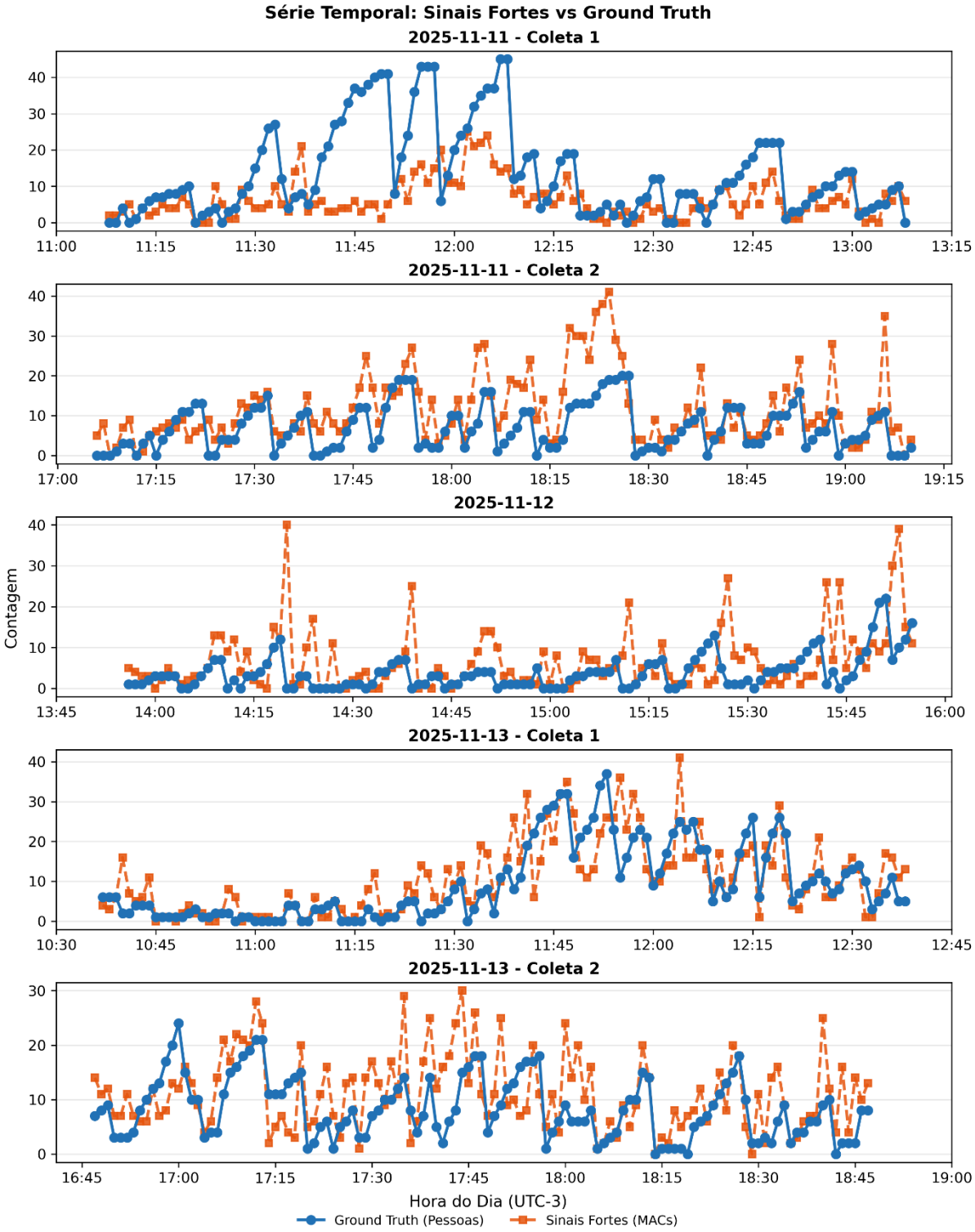


Figura 28: Comparação dos dados de referência, em azul, com os sinais fortes detectados pelo sensor, em laranja.

Pelas Figuras 4.13 e 4.14 apresentadas, observa-se o descompasso entre os sinais fortes e o número real de pessoas na parada. A Figura 4.13 matematicamente aponta a falta de relação destas variáveis pelo valor negativo do coeficiente de determinação (R^2 = -0.0475), que comprova que este modelo que assume que os sinais fortes estão linearmente relacionados com o número de pessoas na parada tem um desempenho pior do que se considerar o número médio histórico de pessoas na parada; ademais, os coeficientes da reta apontam uma impossibilidade de uma sensibilidade baixa ao número de pessoas na parada, uma vez que ele começa em 5,94, que são aproximadamente seis pessoas, quando o número de sinais fortes detectados é zero, e o coeficiente angular de 0.41, menor que 1, aponta a hipersensibilidade a aumentos pequenos no número de sinais fortes. Ademais, pelo gráfico de dispersão observa-se que não há uma bijeção clara entre as duas variáveis, já que um número observado de pessoas pode ter um grande número de sinais fortes, e inversamente um número observado de sinais fortes pode ter várias possibilidades de pessoas; em ambos os casos, é possível observar a ocorrência simultânea de valores extremos, como é o caso de zero pessoas observadas, que podem ter tanto 0 sinais fortes quanto 40, e da mesma forma aproximadamente a quantidade de um sinal forte detectado pelo sensor pode representar 0 pessoas e também mais de 40 (bem como seus valores intermediários).

Já a Figura 4.14 evidencia de forma mais clara a dinâmica entre essas duas variáveis; pode-se ver que, no primeiro período, geralmente há menos sinais fortes do que pessoas, o que pode indicar que as pessoas estão mais longe do sensor, enquanto que no segundo período do mesmo dia, de forma geral há mais sinais fortes do que pessoas, o que pode indicar que as pessoas estão mais próximas do sensor, ou que um número maior de transeuntes passou próximo ao dispositivo sensor. A Figura 4.12 pode ajudar a corroborar essa primeira afirmação, ao evidenciar uma mudança significativa no perfil de potência de sinal, com o visível aumento do número de sondas médias em relação às fracas na amostra analisada. Contudo, o motivo ou mecanismo pelo qual isso acontece ainda permanece uma incógnita.

4.2.4 Otimização dos Hiperparâmetros do Modelo SVM (*Grid Search*)

Considerando as métricas de R^2 de teste conforme discutido no referencial teórico, chegou-se a três grupos principais de resultados com bom desempenho, em ordem decrescente de ajuste do modelo aos dados de teste; ou seja, os que se desempenharam melhor no conjunto de dados no

qual eles não foram treinados foram atribuídos em um grupo com numeração ordinária maior. Além disso, identificaram-se dois outros grupos que tiveram desempenho insatisfatório.

O grupo 1 é caracterizado por uma pequena diferença entre os resultados de teste e total, com ambos próximos a $R^2 = 0,30$, mas não passando da contagem de 15 pessoas. O seu representante com melhor desempenho é encontrado na Figura 4.15, o qual possui R^2 de teste igual a 0,224 e R^2 de treino igual a 0,339, resultando em um R^2 total de 0,329.

O grupo 2, no entanto, é caracterizado por uma relativamente maior diferença entre os resultados de teste e total, com um aumento do coeficiente de treino e decréscimo do coeficiente de teste, já atingindo a contagem de 30 pessoas no treino e quase 20 no teste. O seu representante com melhor desempenho é encontrado na Figura 4.16, o qual possui R^2 de teste igual a 0,174 e R^2 de treino igual a 0,676, resultando em um R^2 total de 0,634.

Já o grupo 3 teve uma severa discrepância entre os resultados de treino e teste. No treino, ele conseguiu se moldar a todos os dados de referência, atingindo um coeficiente perfeito e igual a 1; no entanto, esse aumento em sua capacidade de previsão dos dados de treino não se traduziu em uma melhora nos dados de teste, que viram seu indicador decrescer ainda mais em relação ao grupo anterior. Quanto à capacidade de previsão, ele conseguiu explicar mais de 40 pessoas no treino, mas teve um resultado também pior em relação ao grupo 2, passando apenas ligeiramente de 15 pessoas. O seu representante com melhor desempenho é encontrado na Figura 4.17, o qual possui R^2 de teste igual a 0,122 e R^2 de treino igual a 1,000, resultando em um R^2 total de 0,926.

Além destes melhores resultados, identificaram-se também dois outros grupos de modelos que não tiveram um desempenho satisfatório. Estes são os grupos denominados de sobreajuste, e o dos insensíveis. O grupo de sobreajuste se caracterizou por um ajuste perfeito aos dados de treino, com R^2 de treino igual a 1, mas cujo resultado no teste resultou em um coeficiente negativo, que se traduz em um desempenho pior que assumir que os passageiros se mantêm constantes e iguais à média histórica ao longo do tempo. Similarmente, o grupo dos insensíveis teve um desempenho insatisfatório tanto no teste quanto no treino, caracterizando-se por R^2 negativo em ambos, pois o modelo de demanda se comportou como uma constante; o pior de tudo é que o modelo não conseguiu nem acertar a média histórica, pois, fazendo isso, teria conseguido um desempenho zero, que é melhor do que o negativo. Um exemplar do grupo de sobreajuste se encontra na Figura 4.18 e um exemplar do grupo dos insensíveis se encontra na Figura 4.19.

