

Uma Metodologia para Detecção de Anomalia para Sensoriamento Participativo em Redes Veiculares

Carlos Henrique de O. M. André, Dianne S. V. de Medeiros e Miguel Elias M. Campista
Universidade Federal do Rio de Janeiro - PEE-COPPE/Del-Poli/GTA
{choma,dianne,miguel}@gta.ufrj.br

Abstract—Today, users can actively participate of sensing tasks as the number of built-in sensors on mobile devices substantially grows. This new configuration opens venue to the new paradigm of Participatory Sensing (PS), where users contribute by collecting and possibly analyzing environmental data from pre-determined Regions of Interest (RoI). The idea is to achieve a scalable sensing system with low deployment costs. In this scenario, however, it is fundamental to ensure that the data collected is reliable and that anomalous data are not able to influence the final result. This work proposes a methodology to identify the existence of inconsistency or unreliable data, collected by malicious or faulty sensors. The methodology is composed of three main steps: data validation, anomalous detection, and detection system update. The first step verifies if the sampled data is consistent, while the second checks if this data is within the range expected values. If the data is not rejected, the third step updates the system. We evaluate the proposed methodology using the dataset of the bus fleet in the city of Seattle, which records bus movements. We assume that users inside buses collect data regarding the vehicle speed. Results show that the proposed methodology can identify the contribution threshold from which anomalous data can be detected. For instance, in a higher speed region, we can already identify the presence of anomalous data with 5% of additional contributions.

I. INTRODUÇÃO

Nos últimos anos, a indústria de comunicações móveis tem proporcionado uma enorme evolução tecnológica tanto em hardware quanto em software. Nesse contexto, novos serviços e aplicações vêm surgindo, sendo o recente paradigma de Internet das Coisas (*Internet of Things* – IoT) [1] um dos principais motivadores dos avanços atuais. Com dispositivos móveis em mãos, os usuários se tornam fontes de dados para muitas das aplicações IoT, podendo sensoriar e enviar dados até algum ponto concentrador, possivelmente na Internet. Esse ponto é responsável por analisar e transformar os dados recebidos em informação útil. Com o crescimento do número de participantes, surge o conceito de sensoriamento participativo (*Participatory Sensing* – PS) [2], referente à participação dos usuários no processo de sensoriamento de dados. O maior benefício é o aumento do volume de dados coletados e da área coberta pelos sensores. As aplicações inseridas nesse conceito são capazes de aumentar a escala de dados e serviços devido à participação dos usuários, de forma gratuita. Um exemplo dessa abordagem é a operação do aplicativo Waze, que oferece informações sobre as condições do trânsito usando como dados de entrada as contribuições dos próprios usuários [3].

Muitos sistemas de sensoriamento já foram propostos na literatura com intuítos diversos. Por exemplo, Cruz et al. [4]

apresentaram o sistema *Sensing Bus*, que propõe embarcar sensores em ônibus urbanos para expandir a área monitorada. Mohan et al. [5] apresentaram o *Niricell* que utiliza a localização por GPS (*Global Positioning System*) de *smartphones* para o monitoramento de trânsito e estradas. Já Zhou et al. [6] propuseram um sistema baseado em esforço colaborativo para prever o posicionamento dos ônibus em uma viagem utilizando a potência entre torres de celular e dispositivos. Uma característica em comum de todos esses trabalhos é a quantidade e diversidade dos nós sensores.

Este trabalho propõe uma metodologia para a avaliação da consistência dos dados provenientes de usuários que contribuem com o sistema de sensoriamento. A ideia é detectar possíveis dados anômalos produzidos pelos usuários, garantindo a qualidade e precisão da informação. Esses dados podem ser provenientes de falhas no sensor, erros durante a transferência dos dados, ou ainda, serem intencionalmente alterados por usuários maliciosos. A metodologia é composta por três etapas principais: validação da amostra, detecção de anomalia da amostra e atualização do sistema.

