SENSORIAMENTO PARTICIPATIVO COM DESCRIÇÃO ADAPTATIVA DAS TAXAS DE AMOSTRAGEM E CONSISTÊNCIA DOS DADOS

Carlos Henrique de Oliveira Monteiro André

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Miguel Elias Mitre Campista

Rio de Janeiro
Março de 2018

SENSORIAMENTO PARTICIPATIVO COM DESCRIÇÃO ADAPTATIVA DAS TAXAS DE AMOSTRAGEM E CONSISTÊNCIA DOS DADOS

Carlos Henrique de Oliveira Monteiro André

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Miguel Elias Mitre Campista, D.Sc.

Prof. Diego Gimenez Passos, D.Sc.

Prof. Luís Henrique Maciel Kosmalski Costa, Dr.

Prof. Dianne Scherly Varela de Medeiros, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2018

André, Carlos Henrique de Oliveira Monteiro

Sensoriamento Participativo com Descrição Adaptativa das Taxas de Amostragem e Consistência dos Dados/Carlos Henrique de Oliveira Monteiro André. – Rio de Janeiro: UFRJ/COPPE, 2018.

XII, 54 p.: il.; 29,7cm.

Orientador: Miguel Elias Mitre Campista

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2018.

Referências Bibliográficas: p. 50 – 54.

1. Sensoriamento Participativo. 2. Redes Veiculares. 3. Consistência dos Dados. I. Campista, Miguel Elias Mitre. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Dedico este trabalho a minha esposa Lívia que me deu o apoio necessário para realizar mais esta fase de minha vida.

Agradecimentos

Primeiramente, gostaria de agradecer a Deus por me guiar, cuidar e dar a oportunidade de concretizar mais um ciclo na minha formação profissional. Agradeço também pela tranquilidade e pela força dada para seguir meus objetivos. Agradeço minha esposa Lívia pelo companheirismo nessa jornada. Aos meus pais e irmãos que entenderam a distância nesse período.

Agradeço meu orientador Miguel Campista pela oportunidade dada, pela atenção, pela orientação, pelo auxílio nos artigos submetidos e pela dedicação em sala de aula.

Agradeço, também, a todos os amigos que fiz no Grupo de Teleinformática e Automação (GTA), pois sempre contribuíram positivamente para a conclusão desse trabalho. Em especial, Victor Ribeiro, Fernando Molano, Hugo Sadok, Martin e Victor Fonseca, pelas sugestões técnicas, trocas de experiências e sugestões. Agradeço a toda equipe do laboratório GTA onde trocamos experiências visando alcançar um maior amadurecimento coletivo.

A Dianne Scherly pelo incentivo, auxílios e troca de experiências. Obrigado também pela ajuda nos artigos submetidos e auxílios prestados nas minhas diversas dúvidas ao longo da escrita.

Aos também amigos Daniel Tinôco e Caio Pazzini pelas conversas e incentivos no decorrer do mestrado.

Agradeço, ainda, a todos os professores que participaram da minha formação. Gostaria de agradecer também aos professores Luís Henrique, Pedro Velloso, Aloysio e Otto do GTA/UFRJ, ao professor Henrique Cukierman do PESC/UFRJ. Ao José Geraldo Ribeiro Junior, que me apresentou o GTA e me incentivou a ingressar no laboratório. Não poderia deixar de agradecer ao incentivo dos colegas Filippe Jabour e Alexsander dos Santos.

Agradeço aos funcionários do Programa de Engenharia Elétrica da COPPE/UFRJ, Maurício Machado, Daniele da Silva, Roberto Calvet e Marco Salgueiro pela presteza no atendimento na secretaria do Programa.

A todos que participaram de forma direta e indireta da minha formação. Por fim, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES pela bolsa de estudo concedida para o desenvolvimento da pesquisa.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

SENSORIAMENTO PARTICIPATIVO COM DESCRIÇÃO ADAPTATIVA DAS TAXAS DE AMOSTRAGEM E CONSISTÊNCIA DOS DADOS

Carlos Henrique de Oliveira Monteiro André

Março/2018

Orientador: Miguel Elias Mitre Campista

Programa: Engenharia Elétrica

O Sensoriamento Participativo (*Participatory Sensing* – PS) é um paradigma de redes colaborativas que incentiva a participação dos usuários no sensoriamento de uma Região de Interesse (*Region of Interest* – RoI). Em cenários colaborativos, um dos principais desafios é lidar com os usuários participantes móveis que devem calibrar a quantidade de dados sensoriados para que a carga imposta à rede e a eles próprios não se torne excessiva. O sistema deve garantir que os dados coletados sejam confiáveis e que possíveis dados anômalos não influenciem o resultado final. Nessa direção, este trabalho propõe um sistema centralizado capaz de adaptar a taxa de amostragem a ser atribuída a cada sensor participante, identificando a existência de inconsistência ou dados não confiáveis. Assume-se que parte dos dados pode ser coletada por sensores mal-intencionados ou defeituosos. Dessa forma, é necessário avaliar a presença de dados inconsistentes, baseado na média e desvio padrão das amostras. Em seguida, a taxa de amostragem é calculada em função da variabilidade das amostras coletadas em uma dada RoI a partir das amostras recebidas e validadas no último intervalo de tempo. O sistema proposto é avaliado usando um conjunto de dados que registra a mobilidade da frota de ônibus de Seattle, WA - EUA. Em um primeiro momento, avalia-se através de simulações a robustez do sistema caso haja a presença de dados inconsistentes. Os resultados mostram que o sistema é imune a dados inconsistentes até 70% dos nós. Os resultados também mostram o compromisso entre a taxa de amostragem e o número de sensores participantes. Quanto mais nós participantes, menor pode ser a taxa de amostragem e menor pode ser a quantidade de dados a serem transferidos individualmente. Além disso, é possível uma redução de aproximadamente 67% na carga de dados dos participantes do sensoriamento.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

PARTICIPATORY SENSING WITH ADAPTIVE DESCRIPTION OF SAMPLING RATES AND DATA CONSISTENCY

Carlos Henrique de Oliveira Monteiro André

March/2018

Advisor: Miguel Elias Mitre Campista

Department: Electrical Engineering

Participatory Sensing (PS) is a paradigm of collaborative networks which provides incentives for users to participate of sensing tasks on a Region of Interest (RoI). In collaborative scenarios, one of the major challenges is to deal with mobile participating users who must balance the amount of data collected by each user so as not to impose an excessive load to the network and to them. The system must ensure that the data collected is reliable and that possible anomalous data does not influence the final result. In this direction, this work proposes a centralized system to adapt the sample rate assigned to each participating sensor, identifying the existence of inconsistency or unreliable data. It is assumed that part of the data can be collected by malicious or faulty sensors. Thus, it is necessary to evaluate the presence of inconsistent data, based on the mean and standard deviation of the samples. In the following, the sampling rate is calculated as a function of the variability of the samples collected in a given RoI taking into account the samples received and validated in the last time interval. The proposed system is evaluated using the dataset of the bus fleet of the city of Seattle, WA - USA, which records bus movements. At first, the robustness of the system is evaluated through simulations in the presence of inconsistent data. Results show that the system is robust to inconsistent data up to 70% of the nodes. In addition, results show the tradeoff between sampling rate and number of participating sensors. The more participating users, the lower the individual sample rate and the lower the amount of data transferred. It is possible to reduce approximately 67% of data load of the participants.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O paradigma de sensoriamento participativo (*Participatory Sensing* – PS) [3] vem surgindo como uma forma de escalar sistemas de sensoriamento usando recursos de usuários, como *smartphones*. Para isso, é necessário que os usuários cedam os seus dispositivos para a realização de tarefas de sensoriamento, mediante incentivos. Essas tarefas são escalonadas por alguma entidade capaz de gerenciar uma Região de Interesse (*Region of Interest* – RoI), identificar o potencial de contribuição dos usuários para a coleta de dados e posteriormente disparar a coleta. O sensoriamento participativo tem como principal benefício o aumento do volume de dados amostrados de uma dada RoI a um baixo custo de operação. As aplicações inseridas neste paradigma podem, portanto, aprimorar os seus serviços a partir de um número grande de fontes de dados, que se torna mais viável através da participação dos usuários. Um exemplo de aplicação inserida no contexto do sensoriamento participativo é o Waze [4], que oferece informações sobre as condições do trânsito usando dados dos próprios usuários [5]. Tamanho interesse por dados vem do enorme apelo atual por serviços e aplicações dentro do contexto de Internet das Coisas (*Internet of Things* – IoT) [6, 7] e Cidades Inteligentes (*Smart Cities*) [8], nas quais os usuários deixam de ser meros consumidores de dados e passam a ser também potenciais produtores.

O número elevado de fontes de dados pode representar muitos desafios em termos de largura de banda, devido ao grande volume de dados gerados; acessibilidade, já que os dispositivos devem ser capazes de acessar a Internet mesmo estando espalhados em regiões geo-distribuídas; e mobilidade, já que as fontes podem pertencer a usuários móveis distintos. Todos esses desafios podem ser encontrados simultaneamente nas redes *ad hoc* veiculares (VANETs) [9] que, na era da Internet das Coisas, evoluíram para o novo paradigma da Internet dos Veículos (*Internet of Vehicles* – IoV) [10]. Na IoV, além dos usuários, sensores inteligentes e atuadores podem ser instalados na infraestrutura veicular [11], aumentando ainda mais o ecossistema de sensoriamento. Assim, os serviços associados dependem de uma grande coleção de

nós geo-distribuídos, sejam eles móveis ou não, que são responsáveis por sensoriar e enviar os dados coletados para uma entidade centralizadora, possivelmente na Internet, ou até mesmo por processar e agregar dados com o intuito de realizar análises em tempo real e tomar decisões operacionais rápidas. Essa pluralidade, além de aumentar o volume de tráfego da rede, pode ainda exigir que as aplicações IoV tenham condições de lidar com informações provenientes de várias fontes de dados com características e objetivos diversos [12]. Tendo em vista a participação dos usuários, tais fontes podem inclusive inserir informações errôneas de maneira proposital ou não. Além disso, a quantidade de dados gerados, as restrições dos dispositivos e as dificuldades para transferir esses dados para a Internet representam um compromisso que pode se tornar um importante obstáculo do sensoriamento participativo. É fundamental, portanto, utilizar técnicas para diminuir o volume de dados a serem enviados [13]. Nessa direção, a compactação, a agregação ou a adaptação da taxa de amostragem de cada nó participante são alguns dos mecanismos que tornam o sensoriamento mais inteligente e que ainda não são encontrados na literatura.

Este trabalho propõe um sistema de sensoriamento participativo com taxas de amostragem adaptativas e consistência dos dados que, diferente de trabalhos propostos anteriormente [14, 15], adapta a taxa de amostragem considerando tanto aspectos espaciais quanto temporais. Para isso, assume-se a existência de um nó centralizador, localizado na nuvem, que recebe os dados obtidos pelos usuários participantes do sistema de sensoriamento participativo. Esses dados são utilizados para calcular a variação das métricas coletadas em uma dada região de interesse (RoI) ao longo de um intervalo de tempo. Neste trabalho, assume-se que o sistema é usado para avaliação das condições de trânsito de uma cidade e, portanto, as medidas coletadas são as velocidades dos veículos participantes. Caso haja uma variação elevada nas velocidades coletadas em um dado trecho (ruas ou estradas) da RoI, considera-se que a probabilidade de medidas errôneas é maior nesse trecho. Portanto, o sistema deve aumentar a taxa de amostragem para que sejam coletadas mais amostras apenas desses trechos. Semelhantemente, caso a variação no trecho seja baixa, considera-se a possibilidade de redução da taxa de amostragem visando uma diminuição de dados coletados durante o processo, ou seja, manipular o tamanho amostral para alcançar um intervalo de confiança satisfatório. De maneira geral, caso se consiga uma redução da taxa de amostragem por nó participante, é possível reduzir o consumo de energia dos nós.

Durante o processo de amostragem, no entanto, é possível que dados anômalos sejam produzidos pelos usuários. Dessa forma, antes mesmo de considerar os dados amostrados, o servidor central deve detectar e garantir a qualidade e precisão da informação recebida. Esses dados podem ser provenientes de falhas nos sensores, erros durante a transferência dos dados, incapacidade de armazenamento dos dados

pelos participantes, ou podem, ainda, ser intencionalmente alterados por usuários maliciosos [16]. A metodologia é composta por três etapas principais: validação da amostra, detecção de anomalia da amostra e atualização do sistema. A etapa de validação evita que amostras com valores fora do intervalo de definição prossigam na análise, por exemplo, uma amostra contendo uma velocidade veicular absoluta negativa é descartada imediatamente. Já a etapa de detecção compara se a amostra possui valores próximos dos valores esperados ou não; caso não tenha, a amostra é descartada. Por fim, as amostras com valores próximos ao esperado são usadas para atualização do procedimento de análise. A ideia geral é evitar que o sistema esbarre no compromisso entre o número de amostras coletadas válidas, a carga de dados gerada e as limitações dos dispositivos dos usuários que podem ser, por exemplo, função da quantidade de energia disponível e da capacidade de armazenamento nos dispositivos.

As análises do sistema proposto e da metodologia para verificação da consistência de dados utilizam o conjunto de dados de movimentação da frota de ônibus da cidade de Seattle, nos Estados Unidos [17]. Desse conjunto de dados, analisa-se um dia inteiro, assumindo que o valor sensoriado é a velocidade dos ônibus. Inicialmente, o impacto de amostras inconsistentes em cada cenário é avaliado considerando a existência de usuários cujos sensores contribuem com dados inconsistentes ou com inconsistência intermitente. Os resultados mostram que é possível identificar sempre a presença de dados anômalos contendo valores de velocidade máxima, mesmo usando um método de detecção simples. No entanto, a presença de dados anômalos contendo valores de velocidade mínima nem sempre é detectada. Isso ocorre porque a velocidade média legítima é normalmente baixa em centros urbanos, mesmo descontando a possibilidade de desvio padrão. Ao manipular a média e os valores de desvio padrão, o impacto na detecção de anomalia não é importante mesmo adicionando 100% de amostras falsas [18]. O sistema de sensoriamento proposto consegue reduzir a carga de dados interposta na rede, além disso, o erro médio quadrático é reduzido em $\approx 35\%$. Maiores cargas de dados só são possíveis de serem entregues com poucos nós competindo pelo meio e, caso mais nós participem do sistema, é recomendável reduzir a carga de dados individuais para que maiores taxas de entrega sejam alcançadas. Os resultados ainda indicam que a amostragem adaptativa é capaz de reduzir os erros sem impactar na no aumento de dados no sistema.

1.1 Contribuições

Em resumo, as principais contribuições deste trabalho para o sensoriamento participativo são:

- redução do volume de dados inconsistentes: através do método de detecção de consistência dos dados;
- ajuste da melhor taxa de amostragem de dados por participante de uma rede colaborativa em cenários veiculares: a taxa de amostragem é adaptada de acordo com o erro das medidas nos diferentes trechos da mesma Região de Interesse;
- redução no uso de recursos dos usuários participantes: ajuste da taxa de amostragem dos dados;
- maior cobertura da região sensoriada: distribuição dos usuários em regiões com pouca cobertura;
- redução do custo financeiro dos participantes: é observada, em alguns casos, a redução os dados coletados, resultando em uma economia de recursos computacionais.

1.2 Organização do Texto

Esta dissertação está organizada da seguinte forma. No Capítulo 2 são apresentadas aplicações em redes veiculares, introduzidos noções de sensoriamento participativo e descritos os trabalhos relacionados. O Capítulo 3 descreve a arquitetura de rede assumida neste trabalho, e propõe um sistema de sensoriamento com taxa de amostragem adaptativa e consistência de dados. O Capítulo 4 apresenta o conjunto de dados utilizado e o caracteriza. Já o Capítulo 5 apresenta a análise da consistência dos dados. A verificação do sistema de adaptação da taxa de amostragem é apresentada no Capítulo 6. Por fim, o Capítulo 7 conclui este trabalho e apresenta as possíveis direções futuras.

Capítulo 2

Redes de Sensoriamento Colaborativo usando Redes Veiculares

Este capítulo está dividido em três seções: aplicações em redes veiculares, sensoriamento participativo e trabalhos relacionados. A primeira seção apresenta as categorias de redes veiculares e suas aplicações, trazendo exemplos de sistemas e aplicações. A segunda seção descreve sobre sensoriamento participativo e apresenta sua estrutura básica. A terceira seção, por sua vez, apresenta trabalhos relacionados sobre sensoriamento participativo e seus desafios.

2.1 Aplicações em redes veiculares

As aplicações das redes veiculares podem ser divididas em três categorias, a saber: *(i) segurança no transporte, (ii) gerenciamento e eficiência do tráfego, (iii) infotainment* [2, 12, 19].

***Segurança no transporte*:** As aplicações dessa categoria visam melhorar a segurança dos usuários nas via, através de notificações sobre situações adversas. Os dois principais objetivos são a redução de acidentes e a melhoria do fluxo de veículos nas vias [20, 21]. Aplicativos mais comuns trazem alerta de colisão, assistência no trânsito, aviso de quedas de barreiras, avisos de incidentes na pista. Normalmente a restrição de tempo é primordial para esse tipo de aplicação, requer uma menor latência e normalmente atinge curto alcance entre os veículos [22, 23].

***Gerenciamento e eficiência do tráfego*:** As aplicações dessa categoria visam um melhor gerenciamento e eficiência do tráfego, trabalham com navegação e controle de velocidade. Essas aplicações normalmente trocam informações entre si, para gerar um histórico de condições das vias, visando um melhor fluxo do trânsito ao

redor do usuário. É possível usar as informações para criar uma melhor rota e evitar acesso às vias congestionadas. Normalmente essas aplicações necessitam de um nó centralizador para troca de informações. A prioridade de acesso é baixa e normalmente sofre maior latência [24].

***Infotainment*:** As aplicações dessa categoria visam o entretenimento dos motoristas e passageiros, trazendo um maior conforto nas viagens. Pode-se disseminar conteúdos multimídia, entregar informações de interesse local, dados climáticos da região ou até mesmo prover Internet para os usuários [20]. Normalmente essa categoria depende da comunicação com redes estruturadas. Existem trabalhos que consideram a distribuição de conteúdo primordial para diminuição do tempo de acesso do motorista à informação [22, 25].

A Tabela 2.1 lista algumas aplicações bem como alguns de seus requisitos onde cada aplicação está relacionada à categoria, ao tipo de comunicação, ao tipo da mensagem, ao tempo da mensagem e à maior latência [2].

Tabela 2.1: Características de aplicações veiculares (adaptado de [2]).