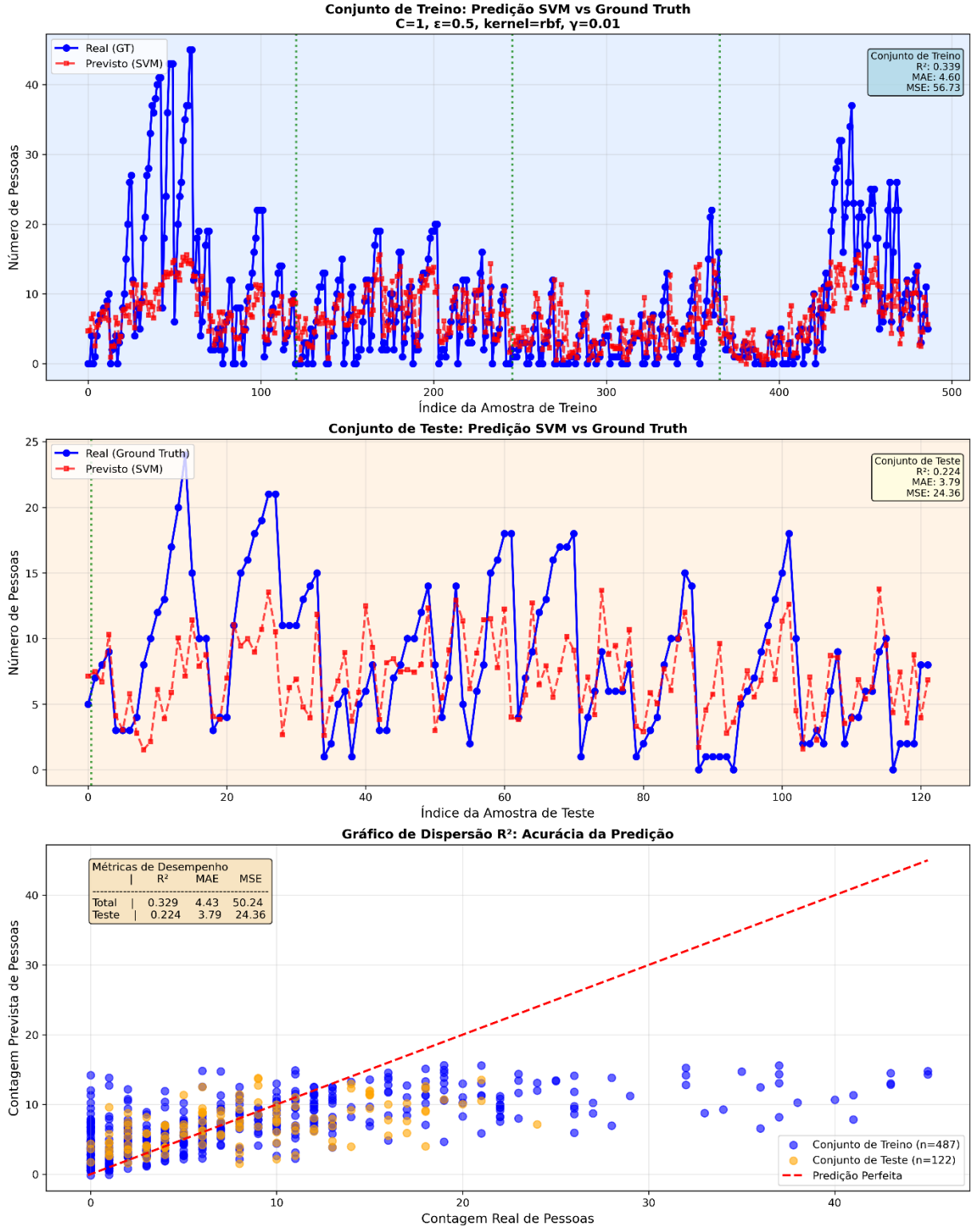


Figura 29: Análise da previsão do modelo SVM, enquadrada como pertencendo ao grupo 1.

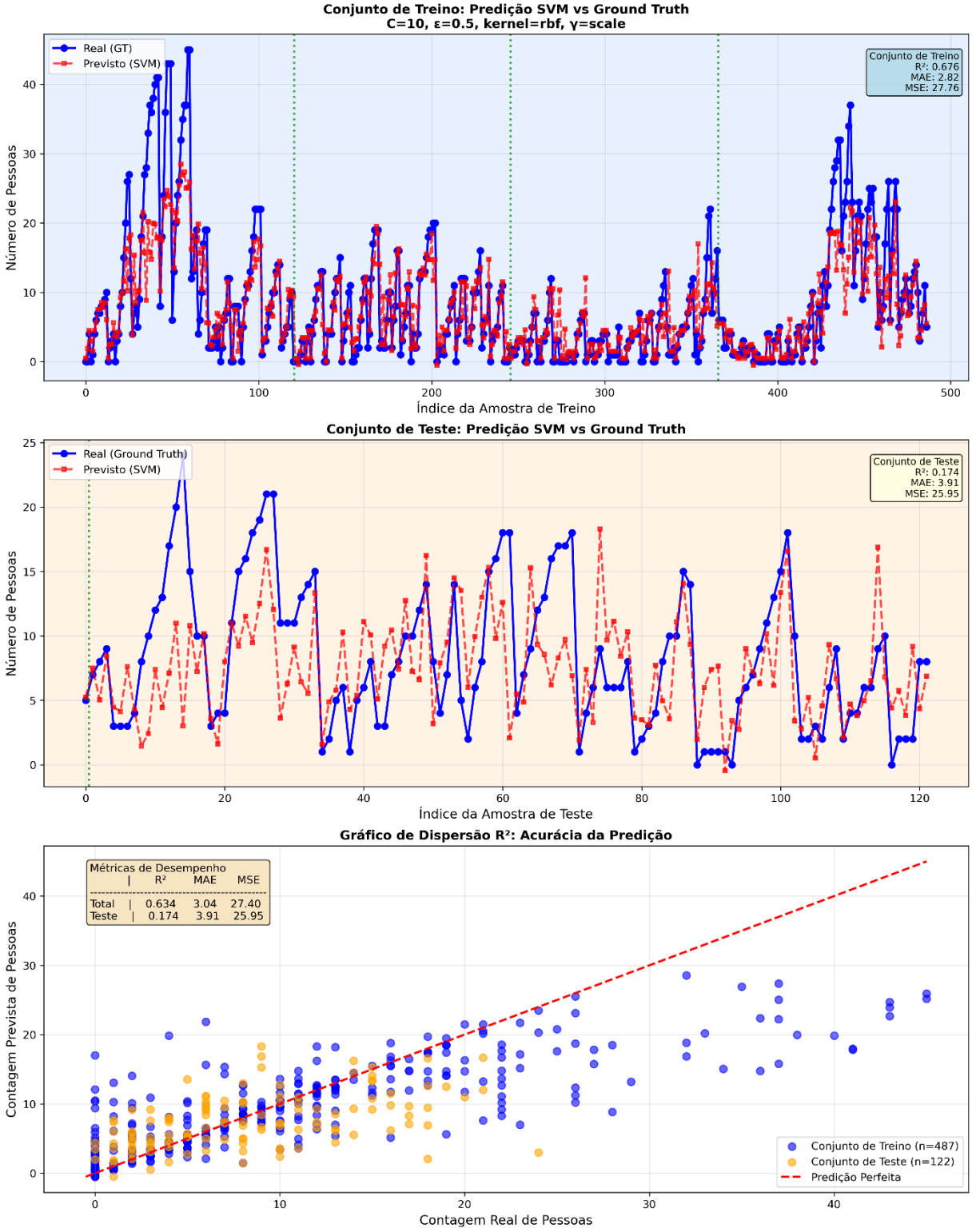


Figura 30: Análise da previsão do modelo SVM, enquadrada como pertencendo ao grupo 2.

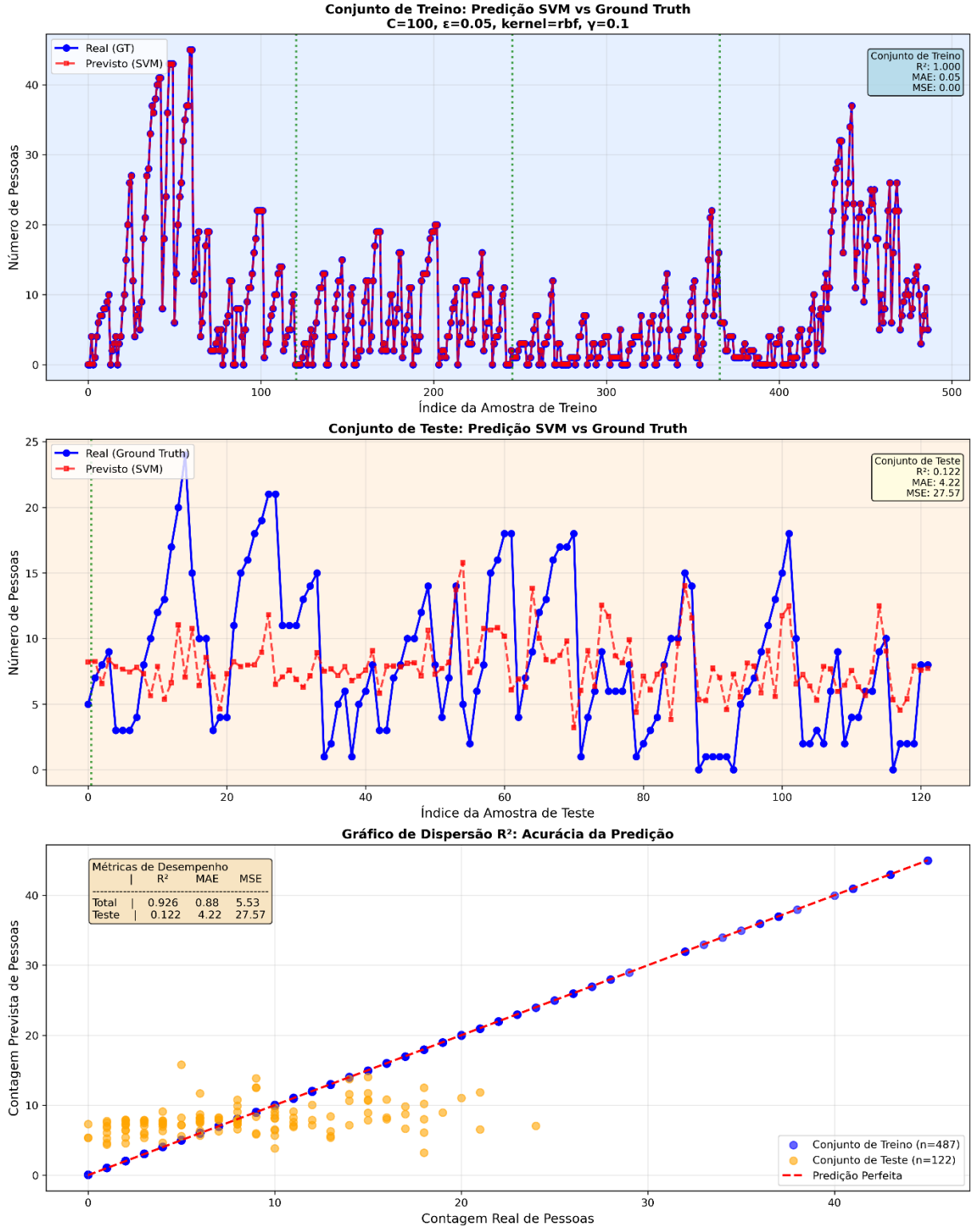


Figura 31: Análise da previsão do modelo SVM, enquadrada como pertencendo ao grupo 3.