Uma característica importante para a detecção de amostras anômalas é a variabilidade das amostras em relação ao valor esperado. Quanto maior a variação, menor é a capacidade de identificação de anomalias. Este trabalho, portanto, quantifica o aumento do volume de dados coletados em uma Região de Interesse (*Region of Interest* – RoI) e avalia sua influência na validade e precisão das informações consolidadas. Para tal, utiliza-se um conjunto de dados com traços de um mês de movimento da frota de ônibus da cidade de Seattle, nos EUA [7]. Desse conjunto, analisa-se um dia inteiro e assume-se que o valor sensoriado é a velocidade dos ônibus. Vale ressaltar que a metodologia proposta não perde generalidade caso outro sensor seja usado. Além disso, a metodologia pode ser reproduzida para outros conjuntos de dados. A avaliação da metodologia considera a existência de três RoIs, divididas conforme a velocidade média dos ônibus: velocidade alta, baixa e mista. Verifica-se o impacto que uma possível amostra anômala pode gerar nos três cenários e considera-se a existência de usuários que contribuem enviando amostras anômalas de maneira permanente ou intermitente. A quantidade de amostras adicionais é calculada como uma proporção do número máximo de amostras no período. Os resultados mostram que a metodologia proposta é capaz de começar a identificar anomalias quando os usuários inserem pelo menos 5% a mais de amostras anômalas na região de

baixa velocidade. Nas regiões de velocidade mista e alta essa proporção aumenta para respectivamente, 10% e 15% a mais de amostras para que seja possível começar a detectar as anomalias durante parte do período investigado. Mesmo com esse volume de amostras adicionais, a carga total de dados gerados não é suficiente para provocar problemas de rede.

O restante deste trabalho está organizado da seguinte forma. A Seção II introduz noções de sensoriamento participativo. A Seção III apresenta a metodologia de detecção de anomalia proposta. A Seção IV apresenta o conjunto de dados utilizado; enquanto a Seção V caracteriza a principal rota utilizada na análise do trabalho. A Seção VI apresenta os resultados de detecção de anomalia. Por fim, a Seção VII conclui este trabalho e lista direções futuras.

II. SENSORIAMENTO PARTICIPATIVO

A evolução da tecnologia móvel e a miniaturização de componentes abrem espaço para um número crescente de sensores embarcados em veículos e em dispositivos pessoais como *smartphones*. Atualmente, os sensores mais comuns são o acelerômetro, o GPS, o giroscópio, a bússola, 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. Quando os dispositivos são usados de maneira coletiva, ou seja, extraindo informações do ambiente juntamente com outros dispositivos, é possível construir um sistema de sensoriamento escalável e de baixo custo. Os dados coletados, ao serem devidamente analisados, oferecem oportunidade para o desenvolvimento de novas aplicações [8]. O sensoriamento através da participação de usuários e seus dispositivos são o foco do novo paradigma de pesquisa denominado Sensoriamento Participativo (*Participatory Sensing* – PS).

Uma das principais vantagens desse tipo de sensoriamento é o aumento do conhecimento sobre a região monitorada, como consequência do compartilhamento dos dados sensorizados. A participação de usuários pode gerar alguns problemas para o processamento e interpretação de diferentes dados coletados durante o trajeto, além do volume de dados que deve ser transferido possivelmente em pouco tempo de contato entre Unidades de Bordo (*On Board Units* – OBUs) e Unidades de Acostamento (*Road Side Units* – RSUs).