Aplicações	**Categoria**	**Comunicação**	**Tipo**	**Tempo**	**Latência**
Alerta de veículo lento	segurança e transporte	*ad hoc*, V2V	*broadcast* permanente	500ms	100ms
Alerta de colisão em cruzamento	segurança e transporte	*ad hoc*, infraestrutura, V2V, V2I	*broadcast* permanente	100ms	100ms
Assistência para direção ecológica	*infotainment*	infraestrutura, *ad hoc*, V2I, V2V e rede de telefonia celular	*unicast*, *broadcast*, sobdemanda	1000ms	500ms
Download/update de mapas	gerenciamento e eficiência do tráfego	infraestrutura, *ad hoc*, V2I, V2V e rede de telefonia celular	*broadcast*	1000ms	500ms
Download de mídia	*infotainment*	infraestrutura, rede de telefonia celular, etc	*unicast*, *broadcast*, sobdemanda	n/d	500ms
Gerenciamento de cruzamento	gerenciamento e eficiência do tráfego	infraestrutura, *ad hoc*, V2I, V2V	*broadcast* periódico, *unicast*	1000ms	500ms
Pré-colisão	segurança e transporte	*ad hoc*, V2V	*broadcast* periódico, *unicast*	100ms	50ms

Este trabalho está associado ao desempenho de duas categorias, segurança no transporte, e gerenciamento e eficiência do tráfego. Isso porque são abordados tanto possíveis usuários que podem influenciar no desempenho do sistema quanto uma adaptação na taxa de amostragem, buscando uma maior integridade dos dados e um menor volume de informações na rede. Muitas das aplicações em redes veiculares se

beneficiam do sensoriamento participativo, por exemplo, Waze e Google Maps [5, 23]

2.2 Sensoriamento participativo

Atualmente, o poder de sensoriamento embarcado em veículos ou mesmo em dispositivos móveis pessoais, como *smartphones*, vem aumentando a passos largos. Tal aumento é consequência da evolução da tecnologia móvel e da miniaturização e barateamento dos componentes eletrônicos. Dentre os sensores mais comuns estão o acelerômetro, o GPS, o giroscópio, o detector de luminosidade, o microfone e a câmera. Todos esses sensores são capazes de aumentar o poder de sensoriamento de cada dispositivo, proporcionando um reconhecimento crescente do ambiente. Por isso, a capacidade de detecção de cada dispositivo aumenta, permitindo a aquisição de conhecimento local e compreensão do contexto. A partir dos dados coletados por esses sensores é possível vislumbrar uma série de aplicações que oferecem noções de contexto ou do ambiente em que o usuário se encontra [26]. Quando os dispositivos são usados em conjunto, ou seja, combinando dados coletados, é possível construir um sistema de sensoriamento mais poderoso, escalável e de baixo custo. Como consequência as aplicações podem ser mais sofisticadas, resultando inclusive em desenvolvimento de novas aplicações das chamadas Cidades Inteligentes [26]. O sensoriamento através da participação coletiva dos usuários e seus dispositivos é o foco do recente tema de pesquisa chamado de Sensoriamento Participativo (*Participatory Sensing* – PS). Vale ressaltar que existem vários nomes diferentes associados à definição de redes de sensoriamento participativo encontrados na literatura [27], por exemplo, *Humans as Data Sources*, *Ubiquitous Crowdsourcing* e *Urban Computing* são também utilizados.

O sensoriamento participativo é composto de tarefas implementadas nos dispositivos móveis que possibilitem a coleta, a análise e o compartilhamento do conhecimento gerado em uma Região de Interesse. Uma das principais vantagens desse tipo de sensoriamento é o aumento do conhecimento sobre a região monitorada, como consequência do compartilhamento dos dados sensoriados. Além disso, é possível ampliar o raio de cobertura da RoI, devido à mobilidade de alguns sensores. Os dados obtidos pelo coletivo precisam ser processados, possivelmente em uma entidade externa. Assim, existe um crescente uso da capacidade da nuvem para realização tanto do processamento quanto do armazenamento do conhecimento obtido para a RoI [28, 29]. Várias aplicações podem ser desenvolvidas segundo este paradigma e em diferentes áreas como: gestão de recursos naturais [28, 30], planejamento e monitoramento urbano [28, 31, 32] e saúde pública [28, 33]. Dado o aumento do número de sensores capazes de integrar um cenário, as abordagens de detecção participativa devem ser escaláveis.

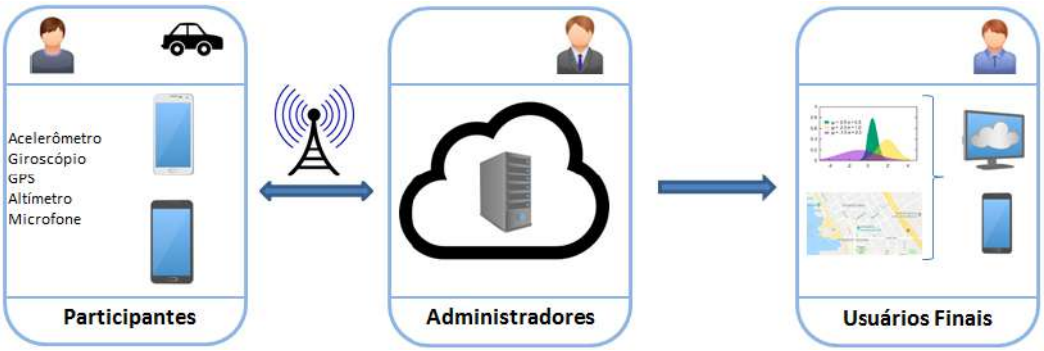


Figura 2.1: Arquitetura de um sistema de sensoriamento participativo, composta por três grupos: participantes, administradores e usuários finais (adaptado de [1]).

A Figura 2.1 representa a arquitetura de um sistema de sensoriamento participativo, composta por três grupos: os participantes, os administradores e os usuários finais [1, 29, 34]. O papel de cada grupo está descrito a seguir.

***Participantes*:** São os usuários que colaboram com os dados coletados. Suas funções são sensoriar, processar, armazenar e transferir os dados para os administradores. Os usuários podem consumir seus próprios dados além dos dados finais, juntamente com os dados agregados dos demais participantes. Os participantes têm o intuito de aumentar o poder de sensoriamento da RoI. Este trabalho está focado neste grupo, sendo importante a participação de usuários confiáveis que não tenham seus recursos esgotados.

***Administradores*:** São os responsáveis pela administração e coordenação do sistema. Normalmente são membros de uma organização que viabiliza o sistema, cuida da manutenção das aplicações e dos servidores. Suas funções são buscar e receber informações dos participantes, processar, armazenar e preparar as informações para os usuários finais.

***Usuários Finais*:** São os usuários que utilizam o sistema para diversos fins, como por exemplo, motoristas que definem o melhor trajeto de deslocamento até um destino final. Esse grupo apenas visualiza as informações de seu próprio interesse. Alguns usuários podem ter características de participante, podendo não apenas consumir, mas também inserir contribuições no sistema.

Os três grupos, participantes, administradores e usuários finais, utilizam dispositivos que são as ferramentas necessárias para sensoriar, processar e armazenar os dados coletados ao longo do tempo. Vale ressaltar que geralmente o grupo de administradores utiliza dispositivos mais robustos que estão localizados em nuvem, e possuei recursos de processamento e armazenamento com características mais robustas, já o grupo de participantes e usuários finais em geral utilizam dispositivos pessoais menos robustos do que os dispositivos usados pelos administradores [1].

Com o sensoriamento participativo pode-se enriquecer o conhecimento de uma RoI a um baixo custo usando dados de usuários. Essa participação, no entanto, traz desafios devido à heterogeneidade e ao volume de dados. Quando os usuários participantes pertencem a uma rede veicular, os usuários devem processar e interpretar diferentes tipos de dados coletados ao longo de uma viagem inteira e transferir esses dados para outros nós, por exemplo, para um *gateway*, durante tempos de contato possivelmente curtos.

A participação do usuário na coleta e transferência de dados deve ser estimulada com formas de incentivo. Nos sistemas de sensoriamento participativo, dependendo da recompensa, um usuário pode se sentir motivado a coletar um grande volume de dados em um curto intervalo de tempo. Por exemplo, um usuário que informe a velocidade do seu veículo pode receber em troca descontos em serviços ou até mesmo anúncios específicos de promoções conforme seu perfil. O benefício pode ser também o retorno do conhecimento coletivo, como a informação de trânsito no Google Maps, ou ainda, benefícios no aplicativo de sensoriamento, como no Waze. Para não exaurir os recursos do usuário, o que poderia levar à exclusão ou a não obtenção de um dado importante para o sistema, deve existir um compromisso entre a quantidade e a qualidade dos dados coletados. Usuários participantes anônimos podem enviar dados incorretos, de baixa qualidade ou até mesmo falsos [19, 35]. Além disso, dados oriundos de sensores diferentes, modelos e calibrações distintas, podem ser redundantes ou inconsistentes. Até o mesmo sensor pode aferir o mesmo evento sob condições diversas, por exemplo, detectar um ruído com o ambiente do carro livre ou com o som ligado.

Este trabalho aborda o sensoriamento participativo em redes veiculares, explorando as áreas de planejamento e monitoramento urbano que estão dentro da categoria gerenciamento e eficiência. Em redes veiculares, o sensoriamento participativo pode ser utilizado, por exemplo, para planejar uma nova rota de deslocamento, instalação de novos pontos de parada e planejamento de expansões viárias. Os dados são inseridos no sistema através de seus próprios usuários. Logo, a participação dos usuários é fundamental, mas pode esbarrar em problemas para a rede e para o próprio sistema. Para a rede, se o volume de dados coletados durante o trajeto em uma região de interesse for muito grande, pode-se transferi-los durante contatos curtos entre as Unidades de Bordo (*On Board Units* – OBUs) e as Unidades de Acostamento (*Road Side Units* – RSUs) através de redes de acesso usando IEEE 802.11, por exemplo, IEEE 802.11p ou outra tecnologia móvel, por exemplo, 4G. O último, no entanto, pode adicionar altos custos aos usuários. Caso os dados não sejam recebidos, ou até mesmo dados inconsistentes sejam processados, o próprio sistema estará comprometido.

O sensoriamento participativo pode ser utilizado em diversos cenários e conjunto

de dados distintos, no entanto possui diversos desafios que são abordados na próxima seção.

2.3 Trabalhos relacionados

Muitos sistemas de sensoriamento já foram propostos na literatura com intuitos diversos. Cruz *et al.*, por exemplo, propuseram o *SensingBus*, que refina o monitoramento das condições climáticas de uma cidade a partir de sensores embarcados em ônibus urbanos [36, 37]. A ideia é um contraponto às soluções mais triviais que seriam aumentar o número de estações meteorológicas ou o poder de coleta de cada uma delas. Como as duas opções implicam investimentos, os autores propõem adicionar aos dados já existentes aqueles coletados pelos ônibus. A consequência é a expansão da área monitorada gratuitamente através da mobilidade dos ônibus. Os nós sensores, embarcados nos ônibus, coletam dados sobre a cidade. Os dados são entregues para o sistema através de pontos de acesso nas paradas de ônibus da cidade. Posteriormente, os dados são apresentados aos usuários que usam um serviço na nuvem.

O *SensingBus* é um sistema composto por três camadas: coleta, recepção e publicação. A camada de coleta é responsável por coletar os dados de interesse pela cidade através de nós sensores sem fio embutidos em ônibus urbanos de transporte público. A camada de recepção recebe os dados coletados diretamente da camada detecção e envia para a camada de publicação, através da Internet. Essa camada é composta por pontos de acesso, montados em paradas de ônibus. Já a camada de publicação recebe dados de todos os nós na camada de recepção e os entrega para os usuários, usando uma API e uma interface web.

Mohan *et al.* [32] propuseram o Nericell, que usa sensores embarcados em *smartphones*, tais como acelerômetros, microfones e GPS, além do sinal de rádio GSM, para determinar as condições da via (ruas e estradas) e do tráfego. O acelerômetro é utilizado para detecção de trepidações e frenagens bruscas. O microfone monitora o ambiente, principalmente para detectar as buzinas. O sensoriamento participativo é utilizado para troca de informações entre os participantes, proporcionando a economia de energia. O principal objetivo do projeto é proporcionar mais conforto, melhorando a condução do motorista.

Zhou *et al.* [38] propuseram um sistema baseado em colaboração para prever a posição dos ônibus ao longo do tempo durante o percurso, utilizando a potência entre as torres de celular e os dispositivos móveis. O sistema é baseado exclusivamente no esforço dos usuários participantes, sendo composto por três principais componentes: usuário distribuidor, servidor *backend* e usuário solicitante. O usuário distribuidor contribui com a informação de detecção para o sistema. Após entrar em um ônibus,

o módulo de coleta de dados começa a coletar uma sequência de IDs de torres de celular próximas. Os dados coletados são transmitidos ao servidor *backend* por meio de redes celulares. O servidor é responsável por pré-processar todos os dados recebidos e todas as consultas solicitadas. Ele contém um banco de dados com a sequência das torres por rota e sua localização física. Com essas informações ele deixa uma previsão para ser consultada pelo usuário solicitante. Por fim, o usuário solicitante verifica o horário de chegada do ônibus enviando a solicitação para o servidor *backend*, indicando a rota e a parada para a qual ele quer prever a chegada do ônibus.

Ganti *et al.* [39] desenvolveram um serviço de navegação, chamado *GreenGPS*, que usa dados de sensoriamento participativo para mapear o consumo de combustível nas ruas da cidade, permitindo que os motoristas encontrem as rotas mais eficientes e econômicas para os veículos entre os pontos de abastecimento. A interface padronizada de comunicação dos veículos, chamada OBD-II, é utilizada para prover acesso a diversos medidores e instrumentação do motor. Além disso, os dados dos sensores são exportados para o dispositivo *GreenGPS*, sabendo que nem sempre a rota mais rápida ou curta é a que consome menos combustível. Nesse contexto, um sistema de sensoriamento participativo pode influenciar as decisões de deslocamento de usuários individuais, pois as informações oferecidas pelo sistema auxiliam na tomada de decisão dos motoristas.

Todos os trabalhos citados anteriormente podem ser visualizados sob o paradigma de sensoriamento participativo. Nenhum deles, no entanto, se preocupa com o controle de dados inconsistentes e com a taxa de amostragem utilizada.

A quantidade de dados é um dos problemas mais importantes que afetam o desempenho do sistema. Zeng *et al.* propuseram a aplicação As-air [14] que adapta a taxa de amostragem de acordo com as características do ambiente, que pode variar em função do dia de utilização, sejam eles dias de semana, finais de semana ou feriados. O As-air é usado para medir a qualidade do ar que pode variar, necessitando de taxas de amostragem mais elevadas em dias de atividades mais intensas. As condições padrão da qualidade do ar são encontradas de acordo com os dados históricos contribuídos pelos participantes. Com base no histórico, a qualidade do ar atual é antecipada e o processo de aprendizagem para escolha dos melhores parâmetros de amostragem são acelerados. A aplicação fornece uma estratégia de amostragem eficiente em termos de energia, que é adaptável ao ambiente de ar exterior, Harald Weinschrott *et al.*, diferente de Zeng *et al.*, aumentam a taxa de amostragem por questões de robustez [15]. Taxas de amostragem maiores são usadas em regiões onde se deseja aumentar a probabilidade de detecção de um evento de interesse.

Apesar de trabalhos que variam as taxas de amostragem já existirem na litera-

tura, até onde se sabe, nenhum adapta a taxa de amostragem considerando tanto aspectos espaciais quanto temporais do cenário. O trabalho proposto nesta dissertação adapta as taxas de amostragem por participante do sistema em função da região onde o nó participante se encontra e da janela de tempo atual.

Além dos aspectos espaciais e temporais da amostragem, a consistência dos dados também é levada em conta nesta dissertação. Sabe-se que a segurança em redes veiculares (*Vehicular Network* – VANET) é um fator imprescindível que necessita ser observado, pois uma informação irreal pode afetar uma decisão, por exemplo, impedir o aviso de um congestionamento, evitando que os veículos busquem caminhos alternativos e os obrigando a entrar no congestionamento [40]. Outro exemplo seriam veículos maliciosos se recusam a transmitir a mensagem para outro veículo sobre um acidente [41]. Em qualquer rede de computadores cabeada ou sem-fio, a presença de usuários ou nós maliciosos que buscam o mau funcionamento da rede deve ser avaliada. Dentre os possíveis ataques, pode-se destacar os ataques de negação de serviço, modificação de mensagens e análise de tráfego. No entanto, as características das redes veiculares, tais como a alta mobilidade dos nós, desconexões frequentes e densidade variável dos nós impõem novos desafios à segurança. Este trabalho considera que usuários que tenham comportamento indesejável ou contribuam com valores irreais possam ser ignorados a amostragem. Além disso, pode-se determinar qual taxa de amostragem é mais adequada para um trecho da via e quantos veículos serão utilizados. Juntamente é proposto um sistema de sensoriamento participativo que consegue ajustar a taxa de amostragem de seus veículos conforme a necessidade.

Capítulo 3

Sistema de Sensoriamento Participativo Proposto

Este capítulo apresenta a arquitetura da rede usada neste trabalho, descreve a operação do sistema de sensoriamento participativo com amostragem de taxas adaptativas e uso da consistência dos dados em redes veiculares.

3.1 Arquitetura da rede veicular

A arquitetura da rede veicular usada neste trabalho é composta por um servidor central, por pontos de acesso (*RoadSide Units* – RSUs) e por usuários do sistema de sensoriamento participativo equipados com unidades de bordo (*OnBoard Units* – OBUs). O servidor central é alcançado pela Internet, acessível pelas RSUs através de uma rede de acesso normalmente cabeada. Já os usuários participantes se conectam às RSUs através de redes sem-fio IEEE 802.11p. O servidor central se serve dos dados coletados pelos usuários para oferecer serviços a usuários externos. Por exemplo, como é o caso de uso deste trabalho, o servidor central oferece informação das condições do trânsito a usuários externos. Já os usuários coletam suas velocidades médias durante o deslocamento e entregam a uma RSU. A Figura 3.1 ilustra a arquitetura de uma rede veicular.

Neste trabalho, utiliza-se o conceito de Região de Interesse (*Region of Interest* – RoI), onde os usuários participantes têm a oportunidade de receber solicitações para contribuição com o sistema de sensoriamento participativo. Tal solicitação é enviada pelo servidor central e entregue ao usuário através de uma RSU na entrada da RoI. Note que por questões de simplicidade, uma RoI é formada por trechos de ruas e estradas delimitadas por RSUs. Ao sair da RoI, o usuário que se propôs a participar do sistema de sensoriamento deve entregar os dados coletados ao servidor central. Para isso, a entrega é realizada através de uma RSU na saída da RoI. Munido

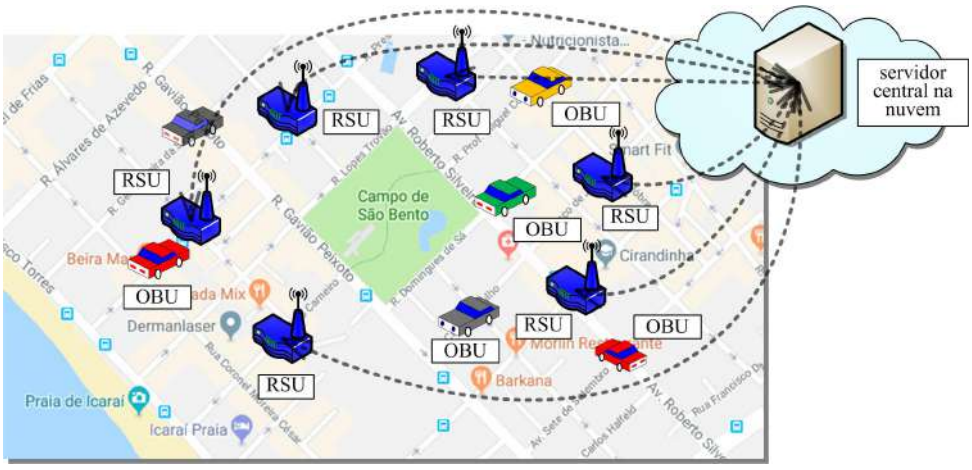


Figura 3.1: Arquitetura de uma rede veicular composta por RSUs e OBUs embarcadas em veículos. As RSUs podem estar conectadas a um servidor central na Internet.

de diferentes contribuições da mesma RoI, o servidor central pode consolidar os dados recebidos para possível oferta de serviços a usuários externos além de poder mesclar dados correlacionados da RoI. Neste trabalho, a questão do incentivo não é abordada, mas assume-se que a participação no sistema de sensoriamento seja de alguma forma recompensada para os usuários. A Figura 3.2 mostra um nó participante que, ao entrar na Região de Interesse, tem a oportunidade de participar do sistema de sensoriamento, conforme utilizado neste trabalho.