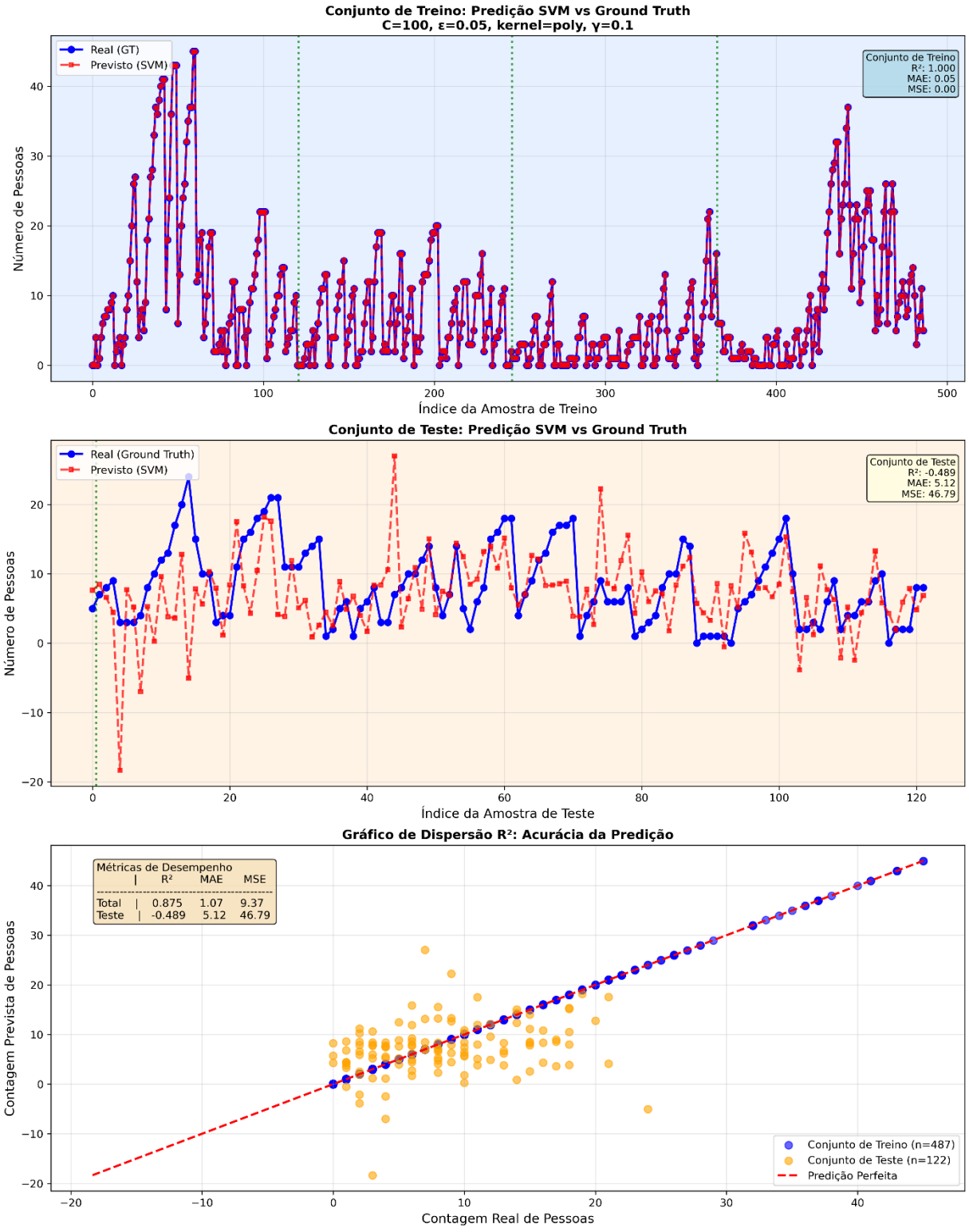


Figura 32: Análise da previsão do modelo SVM, enquadrada como pertencendo ao grupo do sobreajuste.

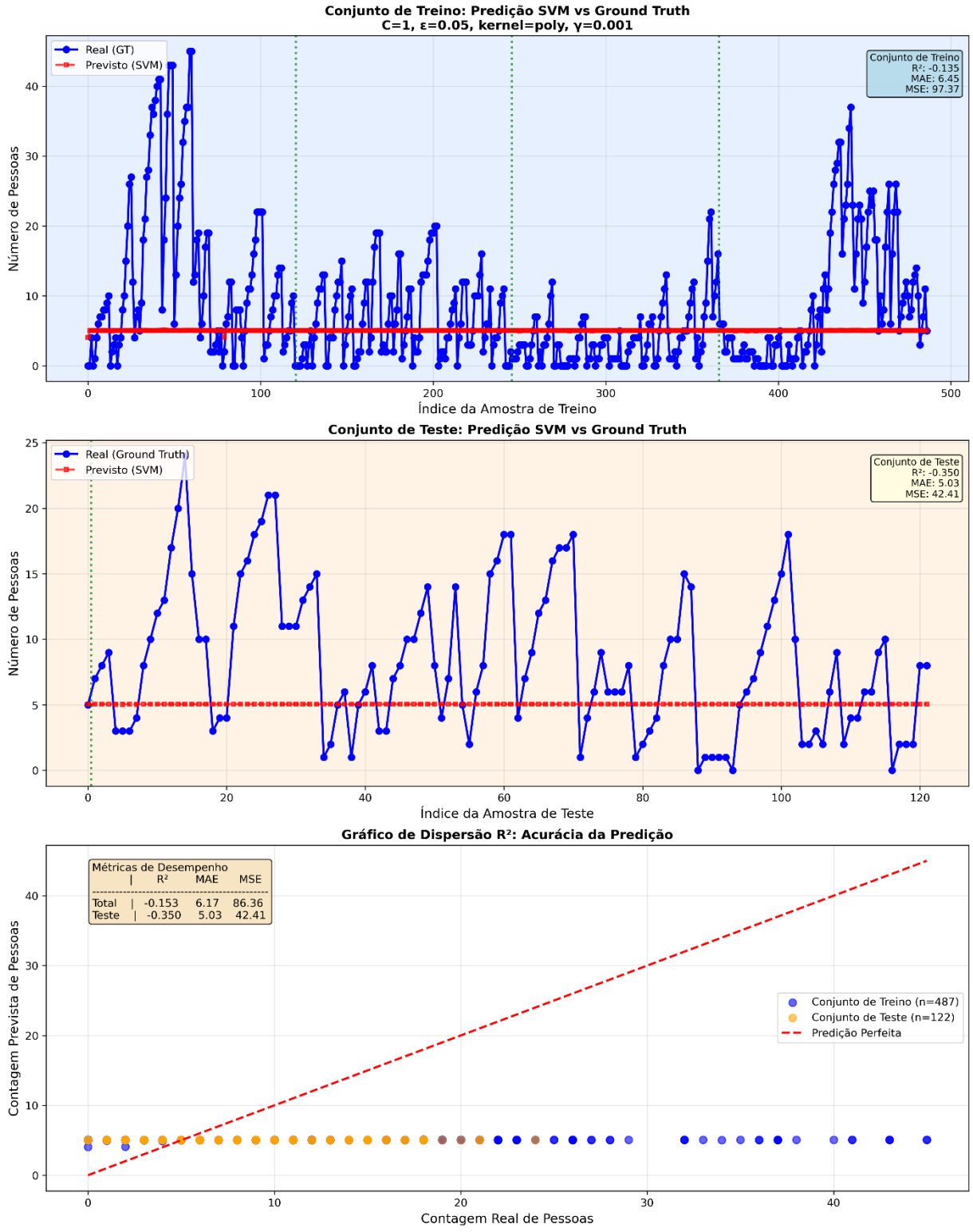


Figura 33: Análise da previsão do modelo SVM, enquadrada como pertencendo ao grupo dos insensíveis.

Portanto, deixando de lado os resultados insatisfatórios, a questão se torna qual dos três grupos escolher como a melhor opção. A resposta racional é o grupo 1, por motivos que são

apresentados em seguida, mas essa resposta a princípio não é óbvia, porque inicialmente o grupo 3 se mostra tentador para ser aceito, por seu treino perfeito (e consequentes elevadas métricas), e seu razoável desempenho no teste. Contudo, o valor mensurável de desempenho atingido pelo grupo 1, simbolizado por seu coeficiente R^2 de teste, não é em vão: vê-se sua maior sensibilidade aos padrões de dados de entrada especialmente no intervalo entre os índices de amostra de teste 20 e 40, nos quais o modelo do grupo 1 alcançou uma maior amplitude de valores e que conseguiu imitar de forma mais fiel o comportamento do gráfico acima dele.