III. DETECÇÃO DE ANOMALIA PARA SENSORIAMENTO PARTICIPATIVO EM REDES VEICULARES

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

A. Metodologia proposta

A metodologia proposta é dividida em três etapas principais: validação da amostra, detecção de anomalia da amostra e atualização do sistema. É importante destacar que a metodologia pode ser tanto executada em um ponto concentrador,

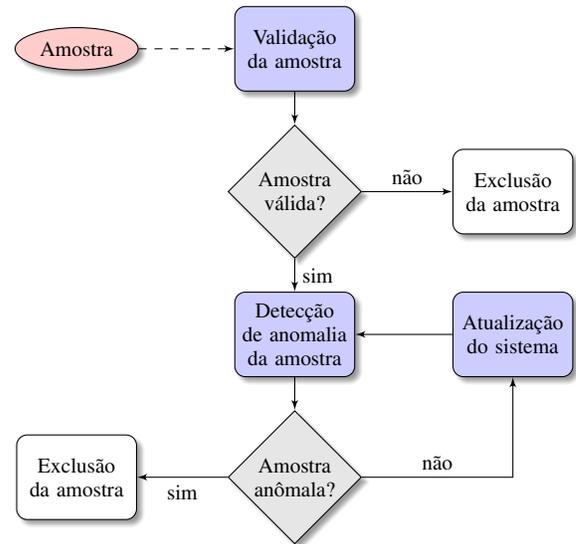


Figura 1. Metodologia proposta para detecção de anomalia no sensoriamento participativo em redes veiculares. A metodologia é dividida em três etapas principais: validação da amostra, detecção de anomalia da amostra e atualização do sistema.

possivelmente um servidor na Internet; ou localmente em um dispositivo de sensoriamento. Configurações híbridas também são possíveis nas quais pode ser parcialmente executada nos dispositivos e também em um servidor na nuvem. O funcionamento da metodologia proposta está ilustrado na Figura 1. A seguir, cada uma das três etapas principais é detalhada.

Validação da amostra: Nessa etapa, o objetivo é eliminar através de filtros amostras que contenham valores não-válidos para os dados sensorizados. Sendo assim, evita-se o desperdício de recursos de processamento e armazenamento com amostras que não contribuem com o sistema participativo. Considerando uma amostra x_i composta por um vetor de propriedades $\langle p_{i,1}, \dots, p_{i,n} \rangle$, 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 sensorizada são usados como propriedades das amostras. Assume-se que a localização deva pertencer à cidade onde a medida foi coletada e que a velocidade deva ser sempre positiva. Além disso, essa velocidade não pode ser maior do que um limite, onde 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, dentre outros fatores.

Detecção de anomalia da amostra: Nessa 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. 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. Apenas a propriedade relacionada com o valor da medida

sensorizada é avaliada.

Neste trabalho, a metodologia proposta usa um mecanismo simples de detecção. Seja x_i a amostra a ser analisada, considera-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 descartada.

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.

Neste trabalho, uma nova amostra válida é usada para recalcular a média, μ , e o desvio padrão, σ , usados no bloco de detecção de anomalia da amostra.

B. Perfis dos usuários

Neste trabalho, assume-se a presença de usuários que contribuem com o sistema de maneira correta, ou seja, de fato contribuem colaborativamente com amostras válidas e não-anômalas; e assume-se 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. 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.

Neste trabalho, o principal objetivo é analisar a detecção de sensores que estejam inserindo dados de velocidade acima 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 não foram encontradas na análise apresentada na Seção VI. Observe porém que a metodologia proposta é agnóstica a essa característica. Ainda, que a anomalia é tão impactante ao sistema quanto for diferente dos valores esperados. Portanto, amostras anômalas com valores dentro do intervalo de amostras esperadas podem não ser descartadas.

Assume-se a existência de dois tipos de usuários que introduzem amostras anômalas. O primeiro tipo introduz amostras com a presença de *anomalia permanente*, ou seja, todas as

amostras apresentam velocidades elevadas, e são chamados neste trabalho de *usuário com anomalia permanente*. O segundo tipo, por sua vez, introduz amostras que intercalam velocidades elevadas e médias, isto é, *anomalias intermitentes*, e recebem o nome de *usuário com anomalia intermitente*, neste trabalho. Os dois 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 1000% do número máximo de amostras existentes no sistema durante o período analisado. Os resultados para os dois tipos de perfis são apresentados na Seção VI.