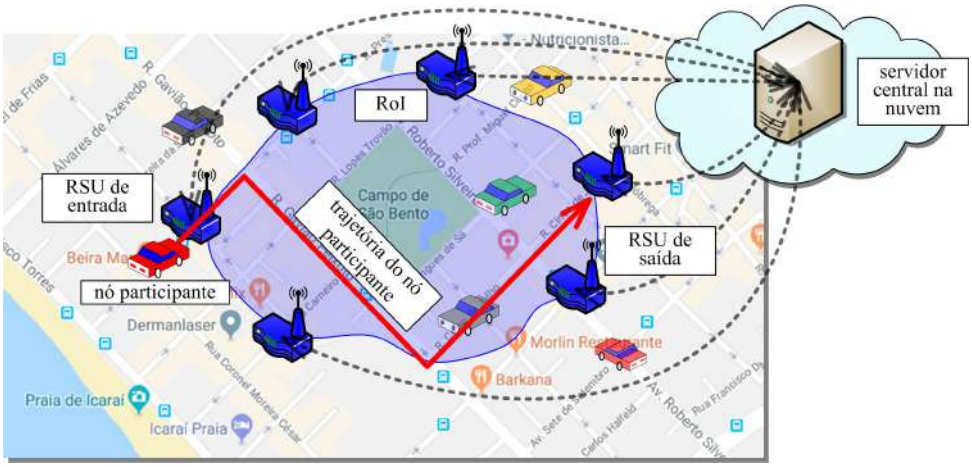


Figura 3.2: Arquitetura da rede veicular usada neste trabalho. Os usuários participantes do sistema de sensoriamento iniciam a coleta de dados ao entrar na RoI e descarregam os dados sensoriados na RSU de saída.

Dois problemas recorrentes no sensoriamento participativo são: a confiança em diversas fontes de dados distintas e o ajuste da taxa de amostragem, já que não necessariamente existe uma única taxa mais apropriada para toda a RoI. Dependendo da variação dos dados amostrados, pode ser necessário aumentar ou diminuir

a taxa de amostragem em diferentes trechos da mesma RoI ao longo do tempo. Isso, obviamente, é importante caso se deseje minimizar em toda a RoI o erro da amostragem obtido a partir dos dados consolidados. A dinamicidade da amostragem deve ser levada em conta para que: *(i) a RoI seja totalmente sensoriada apresentando resultados com baixo erro em comparação ao valor real e que (ii) não haja nem sobrecarga dos usuários participantes do sistema de sensoriamento e nem da rede sem fio ou da rede de acesso ao servidor.*

3.2 Descrição do sistema proposto

Este trabalho é dividido em duas abordagens principais: *(i) descrição adaptativa das taxas de amostragem e (ii) consistência de dados.* Na metodologia de descrição adaptativa das taxas de amostragem, a consistência dos dados a serem analisados também é verificada. Logo, devido à sua importância, a consistência é abordada em separado neste trabalho.

3.2.1 Operação do sistema de sensoriamento participativo com amostragem adaptativa

A taxa de amostragem do sistema de sensoriamento participativo proposto é adaptada ao erro das medidas nos diferentes trechos da mesma RoI, levando em conta tanto características espaciais quanto temporais. Para isso, o servidor central deve, de tempos em tempos, inferir o erro da amostragem para cada trecho dentro da mesma RoI para que os próximos usuários do sistema de monitoramento possam aumentar ou diminuir as taxas de amostragem conforme a necessidade calculada dos trechos. Sendo assim, denota-se o conjunto de usuários participantes do sistema de sensoriamento como $\boldsymbol{\mathcal{P}} = \{p_1, \ldots, p_n\}$, onde p_i é um usuário participante e n é o número total de usuários no sistema. Cada usuário p_i coleta amostras e as registra em um traço de dados T_i. Cada amostra é uma tupla contendo a posição, o instante da coleta e o valor da métrica amostrada. Por exemplo, na rede veicular tratada, cada tupla do traço contém a posição, o instante da coleta e a velocidade do veículo. A Figura 3.3 representa o modelo genérico da tupla de dados suportada pelo sistema. Todos esses dados podem ser obtidos a partir de um GPS (*Global Positioning System*) e do sensor a ser monitorado. Ressalta-se que o sistema de sensoriamento proposto não perde generalidade caso outro sensor ou outros cenários sejam usados.

O servidor central, ao receber o traço $\boldsymbol{T_i}$, verifica se as amostras possuem algum dado inconsistente. Amostras inconsistentes podem ser geradas a partir de falhas do sistema de sensoriamento ou por algum tipo de ação maliciosa [16]. O problema da consistência dos dados será apresentado no Capítulo 5. Assumindo que todas

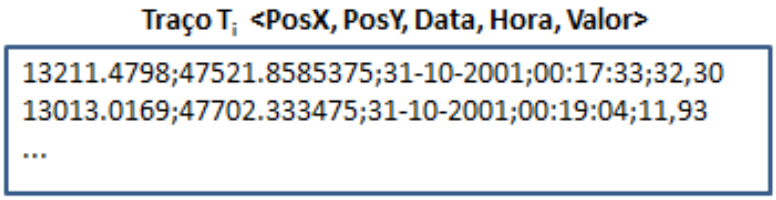


Figura 3.3: Estrutura da tupla de dados no sistema proposto.

as amostras são válidas, o servidor central passa para uma etapa de atualização da variação dos valores contidos nas amostras recebidas por trecho da RoI. A variação é usada como forma de inferir o erro nas medidas em cada trecho, já que este último não pode ser obtido diretamente. Neste trabalho, uma RoI $\mathcal{R}$ é representada por um conjunto de trechos, ou seja, $\boldsymbol{\mathcal{R}} = \{\rho_1, \ldots, \rho_m\}$, onde ρ_j é um trecho e m é o número de trechos em $\mathcal{R}$. Ainda, denota-se como $\boldsymbol{v^{(\Delta t)}} = \langle v_1^{(\Delta t)}, \ldots, v_m^{(\Delta t)} \rangle$ o vetor atualizado da variação das amostras calculadas por trecho ρ_j referente ao intervalo de tempo Δt. Observe que $v_j^{(\Delta t)}$, correspondente à variação no trecho ρ_j, é calculada considerando amostras recebidas em traços anteriores no mesmo intervalo Δt e mais as amostras recebidas no último traço. Portanto, caso exista pelo menos um trecho com variação elevada ou reduzida, significa que a taxa de amostragem ao menos nesse trecho deve ser mais bem adaptada ao cenário. A variação é considerada com base no erro médio da medida em um dado intervalo de tempo. Caso a variação esteja elevada, deve-se aumentar a taxa de amostragem; ou reduzi-la, caso a variação esteja baixa. Define-se então o vetor $\boldsymbol{r^{(\Delta t)}} = \langle r_0^{(\Delta t)}, \ldots, r_m^{(\Delta t)} \rangle$ que contém as taxas de amostragem recomendadas para todos os trechos em $\mathcal{R}$ no intervalo de tempo Δt. Novamente, destaca-se a correspondência entre o trecho ρ_j e a sua respectiva taxa de amostragem $r_j^{(\Delta t)}$. Visando diminuir a troca de informação entre servidor e veículo, a relação $r_j^{(\Delta t)}$ por trecho pode ser simplificada. Ao invés de uma taxa de amostragem por trecho, pode-se ter uma única taxa de amostragem para trechos consecutivos. A Figura 3.4 ilustra todas as etapas descritas, executadas no servidor central para atualização do vetor de variação nos trechos $\boldsymbol{v^{(\Delta t)}}$ e do vetor de taxas de amostragem por trecho $\boldsymbol{r^{(\Delta t)}}$.

Um usuário p_i do sensoriamento participativo, ao entrar na RoI através de uma RSU de entrada, tem a oportunidade de receber o vetor $\boldsymbol{\pi}$ contendo a descrição do sensoriamento a ser realizado em toda RoI no intervalo Δt, como pode ser visto na Figura 3.2. O vetor é composto por uma sequência de tuplas que descrevem os trechos da RoI e as respectivas taxas de amostragem recomendadas. O participante, ciente de sua trajetória, utiliza apenas as taxas de amostragem referentes aos trechos que vai atravessar. Caso o sistema seja ciente da trajetória do participante p_i, o vetor $\boldsymbol{\pi}$ pode ser resumido a apenas os trechos a serem seguidos por p_i. Mesmo nesse caso, não necessariamente todos os trechos da trajetória de p_i precisam estar contidos no

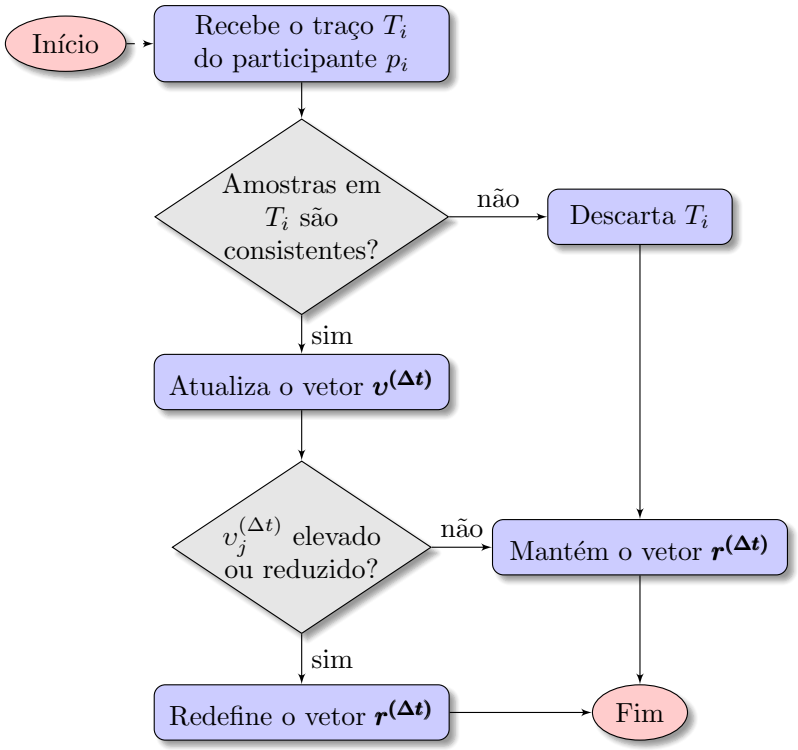


Figura 3.4: Fluxograma do processo executado no servidor central ao receber um novo traço T_i de um participante p_i.

vetor π. Caso a taxa de amostragem de dois trechos consecutivos seja a mesma, o sistema pode agregar dois trechos em uma única tupla. O vetor π é enviado ao usuário p_i através de uma estrutura de dados. Como comentado anteriormente, cada trecho em uma tupla de π pode ser indicado através da sua posição geográfica de início $pos_{inicial}$ e término pos_{final}. Note que um trecho agregado também pode ser representado da mesma forma. As amostras sensoriadas são descarregadas através da RSU de saída da RoI. O servidor, ao receber as novas amostras, atualiza ou não o vetor $r^{(\Delta t)}$, usado no intervalo de tempo Δt. A Figura 3.5 representa o formato da descrição do sensoriamento a ser realizado pelos participantes na RoI.

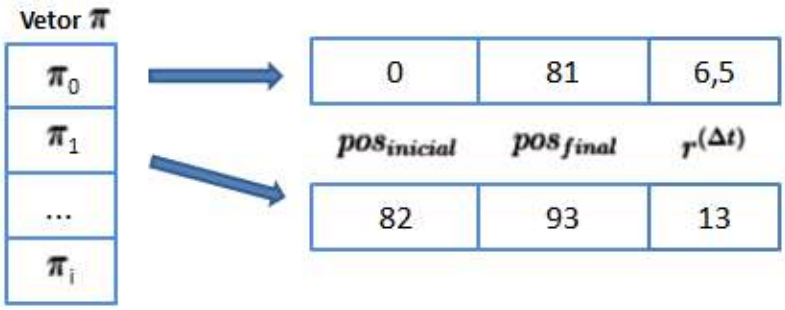


Figura 3.5: Vetor π.

Para um melhor detalhamento do sistema como já descrito no início desta seção, foi necessário um estudo sobre a consistência dos dados. Assim, a Seção 3.2.2 apre-

senta uma metodologia para detecção de dados inconsistentes em redes veiculares.

3.2.2 Consistência de dados em redes veiculares

Esta seção apresenta a metodologia proposta para a detecção de dados inconsistentes provenientes de redes veiculares usadas para sensoriamento participativo. Apresentam-se, ainda, alguns possíveis perfis de usuários que possivelmente inserem informações inconsistentes no sistema.

A metodologia proposta para avaliação da consistência dos dados é dividida em três etapas principais: validação da amostra, detecção de anomalia de cada amostra e atualização do sistema de detecção. É importante destacar que a metodologia pode ser tanto executada em um ponto concentrador de dados sensoriados, possivelmente um servidor na Internet; ou localmente em um dispositivo realizando sensoriamento. Configurações híbridas também são possíveis nas quais a metodologia pode ser parcialmente executada nos dispositivos e também em um servidor na nuvem. O funcionamento da metodologia proposta está ilustrado na Figura 3.6. A seguir, cada uma das três etapas principais é detalhada.

***Validação da amostra*:** Nesta etapa, o objetivo é eliminar através de filtros amostras que contenham valores não-válidos para os dados sensoriados. Sendo assim, evita-se o desperdício de recursos de processamento com amostras que não contribuem com o sistema participativo. Considerando uma amostra x_i composta por um vetor de propriedades $\langle p_{i,1} \ldots, p_{i,n} \rangle$, essa amostra somente é considerada válida caso cada uma de suas propriedades pertença ao intervalo de definição conhecido, ou seja, $p_{i,j} \in \mathcal{D}_j$.

Neste trabalho, a localização do GPS e o valor da métrica sensoriada são usados como propriedades das amostras. Assume-se que a localização deva pertencer à RoI onde a medida foi coletada e que a velocidade deva ser sempre positiva. Além disso, essa velocidade não pode ser maior do que um determinado limite. Esse limite pode variar de um cenário para outro e sua definição pode levar em consideração diversos fatores, como a maior velocidade possível de ser alcançada por um veículo moderno real, movimentação em ambiente urbano ou rural, presença de auto-estradas no cenário.

***Detecção de anomalia da amostra*:** Nesta etapa, as anomalias são de fato detectadas. Para tal, assume-se a existência de um sistema capaz de julgar se uma amostra válida está dentro ou não de um intervalo de valores esperados, de acordo com a variação das amostras. A noção de valores esperados pode ser construída baseada em um histórico de valores recebidos. A metodologia é genérica o suficiente para que diferentes sistemas de detecção possam ser utilizados. Nessa etapa, apenas a propriedade relacionada com o valor da medida sensoriada é avaliada.

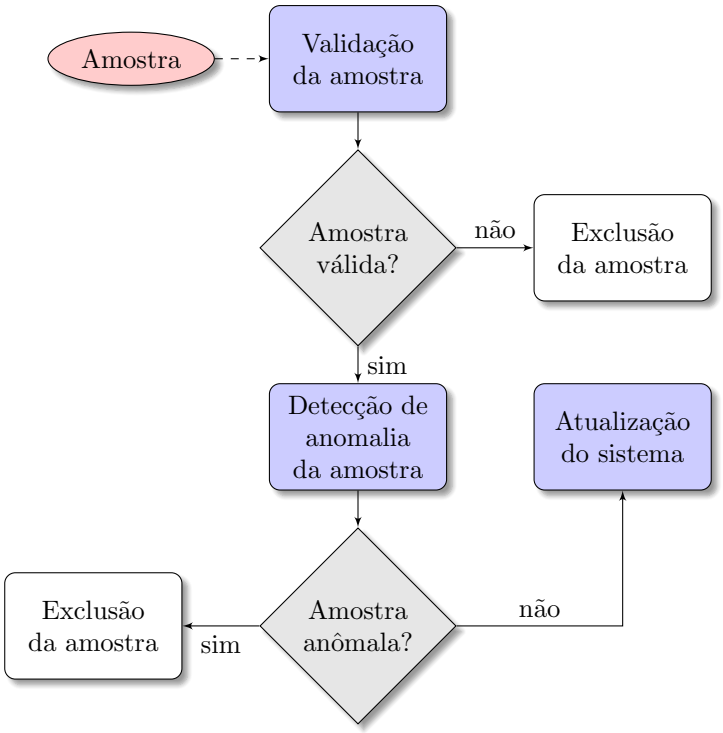


Figura 3.6: Metodologia proposta para detecção de anomalia no sensoriamento participativo em redes veiculares. A metodologia é dividida em três etapas principais: validação da amostra, detecção de anomalia da amostra e atualização do sistema.

Neste trabalho, a metodologia proposta usa um mecanismo simples de detecção. Seja x_i a amostra a ser analisada, consideram-se como anomalia as amostras com valores fora do intervalo $[\mu-3\sigma,\mu+3\sigma]$, onde μ é a média dos últimos valores recebidos e σ é o desvio padrão. Sendo assim, se $x_i \notin [\mu-3\sigma,\mu+3\sigma]$, x_i é considerada uma anomalia e é descartada. Vale lembrar que os dados são classificados conforme o tipo de veículo e cenário, ou seja, caso tenha uma variação em um dos dois parâmetros, o servidor deverá rodar separadamente cada combinação de dados.

***Atualização do sistema*:** Nesta etapa, caso a amostra x_i não tenha sido considerada uma amostra anômala, ela é utilizada para atualizar o sistema de detecção de anomalia da amostra. Note que o processo de atualização pode depender do sistema de detecção empregado. No sistema proposto, a amostra é utilizada para atualizar o valor da média e do desvio padrão.

Neste trabalho, novas amostras válidas são usada para recalcular a média, μ, e o desvio padrão, σ, usados no bloco de detecção de anomalia da amostra.