Neste momento, é propício discutir a representatividade física da diferença entre os resultados dos grupos. A Figura 4.20 mostra o estado da parada de ônibus no instante em que o maior número observado de passageiros simultâneos foi registrado. Pode-se ver pela fotografia que a distância entre o sensor e a última pessoa da fila de mais de 40 pessoas (algumas encobertas pelo ônibus) é grande; de fato, ela foi medida em 25 m, que é bem acima da suposição inicial de um alcance efetivo de 6 m para sinais fortes, conforme a seção 3.6.3. Portanto, é justificável, e até esperável, que o modelo não consiga computar sondas vindas de tal distância, pois elas fisicamente chegam no sensor com um RSSI baixo, o que as torna fáceis de serem confundidas com ruído ambiental.



Figura 34: Vista lateral da parada de ônibus às 11h55 do dia 11 de novembro de 2025, em sua observação mais demandada das amostras do conjunto de dados.

4.3 DISCUSSÃO

Conforme adiantado na seção 2.3 do Referencial Teórico, que caracteriza a tecnologia dos sensores passivos, a aleatorização do endereço MAC traz desafios para a contagem e rastreamento dos usuários. Uma vez que cada dispositivo tem um intervalo diferente para a troca do endereço, embora geralmente agrupado por modelo ou fabricante, e também que cada

dispositivo troca de endereço em um instante diferente, em larga medida pelo tempo em que já está ligado, torna-se funcionalmente inviável tentar compensar esse efeito de aleatorização dos MACs ao prever um tempo médio de troca de endereço. Dessa maneira, o trabalho se orienta muito mais pela densidade de dispositivos, tentando estimar a ocupação da área de estudo, do que com o rastreamento individual, seja espacialmente como uma pesquisa de Origem e Destino, quanto temporalmente, vendo as mudanças desse indivíduo no decorrer dos dias.

Percebeu-se ao longo da aplicação do método a influência da distância do sensor aos dispositivos na qualidade dos dados de entrada, que em torno geram resultados mais assertivos. Inferiu-se que a colocação do dispositivo sensor camuflado em um local acima da linha de visão das pessoas melhorou a captação dos dados, conforme antecipado por Oberli *et al.* (2010) e Minea (2022), e impediu a mudança de comportamento devido ao estranhamento de um objeto estranho. Foram notados endereços MAC repetidos entre algumas coletas de dados, mas não foi possível concluir o que eles representam fisicamente. Percebeu-se uma diferença no perfil de intensidade de sinal entre coletas de dados, também sem explicação aparente, uma vez que o sensor era sempre colocado no mesmo local.

Constatou-se a ineficácia do uso apenas dos sinais fortes para a criação de um modelo de demanda, apontando uma correlação fraca entre esses dados, algo que é explicado visualmente pela Figura 4.13, na qual não há um agrupamento claro dos pontos ao redor de uma reta. Finalmente, estudou-se quais seriam os melhores valores para a definição dos hiperparâmetros do modelo, identificando comportamentos similares entre os resultados e agrupando-os entre insatisfatórios e satisfatórios, e, dentro destes, grupos de maior e menor ajuste aos dados de teste. Percebeu-se a tendência de que o aumento do R^2 de treino era proporcionalmente relacionado a um decréscimo do R^2 de teste, de modo que hiperparâmetros que permitiam um ajuste fino dos dados de treino aos dados de referência tinham um desempenho pior nos dados de teste do que aqueles modelos que tinham um menor ajuste aos dados de treino.

Comparando os resultados com a literatura estudada, especialmente a partir dos artigos selecionados que foram considerados os mais relevantes, verifica-se que as curvas geradas após modelagem dos dados são comparáveis às obtidas por Roriz Junior *et al*. (2022), que, embora com um menor número total de passageiros, também obteve um resultado que se aproximava dos dados de referência, especialmente com um ligeiro atraso entre a entrada do passageiro e sua detecção pelo sensor. O resultado de Santos e Almeida (2021) também foi semelhante, embora com menos dados coletados.

Com o modelo proposto, observou-se um resultado melhor do que Oransirikul *et al*. (2016), que usou apenas filtragem de sondas, embora a tecnologia de captação delas seja similar. No entanto, Xu *et al*. (2025) teve um erro menor e maior correlação entre os dados captados e os de referência utilizando sensores Bluetooth, algo que pode ser devido ao tipo de dado captado, que pode ser ambientalmente mais profícuo com o avanço desta tecnologia, ou ao uso de um modelo diferente para agrupamento dos dados. Madsen *et al*. (2022) também teve um resultado melhor, mas com uma abordagem diferente, que usava mais de um sensor e outro modelo de análise.

Finalmente, é necessário indicar algumas dificuldades encontradas na aplicação do método. Em sua concepção, pensava-se que os ônibus teriam um *headway* de cerca de 20 minutos entre passagens, de modo que seria possível os dispositivos emitirem sondas suficientes para serem percebidos pelo sensor, e diagnosticados como passageiros (mas não tempo suficiente para serem considerados ruído estático). Contudo, não foi este o caso encontrado em campo; para benefício dos usuários e prejuízo da realização da pesquisa, os ônibus passavam a cada 5 ou 10 minutos durante os horários de pesquisa, o que possivelmente tenha prejudicado os dados da coleta e inviabilizado certas análises.

Além disso, os resultados provêm de um único ponto de ônibus, dotado de características geométricas e ambientais específicas. A presença do toldo na proximidade do ponto de parada, que constitui uma facilidade que traz conforto aos usuários em espera, pode induzir uma geometria específica de fila, que não é comum em pontos de ônibus sem esse toldo. Essa característica do local estudado configura um cenário desafiador, já que os passageiros se distribuem por distâncias maiores e com menor densidade espacial, diferente dos pontos de ônibus tradicionais, nos quais os passageiros tipicamente se aglomeram nas áreas de espera, o que é mais favorável à detecção pelo sensor. Portanto, sua validação nesta configuração mais desafiadora sugere uma boa generalização. Contudo, cada ponto de parada pode ser diferente dos demais em sua morfologia urbana e comportamento de usuários, sendo necessária uma validação futura em múltiplos locais para aferir o desempenho do sistema frente a essas diferentes configurações de fila de espera.

Outro aspecto limitante da pesquisa foi a incapacidade de predição do número total de pessoas na parada nos horários de pico, algo que se interpretou ser tanto devido a aspectos físicos de perda de potência na propagação do sinal até chegar ao sensor, quanto pela capacidade do modelo em diferenciar ruído do sinal verdadeiro, conforme antecipado por Minea (2022).

5 CONCLUSÃO

Novas tecnologias trazem oportunidades, dificuldades, e desafios. Sistemas dinâmicos são uma das tecnologias ITS que prometem grandes benefícios aos atores envolvidos, tanto no aspecto de redução de custos quanto aumento de eficiência operacional e qualidade do serviço oferecido aos passageiros. Contudo, eles precisam de uma interface de entrada para sentir o mundo físico com o qual estão se relacionando; para isso, sensores são essenciais. No entanto, os sensores também podem ser usados mesmo sem sistemas dinâmicos, para Planejamento de Transportes.

O método de coleta de dados de demanda usando sensores passivos que são baseados em dispositivo sem imagem é, conforme traz a literatura sobre o tema, capaz de proezas que de outra forma seriam inimagináveis. Especialmente se destacam o conhecimento de ocupação dos pontos de parada e veículos, que constituem os nós e as ligações do sistema de Transporte Público Urbano, cuja atualização pode ser feita de forma automatizada, com baixo custo de implantação, e respeitando a privacidade dos passageiros. Essa informação auxilia tanto na operação quanto no planejamento do sistema, que atualmente é feito com pesquisas intrinsecamente pontuais e localizadas, em contraste com pesquisas que poderiam ser feitas de modo contínuo e integrado usando essa nova tecnologia, que permite uma gestão baseada em dados.

Todos os sensores discutidos na revisão de literatura possuem viabilidade para serem usados na presente pesquisa, embora produzam dados de tipos diferentes, com níveis de qualidade que podem ou não variar em função de aspectos ambientais, e que necessitam de ferramentas de interpretação diferentes. Nesse sentido, os sensores passivos foram apontados como uma alternativa promissora para medir a ocupação dos pontos de parada de ônibus. A partir dessa definição, iniciou-se o processo de construção e teste do equipamento sensor, coleta de dados de campo, e análise dos resultados usando uma técnica simples mas validada pela literatura a fim de provar a utilidade do equipamento para coleta de dados de planejamento de transportes.

5.1 DISCUSSÃO DOS RESULTADOS

Foram realizados testes iniciais que validaram o equipamento para as coletas de dados em campo, e orientaram melhores práticas para a localização do conjunto sensor. Especialmente nos testes, verificou-se a importância de minimizar a distância entre os sensores e os objetos de pesquisa, e também a recomendação de Oberli *et al.* (2010) e Minea (2022) a respeito do posicionamento dos sensores em uma posição com linha de visão clara e desobstruída. Foram realizadas mais de dez horas de coleta de dados em campo, cujos resultados foram analisados

em conjunto, usando as mesmas configurações do modelo para assegurar que a generalização dos parâmetros do modelo fosse verificada.

A distribuição estatística dos dados de intensidade relativa de sinal mostra um histograma assimétrico à direita, com muito mais sinais fracos do que fortes e com alguns intervalos de classe mais frequentes do que outros, o que pode representar dispositivos fixos emitindo seus pacotes constantemente, ou dispositivos de transeuntes que emitiram muitos pacotes de uma só vez. Ao mesmo tempo, a análise UpSet revelou que a ampla maioria dos pacotes registrados como sondas veio de endereços MAC únicos, que não se repetiram em nenhum outro período de coleta; isso reforça o argumento de que a aleatorização do endereço MAC por parte dos dispositivos não pode ser ignorada. Finalizando a análise estatística descritiva, os histogramas temporais por categoria de RSSI revelam que houve períodos de maior e menor movimento, bem como maior e menor proporção de sondas fortes, mesmo sem alteração no posicionamento do conjunto sensor; não foi possível identificar a causa-raiz para tal comportamento.