IV. CONJUNTO DE DADOS

A validação da metodologia proposta é feita através da análise do impacto da inserção de dados anômalos no sistema. Para tanto, utiliza-se o conjunto de dados Ad Hoc City [7], que contém 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. É representado 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 tempo t .

Neste trabalho, analisa-se um dia típico de quarta-feira, 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 registro de mobilidade. Esses registros representam contribuições dos usuários para o sistema de sensoriamento. Assim, no cenário estudado, existem 376.491 amostras, com uma média de 1.595 amostras por rota.

A velocidade máxima dos ônibus na região metropolitana de Seattle é de aproximadamente 56 km/h [9]. Após calcular as velocidades dos ônibus com base nos deslocamentos registrados nos dados originais, surgem alguns valores irreais, que são eliminados pela primeira etapa da metodologia proposta, *validação da amostra*. Para tanto, são executadas duas avaliações. Inicialmente verifica-se se a velocidade é irreal, seja, v_i a velocidade em uma amostra, verifica-se se v_i está no intervalo $[0, v_{max}$ km/h]. Neste trabalho, considera-se $v_{max} = 110$ km/h, que é a velocidade máxima permitida na cidade de Seattle. [9] Além disso, verifica-se o posicionamento do ônibus registrado pelo GPS, isto é, se ele está dentro das coordenadas da cidade de Seattle. Na rota analisada, $id_{route} = 007$, essa validação resulta na exclusão de 55 registros, que representam a uma redução de 0.35% dos dados.

V. CARACTERIZAÇÃO DA ROTA 007

Nesta seção, é detalhada as características da rota escolhida, a fim de facilitar a análise da metodologia proposta. Essa rota

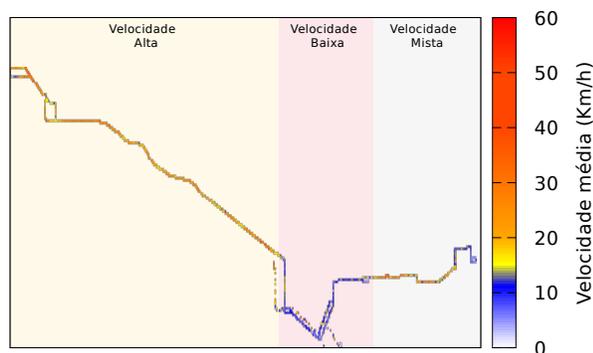


Figura 2. Mapa da velocidade média dos ônibus na Rota 007.

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. A duração média de uma viagem nessa rota em dias de semana é de 44 minutos [10]. No dia analisado, foram registrados 51 ônibus, sendo a quantidade média de amostras por ônibus igual a 305 coletas.

A. Velocidade média dos ônibus

Considerando-se a Rota 007 inteira, a velocidade média no decorrer do dia é de 11,90 km/h. Para melhorar o refinamento dessa medida, permitindo uma avaliação mais detalhada da rota, primeiramente divide-se a rota em pequenas áreas de tamanho $50 \times 50 \text{ m}^2$. Em seguida, para cada pequena área calcula-se a velocidade média dos ônibus durante todo o dia. A Figura 2 mostra a velocidade média obtida para cada área. Nessa figura, quanto mais vermelha a área, maior é a velocidade média, e quanto mais azul, menor é a velocidade média. Observa-se que a velocidade média em alguns pontos da Rota 007 no decorrer do dia pode alcançar até 60 km/h. Isso indica que não existe congestionamento significativo na região, uma vez que os ônibus estão sendo capazes de se moverem com a velocidade máxima permitida para esse tipo de veículo na cidade de Seattle. [9] Destaca-se também na Figura 2, a divisão da rota em regiões com base na velocidade média dos ônibus, a saber: *Velocidade alta*, *Velocidade Baixa*, e *Velocidade Mista*. A Tabela I 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 dividir a rota em segmentos de características semelhantes, para reduzir a variabilidade dos dados sensorizados.