Perfis de usuários

Assuma-se, ainda, a presença de usuários que contribuem com o sistema de maneira correta, ou seja, de fato contribuem colaborativamente com amostras válidas e não-anômalas. Porém, assume-se também a presença de usuários que inserem amostras

anômalas no sistema. Note que tais usuários podem inserir amostras anômalas como consequência de ações maliciosas ou também como consequência de falhas em seus dispositivos de sensoriamento. Essas falhas podem ocorrer, mesmo que o usuário não seja malicioso, como consequência de algum problema de formatação dos dados, algum defeito físico no sensor ou, ainda, devido a alguma falha durante o processo de sensoriamento e transmissão. Já os usuários maliciosos podem enxergar algum tipo de vantagem ao inserir dados falsos ou apenas podem estar interessados em prejudicar a operação dos sistemas de sensoriamento participativo. Dessa forma, eles coletam dados inconsistentes e irreais, inserindo-os no sistema e modificando o resultado final do sensoriamento.

O principal objetivo do sistema proposto é identificar a presença de sensores que estejam inserindo dados de velocidade acima e abaixo do valor esperado. Esses dados poderiam ser provenientes de usuários maliciosos com o objetivo de burlar o sistema de trânsito, impedindo seu bom funcionamento e prejudicando os usuários. Por exemplo, caso o usuário consiga aumentar o valor esperado para a velocidade de um determinado trecho na cidade, os efeitos poderiam influenciar a tomada de decisão e interferir no escoamento do fluxo de veículos, aumentando, por exemplo, o congestionamento e o tempo de deslocamento dos usuários. Como as velocidades médias não são muito elevadas, velocidades baixas que possam ser consideradas anômalas foram encontradas em apenas alguns instantes na simulação e são apresentadas no Capítulo 5. Observe que a metodologia proposta é agnóstica a essa característica e que a anomalia é tão impactante ao sistema quanto mais distante dos valores esperados estiver. Portanto, amostras anômalas com valores dentro do intervalo de amostras esperadas podem não ser descartadas.

Assume-se a existência de quatro tipos de usuários que introduzem amostras anômalas. O primeiro e segundo tipo introduzem amostras com a presença de *anomalia permanente*, podendo introduzir apenas amostras com velocidades elevadas ou apenas amostras com velocidades baixas. Esses dois tipos de usuários inserem amostras que apresentam sempre a consistentemente velocidade e, por isso mesmo, são chamados de *usuários com anomalia permanente*, divididos em: *usuários com anomalia permanente em máxima* e *usuário com anomalia permanente em mínima*. O terceiro tipo de usuário introduz amostras que intercalam velocidades elevadas e velocidades médias, isto é, eles introduzem *anomalias intermitentes* e recebem o nome de *usuário com anomalia intermitente em máxima*. Por fim, o quarto tipo de usuário introduz amostras que intercalam velocidades baixas e velocidades médias e, consequentemente, recebem o nome de *usuário com anomalia intermitente em mínima*. Os quatros tipos de usuários com anomalia podem introduzir pequenas contribuições ao longo da rota. No cenário estudado, os usuários com anomalia atuam separadamente. O valor de velocidade anômala, $v_{anomala}$, inserido em cada

contribuição pode estar no intervalo $0 \leq v_{anomala} \leq v_{max}$, onde v_{max} pode ser a velocidade máxima permitida na cidade. O número de amostras inseridas varia gradualmente entre 0% e 100% do número máximo de amostras existentes no sistema durante o período analisado. Os resultados para os quatros tipos de perfis são apresentados no Capítulo 5.

Capítulo 4

Conjunto de Dados

Este capítulo descreve e analisa o conjunto de dados utilizado e suas principais características. A análise visa aumentar a compreensão e caracterizar a rota utilizada para o estudo.

4.1 Descrição do conjunto de dados

A avaliação da metodologia proposta é feita através das análises do impacto da variação das taxas de amostragem e do impacto da inserção de dados anômalos no sistema. Para tanto, utiliza-se o conjunto de dados Ad Hoc City [17], que contém diversos traços diários de um mês de movimento da frota de ônibus de passageiros da cidade de Seattle, no estado de Washington, nos EUA. A coleta foi realizada entre 31/10/2001 e 02/12/2001, totalizando 125 MB de dados. O conjunto de dados representa o movimento real dos ônibus, sendo os eventos de mobilidade de um ônibus id_{bus} que percorre uma rota id_{route} registrados no formato $\langle d, t, id_{bus}, id_{route}, x, y \rangle$, onde x,y são as coordenadas cartesianas (em pés) da posição do ônibus na data d no horário t. Existe aproximadamente 1 amostra por rota a cada 20 segundos. O traço dos ônibus foi escolhido, pois é um dos poucos que oferece informações coletadas de um ambiente urbano, tipicamente utilizado em aplicações de sensoriamento colaborativo, além de conter a classificação das rotas, que permite a classificação e definição de uma determinada RoI. Outros traços que coletem tipos diferentes de dados poderiam ser utilizados na análise. Porém por questões de disponibilidade, até onde se sabe, apenas o dos ônibus preenchia os requisitos desejados.

Neste trabalho, analisa-se uma quarta-feira típica, 31/10/2001, que inclui informações de mobilidade de 236 rotas de ônibus. As velocidades dos ônibus não são fornecidas no conjunto de dados original e, portanto, são previamente calculadas e anexadas a cada tupla de mobilidade. Para realizar isso, considera-se a posição do ônibus e o intervalo de tempo entre duas entradas consecutivas do mesmo ônibus. Assim, tendo em conta duas entradas consecuti-

vas do ônibus id_{bus} nos instantes t_1 e t_2 no mesmo dia d, respectivamente, $\langle d, t_1, id_{bus}, id_{route}, x_1, y_1\rangle$ e $\langle d, t_2, id_{bus}, id_{route}, x_2, y_2\rangle$, a velocidade do ônibus é calculada como $\frac{\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}}{t_2-t_1}$. Essa velocidade é anexada à segunda entrada, a do instante t_2, e o mesmo procedimento é repetido considerando instantes t_2 e t_3 e assim por diante até a última amostra para o ônibus id_{bus}. O conjunto de dados é totalmente enriquecido depois de computar as velocidades de todos os ônibus. Assuma-se que todas as entradas no conjunto de dados representam as contribuições dos usuários para o sistema de detecção participativa. Assim, no cenário estudado existem 376.491 amostras, com uma média de 1.595 amostras por rota. A Figura 4.1 mostra a função de distribuição cumulativa (*Cumulative Distribution Funcion* – CDF) do número de amostras por rota existentes no conjunto de dados. Observa-se que 90% das rotas apresentam até 4.000 amostras, aproximadamente. Este trabalho considera apenas a rota $id_{route} = 007$ que é a rota que possui a maior quantidade de amostras, totalizando 15.601 amostras.

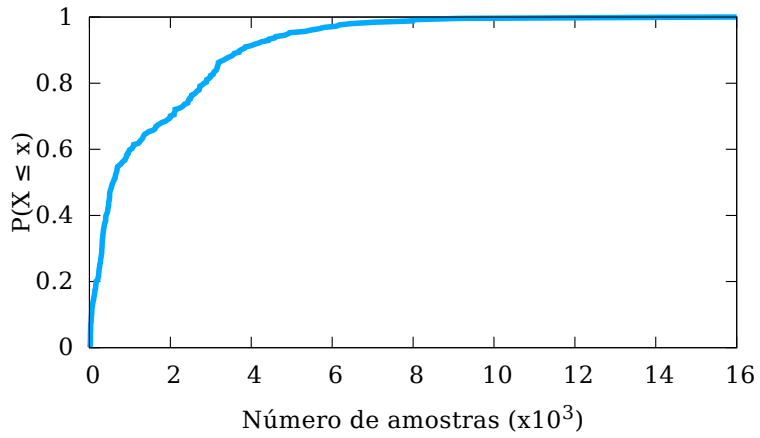


Figura 4.1: CDF do número de amostras por rota do dia 31/10/2001 existentes no conjunto de dados usado neste trabalho.

A velocidade máxima dos ônibus na região metropolitana de Seattle é de aproximadamente 56 km/h [42]. Os cálculos realizados das velocidades dos ônibus com base nos deslocamentos registrados nos dados originais revelam alguns valores irreais, que não são contabilizados (Figura 3.6). Para tanto, são executadas duas avaliações. Primeiramente verifica-se se a velocidade de uma amostra, v_i, está no intervalo [0,v_{max} km/h] que define uma faixa de velocidades plausíveis para o cenário. Neste trabalho, considera-se $v_{max} = 110$ km/h, que é a velocidade máxima permitida na cidade de Seattle [42]. Além disso, verifica-se se o posicionamento do ônibus registrado pelo GPS está dentro das coordenadas da cidade de Seattle. A rota $id_{route} = 007$ escolhida também passa pelo procedimento de validação, que resulta na exclusão de 55 amostras. As 55 amostras equivalem a 0,35% do total de

amostras da rota, o que é considerado uma quantidade desprezível.

4.2 Características da rota $id_{route} = 007$

Nesta seção, as características da rota escolhida são detalhadas, a fim de facilitar a análise da metodologia proposta. Essa rota se inicia na *Prentice Street/Rainier Beach*, localizada ao sul de Seattle, e termina em *Downtown Seattle*, ao norte, totalizando aproximadamente 15,45 km de extensão da rota, sendo a distância cartesiana de ponto a ponto aproximadamente 17 km. A duração média de uma viagem nessa rota em dias de semana é de 44 minutos [43]. A Figura 4.2 mostra a quantidade de ônibus utilizados na Rota 007 e a quantidade de amostras por ônibus dessa rota. Destaca-se também na figura a quantidade média de amostras por ônibus. Nessa figura, o eixo-x representa a identificação do ônibus, enquanto o eixo-y define a quantidade de amostras coletadas por ônibus. No dia analisado, foram registrados 51 ônibus, sendo a quantidade média de amostras por ônibus igual a 305 coletas. Os ônibus 2391, 3562 e 5055 apresentam o menor número de amostras, participando com 9, 19 e 22 amostras, respectivamente. Por sua vez, os ônibus 4004, 4026 e 4042 realizam o maior número de amostras, participando com 1304, 931 e 848 amostras, respectivamente.

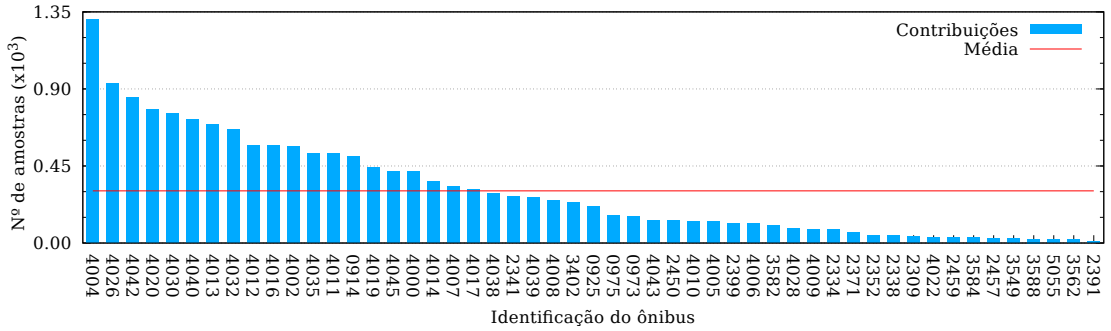


Figura 4.2: Total de contribuições por ônibus da Rota 007, com maior contribuição no dia 31/10/2001.

4.2.1 Velocidade média dos ônibus

Considerando-se a rota $id_{route} = 007$ inteira, a velocidade média no decorrer do dia é de 19,91 km/h. Para melhorar o refinamento dessa medida, permitindo uma avaliação mais detalhada da rota, primeiramente divide-se a rota em trechos de tamanho 100 m. Em seguida, para cada trecho, calcula-se a velocidade média dos ônibus durante todo o dia. A Figura 4.3 mostra a velocidade média obtida para cada trecho. Nessa figura, observa-se que a velocidade média em alguns pontos da

rota id_{route} = 007 no decorrer do dia pode alcançar até 34 km/h. Isso indica que há congestionamentos na região ou que há mais pontos de parada, como sinais de trânsito ou pontos de ônibus. Essa conclusão é possível uma vez que os ônibus não se movem na velocidade máxima permitida para esse tipo de veículo na cidade de Seattle [42] em nenhum trecho por completo por todo o tempo. Porém, nota-se na figura que há trechos onde a velocidade média é mais elevada e trechos onde a velocidade média é menos elevada.

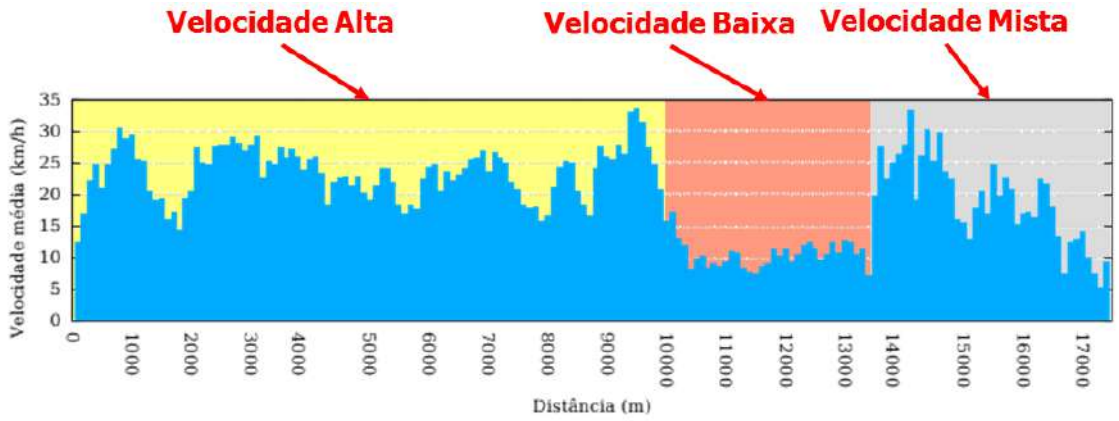


Figura 4.3: Velocidade média dos ônibus da rota $\boldsymbol{id_{route} = 007}$ obtida a partir da subdivisão da rota em trechos de **100** m.

Tabela 4.1: Características das regiões definidas com base na velocidade média dos ônibus.

Região	**Descrição**	**# de amostras**	**# de ônibus**	**Velocidade média (km/h)**
1	Velocidade Alta	6.809	50	17,17
2	Velocidade Baixa	7.172	50	7,18
3	Velocidade Mista	1.565	32	10,63

Analisando a Figura 4.3, é realizado o agrupamento dos trechos em regiões com base na velocidade média dos ônibus, a saber: *Velocidade Alta*, *Velocidade Baixa*, e *Velocidade Mista*. Usando o sistema de coordenadas da Figura 4.3, a Região de Velocidade Alta é de 0 $\approx$ 9.700, a Região de Velocidade Baixa é de $\approx$ 9.700 a $\approx$ 13.700, e a Região de Velocidade Mista é de $\approx$ 13.700 a $\approx$ 17.300. A Tabela 4.1 apresenta a nomenclatura adotada para cada região, além do número total de amostras, a quantidade de ônibus em cada uma e a velocidade média dos ônibus que fazem parte dela. A divisão em regiões permite classificar a rota em segmentos de características semelhantes, para reduzir a variabilidade dos dados sensoriados. Nota-se que essa divisão pode ser realizada em todas as rotas do conjunto de dados.

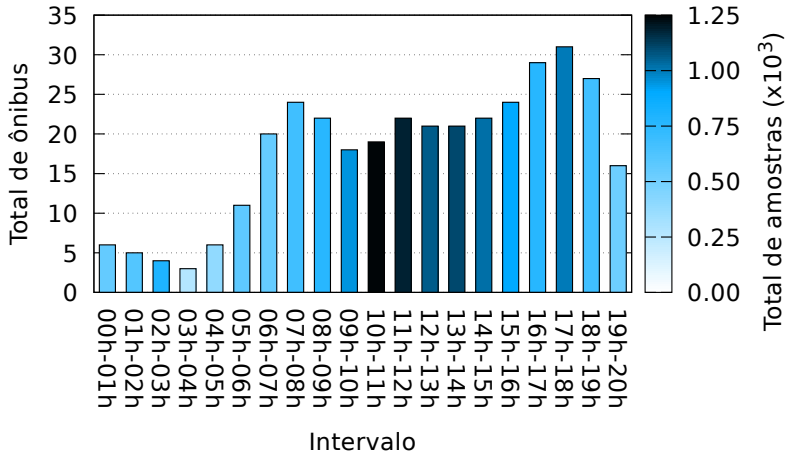


Figura 4.4: Número de ônibus e de amostras por hora na rota $\boldsymbol{id_{route} = 007}$ no dia 31/10/2001.

4.2.2 Total de ônibus e amostras

A Figura 4.4 apresenta o número de ônibus e de amostras por hora na rota $\boldsymbol{id_{route} = 007}$ no dia 31/10/2001. Nessa figura, o eixo-x representa o intervalo de hora analisado, enquanto o eixo-y diz respeito ao número total de ônibus existentes naquele intervalo. A escala de cor mostra o número total de amostras em cada intervalo de tempo, sendo que quanto mais escuro for o tom de azul, maior é o número de amostras existentes. Essa figura mostra que a quantidade de ônibus que contribuem ao longo do dia varia entre 3 e 31 ônibus, dependendo da hora analisada do dia. Contudo, apesar disso, a quantidade de amostras obtidas em cada hora não é proporcional ao número de ônibus participantes da coleta. Por exemplo, o maior número de ônibus participantes ocorre entre 17 e 18 horas, totalizando 31 ônibus, que são capazes de coletar 998 amostras. Já o maior número de amostras é encontrado entre 10 e 11 horas da manhã, alcançando um total de 1.240 amostras, realizadas por 19 ônibus. O menor número de amostras ocorre entre 3 e 4 horas da manhã, totalizando 265 amostras realizadas por apenas 3 ônibus. A média de ônibus por hora nesse dia é igual a 19 e a média de amostras por hora é de 777.