Em seguida, realizou-se um estudo para avaliar a correlação entre os sinais fortes percebidos pelo sensor e o número efetivo de pessoas na parada de ônibus naquele minuto. Além da falta de correlação clara tanto visualmente quanto na análise de regressão linear, observou-se também uma falta de bijeção extrema, onde um mesmo dado de sinal forte pode significar nenhuma pessoa ou mais de 40 pessoas, e similarmente um mesmo número de pessoas pode ser representado por nenhum sinal forte, ou 40 sinais fortes. Observou-se também que não havia uma correlação monotônica entre os sinais fortes e o número de passageiros na parada de ônibus, já que por vezes havia mais sinais fortes do que passageiros, e por outras menos sinais que passageiros.

Para resolver este problema já antecipado pela literatura, adotou-se um dentre os modelos de interpretação de dados validados por artigos científicos selecionados, denominado Máquina de Vetores de Suporte (SVM), categorizado como aprendizado de máquina supervisionado, e que permite a otimização de parâmetros a fim de aproximar as predições dos dados de referência observados no mundo real, que foram registrados manualmente em simultâneo aos dados do sensor. Dessa maneira, procedeu-se a um breve ajuste das configurações do modelo, chamadas de hiperparâmetros, discutindo então os cinco grupos de comportamento encontrados, em função do ajuste das predições à curva real observada. Três grupos tiveram resultados satisfatórios, com discrepâncias crescentes entre os coeficientes de determinação de treino e de teste à medida em que o ajuste dos dados de treino era permitido, chegando ao caso extremo do quarto grupo, em que o sobreajuste perfeito do treino impediu a generalização para os casos de

teste. Um quinto grupo foi insensível aos dados de entrada, resultando em uma predição invariável e não representativa dos dados reais.

A arquitetura do sistema proposto oferece flexibilidade para aplicações além do transporte rodoviário por ônibus. Devido à natureza passiva e compacta do conjunto sensor, ele pode ser instalado em diversas infraestruturas de transporte, como estações de trem, plataformas de metrô ou terminais de integração. A capacidade de detecção de dispositivos móveis independe do modo de transporte, permitindo que o mesmo sistema seja adaptado para monitorar a demanda em diferentes cenários de mobilidade urbana.

Concluiu-se, baseado na análise do modelo, que o sensor é efetivo para captar os dados de pacotes Wi-Fi de dispositivos de passageiros localizados a até cerca de 6 metros de distância do equipamento, sendo portanto útil na análise de pontos de parada de ônibus urbanos.

5.2 LIMITAÇÕES

Por mais promissores que sejam estes sensores, eles também têm desafios e limitações próprias, como por exemplo a aleatorização dos endereços MAC dos dispositivos, a correlação entre o número de dispositivos e o número de pessoas, e o alcance de captura de sinais. Adotaram-se estratégias para contornar estes dois primeiros desafios, e tentar melhor descrever esta limitação; a saber, já preparando o modelo para a presença de dispositivos com endereço aleatorizado, bem como adotando um modelo que pudesse correlacionar os dados captados com os dados de referência por meio de aprendizado de máquina supervisionado. Sobre a limitação de alcance do sensor, estudou-se a correlação entre a acurácia de predição do modelo com o número de passageiros da parada, resultando em um número aproximado de passageiros em fila que podem ser contados, bem como uma correlação com a distância para tal.

Portanto, o modelo tal como se encontra neste trabalho é incapaz de rastrear o mesmo usuário ao longo do tempo, o que seria ótimo para fins de planejamento de transportes e matrizes Origem/Destino mas que ao mesmo tempo levantam preocupações de privacidade. Então a tecnologia de endereços MAC aleatorizados permite somente a mensuração de demanda agregada instantânea no ponto de parada, ainda que possa ser registrada ao longo do tempo.

Além disso, notaram-se dois fatores como potencialmente indutores de erros, ruído ou dificuldades de identificação de dispositivos: a passagem de transeuntes por dentro da parada de ônibus, e o tempo entre passagens consecutivas de ônibus na parada ser menor que o previsto.

A primeira questão se refere à captação de pacotes com alta intensidade de sinal, oriundos de dispositivos que muitas vezes podem inclusive estar fisicamente ao lado de um passageiro, mas que não deveriam ser contados como tal. Esta é uma diferença deste trabalho para outros que medem o número de passageiros já dentro do veículo; um escopo diferente que não tem esse problema em específico, e que portanto pode usar modelos binários de decisão. Já quanto à passagem frequente de ônibus, isto por um lado é bom para o usuário, mas por outro pode impedir que um dispositivo seja detectado pelo sensor, reduzindo a efetividade da coleta de dados.

Outro aspecto que pode ter introduzido viés à pesquisa é a presença da rede Wi-Fi gratuita da UnB, que poderia induzir os usuários a deixarem a comunicação Wi-Fi ligada em seus telefones, algo que em outras ocasiões eles não fariam. Contudo, uma pequena amostra não-representativa mostrou que nem sempre este é o caso; mesmo com a rede gratuita, alguns usuários permanecem com a função Wi-Fi desligada em seus telefones.

Finalmente, há também intrínseca ao método a dificuldade de saber o instante exato de partida do dispositivo; isto é, o momento em que o usuário entrou no ônibus e não deve mais ser contado na parada. O aspecto mais relevante desse problema é relativo à diferença conceitual e prática entre o último avistamento, que é o instante em que o sensor capturou o último pacote oriundo do dispositivo em questão, e o momento de fato em que ele partiu, já que o envio de sondas é periódico, e não contínuo.

5.3 TRABALHOS FUTUROS

As investigações sobre o monitoramento passivo por Wi-Fi para fins de planejamento de transportes deste trabalho demonstram apenas parte do potencial da ferramenta estudada. Além das oportunidades de melhoria instrumental, serão também sugeridos trabalhos futuros no âmbito de melhor descrever o comportamento do usuário e superação de obstáculos encontrados. Propõem-se algumas ideias principais, divididas em quatro temas: melhoria do sensor, comparação de modelos alternativos, análises suplementares, e investigações longitudinais.

Quanto ao conjunto sensor, aqui compreendido tanto pela parte física quanto programática, há sempre a alternativa de tentar baixar ainda mais o custo do sensor, usando microcontroladores mais compactos e menos potentes, mas com graus similares de aquisição de sondas. Contudo,

a principal recomendação é relativa à integração de sensores, conforme destacado por Madsen *et al*. (2022): investigar usar dois ou mais sensores diferentes em paralelo; apesar do aumento de custo intrínseco à integração de outras tecnologias, isso geraria um equipamento com ortogonalidade de tipos de sensores e possível redução da influência dos ruídos ambientais. Em especial, destaca-se a integração com o sensor FLIR proposto por Lin e Zhao (2025).

É possível também integrar sensores do mesmo tipo, formando uma rede de modo a cobrir uma área maior, como em estações de integração, ou ver se há alguma outra alternativa de baixo custo que se destaca, como o Bluetooth, haja vista o aumento do uso de dispositivos deste tipo no cotidiano, ou incorporar outras informações aos dados de entrada, usando Séries Temporais Multivariadas, que contêm variáveis primárias ou derivadas, como por exemplo temperatura, e pressão atmosférica para avaliar se está ou não chovendo.

Após a aquisição de dados, eles são processados por um modelo para gerar um resultado. Portanto, estudar modelos alternativos é de suma importância. Existem alguns caminhos possíveis, principalmente relacionados às Redes Neurais, sejam elas Recorrentes (RNNs), de Memória de Longo e Curto Prazo (LSTMs), ou Convolucionais (CNNs). Estas podem ser superiores ao modelo estático aqui utilizado por poderem interpretar interdependências temporais entre os dados.

Também pode-se pensar em usar estratégias de otimização diferentes em oposição à *Grid Search*, como Algoritmos Genéticos (GAs), ou Otimização por Enxame de Partículas (PSOs). Estas provavelmente convergirão mais rapidamente em soluções ótimas das configurações do modelo selecionado, seja o SVM ou outro.

Quanto a possíveis melhorias na análise dos dados usando o mesmo modelo, o caminho mais óbvio é incrementar a análise por SVM aplicando o método *GridSearchCV* nos dados, para garantir que de fato o arranjo ótimo de hiperparâmetros do modelo foi encontrado; algo que foi no presente trabalho apenas procurado entre opções presumidas. Também é possível minimizar os efeitos da aleatorização MAC, usando o parâmetro de SSID procurado, registrado pelo sensor, para correlacionar MACs diferentes mas que procuram o mesmo sinal, um indicativo que estes endereços poderiam de fato se referir ao mesmo dispositivo, caso seja identificado o término de transmissões de um endereço e início de outro.

Também seria possível trabalhar com dados desagregados usando o modelo SVM, da mesma forma que Roriz Junior *et al*. (2022). Contudo, para isso seria necessário acrescentar ao filtro usado pelos autores algum parâmetro que conseguisse identificar se o dispositivo identificado

pertenceria a um transeunte ou a um passageiro efetivo, tal como uma medição de tempo de permanência antes do registro como passageiro. Caso tal procedimento fosse bem-sucedido nos testes, seria possível medir o tempo de permanência de cada passageiro na parada, bem como a sua posição física aproximada em relação ao sensor, usando a correlação entre distância e RSSI apontada por Minea (2022).