B. Total de ônibus e amostras

A quantidade de ônibus que contribuem ao longo do dia varia entre 3 e 31 ônibus, dependendo da hora do dia analisada. 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 totalizam 31 ônibus, que são capazes de coletar 998 amostras. Já o maior número de amostras é de 1.240 amostras, realizadas por 19 ônibus. O menor número

Tabela I
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

Tabela II
CARACTERÍSTICAS DAS INSERÇÕES ANÔMALAS.

Região	# de amostras por usuário anômalo	Valor da contribuição (km/h)	
		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

de amostras totalizam 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. Já a média de amostras por ônibus/hora 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.

VI. ANÁLISE DA METODOLOGIA PROPOSTA

O foco principal desta análise é verificar a viabilidade da detecção das amostras anômalas. Para tanto, utiliza-se os dois modelos de usuários com anomalia definidos na Seção III. Na simulação realizada, foram executadas 201 rodadas nas quais os usuários com anomalia inserem uma determinada quantidade de amostras no sistema. Assume-se que, inicialmente, o sistema está livre de amostras anômalas. A cada rodada, os usuários com anomalia contribuem com 5% da maior contribuição por hora (original) de cada região analisada. Assim, ao fim de todas as rodadas chega-se a 1000% de amostras anômalas adicionais. As amostras inseridas pelos usuários com anomalia permanente possuem a maior velocidade permitida para os ônibus na cidade Seattle. Os usuários com anomalia intermitente, por sua vez, contribuem com amostras que carregam a velocidade média na região e a velocidade máxima permitida por lei. O número de amostras e seus valores estão sumarizados na Tabela II. Na tabela, a média é variável e se refere à média na hora, na região.

Nas seções subsequentes, investiga-se o número de amostras anômalas necessárias para influenciar o resultado final e o impacto dessas amostras 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 e utilizando-se quaisquer outros tipos de medidas, como poluição, luminosidade, pressão, dentre outras.

A. Impacto das amostras anômalas no sensoriamento

Inicialmente, assume-se que o usuário anômalo pode inserir amostras capazes de alterar a média de velocidades da região