A Figura 4.5 mostra a quantidade média de amostras por ônibus em cada hora do dia. A média total é de 58 amostras por veículo, sendo o intervalo de 2 às 3 horas da manhã o que possui a maior média por veículo, que contém 199 amostras por veículos. Isso ocorre devido ao número pequeno de ônibus existentes nessa hora comparados à quantidade de contribuições que eles conseguiram realizar. No intervalo de 10 às 11 horas da manhã, que apresenta o maior número de amostras (Figura 4.4), a média de amostras por ônibus é igual a 59. É interessante que essa média não seja muito alta para que a quantidade de dados enviados não seja muito elevada, mas que também não seja muito baixa para que seja representativa para o

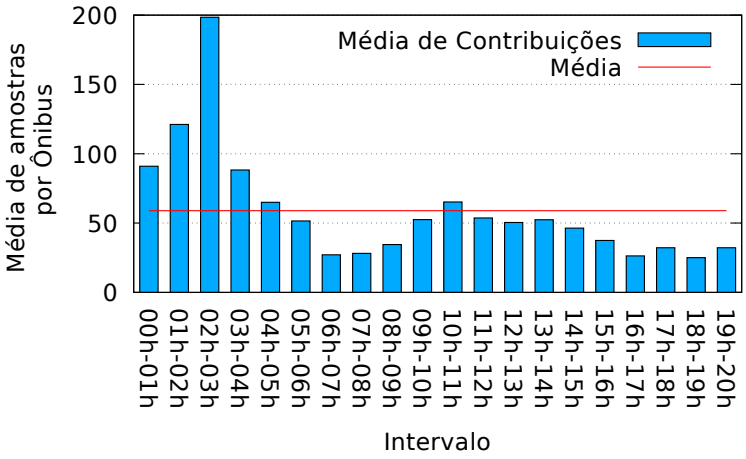


Figura 4.5: Quantidade média de amostras por ônibus em cada hora do dia.

conjunto de dados do sensoriamento.

4.2.3 Densidade das Amostras

A avaliação da densidade dos veículos na rota $id_{route} = 007$ é feita utilizando-se a divisão em áreas 50×50 m^2. A Figura 4.6 mostra o resultado obtido. Observa-se que ocorre uma maior concentração de veículos em pontos distintos da rota, sendo a densidade média de toda a rota igual a 0,0033 veículos por 50×50 m^2, isto é, existem aproximadamente 8 ônibus por área de 50×50 m^2.

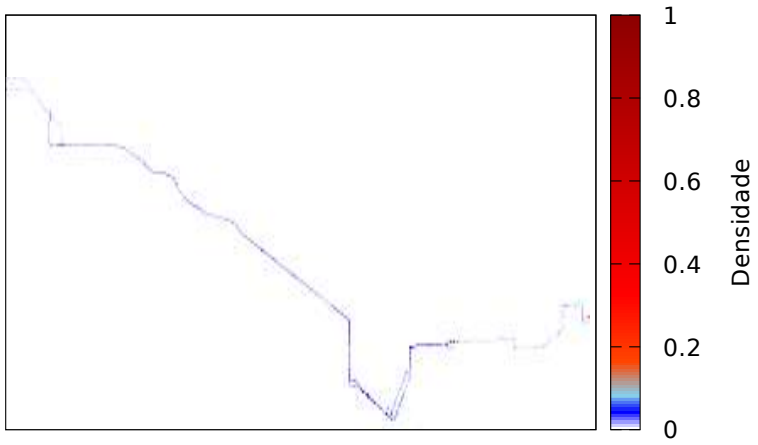


Figura 4.6: Mapa da densidade dos ônibus na rota $id_{route} = 007$.

Capítulo 5

Detecção de Anomalia

O foco principal desse capítulo é verificar a viabilidade da detecção de dados anômalos. Para realizar isso, os quatro perfis de usuários atacantes definidos na Seção 3.2.2 são utilizados em simulações distintas. Em cada um dos cenários, os usuários anômalos injetam diferentes quantidades de amostras no sistema.

Nas seções subsequentes, investiga-se o número de amostras anômalas necessárias para influenciar o resultado final e o impacto dessas amostras adicionais na vazão da rede. Note que, apesar de a análise ter sido feita para amostras de velocidade, a metodologia proposta é genérica o suficiente para ser reutilizada em outros cenários, utilizando quaisquer outros tipos de medidas, como poluição, luminosidade, pressão, dentre outras.

5.1 Impacto das amostras anômalas nas velocidades médias

Esta seção investiga o número de amostras anômalas necessárias para alterar o resultado final do sensoriamento. Assume-se que cada usuário anômalo pode injetar amostras para modificar a velocidade média da região, aumentando ou diminuindo o valor médio real. Por um lado, para reduzir a velocidade média da região, o usuário deve contribuir com um valor de velocidade menor do que a média, e o menor possível é 0 km/h. Por outro lado, para aumentar a velocidade média da região, o usuário deve injetar amostras com valores de velocidade mais altos e o mais alto possível é 57 km/h, que é a velocidade máxima dos ônibus na cidade de Seattle [42], no cenário avaliado. Assume-se que esse tipo de usuário utiliza os valores estipulados nas Tabelas 5.1 e 5.2. Ambas as tentativas de alterar a velocidade média podem causar problemas no sistema se não forem detectadas. Isso corrobora a necessidade da etapa de *detecção de anomalia* da metodologia proposta na Seção 3.2.2.

A Figura 5.1 mostra a velocidade média dos ônibus em cada intervalo de 1 hora.

Tabela 5.1: Características das inserções anômalas dos atacantes com carga máxima.

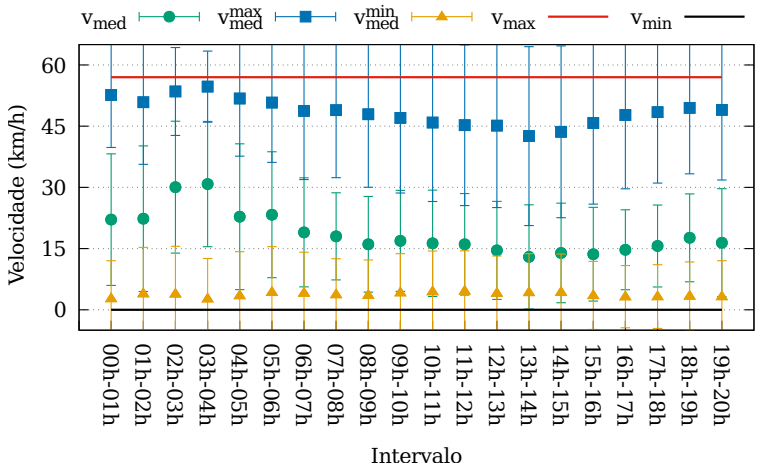
		Valor da contribuição (km/h)	
Região	**# de amostras por usuário anômalo**	**Anomalia permanente**	**Anomalia intermitente**
Velocidade Alta	29	57	Média e 57
Velocidade Baixa	33	57	Média e 57
Velocidade Mista	6	57	Média e 57

Tabela 5.2: Características das inserções anômalas dos atacantes com carga mínima.

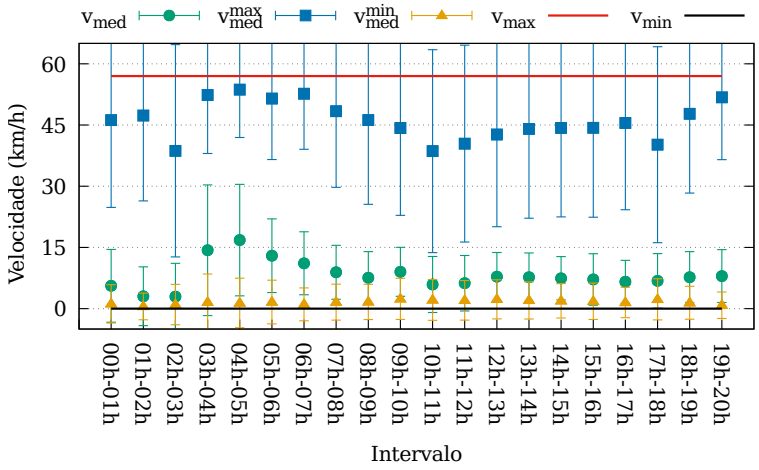
		Valor da contribuição (km/h)	
Região	**# de amostras por usuário anômalo**	**Anomalia permanente**	**Anomalia intermitente**
Velocidade Alta	29	0	0 e Média
Velocidade Baixa	33	0	0 e Média
Velocidade Mista	6	0	0 e Média

Cada tipo de ponto representa um cenário distinto. Os pontos v_{med} representam os resultados quando o sistema está livre de dados anômalos. Os pontos v_{med}^{max}, por sua vez, representam a velocidade média considerando amostras anômalas contendo velocidades máximas. Finalmente, os pontos v_{med}^{min} representam a velocidade média considerando amostras anômalas contendo velocidades mínimas. Esses dois últimos resultados foram obtidos utilizando-se um adicional de 777 amostras anômalas, pois este é o número médio de amostras por hora (Seção 4.2.2). As barras verticais representam o valor médio $\pm 3\sigma$. Ainda podemos notar que as linhas horizontais v_{min} e v_{max} representam, respectivamente, o menor e maior valor possível para o cenário avaliado. Observe que no intervalo 2-4h, a região da Velocidade Mista não possui amostras (Figura 5.1(c)).

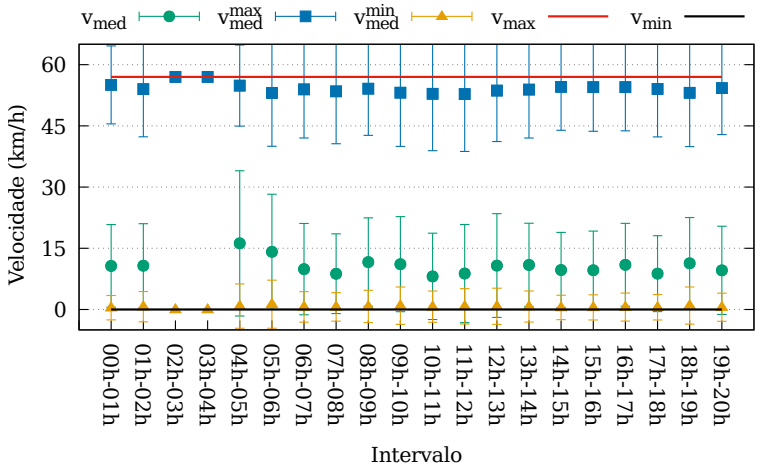
Os resultados indicam que as amostras anômalas que carregam a velocidade mínima (v_{min}), se usadas, podem diminuir a velocidade média de cada região. Considerando que essas amostras anômalas são raramente detectáveis usando a abordagem simples de $v_{med} - 3\sigma$, o impacto seria a presença de mais falsos positivos e falsos negativos. Isso ocorre porque a velocidade média de cada região não é alta, de modo que amostras anômalas com velocidades negativas, que não são válidas, seriam necessárias para diminuir a velocidade média de cada região a ponto de ficar menor que $v_{med} - 3\sigma$. Alternativamente, seria necessário utilizar métodos mais sofisticados para detectar como anomalia o deslocamento para baixo da velocidade



(a) Região de Velocidade Alta.



(b) Região de Velocidade Baixa.



(c) Região de Velocidade Mista.

Figura 5.1: Impacto das amostras anômalas no sistema de detecção em cada região considerando a injeção de 777 amostras anômalas por hora.

média real.

Considerando agora as amostras contendo a velocidade máxima (v_{max}), todas

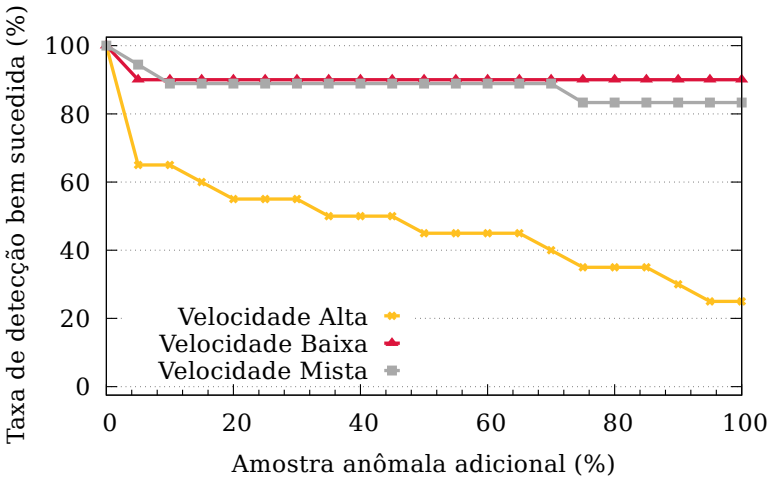
elas já são detectáveis em todas as regiões usando o método $v_{med}+3\sigma$. Em oposição às amostras de velocidade mínima, as amostras de velocidade máxima são substancialmente maiores do que a velocidade média real. Se, no entanto, essas amostras são usadas para calcular a velocidade média (v_{med}^{max}), o resultado final mudaria completamente. Todas as amostras anômalas ficariam dentro do intervalo de expectativas e quase todas as amostras legítimas seriam anômalas. Claro, o impacto depende do cenário e do número de amostras anômalas, mas a tendência observada ainda seria verdadeira. A detecção de amostras de velocidades máximas anômalas e de velocidades mínimas anômalas, usando os perfis de usuário previamente definidos conforme as Tabelas 5.1 e 5.2 ainda será investigada.

5.2 Impacto das amostras anômalas no método de detecção

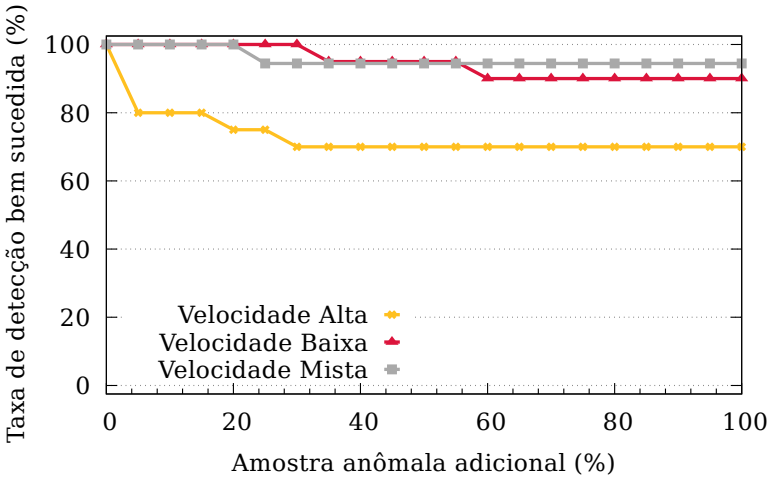
Foram realizadas 21 execuções de simulação para cada combinação de usuário, assumindo que o sistema está inicialmente livre de dados anômalos. Em cada execução, adiciona-se às amostras legítimas de cada hora, 5% de amostras anômalas para recalcular a média e o desvio padrão. Estes 5% são proporcionais ao maior número original de amostras por hora em cada região. Na última execução de simulação, tem-se a mesma quantidade de amostras legítimas e anômalas, o que resulta em 100% de amostras anômalas adicionais.

A quantidade de amostras anômalas, isto é, os valores das amostras anômalas por hora e região estão resumidos nas Tabelas 5.1 e 5.2, onde representam, respectivamente, os usuários com carga máxima e mínima. Note que as velocidades utilizadas nas amostras anômalas são diferentes conforme o perfil do usuário. A velocidade média real pode mudar de acordo com a região e no intervalo atual de uma hora. Vale observar que os dados anômalos inseridos vieram do próprio conjunto de dados, ou seja, não foi criado nenhum dado falso, todos os dados pertencem ao próprio conjunto de dados, foi retirado e acrescentado de forma aleatória no mesmo intervalo de tempo e espaço do dado original, alterando apenas o valor da velocidade.

A Figura 5.2 apresenta a influência das amostras anômalas adicionadas no sistema de detecção. O eixo-y representa a fração de horas do período analisado de um dia em que as anomalias foram detectadas com sucesso. Cada curva nesta figura representa uma das regiões previamente definidas. Observe que detectar uma amostra de velocidade máxima anômala depende exclusivamente da média e do desvio padrão calculado na hora atual. Por um lado, no caso de anomalia máxima, permanente ou intermitente, se $v_{max} > v_{med}^{max} + 3\sigma$, é possível garantir que a amostra anômala é detectada. Por outro lado, no caso de anomalia mínima, também permanente ou



(a) Usuários com inconsistência permanente com carga máxima.



(b) Usuários com inconsistência intermitente com carga máxima.

Figura 5.2: Influência das amostras anômalas adicionais, usadas para recalcular a velocidade média a cada tempo, na taxa de detecções bem sucedidas em cada região usando o método de detecção simples empregado na metodologia proposta.

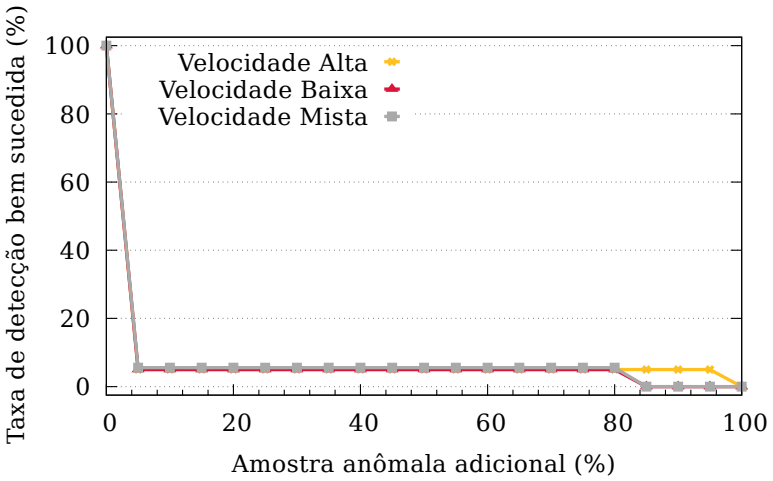
intermitente, se $v_{min} < v_{med}^{min} - 3\sigma$, é possível garantir que a amostra anômala é igualmente detectada.

A Figura 5.2(a) mostra a taxa de detecções bem sucedidas em função do número de amostras anômalas injetadas no sistema para usuários com anomalia permanente em máxima. Observa-se que a taxa de detecção bem sucedida da região de alta velocidade cai mais rapidamente do que as demais. Isso ocorre porque a velocidade média está mais próxima da velocidade máxima anômala. Assim, com algumas amostras anômalas, já temos uma redução nas detecções anômalas. Em todas as regiões, o sistema é capaz de detectar todas as amostras anômalas usando apenas as amostras legítimas para calcular a média e os valores de desvio padrão, ou seja, quando $x = 0\%$. A região de baixa velocidade é a que tem o melhor resultado, como esperado, já que a velocidade média está mais distante do valor anômalo injetado.

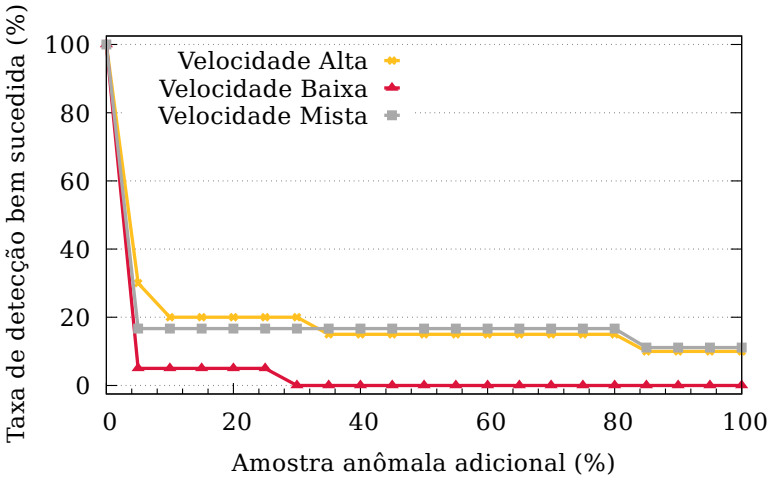
Mesmo na região de melhor resultado, porém, o sistema já reduz a taxa de detecções bem sucedidas para 90%, quando os valores de desvio padrão e média começam a variar utilizando 5% adicionais de amostras anômalas. Essa redução no início é uma consequência da alta média e desvio padrão de valores legitimamente encontrados durante o intervalo entre $3 - 5h$. A região mista exibe um comportamento intermediário considerando as duas regiões anteriores. Uma observação importante é que a taxa de detecção bem sucedida nunca atinge 0%, mesmo quando a média e os valores de desvio padrão são altamente manipulados, ou seja, quando $x = 100\%$.