Finalmente, pode-se incrementar o conhecimento sobre o tema utilizando o conjunto sensor em aplicações de medição contínua do fluxo de passageiros, por um intervalo de tempo maior do que as duas horas ininterruptas neste trabalho utilizado. Contudo, para isso seriam necessários uma bateria de maior capacidade, um sensor mais eficiente no uso de energia, ou uma infraestrutura com ponto de energia dedicado.

Além disso, pode-se investigar possíveis formas de transmissão de dados, seja por redes sem fio de curto alcance ou de telefonia móvel, a fim de implementar o monitoramento em tempo real da demanda, para uma integração com os sistemas de transporte público e início dos testes reais de sistemas dinâmicos verdadeiramente baseados na demanda real do ponto de parada, que é o estado da arte há muito imaginado.

REFERÊNCIAS BIBLIOGRÁFICAS

ASAKURA, Yasuo; HATO, Eiji. Behavioral monitoring of public transport users through a mobile communication system. **Journal of Advanced Transportation**, v. 35, n. 3, p. 289–304, 2001.

BELL, Daniel Steven; JAMES, Philip; LÓPEZ-GARCÍA, Martín. Social Distance Approximation on Public Transport Using Stereo Depth Camera and Passenger Pose Estimation. **Sensors**, v. 23, n. 24, 1 dez. 2023.

CEDER, Avishai. **Public Transit Planning and Operation**. *[S.l.]*: Elsevier Ltd., 2007.

CHEN, Ting Zhao; CHEN, Yan Yan; LAI, Jian Hui. Estimating Bus Cross-Sectional Flow Based on Machine Learning Algorithm Combined with Wi-Fi Probe Technology. **Sensors (Switzerland)**, v. 21, n. 3, p. 1–16, 1 fev. 2021.

DANIELIS, Peter; KOUYOUMDJIEVA, Sylvia T.; KARLSSON, Gunnar. UrbanCount: Mobile Crowd Counting in Urban Environments. *In*: IEEE, 2017.

FERRAZ, Antonio Clóvis Pinto *et al.* **Transporte Público Coletivo Urbano**. São Carlos, SP: RiMa Editorial, 2023.

GONZÁLEZ, Ana Belén Rodríguez *et al.* Characterization of COVID-19's impact on mobility and short-term prediction of public transport demand in a mid-size city in Spain. **Sensors**, v. 21, n. 19, 1 out. 2021.

HSU, Ya Wen; CHEN, Yen Wei; PERNG, Jau Woei. Estimation of the number of passengers in a bus using deep learning. **Sensors (Switzerland)**, v. 20, n. 8, 2 abr. 2020.

IEEE Standard for Information Technology--Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks--Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 1: Operation with Randomized and Changing MAC Addresses. Piscataway, NJ, USAIEEE, 26 set. 2024.

IEEE STANDARDS ASSOCIATION. **The Evolution of Wi-Fi Technology and Standards**. Disponível em: <https://standards.ieee.org/beyond-standards/the-evolution-of-wi-fi-technology-and-standards/>. Acesso em: 10 jul. 2025.

LIN, Yukai; ZHAO, Qiangfu. Human Occupancy Monitoring and Positioning with Speed-Responsive Adaptive Sliding Window Using an Infrared Thermal Array Sensor. **Sensors**, v. 25, n. 1, 1 jan. 2025.

LITMAN, Todd. **Evaluating Public Transportation Health Benefits**. Canadá: Victoria Transport Policy Institute. Disponível em: <https://www.vtpi.org/tran_health.pdf>. Acesso em: 30 jun. 2025.

MADSEN, Tatiana *et al.* Comparison of WLAN Probe and Light Sensor-Based Estimators of Bus Occupancy Using Live Deployment Data. **Sensors**, v. 22, n. 11, 1 jun. 2022.

MAGRO, Diogo Dal; FORTES, Vinícius Borges. O reconhecimento facial nas smart cities e a garantia dos direitos à privacidade e à proteção dos dados pessoais. **Revista de Direito Internacional**, v. 2, n. 18, 2021.

MANDAL, Ratna *et al.* Exploiting Multi-modal Contextual Sensing for City-bus's Stay Location Characterization: Towards Sub-60 Seconds Accurate Arrival Time Prediction. **ACM Transactions on Internet of Things**, v. 4, n. 1, 24 maio 2023.

MARTIN, Samuel Seixas; ARRUDA, Fabiana Serra de. Proposição de ferramenta de baixo custo para coleta de dados em faixas de pedestres e identificação de conflitos no trânsito. **30º Congresso de Iniciação Científica da Universidade de Brasília**, 2024.

MINEA, Marius. An Experimental Assessment of People's Location Efficiency Using Low-Energy Communications-Based Movement Tracking. **Sensors**, v. 22, n. 22, 1 nov. 2022.

MOLINA, Oscar Amador *et al.* Protection of Vulnerable Road Users using Hybrid Vehicular Networks. *In*: Institute of Electrical and Electronics Engineers Inc., 2022.

MURDAN, Anshu Prakash *et al.* Low-cost bus seating information technology system. **IET Intelligent Transport Systems**, v. 14, n. 10, p. 1303–1310, 1 out. 2020.

NUNES, EDSON HILIOS MARQUES. **Estimativa da ocupação utilizando sensoriamento Wi-Fi de telefones móveis: Uma aplicação no transporte público urbano de passageiros por ônibus**. Dissertação (Mestrado)—São Paulo: Escola Politécnica da Universidade de São Paulo, 2019.

NTU. **Anuário NTU 2023-2024**. Brasília: Associação Nacional das Empresas de Transportes Urbanos. Disponível em: <https://ntu.org.br/novo/upload/Publicacao/Pub638573500081945042.pdf>. Acesso em: 17 jul. 2025.

NTU. **Os Grandes Números da Mobilidade Urbana**. Brasília: Associação Nacional das Empresas de Transportes Urbanos. Disponível em: <https://www.ntu.org.br/novo/AreasInternas.aspx?idArea=7>. Acesso em: 15 jul. 2025.

OBERLI, Christian; TORRES-TORRITI, Miguel; LANDAU, Dan. Performance evaluation of UHF RFID technologies for real-time passenger recognition in intelligent public transportation systems. **IEEE Transactions on Intelligent Transportation Systems**, v. 11, n. 3, p. 748–753, set. 2010.

ORANSIRIKUL, Thongtat *et al.* Feasibility of analyzing Wi-Fi activity to estimate transit passenger population. *In*: Institute of Electrical and Electronics Engineers Inc., 19 maio 2016.

PARADEDA, Diego Benites; KRAUS JUNIOR, Werner; CARLSON, Rodrigo Castelan. Bus passenger counts using Wi-Fi signals: some cautionary findings. **TRANSPORTES**, v. 27, n. 3, p. 115–130, 13 nov. 2019.

PEREIRA, Rafael H. M. *et al.* **Texto para discussão 2673 - Tendências e desigualdades da mobilidade urbana no Brasil I: o uso do transporte coletivo e individualTexto para Discussão**. Rio de Janeiro: Instituto de Pesquisa Economica Aplicada - IPEA, jul. 2021.

RORIZ JUNIOR, Marcos Paulino; MEDRANO, Ronny Marcelo Aliaga; ALMEIDA, Cristiano Farias. UAI-FI: using artificial intelligence for automatic passenger counting through Wi-Fi and GPS data. **TRANSPORTES**, v. 30, n. 2, p. 2555, 6 ago. 2022.

SANTOS, Caio Vinicius Castro dos; ALMEIDA, Madson Cortes de. **Sistema para Contagem de Passageiros e Descrição de Microrrotas em Transporte Coletivo**. Campinas, SP: *[S.n.]*, 2021. Disponível em: <https://proceedings.science/p/137302 >.

SPAETH, Utz *et al.* Potential of Pressure Sensor Based Mass Estimation Methods for Electric Buses. **Electronics**, v. 9, n. 5, p. 711, 26 abr. 2020.

SUN, Wenzhe; SCHMÖCKER, Jan-Dirk; FUKUDA, Koji. Estimating the route-level passenger demand profile from bus dwell times. **Transportation Research Part C: Emerging Technologies**, v. 130, set. 2021.

TECA, Geovani; NATKANIEC, Marek. A Novel Covert Channel for IEEE 802.11 Networks Utilizing MAC Address Randomization. **Applied Sciences**, v. 13, n. 14, p. 8000, 8 jul. 2023.

TSIKTSIRIS, Dimitris *et al.* A complete in-cabin monitoring framework for autonomous vehicles in public transportation. **IET Intelligent Transport Systems**, v. 19, n. 1, 1 jan. 2025.

UNIVERSIDADE DE BRASÍLIA. Relatório Síntese da Consulta Pública: Avaliação da Qualidade do Transporte Público e dos Pontos de Parada do Campus Darcy Ribeiro. Brasília, DF, mar. 2023.

VIEIRA, Túlio *et al.* Public Transport Occupancy Estimation using WLAN Probing and Mathematical Modeling. *In*: Elsevier B.V., 1 jan. 2020.

VUCHIC, Vukan. **Urban Transit Operations, Planning, and Economics**. Hoboken New Jersey: John Wiley & Sons, Inc., 2005.

WU, Pan *et al.* Recognizing Real-Time Transfer Patterns between Metro and Bus Systems Based on Spatial–Temporal Constraints. **Journal of Transportation Engineering, Part A: Systems**, v. 148, n. 9, set. 2022.

XU, Junming *et al.* Leveraging Bluetooth and GPS Sensors for Route-Level Passenger Origin–Destination Flow Estimation. **Sensors**, v. 25, n. 8, 1 abr. 2025.

YU, Haibin; HE, Zhiwei; LIU, Jilin. A Vision-Based Method to Estimate Passenger Flow in Bus. *In*: IEEE, 2007.

ZHANG, Cheng *et al.* Estimating Bus Passenger Origin-Destination Flow via Passenger Reidentification Using Video Images. **IEEE Transactions on Intelligent Transportation Systems**, 2025.