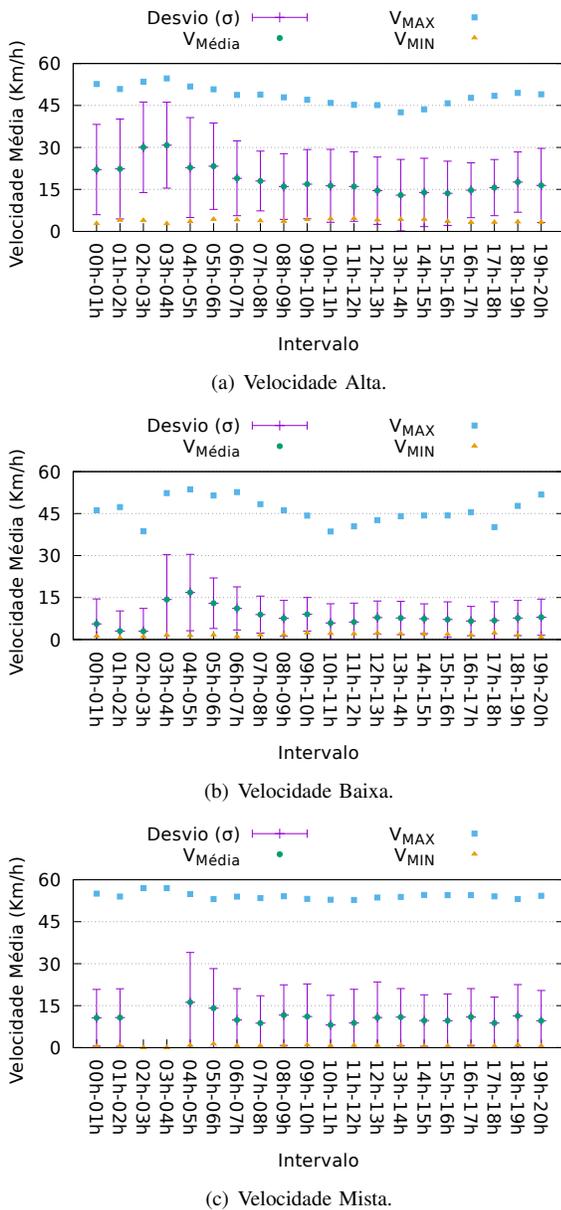


Figura 3. Comportamento das regiões com base na inserção de 777 amostras anômalas por hora.

tanto para cima quanto para baixo da média real. Por um lado, para reduzir a média da região, o usuário deve contribuir com uma velocidade inferior à média, sendo o valor mínimo possível igual a 0 km/h. Por outro lado, para aumentar a média de velocidades da região, deve-se inserir amostras com valores elevados de velocidade. Assume-se que esse tipo de usuário utiliza os valores estipulados na Tabela II. Ambas as tentativas de alteração da média podem gerar transtorno para o sistema caso não sejam previamente detectadas. É nesse momento que os dados amostrados são introduzidos na segunda etapa da metodologia proposta, *deteção de anomalia da amostra*.

A Figura 3 mostra a velocidade média dos ônibus em cada intervalo de 1 hora. Cada tipo de ponto representa um

cenário distinto. Os pontos V_{MAX} representam a velocidade média de amostras com velocidades anômalas elevadas. Já os pontos V_{MIN} são o resultado para amostras anômalas com velocidades baixas. Por fim, os pontos $V_{Média}$ mostram o resultado livre de influência anômala. Esses resultados foram obtidos utilizando-se um adicional de 777 amostras anômalas. Fica claro que amostras com velocidades mais baixas reduzem a média de cada região, porém não são detectáveis por não estarem fora do intervalo $[\mu - 3\sigma, \mu + 3\sigma]$. Isso ocorre porque a velocidade média em cada região não é tão elevada, de forma que seria necessária a existência de amostras com velocidades negativas ou um número enorme de amostras com velocidades nulas para que fosse possível detectar a anomalia. Assim, por um lado, para detectar a alteração da média para baixo no cenário estudado seria necessário utilizar métodos de detecção mais elaborados. Por outro lado, a detecção da alteração da velocidade média para cima é possível, ficando mais evidente conforme a velocidade média da região diminui, mesmo com um método de detecção simples. O restante deste trabalho concentra-se, portanto, na detecção de alterações para cima com os dois tipos de usuários anômalos já definidos.

1) *Usuário com anomalia permanente*: A Figura 4(a) mostra a variação da taxa de detecções bem sucedidas em função do adicional de amostras anômalas inseridas no sistema. Os valores do eixo-y representam a fração de horas do período analisado de um dia em que a detecção da anomalia foi bem sucedida e o sensor anômalo não conseguiu influenciar o resultado final do sistema. Cada curva nessa figura representa uma das regiões previamente definidas. Observa-se que para a região de velocidade baixa o início da detecção ocorre mais rapidamente do que nas outras regiões. A partir de um adicional de 5% de amostras anômalas já é possível realizar a detecção em 10% do período de tempo analisado. Quando o adicional chega a 85%, a detecção é bem sucedida em 100% do tempo. As regiões de velocidade mista e alta apresentam um comportamento semelhante. No entanto, a detecção já se torna possível na região mista a partir de 5% de amostras anômalas. Em ambas as regiões mista e alta, são necessárias um grande número de amostras adicionais para que sempre ocorra detecções bem sucedidas durante o dia. Apesar disso, a partir de 115% de amostras adicionais na região mista, e de 135% na região alta, já é possível realizar detecções bem sucedidas em 85% do tempo.