A Figura 5.2(b) mostra um resultado semelhante, mas para anomalias intermitentes. Como esperado, em geral, é necessário um maior número de amostras anômalas para reduzir a taxa de detecções bem sucedidas. Isso ocorre porque algumas amostras anômalas têm o valor da velocidade média legítimo. Uma diferença notável, no entanto, entre as Figuras 5.2(a) e 5.2(b) é que, mesmo na presença de amostras anômalas adicionais de 100%, a taxa de detecções bem sucedidas nunca atinge valores inferiores a 70%. As duas Figuras representam o comportamento quando os usuários inserem dados com carga máxima.

As Figuras 5.3(a) e 5.3(b) apresentam a influência das amostras anômalas adicionadas ao sistema de detecção, sendo os usuários permanente em mínimo e intermitente em mínimo, respectivamente. Os resultados mostram que a queda da taxa bem sucedida de detecção acontece mais rapidamente em ambos os casos em comparação aos resultados das Figuras 5.2(a) e 5.2(b). Isso ocorre mais uma vez porque a velocidade anômala, 0 km/h, está mais próxima da velocidade média e, portanto, a condição $\mathrm{v}_{\mathrm{min}} < \mathrm{v}_{\mathrm{med}}^{\mathrm{min}} - 3\sigma$ deixa de ser atendida mais rapidamente. Na Figura 5.3(a), a partir de 80%, observa-se que a taxa de detecção bem sucedida da região de baixa e mista velocidade cai um pouco mais rapidamente do que a da região de alta velocidade. Isso ocorre porque a velocidade média está mais próxima da velocidade mínima anômala. Assim, similar ao usuário com anomalia permanente em máxima, com algumas amostras anômalas, já é possível perceber uma redução nas detecções anômalas. Em todas as regiões, o sistema é capaz de detectar todas as amostras anômalas usando apenas as amostras legítimas para calcular a média e os valores de desvio padrão, ou seja, quando $x = 0\%$. No entanto, a taxa de detecções bem sucedidas com $x \geq 5\%$ é muito baixa, alcançando na maior parte do experimento 5% de taxa de sucesso. A Figura 5.3(b) mostra um resultado semelhante para anomalia intermitente, porém, com um desempenho melhor que o dos usuários com anomalia permanente. É necessário um maior número de amostras anômalas para reduzir a taxa de detecções bem sucedidas. O sistema tem as regiões de velocidade alta e mista com um desempenho melhor e começam a ter uma queda de desempenho em $x = 5\%$ que representa, respectivamente, 30% e 20% de taxa de detecções bem sucedidas. Quando $x = 80\%$ o desempenho do sistema cai para 10%



(a) Usuários com inconsistência permanente com carga mínima.



(b) Usuários com inconsistência intermitente com carga mínima.

Figura 5.3: Influência das amostras anômalas adicionais, usadas para recalcular a velocidade média a cada tempo, na taxa de detecções bem sucedidas em cada região usando o método de detecção simples empregado na metodologia proposta.

nas duas regiões. Já para a região de velocidade baixa, ocorre a não detecção dos dados em $x = 25\%$. Vale lembrar que as duas figuras representam o comportamento quando os usuários inserem dados com carga mínima.

5.3 Impacto das amostras anômalas na carga da rede

Esta análise considera que os dados são transmitidos usando o padrão IEEE 802.11p, que é usado em redes veiculares para fornecer comunicações entre OBUs e entre OBUs e RSUs. A ideia é avaliar o impacto das amostras adicionais na vazão da rede. As RSUs podem ser conectadas a um controlador central para ter uma visão completa da rede. A camada física do IEEE 802.11p opera a uma taxa de transmissão

máxima de 27 Mb/s, e mínimo de 3 Mb/s, usando uma frequência de operação de 5,890 GHz e largura de banda de 10 MHz [44].

Considera-se que cada amostra tenha 480 Bytes [17] e o número médio de amostras inseridas por hora é 777, totalizando uma carga adicional de 372,96 kBytes. As amostras anômalas adicionais podem introduzir problemas de rede, pois a carga total atingiria 745,92 kBytes para 100% de amostras extras. Em uma hora, o rendimento agregado da rede precisaria de 1,67 Mb/s, o que pode exceder a capacidade disponívelx dependendo da configuração da rede. Os resultados obtidos são mostrados na Figura 5.4. Nessa figura, o eixo-x representa o adicional de amostras anômalas inseridas no sistema, enquanto o eixo-y diz respeito à carga total extra devido a essas amostras adicionais.

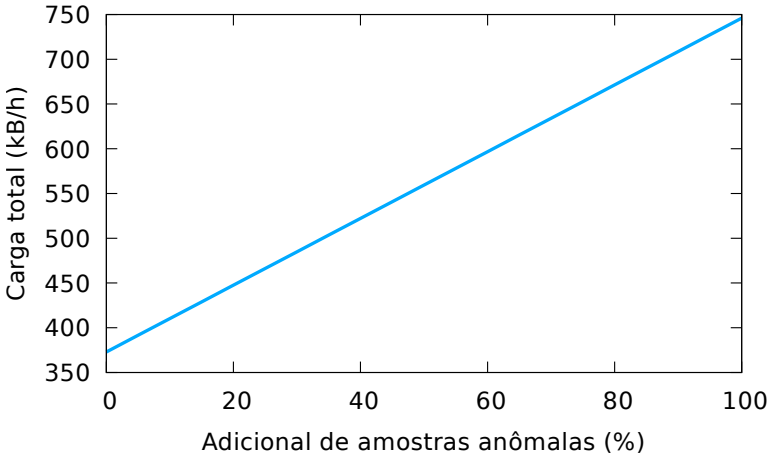


Figura 5.4: Influência das amostras anômalas na carga da rede.

Capítulo 6

Adaptação da Taxa de Amostragem

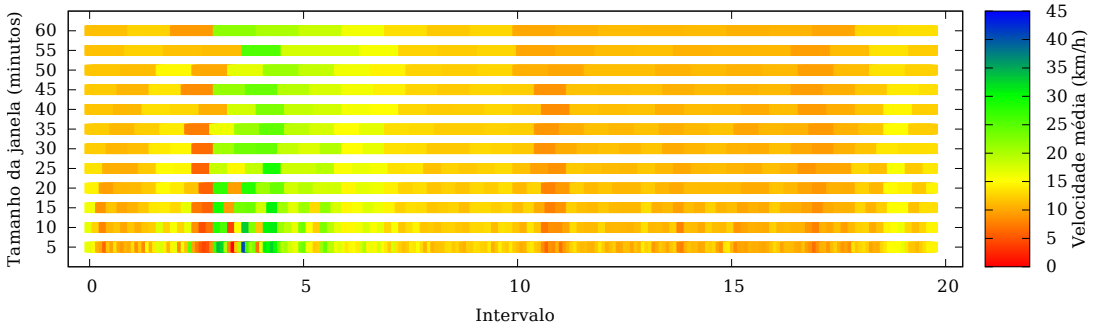
Este capítulo apresenta a avaliação do sistema proposto através da definição do tamanho para a janela de tempo Δt. Em seguida, usando o intervalo de tempo selecionado, prossegue-se com a análise da proposta.

6.1 Análise do tamanho da janela de tempo Δt

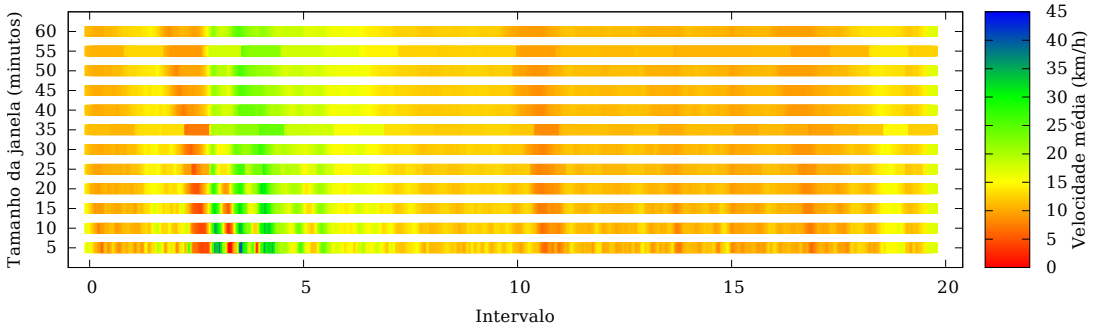
A janela de tempo Δt é utilizada para obter resultados como a velocidade média dos veículos ou o número de amostras sensoriadas por período. O tamanho ideal dessa janela é investigado utilizando-se duas abordagens: janela de tempo deslizante e janela de tempo fixo. Em ambos os casos, considera-se que a primeira janela se inicia no instante de tempo 00:00 e a última janela é finalizada no instante de tempo 20:00, que é a duração total do traço do dia 31/10/2001 do conjunto de dados de Seattle (Capítulo 4).

Diversos valores para o tamanho da janela de tempo Δt, tal que $5 \leq \Delta t \leq 60$ min com passos iguais a 5 min são analisados. A Figura 6.1(a) mostra o resultado para janelas de tamanho fixo no tempo e a Figura 6.1(b), os resultados para janela deslizante no tempo. A cor de cada par (x,y) nessas figuras representa a velocidade média calculada na hora x usando um intervalo de y minutos. Em ambas as abordagens, observa-se que a utilização de janelas menores captura maiores variações nas medidas de velocidade média. As diferenças entre os resultados para cada tipo de janela são sutis, sendo mais perceptíveis ao se comparar o momento de transição entre duas janelas fixas consecutivas e o intervalo correspondente utilizando a janela deslizante. No entanto, nem a redução na granularidade da informação ao aumentar o tamanho da janela até 60 min, nem a perda de informação no momento de transição entre duas janelas consecutivas ao se utilizar janelas fixas, alteram signi-

ficativamente a interpretação do estado momentâneo da rota. Por essa razão, no restante deste trabalho utiliza-se uma janela fixa de 60 min para obtenção e análise dos resultados. Essa janela é a ideal pois reduz o número de processamentos a serem realizados pelo servidor.



(a) Janela de tamanho fixo no tempo.



(b) Janela deslizante no tempo.

Figura 6.1: Análise do tamanho da janela de tempo $\boldsymbol{\Delta t}$ comparando o uso de janelas de tamanho fixo e janelas deslizantes.

6.2 Desempenho do sistema de sensoriamento proposto

A metodologia proposta é avaliada através da análise da taxa de entrega e do erro da amostragem, considerando que a RoI é a rota $id_{route} = 007$ e a medida a ser sensoriada é a velocidade média dos ônibus em trechos pré-definidos. Esses trechos são definidos pelo administrador do sistema, e para nosso cenário, foi dividido em trechos de 100 m. A mobilidade e a comunicação no cenário são simulados utilizando o simulador de redes NS-3.26 (*Network Simulator version 3.26*) [45].

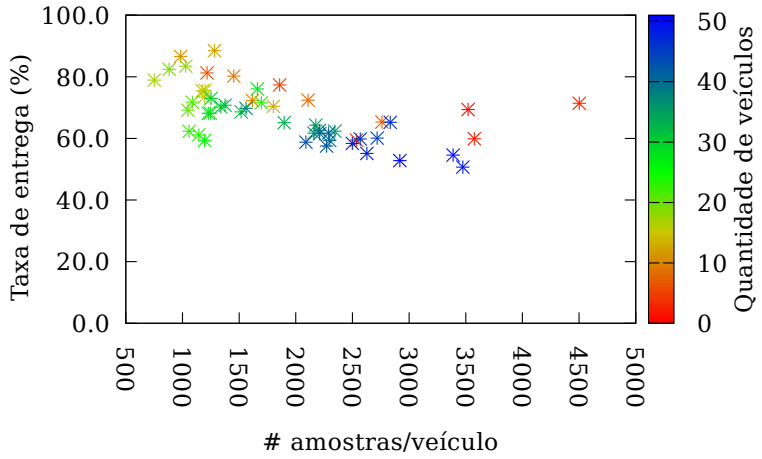
6.2.1 Avaliação da taxa de entrega

Essa análise verifica o compromisso entre o número de nós e a quantidade de dados transferidos por nó, de forma que a RoI seja coberta e as RSUs sejam capazes de

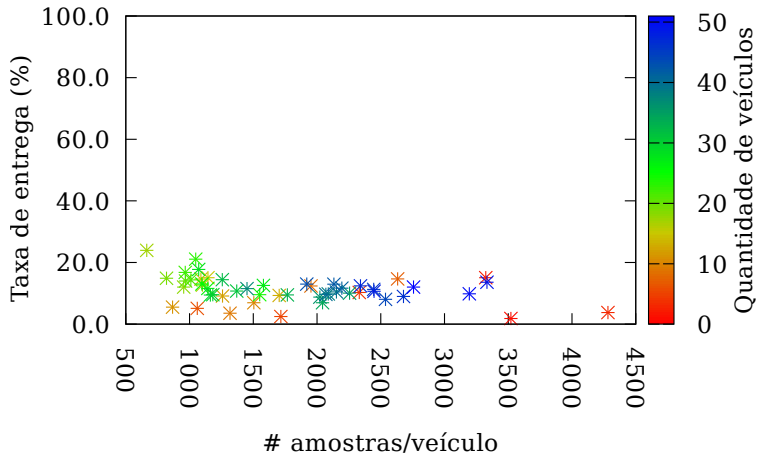
receber por completo todos os dados descarregados. Dessa forma, é possível fazer uma análise de capacidade da rede. Para isso, foram realizadas simulações variando o número de veículos e o número de amostras por veículo a serem entregues às RSUs #1 e #2. Ao final, calcula-se a taxa de entrega.

A Figura 6.2 apresenta a relação entre a taxa de entrega, a carga média transmitida dos ônibus para as RSUs e o número de veículos que competem pelo meio durante uma transferência de dados. As RSUs estão localizadas nas extremidades da rota $id_{route} = 007$, o que significa que há uma RSU (RSU #1) na extremidade direita e outra RSU (RSU #2) na extremidade esquerda da rota, conforme visto na Figura 4.3. A taxa de entrega na RSU #1 é alta para uma carga média por veículo mais baixa, por exemplo, 1.000 amostras por veículo, e diminui à medida que a carga média aumenta, chegando a $\approx 60\%$ para uma carga de ≈ 2.500 amostras por veículo. Isso ocorre devido à curta duração do contato com a RSU, sendo insuficiente para descarregar os dados. A quantidade de veículos competindo pelo meio também reduz a taxa de entrega. Por exemplo, para uma mesma carga, 1.500 amostras por veículo, a taxa de entrega é de $\approx 65\%$ com ≈ 30 veículos, e $\approx 80\%$ para ≈ 10 veículos. Essa redução ocorre devido ao aumento do número de colisões quando mais de um veículo tenta descarregar os dados ao mesmo tempo na mesma RSU. Note que para cargas elevadas como maiores que 3.500 amostras por veículo, apenas as configurações com um número menor de veículos competindo pelo meio conseguem entregar dados.

Na RSU #2 ambas as variáveis, quantidade de veículo e carga por veículo, influenciam a taxa de entrega de forma semelhante. Porém, mesmo com cargas mais baixas e menor quantidade de veículos, a taxa de entrega é pequena. A diferença entre os comportamentos das duas RSUs se deve ao menor tempo de contato entre os veículos e a RSU #2, que está localizada em uma região de maior velocidade da rota (Figura 4.3). Observa-se novamente que há alguma entrega de dados para cargas mais elevadas somente quando o número de nós competindo pelo meio é baixo. Isso significa que cargas mais elevadas são apenas possíveis de serem entregues em cenários com poucos carros competindo pelo meio. Para cargas baixas, porém, a taxa de entrega pode ser mais alta mesmo quando há mais carros competindo pelo meio. Essa última observação é interessante pois permite a entrega de maior carga de dados agregados, caso sejam utilizados vários veículos ao mesmo tempo com cargas baixas individuais. Sobrecarregar um único veículo pode não ser atraente sob o ponto de vista do usuário, já que pode levar mais rapidamente à exaustão de recursos do próprio usuário.



(a) RSU #1.



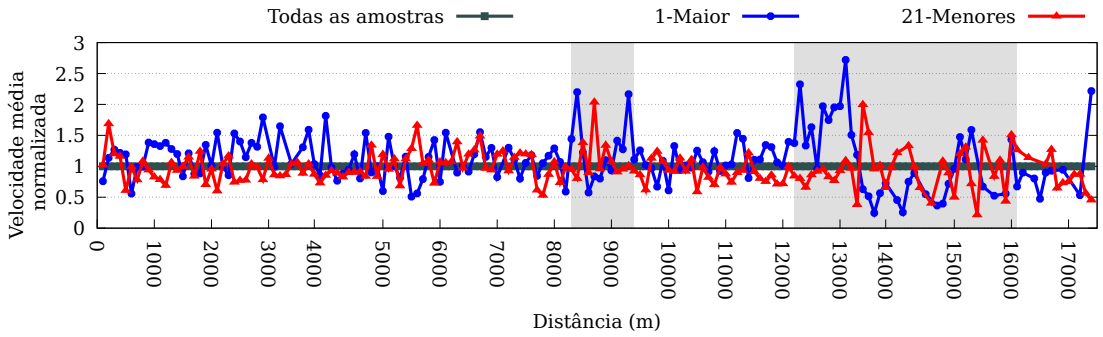
(b) RSU #2.

Figura 6.2: Taxa de entrega nas duas RSUs nas extremidades da rota $id_{route} = 007$ variando a quantidade de veículos que participam do sistema de sensoriamento e a carga de dados por veículo.