APÊNDICE A – Discussões Suplementares sobre o Método

Algoritmo FNV

A respeito do algoritmo FNV (Fowler–Noll–Vo) mencionado no método, importa caracterizá-lo como uma função de dispersão (*hash*) não criptográfica projetada para alta velocidade e baixa taxa de colisão em chaves curtas. Antes de tudo, é importante trazer à luz dois motivos por que decidiu-se implementar esse algoritmo sobre os dados de entrada antes mesmo do salvamento das informações, e outros motivos relativos à escolha do algoritmo em si.

Quanto à implementação ou não do algoritmo, existia durante a prototipagem inicial do equipamento a preocupação de que o nome da rede pudesse ocupar muito espaço dentro da memória do módulo ESP32, quando ainda não havia sido instalado o cartão SD, de modo que foi feito um esforço de tentar compactar ao máximo as informações que seriam salvas no arquivo. Quanto ao espaço ocupado por cada variável, conforme a alocação de memória do código-fonte, o RSSI ocupa 2 bytes, a marca temporal, 4, o endereço MAC, 6, e o SSID em sua forma original, até 32 bytes; como o ganho de redução do SSID era o maior dentre as quatro variáveis, reduzindo de 32 para 4 bytes, a função FNV foi implementada nos valores da variável SSID antes de serem registradas na memória.

Além disso, à época havia a preocupação de se adequar às boas práticas de anonimização dos dados, evitando potenciais vazamentos do nome de redes privadas ventiladas por telefones celulares; mas, por outro lado, o programa armazena endereços MAC físicos que foram enviados pela rede, então essa justificativa não é completamente coerente, e pode adicionar tempo desnecessário ao processo. Isso, no entanto, não é um problema muito grande porque a captura dos pacotes pela antena é desacoplada do processamento pelo microprocessador, que pode armazenar uma fila de pacotes a serem processados, tornando-o robusto a picos de tráfego que não sejam extremos e contínuos devido às limitações físicas do *buffer*.

Quanto aos motivos pela escolha do algoritmo FNV-1a em particular dentre os diversos algoritmos *hash* existentes, destaca-se seu baixo custo computacional, em oposição à alta complexidade de algoritmos criptográficos que também poderiam ser usados para esta mesma função, e também por sua alta dispersão, o que significa que pequenas mudanças no SSID produzem resultados completamente diferentes e inequívocos; algo importante para diferenciar redes que podem ter apenas um prefixo diferente em relação ao uso de 2,4 ou 5 GHz, por

exemplo. Também é importante mencionar a sua diferença para algoritmos que trabalham com tamanho fixo e necessitam de *padding*; como o FNV trabalha byte-a-byte, o resultado é atualizado a cada iteração do dado de entrada e resulta em um custo computacional linear, proporcional apenas ao tamanho do nome da rede, o que é excelente.

Hiperparâmetros do Modelo SVM

O hiperparâmetro C também é conhecido como parâmetro de regularização, e define a tolerância do modelo a erros. Quando o C é baixo, o modelo é menos rígido, aceitando errar alguns pontos de treino para ter uma fronteira de decisão mais suave, o que evita o sobreajuste (*overfitting*), mas pode ter um viés alto, enquanto que um C alto gera um modelo rígido e punitivo, que tenta acertar todos os pontos de treino, criando fronteiras complexas e tortuosas, gerando um risco alto de sobreajuste.

Já o hiperparâmetro Gama define o coeficiente do kernel, caso ele seja do tipo RBF, e controla o alcance da influência de cada ponto do GT no treinamento. Quando o Gama é baixo, o alcance da influência é longo, permitindo a generalização com pontos distantes no cálculo da fronteira de decisão, o que resulta em uma separação mais suave e menos propensa ao sobreajuste mas que pode ter um viés alto, enquanto que um Gama alto restringe a influência de cada ponto a uma área imediata, o que força o modelo a se ajustar a cada detalhe local, gerando um risco alto de sobreajuste ao criar fronteiras fragmentadas e isoladas ao redor dos pontos.

O hiperparâmetro Epsilon define a margem de tolerância (ou a “largura do tubo,” como é popularmente conhecido) dentro da qual não há penalidade para erros na função de perda. Quando o Epsilon é baixo, o modelo é mais estrito, exigindo que as predições estejam extremamente próximas dos valores reais observados, o que aumenta a precisão nos dados de treino, mas pode levar ao sobreajuste por capturar ruídos irrelevantes. Um Epsilon alto, no entanto, cria uma zona de insensibilidade mais larga, permitindo maiores desvios sem punição, resultando em um modelo mais simplificado e com menos vetores de suporte, porém com risco de subajuste (*underfitting*) ao ignorar tendências importantes.

O Kernel determina a função matemática utilizada para projetar os dados de entrada em um espaço dimensional superior, viabilizando a separação das classes. Quando se utiliza um Kernel Linear, o modelo busca uma fronteira de decisão plana no espaço original, sendo uma abordagem conservadora e computacionalmente leve, ideal para dados com separabilidade

óbvia, o que evita sobreajuste, mas falha em capturar padrões complexos. Estes são possíveis de serem identificados com Kernels não-lineares (como RBF ou Polinomial), que permitem que o modelo dobre e distorça o espaço de características, conseguindo separar dados intrincados e sobrepostos, com a consequência de aumentar a complexidade do modelo e gerar um risco alto de sobreajuste se não forem devidamente regularizados. O Kernel polinomial distorce o espaço inteiro de forma rígida e estrutural, enquanto que o Kernel de Base Radial (RBF) cria zonas de influência local ao redor dos aglomerados de dados, como se fossem ilhas independentes. Por simplificação, utilizou-se o Kernel polinomial de grau três.

APÊNDICE B – Lista de Atributos

A base de dados contém 70 colunas, das quais 68 são atributos e duas são colunas auxiliares, a saber: a primeira coluna, referente à janela temporal em intervalos de um minuto, sendo única na base de dados e servindo como chave primária, e a última coluna, que se refere ao instante em que o programa foi executado e aquela janela temporal foi adicionada à base. Este esquema da base de dados com a descrição individual das colunas se encontra ao final deste apêndice.

A seguir, apresentam-se os grupos de atributos com uma breve explicação de cada um, bem como a descrição dos atributos individuais, o que eles representam, em que unidade são medidos, qual o intervalo esperado de valores, e se são separados por categoria de sinal.

Estatísticas Agregadas Básicas

Estes seis atributos são as métricas estatísticas principais dos dados agregados dentro de uma mesma janela de tempo. Contagens mais altas de sondagens e dispositivos únicos indicam maior atividade de dispositivos com Wi-Fi habilitado, o que sugere maior movimento, enquanto que estatísticas de RSSI podem indicar padrões de proximidade, já que dispositivos próximos geralmente geram intensidades relativas de sinal mais fortes.

Tabela 101: Estatísticas agregadas para cada janela de tempo.

Atributo	Descrição	Unidade	Intervalo	Quantidade
probe_count	Número total de requisições de sondagem	-	[0, ∞)	1
unique_devices	Número de endereços MAC únicos	-	[0, ∞)	1
avg_rssi	Intensidade média de sinal	dBm	[-100, 0]	1
max_rssi	Sinal mais forte detectado	dBm	[-100, 0]	1
min_rssi	Sinal mais fraco detectado	dBm	[-100, 0]	1
std_rssi	Desvio-padrão da intensidade de sinal	dBm	[0, ∞)	1

Análise de Sinal

Além das análises globais de intensidade de sinal, foram também descritas estatisticamente as distribuições de sinal para cada uma das três categorias de RSSI. Considerando que foram consideradas seis medidas descritivas para cada uma dessas categorias, este grupo gera 18 atributos, seguindo o padrão {categoria}_{atributo}, onde {categoria} ∈ {strong, medium, weak}. Sinais fortes indicam pedestres nas imediações, enquanto sinais médios/fracos podem representar dispositivos mais distantes ou obstruções de sinal, conforme Minea *et al*. (2020). Estatísticas dentro da categoria podem revelar padrões de agrupamento de sinal e consistência de posicionamento de dispositivos.

Tabela 11. Estatísticas de análise de sinal para cada janela de tempo.

Padrão de Atributo	Descrição	Unidade	Intervalo	Quantidade
{categoria}_signal_count	Número de requisições de sondagem na categoria	-	[0, ∞)	3
{categoria}_signal_macs	Endereços MAC únicos na categoria	-	[0, ∞)	3
{categoria}_rssi_avg	RSSI médio dentro da categoria	dBm	[-100, 0]	3
{categoria}_rssi_max	RSSI máximo dentro da categoria	dBm	[-100, 0]	3
{categoria}_rssi_min	RSSI mínimo dentro da categoria	dBm	[-100, 0]	3
{categoria}_rssi_std	Desvio padrão de RSSI dentro da categoria	dBm	[0, ∞)	3

Detecção de Aleatorização de Endereço MAC

Endereços MAC administrados localmente (com bit 1 do primeiro octeto definido) indicam dispositivos com a nova tecnologia de aleatorização de endereço MAC, focados em privacidade. Para tal, o algoritmo usa como método de deteção a identificação do segundo caractere hexadecimal do endereço MAC.

Consequentemente, todo segundo caractere hexadecimal do endereço MAC será sempre 2, 6, A, ou E, devido à aritmética hexadecimal. Por exemplo, o endereço x2:xx:xx:xx:xx:xx é

randomizado, enquanto o x0:xx:xx:xx:xx:xx é globalmente único (atribuído fisicamente pelo fabricante). Caracteres 3, 7, B, F, embora tenham o segundo caractere igual a 1, ativam também o "primeiro bit" (o bit 0, I/G), que sinaliza que o endereço é um endereço de grupo (*multicast*). Protocolos de rede proíbem que um endereço de grupo seja usado como endereço de origem (*source* MAC), o que faz muitos roteadores e firewalls descartarem o pacote por considerá-lo malformado ou uma tentativa de ataque (*spoofing*).