2) *Usuário com anomalia intermitente*: A Figura 4(b) mostra o mesmo resultado anterior, porém para a inserção de anomalias intermitentes no sistema. Como era de se esperar, de forma geral, é necessário um número muito maior de amostras para que a detecção da anomalia seja bem sucedida. A região de baixa velocidade continua a apresentar o melhor resultado, sendo ele semelhante ao resultado da Figura 4(a). Com esse tipo de usuário, torna-se necessário 720% de amostras anômalas adicionais para que seja possível detectar a anomalia em 100% do tempo na região de baixa velocidade. Nas outras duas regiões, mesmo com um adicional de 1000% a detecção em 100% do tempo não é possível. Para esse perfil de usuário é necessário utilizar um método de detecção mais sofisticado.

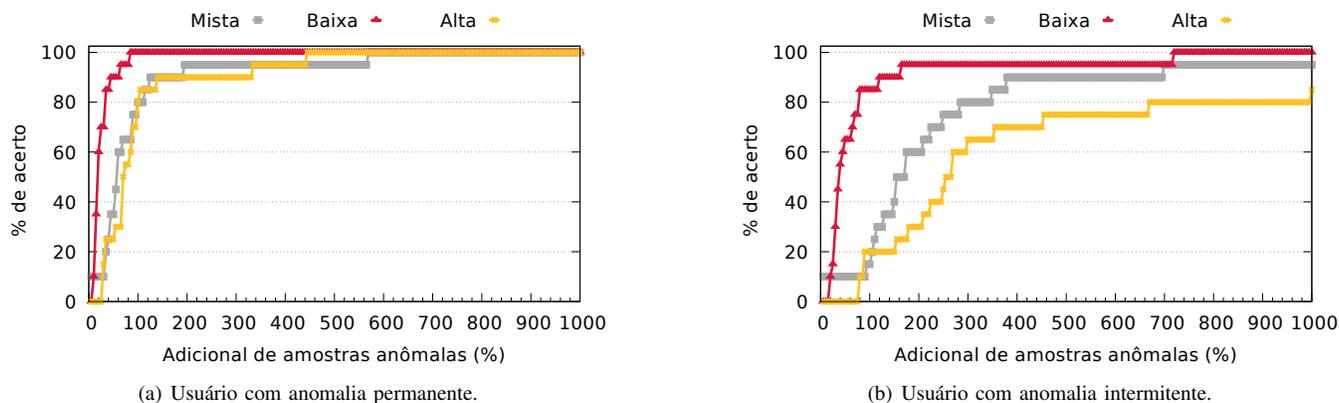


Figura 4. Quantidade de amostras anômalas necessárias para que o sistema seja capaz de realizar detecções bem sucedidas em cada região, utilizando um método de detecção simples.

B. Impacto das amostras anômalas na capacidade da rede

Esta análise considera que os dados são transmitidos utilizando-se o padrão IEEE 802.11p. Essas redes são usadas em redes veiculares para prover comunicação entre OBUs e RSUs, e entre OBUs. As RSUs podem estar interligadas a um controlador central. A camada física do 802.11p opera a uma taxa de transmissão máxima de 27 Mb/s, e mínima de 3 Mb/s, na frequência de 5,890 GHz e com banda de 10 MHz [11].

Para o conjunto de dados analisado, considera-se que o tamanho dos dados armazenados por amostra possui 480 Bytes, e a quantidade média de amostras inseridas durante uma hora é igual a 777, totalizando uma carga de dados extra de 372,96 kB. Existe pouca influência das amostras anômalas na carga da rede, de forma que mesmo com um adicional de 1.000% o desempenho da rede não é comprometido, uma vez que a carga total atingiria apenas 173,89 kb/s.