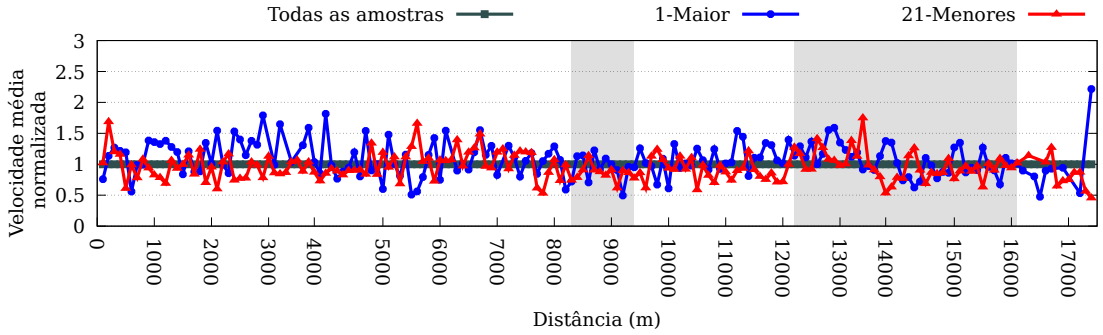
6.2.2 Avaliação do erro da amostragem

Em um cenário ideal, o sistema conhece exatamente a quantidade de dados necessários para reproduzir com baixo erro a informação sensoriada. Isso evitaria o problema de subamostragem, que pode levar a erros na medida; e de superamostragem, que pode representar desperdício de recursos. O sistema proposto neste trabalho não consegue determinar com certeza esses valores, mas é capaz de estimar quais regiões necessitam de mais ou menos coletas, de forma adaptativa, com base na análise da variabilidade dos dados nos diferentes trechos.

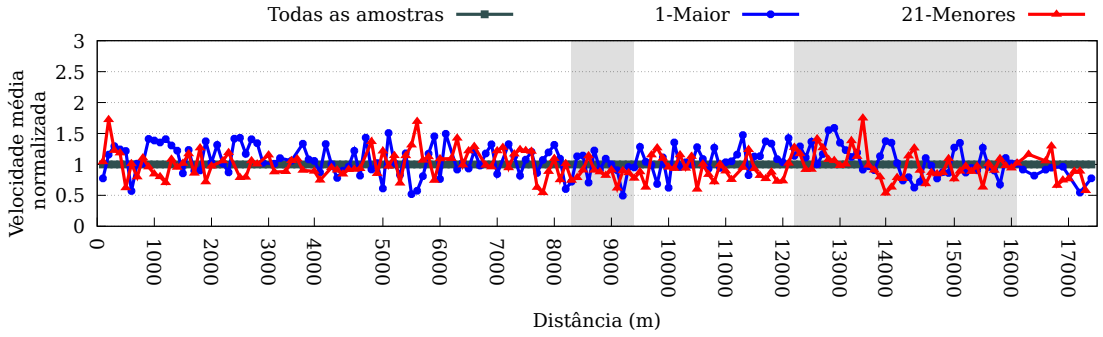
A Figura 6.3 mostra a variabilidade das medidas de velocidade média sem o uso do sistema, com o uso do sistema de sensoriamento proposto apenas na região cinza e com uso do sistema de sensoriamento proposto nas regiões cinza e branca, as regiões com alta variabilidade estão preenchidas em cinza, onde a variação em



(a) Sem o uso do sistema proposto, todos os trechos realizam amostragem média constante igual a R.



(b) Com o uso do sistema proposto, trechos com menor variação com taxa de amostragem igual a R e trechos com maior variação com taxa de amostragem igual a $2R$.



(c) Com o uso do sistema proposto, trechos com menor variação com taxa de amostragem igual a $0{,}5R$ e trechos com maior variação com taxa de amostragem igual a $2R$.

Figura 6.3: Velocidades médias normalizadas usando um único veículo ('1-Maior') e **21** veículos com a mesma carga ('21-Menores') nos trechos de **100** m. A região cinza representa os trechos com maiores diferenças de velocidades médias.

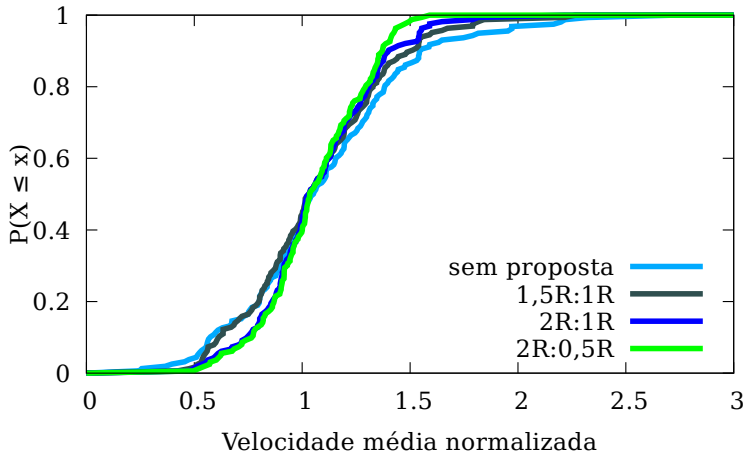
Tabela 6.1: Distribuição dos ônibus dentro dos conjuntos de ônibus '1-maior' e '21-menores'.

Conjunto	**# de identificação dos ônibus**
1-maior	4004
21-menores	2309, 2334, 2338, 2352, 2391, 2399, 2457, 2459, 3549, 3562, 3582, 3584, 3588, 4005, 4006, 4009, 4010, 4022, 4028, 4043 e 5055

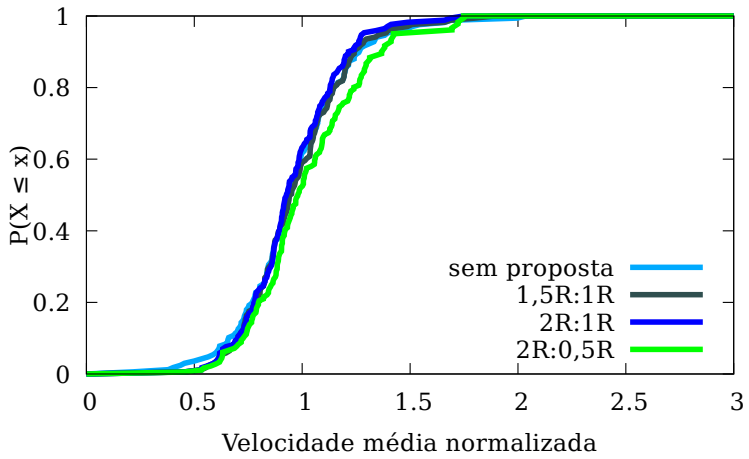
torno da linha horizontal de velocidade média normalizada igual a 1 é maior, já as regiões com baixa variabilidade estão preenchidas em branco. São apresentados os resultados para o sistema usando apenas o ônibus que coleta o maior número de amostras do conjunto de dados, chamado '1-Maior', e um conjunto de 21 ônibus que somados coletam a mesma quantidade de amostras do '1-Maior', chamado '21-Menores'. Essa separação faz necessário para analisar qual a melhor condição de definição dos participantes, ou seja, confiar em apenas 1 participante com muitos recursos ou dividir a contribuição em diversos participantes com recursos limitados. A Tabela 6.1 apresenta a identificação dos números pertencentes a cada um dos conjuntos. As médias estão normalizadas, sendo todas elas divididas pela velocidade média obtida com todas as amostras do conjunto de dados para cada trecho de 100 m. A adaptação da taxa de amostragem ocorre em dois momentos, um quando o sistema determina que mais coletas devem ser realizadas em uma região onde existe elevada variabilidade e outro quando o sistema detecta uma região com baixa variabilidade. A Figura 6.3(a) mostra os resultados considerando uma taxa de amostragem média ($R = \frac{\#\text{ de amostras}}{\#\text{ de trechos}}$) em todos os trechos de 7,5, ou seja, foram coletadas 1.304 amostras ao longo dos 174 trechos da rota $id_{route} = 007$. Note que há uma maior variação das médias nos trechos preenchidos com cinza. Dobrando a taxa de amostragem para $2R = 15$, apenas nos trechos de maior variação (preenchidos com cinza), já há uma melhora, como visto na Figura 6.3(b).

A adaptação da taxa de amostragem pode ser realizada também quando em trechos com baixa variação (região branca). Nesse caso, é possível reduzir a quantidade de dados a serem coletados na região com o objetivo de evitar superamostragem e, consequentemente, economizar recursos dos participantes do sistema colaborativo. A Figura 6.3(c) mostra os resultados considerando uma taxa de amostragem $0{,}5R$, apenas nos trechos de menor variação (região branca) em conjunto com a taxa de amostragem $2R$ na região com maior variação. Pode-se verificar que, mesmo reduzindo a quantidade de dados coletados na região de menor variação, é possível obter uma variação de dados estável e até mesmo, em alguns casos, uma redução significativa do erro da medida. Por conseguinte, verifica-se a possibilidade de efetuar a redução dos dados coletados por cada participante sem perda de informação.

A Figura 6.4 mostra a CDF das velocidades médias normalizadas usando também



(a) 1-Maior.



(b) 21-Menores.

Figura 6.4: CDF das diferenças de toda a rota $\boldsymbol{id_{route} = 007}$.

uma taxa de amostragem intermediária, 1,5R, na região preenchida com cinza para enriquecer os resultados. Note que há uma redução das diferenças tanto para o caso de um único ônibus, '1-Maior', quanto para o caso dos 21 ônibus, '21-Menores'. Tal redução ainda é tímida porque o aumento da taxa foi realizado apenas em alguns trechos. Além da variação das taxas de amostragem da região cinza, a CDF mostra também a taxa de amostragem 0,5R na região branca, sendo que essa taxa está combinada com a taxa 2R para a região cinza. Para uma melhor compreensão dos testes realizados, a Tabela 6.2 apresenta a combinação das taxas utilizadas em cada experimento e como cada uma dessas combinações está denotada nas Figuras 6.4.

Caso se isole a CDF apenas dos trechos de maior variação, como visto na Figura 6.5, é possível perceber com mais clareza a redução das maiores diferenças de velocidade. As velocidades medidas se tornam mais próximas da existente no conjunto de dados conforme a curva da CDF se aproxima de um degrau em $x = 1$. Valores medidos menores que 1 significam velocidades médias menores que a do

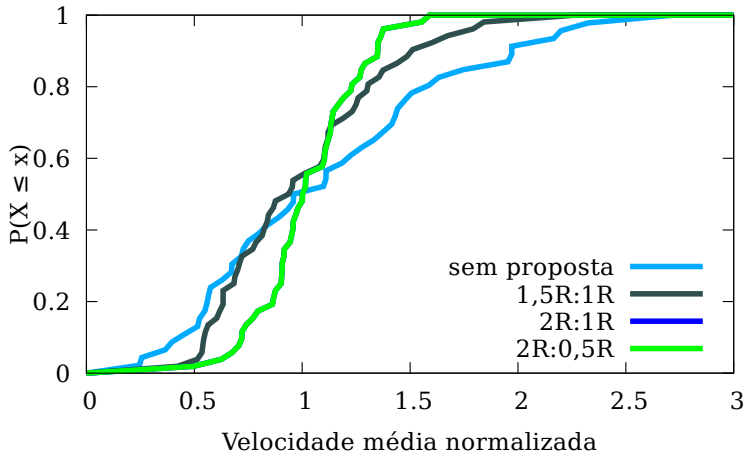
Tabela 6.2: Combinação das taxas de amostragem utilizadas para simulação pelo sistema proposto.

Nome	**Trechos com alta variação (em cinza)**	**Trechos com baixa variação (em branco)**
sem proposta	$1R$	$1R$
$1{,}5R : 1R$	$1{,}5R$	$1R$
$2R : 1R$	$2R$	R
$2R : 0{,}5R$	$2R$	$0{,}5R$

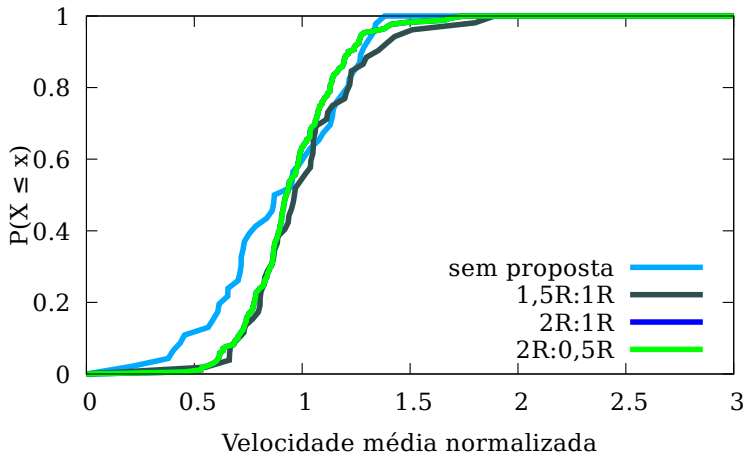
conjunto de dados, enquanto valores medidos maiores que 1 representam valores de velocidades médias maiores que a do conjunto de dados. Ao utilizar 21 ônibus, o impacto é menor porque as amostras utilizadas são aleatórias dentro da região de maior variação e, no momento da amostragem, pode haver trechos que não foram contemplados. Além disso, isolando a CDF apenas dos trechos de menor variação, como visto na Figura 6.6, observa-se um comportamento similar ao da opção sem proposta. Note que as propostas $1{,}5R : 1R$ e $2R : 1R$ não sofrem alteração pois somente modificam a taxa de amostragem da região de alta variação (região cinza). Já a proposta $2R : 0{,}5R$, que tem taxa $0{,}5R$ na região branca, consegue manter o erro da medida baixo para os dois grupos (1 ônibus e 21 ônibus). O comportamento inferior para 21 ônibus é esperado pelo mesmo motivo apontado anteriormente, visto que a retirada das amostras dentro do conjunto de dados é aleatória. É importante ressaltar que, como a variação nessa região é menor, um pouco mais de variação é tolerável tendo em vista os ganhos em economia de recursos para os usuários participantes. A Tabela 6.3 reforça os resultados mostrando a redução do Erro Médio Quadrático (*Root Mean Square Error – RMSE*) das velocidades médias normalizadas. Percebe-se que há uma tendência de redução do erro para toda a rota com o aumento das taxas de amostragem nos trechos de maior variação (região cinza). O RMSE é menor ainda com a redução das taxas de amostragem nos trechos de menor variação (região branca), combinada com a taxa de amostragem $2R$ nos trechos de maior variação (região cinza). Esse resultado é surpreendente, se comparado ao resultado da proposta $2R : 1R$ e deve ser uma consequência da amostras utilizadas.

Tabela 6.3: RMSE das velocidades médias considerando toda a rota $\boldsymbol{id_{route} = 007}$.

Configuração	**sem proposta**	$1{,}5R : 1R$	$2R : 1R$	$2R : 0{,}5R$
1-Maior	0,422	0,342	0,290	0,264
21-Menores	0,441	0,387	0,366	0,350



(a) 1-Maior.



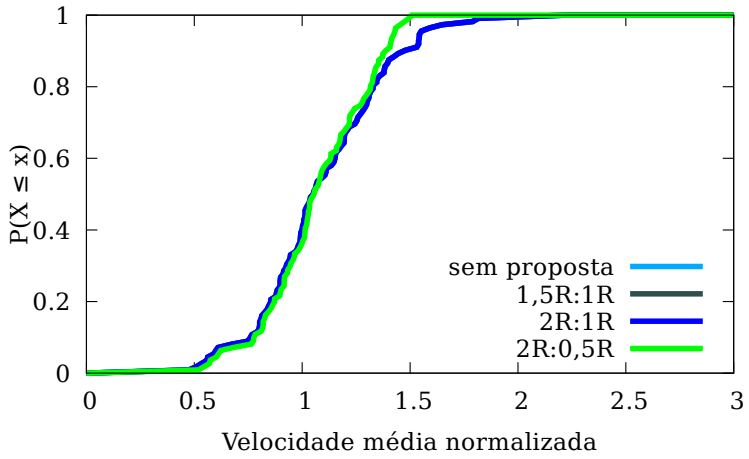
(b) 21-Menores.

Figura 6.5: CDF das diferenças das velocidades médias nos trecho de maior variação (região cinza).

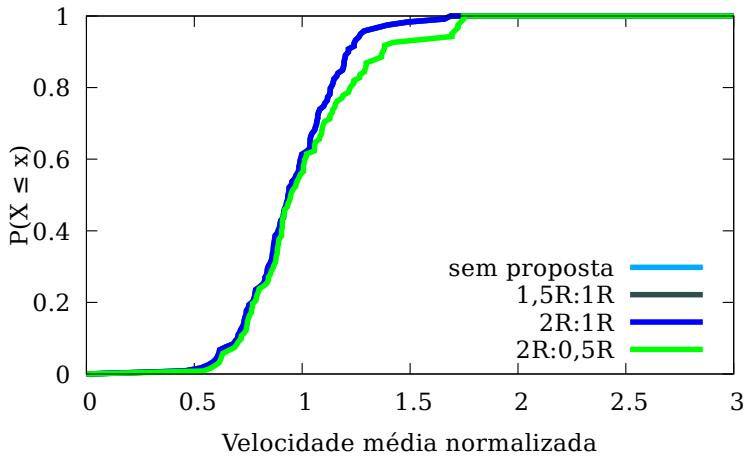
6.3 Carga de dados no sistema

Além da avaliação do sistema proposto, é necessário verificar a quantidade de dados carregada por cada participante, visto que a carga de dados pode aumentar para trechos com grande variação dos dados (região cinza) e diminuir para trechos com pouca variação dos dados. Na rota estudada, tem-se um total de 51 veículos (Figura 4.2) que circulam entre os 174 trechos. Como observado, nem todos os veículos fazem troca de taxa de amostragem e, além disso, nem todos os trechos necessitam de maior taxa de amostragem.

A Figura 6.7 apresenta um comparativo entre os dados sem proposta e com proposta, seguindo os cenários descritos na Tabela 6.2. Os resultados mostram duas comparações, a primeira apresenta o comportamento da carga dos ônibus e a segunda apresenta a carga dos trechos divididos da rota. A Figura 6.7(a) traz o



(a) 1-Maior.

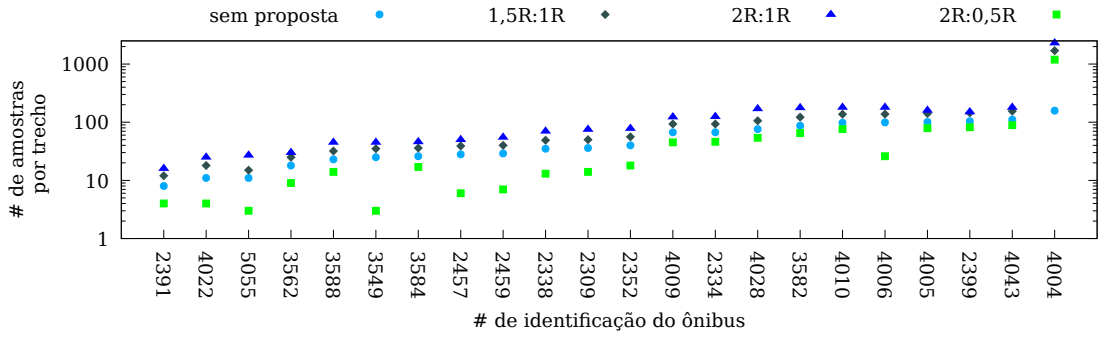


(b) 21-Menores.