Cada categoria gera 1 atributo seguindo o padrão {categoria}_randomized_mac_ratio}, onde {categoria} ∈ {strong, medium, weak}. Altas taxas de aleatorização dificultam a contagem de dispositivos, mas fornecem insight sobre comportamento de usuário consciente da privacidade.

Tabela 12. Atributos de aleatorização de endereço MAC para cada janela de tempo.

Padrão de Atributo	Descrição	Unidade	Intervalo	Quantidade
{categoria} _randomized_mac_ratio	Proporção de MACs randomizados na categoria	-	[0, 1]	3

Atributos Temporais

Estatísticas de tempo entre sondas recebidas, seja do mesmo emissor ou de emissores diferentes, medem a distribuição temporal de requisições de sondagem, computadas separadamente para cada categoria de intensidade de sinal. Por exemplo; strong_inter_arrival_avg = 2.5 significa que sinais fortes chegam a cada 2.5 segundos em média. Existem algumas interpretações possíveis para esses indicadores; intervalos longos entre pacotes podem indicar tráfego esparso de pedestres ou usuários móveis, enquanto que alta variabilidade indica padrões comportamentais mistos.

Tabela 13. Atributos de tempo entre sondas para cada janela de tempo.

Padrão de Atributo	Descrição	Unidade	Intervalo	Quantidade
{categoria} _inter_arrival_avg	Tempo médio entre sondagens consecutivas	segundos	[0, ∞)	3
{categoria} _inter_arrival_max	Maior intervalo entre sondagens	segundos	[0, ∞)	3
{categoria} _inter_arrival_min	Menor intervalo entre sondagens	segundos	[0, ∞)	3
{categoria} _inter_arrival_std	Variabilidade de tempo de sondagem	segundos	[0, ∞)	3

Atributos de Movimento

Esses atributos rastreiam mudanças de RSSI ao longo do tempo para endereços MAC individuais, indicando movimento de dispositivo relativo ao sensor. Ou seja, caso seja detectada uma variação de RSSI superior a 5 dBm entre sondagens consecutivas do mesmo endereço MAC, infere-se que ele está em movimento. Disso, interpreta-se que um alto moving_mac_count para determinada classe indica fluxo de transeuntes a uma determinada distância do sensor, enquanto que um baixo valor para este atributo sugere dispositivos estacionários, mesmo que sejam referentes a passageiros (que podem estar esperando muito tempo, estarem sentados, etc.)

Tabela 14. Atributos de movimentação de dispositivos para cada janela de tempo.

Padrão de Atributo	Descrição	Unidade	Intervalo	Quantidade
{categoria} _avg_rssi_delta_per_mac	Mudança média de RSSI através de MACs com 2+ sondagens	dBm	[0, ∞)	3
{categoria} _moving_mac_count	Contagem de MACs com mudança de RSSI > 5 dBm	-	[0, ∞)	3
{categoria} _max_rssi_delta	Maior mudança de RSSI de qualquer MAC	dBm	[0, ∞)	3

Atributos de Fabricante de Dispositivo

São atributos derivados da propriedade OUI (Identificador Único Organizacional), presente nos primeiros 3 bytes de endereços MAC que identificam fabricantes de dispositivos. Os oito atributos deste grupo são divididos em dois atributos gerais, de contagem geral de fabricantes e proporção de principais marcas, e seis atributos por categoria de sinal, sendo dois atributos por categoria de sinal seguindo o padrão {categoria}_{atributo}. As proporções individuais por marca (apple_device_ratio, samsung_device_ratio, motorola_device_ratio) são calculadas durante a extração, mas NÃO são armazenadas no banco de dados. Apenas a major_brands_ratio combinada, dada pela soma das proporções individuais, é gravada.

Tabela 15. Atributos de fabricante de dispositivo para cada janela de tempo.

Padrão de Atributo	Descrição	Unidade	Intervalo	Quantidade
unique_oui_count	Número de fabricantes distintos detectados	-	[0, ∞)	1
major_brands_ratio	Proporção de dispositivos Apple + Samsung + Motorola	-	[0, 1]	1
{categoria} _major_brands_ratio	Proporção de marca principal dentro da categoria	-	[0, 1]	3
{categoria} _oui_entropy	Diversidade de fabricante (entropia de Shannon)	bits	[0, ∞)	3

Foram pesquisados os OUIs públicos mais comuns para cada uma das fabricantes selecionadas, sendo da Apple prefixos comuns como 00:17:F2, 00:1C:B3, etc., da Samsung 00:12:FB, 00:15:B9, etc., e da Motorola 00:04:56, 00:0C:E5, etc.

Para o cálculo de Entropia, foi utilizada a definição que resulta na ordem de grandeza da quantidade de fabricantes de dispositivos que enviaram sondas, dada pela Equação 1, na qual p_i é a proporção de dispositivos do fabricante i. Uma maior entropia indica fabricantes de dispositivos diversos, enquanto que uma menor entropia sugere dispositivos de mesmo fabricante.

$$H = -\sum (p_i \times \log_2 p_i)$$

Equação 1: Cálculo da Entropia de Shannon para a obtenç

Atributos de Hash de SSID

Hashes de SSID rastreiam quais redes WiFi os dispositivos estão procurando, permitindo rastreamento cruzado de MAC e análise de persistência. Isso permite identificar o mesmo dispositivo através de mudanças de endereço MAC, melhorando a precisão da contagem de pedestres.

Tabela 16. Atributos de SSID para cada janela de tempo.

Padrão de Atributo	Descrição	Unidade	Intervalo	Quantidade
{categoria} _unique_ssid_hash_count	Número de requisições de rede únicas (hashes de SSID)	-	[0, ∞)	3
{categoria} _ssid_hash_device_ratio	Razão de SSIDs únicos para MACs únicos	-	[0, ∞)	3
{categoria} _non_broadcast_probe_ratio	Proporção de sondagens direcionadas	-	[0, 1]	3
{categoria} _ssid_persistence_score	Número médio de janelas em que cada SSID aparece	-	[1, ∞)	3

Razões menores que 1 indicam múltiplos endereços MAC solicitando os mesmos SSIDs, o que sugere aleatorização de MAC. Uma razão próxima de 1 indica que cada dispositivo solicita redes únicas, enquanto que uma razão superior à unitária indica que os dispositivos estão sondando múltiplas redes.

Quanto à pontuação de Persistência, valores altos indicam dispositivos “fixos,” sejam roteadores ou visitantes regulares, enquanto que valores baixos indicam dispositivos transitórios, de visitantes únicos.

Comando para criação da Base de Dados SQLite3

```
CREATE TABLE "features" (
  "window" TIMESTAMP,
  "probe_count" INTEGER,
  "unique_devices" INTEGER,
  "avg_rssi" REAL,
  "max_rssi" INTEGER,
  "min_rssi" INTEGER,
  "std_rssi" REAL,
  "strong_signal_count" REAL,
  "medium_signal_count" REAL,
  "weak_signal_count" INTEGER,
  "strong_signal_macs" REAL,
  "medium_signal_macs" REAL,
  "weak_signal_macs" INTEGER,
  "strong_rssi_avg" REAL,
  "strong_rssi_max" REAL,
  "strong_rssi_min" REAL,
  "strong_rssi_std" REAL,
  "medium_rssi_avg" REAL,
  "medium_rssi_max" REAL,
  "medium_rssi_min" REAL,
  "medium_rssi_std" REAL,
  "weak_rssi_avg" REAL,
  "weak_rssi_max" INTEGER,
  "weak_rssi_min" INTEGER,
  "weak_rssi_std" REAL,
  "strong_randomized_mac_ratio" REAL,
```

```
"medium_randomized_mac_ratio" REAL,
"weak_randomized_mac_ratio" REAL,
"strong_inter_arrival_avg" REAL,
"strong_inter_arrival_max" REAL,
"strong_inter_arrival_min" REAL,
"strong_inter_arrival_std" REAL,
"medium_inter_arrival_avg" REAL,
"medium_inter_arrival_max" REAL,
"medium_inter_arrival_min" REAL,
"medium_inter_arrival_std" REAL,
"weak_inter_arrival_avg" REAL,
"weak_inter_arrival_max" REAL,
"weak_inter_arrival_min" REAL,
"weak_inter_arrival_std" REAL,
"strong_avg_rssi_delta_per_mac" REAL,
"strong_moving_mac_count" REAL,
"strong_max_rssi_delta" REAL,
"medium_avg_rssi_delta_per_mac" REAL,
"medium_moving_mac_count" REAL,
"medium_max_rssi_delta" REAL,
"weak_avg_rssi_delta_per_mac" REAL,
"weak_moving_mac_count" REAL,
"weak_max_rssi_delta" REAL,
"unique_oui_count" REAL,
"major_brands_ratio" REAL,
"strong_oui_entropy" REAL,
"strong_major_brands_ratio" REAL,
"medium_oui_entropy" REAL,
```

```
  "medium_major_brands_ratio" REAL,
  "weak_oui_entropy" REAL,
  "weak_major_brands_ratio" REAL,
  "strong_unique_ssid_hash_count" REAL,
  "strong_ssid_hash_device_ratio" REAL,
  "strong_non_broadcast_probe_ratio" REAL,
  "medium_unique_ssid_hash_count" REAL,
  "medium_ssid_hash_device_ratio" REAL,
  "medium_non_broadcast_probe_ratio" REAL,
  "weak_unique_ssid_hash_count" REAL,
  "weak_ssid_hash_device_ratio" REAL,
  "weak_non_broadcast_probe_ratio" REAL,
  "strong_ssid_persistence_score" REAL,
  "medium_ssid_persistence_score" REAL,
  "weak_ssid_persistence_score" REAL
)
```