VII. CONCLUSÕES

Este trabalho propôs uma metodologia para detecção de anomalia para sensoriamento participativo em redes veiculares. Uma anomalia pode ser uma simples inconsistência ou dados não confiáveis, que pode prejudicar a análise dos dados, levando a conclusões irreais. Foram apresentados dois possíveis usuários que produzem amostras anômalas, o persistente e o intermitente. Os dois tipos de usuários anômalos conseguem ser identificados já com poucas amostras anômalas. Desses dois usuários, foi mostrado que o permanente detecta as anomalias de forma mais rápida a partir de suas primeiras amostras. Já o intermitente necessita de mais amostras, visto que ele contribui com medidas reais e medidas erradas em instantes consecutivos. A avaliação foi realizada em três regiões distintas, definindo situações reais do dia-a-dia. Os resultados mostram que quando os usuários inserem pelo menos 5% a mais de amostras anômalas na região de baixa velocidade, e capaz de serem identificados. Nas regiões de velocidade mista e alta essa proporção aumenta, sendo necessário, respectivamente, 10% e 15% a mais de amostras. Mesmo com esse volume de amostras adicionais, a carga total de dados gerados não é suficiente para provocar problemas de rede.

Como trabalhos futuros, pretende-se avaliar a metodologia proposta em todas as rotas da cidade de Seattle, sendo possível confirmar se o comportamento das três regiões definidas se mantém conforme o atual trabalho. Também pode-se definir novos perfis de usuários para representar novos cenários. Implementar a metodologia para considerar o histórico também é visualizada como uma direção futura.

AGRADECIMENTOS

Os autores gostariam de agradecer o apoio da CAPES, do CNPq e da FAPERJ.

REFERÊNCIAS

- [1] I. I. Reports, "ITU Strategy and Policy Unit (SPU)," International Telecommunication Union (ITU), Tech. Rep., 2005.
- [2] Y. Xiao, P. Simoens, P. Pillai, K. Ha, and M. Satyanarayanan, "Lowering the barriers to large-scale mobile crowdsensing," in *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*. ACM, 2013, p. 9.
- [3] V. Ribeiro Neto, D. S. V. Medeiros, and M. E. M. Campista, "Analysis of mobile user behavior in vehicular social networks," in *International Conference Network of the Future (NOF 2016)*, 2016, pp. 1–5.
- [4] P. Cruz, F. F. da Silva, R. G. Pacheco, R. S. Couto, P. B. Velloso, M. E. M. Campista, and L. H. M. Costa, "Sensingbus: um sistema de sensoriamento baseado em ônibus urbanos," *XXXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, vol. 14, no. 1, 2017.
- [5] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*. ACM, 2008, pp. 323–336.
- [6] P. Zhou, Y. Zheng, and M. Li, "How long to wait?: predicting bus arrival time with mobile phone based participatory sensing," in *Proceedings of the 10th international conference on Mobile systems, applications, and services*. ACM, 2012, pp. 379–392.
- [7] J. G. Jetcheva, Y.-C. Hu, S. PalChaudhuri, A. K. Saha, and D. B. Johnson, "CRAWDAD dataset rice/ad_hoc_city (v. 2003-09-11)," http://crawdad.org/rice/ad_hoc_city/20030911/bus_mobility, Sep. 2003, traceset: bus_mobility.
- [8] P. C. F. Melo *et al.*, "Csvm: Uma plataforma para crowdsensing móvel dirigida por modelos em tempo de execução," 2014.
- [9] "Seattle department of transportation," <http://www.seattle.gov/transportation/sdotfaqs.htm>, 2017, acessado: 20/07/2017.
- [10] K. C. Metro, "Route 7 - king county," <http://kingcounty.gov/depts/transportation/metro/schedules-maps/007.aspx#weekday>, 2017, acessado: 13/06/2017.
- [11] D. O. Basso, "Ieee 802.11 p: análise de desempenho de redes sem fio veiculares," 2013.