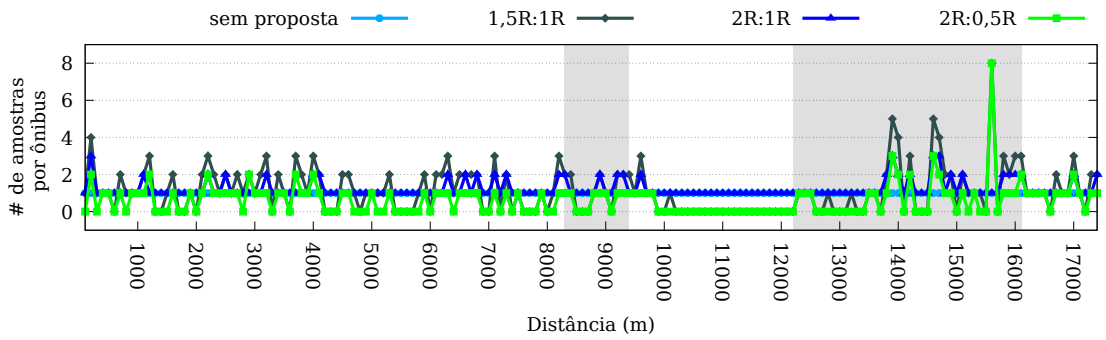
Figura 6.6: CDF das diferenças das velocidades médias nos trecho de menor variação (região branca).

comparativo dos ônibus, no eixo-x está a identificação dos ônibus e no eixo-y está o número de amostras por trecho, que representa a quantidade de contribuições por trecho. Observa-se nos resultados uma redução na carga de dados com a proposta utilizada. Pode-se verificar ainda que as propostas $1{,}5R : 1R$ e $2R : 1R$ apresentam uma carga de amostras maior nos ônibus, isso acontece pois nesse cenário temos apenas o acréscimo da taxa de amostragem na região cinza. Porém, quando aplicamos a proposta $2R : 0{,}5R$ temos uma inversão pois a taxa calculada na região branca ($0{,}5R$) compensou a sobrecarga aplicada na região cinza ($2R$). Vale ressaltar que esse comportamento acontece apenas para o grupo de ônibus 21-menores pois apresentam uma carga de dados distribuída entre todos os participantes, comparado ao grupo 1-maior. Já a Figura 6.7(b) traz o comparativo dos trechos, no eixo-x está a identificação dos trechos e no eixo-y está o número de amostras por ônibus, as medidas estão normalizadas em função do resultado sem proposta. Para o cenário

com proposta $2R : 0{,}5R$ é obtida uma redução mais significativa do erro em grande parte dos trechos, sendo que apenas o trecho 154.000 apresenta uma carga mais elevada, é justamente nesse pronto que exite uma maior concentração de amostras do veículo do conjunto 1-maior além de ser a região que tem maior variação (região cinza), o que significa que foi inserida uma carga de dados elevados para o ônibus 4004. Esse problema pode ser ainda atenuado, caso existisse uma maior quantidade de veículos nesses trechos. Dessa forma, seria possível distribuir a carga imposta no trecho e até mesmo reduzir a taxa $2R$.



(a) Ônibus.



(b) Trechos.

Figura 6.7: Comparativo de desempenho do sistema sem proposta em relação às variações do sistema com proposta.

A Tabela 6.4 apresenta alguns dados observados após a execução do sistema proposto (posposta $2R : 0{,}5R$ – Tabela 6.2). Em relação aos participantes (ônibus), observa-se que 43,14% fizeram a troca das taxas de amostragem, sendo que $\approx$ 95,46% fizeram a redução em suas coletadas de dados. Esse ajuste representa uma redução da carga de dados de $\approx$ 67,49%. Além disso, observando os trechos da RoI, o sistema reduziu a taxa de amostragem em $\approx$ 86% dos trechos, mostrando que a adaptação de taxas de amostragem é bastante importante ao longo de uma rota como a $id_{route} = 007$.

Tabela 6.4: Comparação da evolução dos dados sem o sistema proposto. Totais dos ônibus e trechos com sistema proposto $2R : 0{,}5R$ – Tabela 6.2.

Descrição	**Não fizeram troca de taxa**	**Fizeram troca de taxa**	**Aumentaram a taxa de amostragem**	**Reduziram a taxa de amostragem**	**Redução da carga de dados**
Ônibus	56,86%	43,14%	4,54%	95,46%	67,49%
Trecho	42,53%	57,46%	14,00%	86,00%	67,49%

6.4 Carga da rede

É importante ressaltar que a taxa de amostragem em toda a rota poderia ser aumentada, já que o sistema ainda está distante do número máximo de amostras contidas no conjunto de dados e da capacidade máxima da rede, como visto na Figura 6.2. A rota $id_{route} = 007$ possui 15.546 amostras no total, sendo que nos experimentos com mais amostras (proposta $2R : 0{,}5R$) foram coletas apenas 1.696 amostras o que representa uma redução de 67,49% como apresentado na Tabela 6.4. Esse valor não representa problema, pois está abaixo do máximo contido no conjunto de dados e ainda possui probabilidade de ser entregue com sucesso à RSU durante o contato.

Capítulo 7

Conclusões e Trabalhos Futuros

Este trabalho propôs um sistema de sensoriamento participativo com consistência de dados em redes veiculares que leva em conta características tanto espaciais quanto temporais de regiões de interesse (RoI). O principal objetivo deste trabalho foi apresentar um método de ajuste da taxa de amostragem dos dados e mostrar que é possível ajustar a taxa de amostragem em um sistema de sensoriamento participativo, garantindo uma redução no erro das medidas. A redução do erro é possível com o aumento da taxa de amostragem apenas nos trechos de maior variação das medidas. Já os trechos com menor variação das medidas, o erro pode ser mantido baixo mesmo reduzindo a taxa de amostragem.

Para isso, assume-se que a RoI é dividida em trechos e que as métricas coletadas podem variar dentro de um intervalo de tempo Δt. Os usuários participantes coletam dados obedecendo a taxas de amostragem definidas por um servidor central, sendo que a boa acurácia do sistema depende da taxa de amostragem dos usuários. Esses dados coletados, antes de serem considerados para cálculo no controlador, são verificados quanto a consistência, ou seja, se o dado é anômalo ou não. Dessa forma, atinge-se o segundo objetivo do trabalho, que é assegurar a não existência de dados anômalos, garantindo a consistência dos dados. A possibilidade de dados inconsistentes é explorada a partir da definição de quatro perfis de usuários que podem produzir e injetar dados anômalos no sistema. Esses perfis são compostos por usuários com inconsistências permanentes e intermitentes que contribuem com valores errôneos ao sistema.

As análises realizadas considerando cada tipo de usuário mostraram que inconsistências permanentes são mais fáceis e rápidas de detectar. O intermitente, no entanto, precisa de mais amostras para serem detectadas porque intercala medidas reais e irreais. Os resultados mostram que é possível identificar sempre a presença de dados anômalos contendo valores de velocidade máxima em todas as regiões, considerando os valores legítimos das velocidades média e de desvio padrão, mesmo usando um método de detecção simples. Quando os usuários informam medidas próximas

a zero, tem-se um desempenho inferior, isso ocorre pois as velocidades médias em todas as regiões são relativamente baixas. Mesmo manipulando a média e os valores de desvio padrão, o impacto na detecção de anomalia não é importante, mesmo adicionando 100% de amostras falsas. O volume adicional de amostras, dependendo da configuração da rede, pode resultar em uma carga de rede elevada.

Em relação à taxa de amostragem, quanto maior for a taxa, melhores são os resultados. Esse princípio, porém, esbarra no compromisso entre o número de amostras coletadas, limitações dos dispositivos e carga de dados gerada. Por isso, faz-se necessário identificar regiões de menor variação de medidas e regiões com maior variação de medidas. Caso seja possível tal identificação, as taxas de amostragem podem ser adaptadas para coletar mais dados das regiões com maior variação e menos dados das regiões de menor variação. Mesmo alterando as taxas de $\approx 43\%$ dos veículos, o sistema proposto obteve uma redução de 67,49% na carga de dados da rede.

Considerando o cenário de monitoramento das velocidades dos veículos, levando em conta amostras coletadas por participantes em veículos e a quantidade de trechos existente, existe uma preocupação quanto ao volume de dados a serem coletados por veículo. O sistema proposto ajusta a taxa de amostragem de cada trecho da RoI em função da variação das medidas, no caso, em função da velocidade dos veículos. Os trechos com maior variação nas velocidades devem ter maior taxa de amostragem, enquanto os de baixa variação podem ter taxas menores. Os resultados obtidos com conjuntos de dados reais mostram que maiores cargas de dados só são possíveis de serem entregues com poucos nós competindo pelo meio. Os resultados ainda mostram que a amostragem adaptativa é capaz de reduzir os erros sem impactar na carga de dados uma vez que ao dividir as regiões (cinza e branca), consegue-se um compromisso entre os ajustes das taxas nas duas regiões.

Como trabalhos futuros, planeja-se avaliar o sistema usando mais rotas na cidade de Seattle. É necessário abordar separadamente cada parte da metodologia. Na consistência de dados, faz-se necessário investigar novos comportamento de usuários e trabalhar com novos cenários, além de melhorar o método de detecção, usando métodos mais sofisticados. Em relação à adaptação da taxa de amostragem, sugere-se tornar autônoma a escolha da taxa de amostragem por trecho. Pretende-se também elaborar uma ferramenta capaz consolidar e gerenciar os dados para os administradores do sensoriamento participativo.

Referências Bibliográficas

[1] GANTI, R. K., YE, F., LEI, H. "Mobile crowdsensing: current state and future challenges", *IEEE Communications Magazine*, v. 49, n. 11, 2011.

[2] PAPADIMITRATOS, P., DE LA FORTELLE, A., EVENSSEN, K., et al. "Vehicular communication systems: Enabling technologies, applications, and future outlook on intelligent transportation", *IEEE communications magazine*, v. 47, n. 11, 2009.

[3] XIAO, Y., SIMOENS, P., PILLAI, P., et al. "Lowering the barriers to large-scale mobile crowdsensing". In: *14th Workshop on Mobile Computing Systems and Applications*, pp. 9:1–9:6, 2013.

[4] WAZE. "Rotas, Mapas, GPS, Trânsito e Engarrafamentos - Waze". `https://www.waze.com/pt-BR/about`. Acesso em Agosto de 2017.

[5] RIBEIRO NETO, V., MEDEIROS, D. S. V., CAMPISTA, M. E. M. "Analysis of Mobile User Behavior in Vehicular Social Networks". In: *International Conference Network of the Future (NOF 2016)*, pp. 1–5, 2016.

[6] ITU INTERNET REPORTS. *ITU Strategy and Policy Unit (SPU)*. Relatório técnico, International Telecommunication Union (ITU), 2005.

[7] GUBBI, J., BUYYA, R., MARUSIC, S., et al. "Internet of Things (IoT): A vision, architectural elements, and future directions", *Future generation computer systems*, v. 29, n. 7, pp. 1645–1660, 2013.

[8] CHOURABI, H., NAM, T., WALKER, S., et al. "Understanding smart cities: An integrative framework". In: *Hawaii International Conference on System Science (HICSS)*, pp. 2289–2297, 2012.

[9] LIU, Y., NIU, J., MA, J., et al. "The insights of message delivery delay in VANETs with a bidirectional traffic model", *Journal of Network and Computer Applications*, v. 36, n. 5, pp. 1287–1294, 2013.

[10] MICHE, M., BOHNERT, T. M. "The internet of vehicles or the second generation of telematic services", *ERCIM News*, v. 77, pp. 43–45, 2009.

[11] YU, R., ZHANG, Y., WU, H., et al. "Virtual machine live migration for pervasive services in cloud-assisted vehicular networks". In: *8th International ICST Conference on Communications and Networking in China (CHINACOM)*, pp. 540–545, 2013.

[12] KARAGIANNIS, G., ALTINTAS, O., EKICI, E., et al. "Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions", *IEEE communications surveys & tutorials*, v. 13, n. 4, pp. 584–616, 2011.

[13] ANDRÉ, C. H. O. M., MEDEIROS, D. S. V., CAMPISTA, M. E. M. "Sensoriamento Participativo de Regiões de Interesse com Descrição Adaptativa das Taxas de Amostragem". In: *XXXVI Simpósio Brasileiro de Redes de Computadores (SBRC 2018)*, v. 14, pp. 1–14, Maio 2018.

[14] ZENG, Y., XIANG, K. "Adaptive Sampling for Urban Air Quality through Participatory Sensing", *Sensors (Basel)*, v. 17, n. 11, pp. 1–16, 2017.

[15] WEINSCHROTT, H., WEISSER, J., DURR, F., et al. "Participatory Sensing Algorithms for Mobile Object Discovery in Urban Areas". In: *IEEE International Conference on Pervasive Computing and Communications (PERCOM)*, pp. 128–135, 2011.

[16] ANDRÉ, C. H. O. M., MEDEIROS, D. S. V., CAMPISTA, M. E. M. *A Methodology to Assess Data Consistency in Vehicular Networks Using Participatory Sensing*. Relatório Técnico GTA-17-31, GTA/PEE/UFRJ, 2017.

[17] JETCHEVA, J. G., HU, Y.-C., PALCHAUDHURI, S., et al. "CRAWDAD dataset rice/ad_hoc_city (v. 2003-09-11)". `http://crawdad.org/rice/ad_hoc_city/20030911/bus_mobility`, set. 2003. traceset: bus_mobility.

[18] ANDRÉ, C. H. O. M., MEDEIROS, D. S. V., CAMPISTA, M. E. M. "Uma Metodologia para Detecção de Anomalia para Sensoriamento Participativo em Redes Veiculares". In: *VII Brazilian Symposium on Computing Systems Engineering (SBESC 2017)*, pp. 1–6, 2016.

[19] WANGHAM, M. S., NOGUEIRA, M., FERNANDES, C. P., et al. "Segurança em Redes Veiculares: Inovações e Direções Futuras", *Minicursos do XIV Simpósio Brasileiro em Segurança da Informaçao e de Sistemas Computacionais*.

[20] HOSSAIN, E., CHOW, G., LEUNG, V. C., et al. "Vehicular telematics over heterogeneous wireless networks: A survey", *Computer Communications*, v. 33, n. 7, pp. 775–793, 2010.

[21] SICHITIU, M. L., KIHL, M. "Inter-vehicle communication systems: a survey", *IEEE Communications Surveys & Tutorials*, v. 10, n. 2, 2008.

[22] STAHLMANN, R., FESTAG, A., TOMATIS, A., et al. "Starting European field tests for Car-2-X communication: the DRIVE C2X framework". In: *18th ITS World Congress and Exhibition*, 2011.

[23] AHMED-ZAID, F., BAI, F., BAI, S., et al. *Vehicle Safety Communications–Applications (VSC-A) Final Report: Appendix Volume 3 Security.* Relatório técnico, 2011.

[24] FESTAG, A. "Cooperative intelligent transport systems standards in Europe", *IEEE communications magazine*, v. 52, n. 12, pp. 166–172, 2014.

[25] BALDESSARI, R., BÖDEKKER, B., DEEGENER, M., et al. "Car-2-car communication consortium-manifesto", 2007.

[26] MELO, P. C. F. *CSVM: Uma plataforma para crowdsensing móvel dirigida por modelos em tempo de execução.* Tese de Mestrado, Universidade Federal de Goiás, 2014.

[27] MASHHADI, A. J., CAPRA, L. "Quality control for real-time ubiquitous crowdsourcing". In: *Proceedings of the 2nd international workshop on Ubiquitous crowdsouring*, pp. 5–8. ACM, 2011.

[28] ESTRIN, D., CHANDY, K. M., YOUNG, R. M., et al. "Participatory sensing: applications and architecture", *IEEE Internet Computing*, v. 14, n. 1, pp. 12–42, 2010.

[29] BURKE, J. A., ESTRIN, D., HANSEN, M., et al. "Participatory sensing". In: *World Sensor Web Workshop, ACM Sensys*, pp. 1–5, 2006.

[30] MCCALL, M. K., MINANG, P. A. "Assessing participatory GIS for community-based natural resource management: claiming community forests in Cameroon", *Geographical Journal*, v. 171, n. 4, pp. 340–356, 2005.

[31] D'HONDT, E., STEVENS, M., JACOBS, A. "Participatory noise mapping works! An evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring", *Pervasive and Mobile Computing*, v. 9, n. 5, pp. 681–694, 2013.

[32] MOHAN, P., PADMANABHAN, V. N., RAMJEE, R. "Nericell: rich monitoring of road and traffic conditions using mobile smartphones". In: *ACM conference on Embedded network sensor systems*, pp. 323–336, 2008.

[33] BOULOS, M. N. K., WHEELER, S., TAVARES, C., et al. "How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX", *Biomedical engineering online*, v. 10, n. 1, pp. 24:1–24:14, 2011.

[34] CHRISTIN, D., REINHARDT, A., KANHERE, S. S., et al. "A survey on privacy in mobile participatory sensing applications", *Journal of systems and software*, v. 84, n. 11, pp. 1928–1946, 2011.

[35] ZHANG, X., YANG, Z., WU, C., et al. "Robust trajectory estimation for crowdsourcing-based mobile applications", *IEEE Transactions on Parallel and Distributed Systems*, v. 25, n. 7, pp. 1876–1885, 2014.

[36] CRUZ, P., DA SILVA, F. F., PACHECO, R. G., et al. "SensingBus: um Sistema de Sensoriamento Baseado em Ônibus Urbanos". In: *Salão de Ferramentas do XXXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pp. 1–14, 2017.

[37] CRUZ, P., COUTO, R. S., COSTA, L. H. M. "An algorithm for sink positioning in bus-assisted smart city sensing", *Future Generation Computer Systems*, 2017. (In press, Accepted Manuscript).

[38] ZHOU, P., ZHENG, Y., LI, M. "How long to wait?: predicting bus arrival time with mobile phone based participatory sensing". In: *10th international conference on Mobile systems, applications, and services*, pp. 379–392, 2012.

[39] GANTI, R. K., PHAM, N., AHMADI, H., et al. "GreenGPS: a participatory sensing fuel-efficient maps application". In: *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pp. 151–164. ACM, 2010.

[40] TANGADE, S. S., MANVI, S. S. "A survey on attacks, security and trust management solutions in VANETs". In: *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*, pp. 1–6. IEEE, 2013.

[41] AL-KAHTANI, M. S. "Survey on security attacks in Vehicular Ad hoc Networks (VANETs)". In: *Signal Processing and Communication Systems (ICSPCS), 2012 6th International Conference on*, pp. 1–9. IEEE, 2012.

[42] SEATTLE.GOV. "Seattle Department of Transportation". http://www.seattle.gov/transportation/sdotfaqs.htm, 2017. Acessado: 20/07/2017.

[43] METRO, K. C. "Route 7 - King Country". http://kingcounty.gov/depts/transportation/metro/schedules-maps/007.aspx#weekday, 2017. Acessado: 13/06/2017.

[44] JIANG, D., DELGROSSI, L. "IEEE 802.11 p: Towards an international standard for wireless access in vehicular environments". In: *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*, pp. 2036–2040. IEEE, 2008.

[45] NSNAM. "What is ns-3". https://www.nsnam.org/overview/what-is-ns-3/. Acesso em Julho de 2017.